의미 있는 태그 클러스터 구축을 위한 설계 방안

박병재*, 우종우*
*국민대학교 컴퓨터공학과
e-mail : bejey79@gmail.com

A Design of Building a Meaningful Tag Cluster

Park, Byoung-Jae*, Woo, Chong-Woo*
*School of Computer Science, Kookmin University

요 약

태깅은 웹 2.0 의 핵심 기술 중 하나로, 매우 유연하고 역동적인 분류 체계를 제공한다. 하지만 유연성과 역동성의 확보에 의해 계층 구조나 연관 관계와 같은 태그의 관계성이 부족하거나 존재하지 않는 한계점을 가지고 있는 것 또한 사실이다. 이런 한계점을 보완하기 위한 방법으로 계층 관계를 형성하기 위한 계층 클러스터링 방법과, 연관 관계를 형성하기 위한 협업 필터링 방법이 존재한다. 이 두 가지 방법은 태그의 관계성을 제공하지만, 연관 관계와 계층 관계 중 하나만 제공한다는 단점을 가진다. 본 논문에서는 태그 검색 시 연관 관계뿐 아니라 계층 구조의 탐색을 제공해주기 위한 태그 클러스터링 알고리즘을 설계하였다. 제안한 알고리즘은 사용자 태그셋을 활용하여 태그의 유사성을 계산하는 방법을 제시하고, 기존의 시각화 방법(태그 구름)과 다른 새로운 형태로 시각화할 수 있는 결과 데이터를 제공한다.

1. 서론

웹 2.0 환경을 주도하고 있는 기술 중 하나가 태깅이다. 태깅은 어떤 정보에 대하여 사용자가 직접 작성한 키워드 (메타데이터)를 의미한다[1]. 태깅은 사용자의 자유로운 연상 작용을 통해 떠오른 다수의 키워드를 이용하여 데이터를 분류한다. 따라서 기존의 정해진 분류 체계의 틀을 따를 필요 없이 각 사용자들의 개성을 반영하는 다채로운 분류체계가 생성된다. 또한, 기존의 그 어떤 카테고리에도 속할 수 없는 특이한 분야, 혹은 새롭게 생성되는 분야에 속하는 데이터도 태깅으로 분류될 수 있다. 이는 웹 2.0 환경에서 빠른 속도로 생성되는 새로운 데이터에 대해 유연하게 대처할 수 있다는 점에서 태깅의 큰 장점이라고 할 수 있다.

이렇게 태깅은 매우 유연하고 역동적인 분류체계를 제공한다. 하지만 유연성과 역동성의 확보로 인해 발생하는 근본적인 한계가 있으며, 이를 정리하면 다음과 같다 [2].

첫째, 태깅에는 태그 간 계층구조가 존재하지 않는다는 것이다. 예를 들면, 자바 프로그래밍에 대한 소개를 다루는 문서에 대해 '컴퓨터', '프로그래밍', '자바'라는 세 개의 태그를 기록했다고 하면, 이들 태그들은 실제로 상위-하위 개념이라는 관계를 가지고 있지만 태깅 시스템에서는 이를 반영할 수가 없다.
둘째, 태그간의 연관 관계를 쉽게 파악하기 힘들다. 셋째, 태깅은 동의어나 유의어에 대한 관리를 제공해주지 않는다. 예를 들면, '남자'와 '남성' 혹은, '해변'과 '바닷가' 등과 같은 태그들은 실제로는 서로 의미가 동일하거나 유사함에도 불구하고 서로 다른 태그로 분류된다.

이러한 태깅의 한계점을 한 마디로 요약하면 관계성의 부족이라고 할 수 있으며, 본 논문에서는 앞서 서술한 태깅의 한계점 중 첫 번째와 두 번째 한계점을 보완하고자 한다. 태깅의 연관 관계와 계층 구조를 형성하는 기존 방법으로는 계층 클러스터링과 협업 필터링 방법이 있다. 두 방법은 태깅의 관계성을 제공하고 있지만 계층 구조와 연관 관계 하나만 제공한다는 단점이 있다. 본 논문에서는 계층 구조와 연관 관계 모두 제공하기 위해 협업 필터링과 계층 클러스터링 알고리즘의 장점을 결합한 형태의 알고리즘을 설계하고 제안한다. 제시한 알고리즘은 태그의 한계점을 보완할 뿐 아니라 태그 탐색의 편리성을 제공하기 위한 데이터 구조를 생성한다.

본 논문의 구성은 다음과 같다. 2 장에서는 태그의 한계점을 보완하기 위해 기존 방법인 협업 필터링과 계층 클러스터링에 대해 알아보고, 3 장에서는 설계한 알고리즘의 내용과 구조에 대해서 설명한다. 마지막으로 4 장에는 결론과 향후 과제에 대해 기술한다.

2. 관련 연구

2.1 협업 필터링

태그의 연관 관계를 보여주기 위한 방법으로 협업 필터링 방법이 있다. 협업 필터링은 크게 사용자 기

※ 본 논문은 서울시 산학연 협력사업의 지원을 받아 연구 수행된 논문입니다.

반 협업 필터링과 아이템 기반 협업 필터링[3]으로 나눌 수 있는데, 태그는 하나의 아이템이라고 볼 수 있으므로 아이템 기반 협업 필터링이 적합하다. 아이템 기반 협업 필터링은 크게 다음과 같은 두 가지 절차를 가진다.

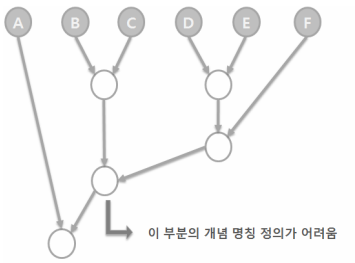
① 아이템 간의 상관관계를 결정하는 아이템 매트릭스(Item-Item matrix)를 만든다.
② 매트릭스를 사용하여 검색한 아이템과 유사한 아이템을 유추하여 추천한다.

아이템 기반 협업 필터링을 사용하여 태그의 연관 관계를 보여주는 사례로는 딜리셔스[4]와 플리커[5]가 있다.

2.2 계층 클러스터링

태그의 상위-하위 개념이 없다는 한계점을 보완하기 위한 방법으로 계층 클러스터링 알고리즘[6]을 이용하여 태그의 계층을 만드는 방법[7]이 있다. 그림 2는 계층 클러스터링 알고리즘의 개요를 도식화한 그림이며, 알고리즘의 절차는 아래와 같다.

① Euclidean 알고리즘[8]이나 Correlation 알고리즘[9]등의 거리 측정(distance metric) 알고리즘을 이용하여 태그 유사도를 계산한다.
② 유사도가 가장 높은 두 개의 태그를 클러스터링한다.
③ 클러스터된 태그와 나머지 태그와 다시 유사도 계산을 하고 유사도가 가장 높은 태그 두 2 개를 클러스터링 한다.
④ 하나로 클러스터 될 때까지 ③을 반복한다.



(그림 1) 계층 클러스터링 개요

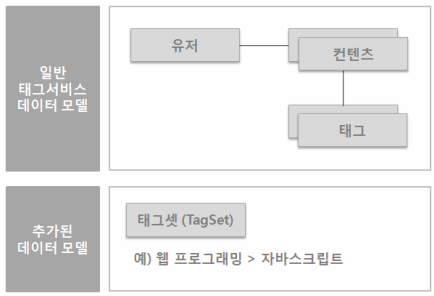
계층 클러스터링 알고리즘은 태그를 계층적으로 클러스터링 하지만, 그림 1 에서 보듯이 b 와 c 의 클러스터에 대한 개념 명칭을 정하기 어렵다는 단점이 있으며, 태그 계층 클러스터링의 사례로는 [10]에서 소개하고 있는 rawsugar[11] 서비스가 있다.

3. 시스템 설계

이장에서는 본 연구의 시스템에서 사용되는 데이터 모델의 구조에 대해서 먼저 기술한 후, 설계원칙 및 태그 클러스터링 알고리즘의 상세내용과 세부 절차를 살펴보도록 한다.

3.1 데이터 모델

본 논문에서 태그 클러스터링 알고리즘을 설계하기 위해 사용하는 데이터 모델은 그림 2 에서처럼 크게 두 부분으로 나뉜다. 첫 번째는 일반적으로 태그를 이용하는 서비스에서 나타나는 데이터 모델로서, 하나의 컨텐츠(페이지)에 여러 개의 태그를 작성할 수 있는 데이터 구조를 가진다. 두 번째는 컨텐츠 검색의 편의성을 제공하기 위한 모델로서, 컨텐츠 검색 시 자주 사용되는 AND 연산 검색 태그들을 저장하기 위한 데이터 구조를 가진다. 본 논문에서는 이 데이터 구조를 태그셋이라 명명하였다. 예를 들면 “웹프로그래밍 ∩ 자바스크립트”와 같이 검색어를 입력했을 경우 “웹프로그래밍”와 “자바스크립트” 태그가 동시에 작성된 컨텐츠를 검색해주는데, 여기서 태그셋은 검색 태그의 집합인 {웹프로그래밍, 자바스크립트}이다. 태그셋은 “웹프로그래밍”과 같은 태그 정보뿐만 아니라 태그가 나타난 위치 정보도 저장한다. 위 예제에서 “웹프로그래밍”이 첫 번째로 나타났으므로 위치 정보는 “1”의 값을 가지며, 자바스크립트는 “2”의 값을 가진다. 이러한 태그의 위치 정보는 태그의 계층 관계 구조를 만드는데 사용된다.



(그림 2) 데이터 모델

3.2 설계의 주안점

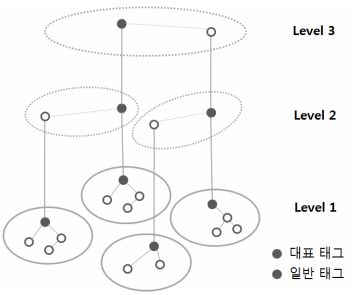
태그 클러스터링 알고리즘 설계는 다음과 같은 주안점을 가지고 설계하였다.

① 태그의 한계점을 보완하기 위해 태그의 상하 관계 및 연관 관계 데이터를 생성해야 한다.
② 생성된 결과 데이터는 태그의 시각화가 용이해야 하며, 태그 시각화 기법 중 많이 사용되고 있는 태그 구름을 이용한 정보 탐색의 문제점을 보완할 수 있어야 한다[7].
③ 태그 검색 시 검색한 태그의 연관 태그 뿐 아니라 하위 개념의 태그에 대한 탐색이 용이해야 한다. 예를 들어 ‘프로그래밍언어’를 검색했을 경우 ‘프로그래밍 언어’와 유사한 태그를 보여줄 뿐 아니라 ‘Java 언어’와 같은 하위 개념의 태그들도 보여줄 수 있어야 한다.

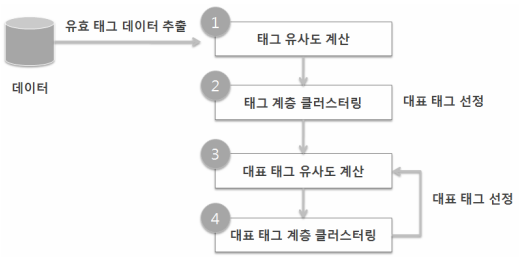
3.3 태그 클러스터링 알고리즘

그림 3 은 알고리즘의 데이터 구조를 도식화한 그림이다. 데이터 구조의 전체적인 모습은 클러스터의 계층을 이루는 Layer 구조의 형태를 가진다. 데이터 구조의 기본 단위는 태그이며, 태그 타입으로 구분했을 때 일반 태그와 대표 태그로 나누어지며, 클러스터 단위로 구분했을 때 일반 태그 클러스터와 대표 태그 클러스터로 나누어진다. 클러스터 내부는 태그간 계층 구조를 이루고 있으며 가장 상위에 있는 태그가 클러스터를 대표하는 대표 태그이다. 이런 태그의 관계성을 나타내는 위한 알고리즘의 절차는 그림 4 에 나타나 있으며, 그 내용은 다음과 같다.

① 태그간 유사도를 계산
② 일반 태그 클러스터링 (대표 태그 선정)
③ 대표 태그 유사도 계산
(클러스터 내 대표 태그선정)
④ 대표 태그 클러스터링
⑤ 3 ~ 5 반복



(그림 3) 알고리즘 데이터 구조



(그림 4) 알고리즘 절차

3.3.1 유사도 계산

태그 유사도 계산을 하기 전에 유효한 태그를 먼저 추출한다. 유효한 태그란 연관 관계가 없는 태그는 유사도 계산에 영향을 안 주므로 연관 가중치(표 1 에서 w)가 10 이상인 태그를 말한다. 태그간 유사도는 기본적으로 표 1 에서 설명한 항목을 사용하여 아래와 같은 수식을 사용하여 계산된다.

$$Sim(T_1,T_2)=\frac{w(T_1,T_2)}{C(T_1\in T)+C(T_2\in T)}+\frac{tw(T_1,T_2)}{C(T_1\in TS)+C(T_2\in TS)}$$

<표 1> 유사도 계산을 위한 항목

항목	설명
T	Tag
TS	Tag Set
C(T)	Count(distinct(T)) 태그가 나타난 빈도수
w (T1, T2)	태그간 연관 가중치 같은 컨텐츠에 두 개의 태그가 동시에 있는 경우의 합
tw (T1, T2)	태그셋내에서의 연관 가중치 같은 태그셋에 두 개의 태그가 동시에 있는 경우의 합

3.3.2 태그 클러스터링

3.3.1 에서 계산된 유사도를 사용하여 태그를 클러스터링한다. 클러스터링의 절차는 다음과 같다.

① 각각의 태그에 대해서 가장 높은 유사도를 가진 태그를 찾는다.
② 표 2 의 수식을 이용하여 두 태그의 상하 관계를 설정한다.
③ 각각의 태그의 관계를 이용하여 태그를 연결하고, 가장 상위 부모 태그를 태표 태그로 설정한다.

<표 2> 태그의 상하 관계 설정 수식

$$CF(T_1,T_2)=\begin{cases}1 & \text{if } T_1\text{'s position is front than } T_2\\ 0 & \text{else}\end{cases}$$

$$(T_1\in TS_1, T_2\in TS_1, TS_1\in TS)$$

$$CH(T_1,T_2)=\sum_{TS_1\in TS}CF(T_1,T_2)$$

$$P(T_1,T_2)=\begin{cases}T_1 & \text{if } CH(T_1,T_2)>CH(T_2,T_1)\\ T_2 & \text{else}\end{cases}$$

T1과 T2를 포함하는 충분한 태그셋이 존재하지 않을 시에는 아래의 수식을 사용하여 상하 관계를 설정한다.

$$P(T_1,T_2)=\begin{cases}T_1 & \text{if } C(T_1)>C(T_2)\\ T_2 & \text{else}\end{cases}$$

3.3.3 대표 태그 클러스터링

추출된 대표 태그 클러스터들을 다시 클러스터링한다. 클러스터링 절차는 유사도 계산 부분과 클러스터링 종료 조건을 제외한 나머지 부분은 3.3.2 의 절차

와 동일하다. 대표 태그 클러스터에 사용되는 유사도 계산은 각각의 태그 클러스터내의 태그 중 대표 태그와 유사한 10 개를 추출하여 아래와 같은 수식을 사용하여 계산한다.

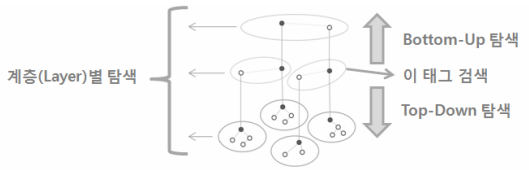
$$SimC(C_1, C_2) = \frac{\sum_{T_1 \in C_1} \sum_{T_2 \in C_2} Sim(T_1, T_2)}{Count(T_1 \in C_1) * Count(T_2 \in C_2)}$$

클러스터링의 종료는 클러스터의 유사도가 특정 임계값을 넘는 클러스터 데이터가 없을 때 종료하게 된다.

3.4 알고리즘의 특징

본 논문에서 제시한 알고리즘은 다음과 같은 특징을 가지고 있다.

① 태그의 한계점을 보완하기 위한 기존의 두 알고리즘(계층 클러스터링과 협업 필터링)이 제공하는 태그의 관계성(계층 구조와 연관관계)을 모두 제공한다.
② 계층 구조를 형성하기 위한 계층 클러스터링 알고리즘의 단점이라고 할 수 있는 상위 개념의 명칭 정의 어려움을 해결하였다.
③ 계층 관계뿐 아니라 연관 관계를 통해 그림 5 와 같이 태그의 Top-Down 탐색과 Bottom-up 탐색을 제공하고, 계층별로 태그를 시각화하여 보여줌으로써 편리하게 태그를 탐색할 수 있다.



(그림 5) 태그 탐색

4. 결론 및 향후 과제

본 논문에서는 웹 2.0 환경의 핵심 기술인 태깅의 한계를 알아보고, 보완할 수 있는 알고리즘을 설계하고 제안하였다. 제시한 알고리즘은 태그의 한계점인 관계성의 부재를 보완할 뿐만 아니라 태그 탐색의 편이성을 제공해줄 수 있다. 그러나 태그간의 관계는 본 논문에서 언급한 연관관계와 계층 관계 외에 동의어, 반의어 등의 다양한 관계가 존재할 수 있다.

향후 과제로는 OWL[12]을 통한 태그 기반 온톨로지 구축이나 협업을 이용한 관계 설정을 통해서 태그 관계가 유연하게 확장될 수 있도록 설계를 수정 보완할 것이다. 또한 본 논문에서 설계한 알고리즘을 구현하고, 딜리셔스의 태그 데이터을 이용하여 알고리즘의 결과 데이터(태그 관계 데이터)의 정확성을 실험할 예정이다.

참고문헌

[1] 이강표, 김두남, 김형주, “웹 2.0 환경에서의 태깅 기술 동향”, 정보과학회지 제 25 권 제 10 호, p36-42, 2007
[2] Ellyssa Kroski, “The Hive Mind Folksonomies and User Based Tagging”, http://infotangle.blogsome.com /2005/12/07/ the-hive-mind folksonomies-and-user-based-tagging/
[3] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, “ItemBased Collaborative Filtering Recommendation Algorithms”, http://wwwusers.cs. umn.edu/~sarwar/sdm.ps
[4] http://del.icio.us
[5] http://www.flickr.com
[6] Jain, Murty and Flynn, “Data Clustering: A Review”, http://www.cs.rutgers.edu/~mlittman/courses/lightai03/jain99data.pdf
[7] Christopher H. Brooks and Nancy Montanez, “Improved Annotation of the Blogosphere via Autotagging and Hierarchical Clustering”, http://www.cs.usfca.edu/ ~brooks/papers/brooks-montanez-www06.pdf
[8] http://en.wikipedia.org/wiki/Euclidean_distance
[9] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient
[10] Grigory Begelman, Philipp Keller and Frank Smadja “Automated Tag Clustering: Improving search and exploration in the tag space”,
[11] http://rawsugar.com
[12] http://www.w3.org/TR/owl-features/