Web2.0 환경에서의 Topic Map 생성을 위한 Tag Clustering에 관한 연구

이시화*, 무효려*, 이만형*, 황대훈*
*경원대학교 전자계산학과
e-mail : leesihwaman@gmail.com, Hwangdh@kyungwon.ac.kr

A Study on Tag Clustering for Topic Map Generation in Web 2.0 Environment

Si-Hwa Lee*, Xiao-Li Wu*, Man-Hyoung Lee*,
Dae-Hoon Hwang*
*Dept of Computer Science, Kyungwon University

요 약

기존의 웹서비스가 정적이고 수동적인데 반해 최근의 웹 서비스는 점차 동적이고 능동적으로 변화하고 있다. 이러한 웹서비스 변화의 흐름을 잘 반영하는 것이 웹 2.0이다. 웹 2.0에서 대부분의 정보는 사용자에 의해 생산되고, 사용자가 붙인 태그(tag)에 의해 분류되어진다. 그러나 현재 태그에 관한 서비스 및 연구들은 태깅(tagging) 방법에 대한 연구를 비롯해 이를 표현하기 위한 tag cloud에 초점이 맞춰져 진행됨에 따라, 다양한 태그 정보자원 간의 체계와 연결 관계인 지식체계를 제공하지 못하고 있다.

이에 본 논문에서는 체계화된 지식표현을 위해 웹상에 편재되어 있는 학습 관련 리소스(resources) 및 태그들를 수집한다. 이를 사용자가 요청한 검색 키워드와 연관성이 있는 태그 정보들을 맵핑 및 클러스터링하여 최적화된 표현 형식인 토픽 맵(topic map)화하기 위한 시스템을 제안하며, 이 중 토픽 맵 생성을 위한 초기 연구 단계로서, 연관 태그들 간의 맵핑 및 클러스터링을 위한 알고리즘 제시를 중심으로 소개한다.

1. 서론

인터넷의 발달과 사용자의 적극적인 참여에 힘입어 웹서비스 환경은 다양하게 변화하고 있다. 기존의 웹 서비스가 정적이고 수동적인데 반해 최근의 웹 서비스는 점차 동적이고 능동적으로 변화하고 있다. 이러한 웹 서비스 변화의 흐름을 잘 반영하는 것이 웹 2.0이다[1].

웹 2.0에서 대부분의 정보는 사용자의 의해 생산되고, 사용자가 붙인 태그에 의해 정보들을 체계화시키고, 이를 공유함으로서 다양한 정보자원간의 체계와 연결 관계를 만들 수 있도록 하는 것이다.

이러한 웹 2.0 환경에서의 현재 태그에 관한 서비스 및 연구들은 태깅 방법에 대한 연구를 비롯해 이를 표현하기 위한 tag cloud에 초점이 맞춰져 진행 중에 있다. 그러나 tag cloud은 한눈에 모든 태그들의 트렌드를 살펴볼 수 있는 큰 장점이 있지만 웹상에 산재되어 있는 수많은 태그들 간의 연관성에 대해 표현하고 체계화하는 것은 부족하다[2].

이에 본 논문에서는 체계화된 지식표현을 위해 수집한 태그들을 연관성에 따라 팹핑하고, 이를 관련성 높은 태그 그룹으로 클러스터링한다. 또한 클러스터링 된 결과를 최적화된 표현 형식인 토픽 맵화 하기 위한 시스템을 제안한다. 이 중 토픽 맵 생성을 위한 초기 연구 단계로서. 연관 태그들 간의 맵핑 및 클러스터링을 위한 알고리즘 제시를 중심으로 소개한다.

2. 관련 연구

2.1 태그의 정의

태그(tag)는 학생들의 이름표, 수하물의 딱지, 제품의 상표를 뜻한다[1]. 인터넷 사용자들은 자유롭게 웹 페이지, 사진, 웹 링크와 같은 다양한 콘텐츠들에

태그를 이용하여 자발적으로 정보들을 체계화시키고, 이를 공유함으로써 다양한 정보자원간의 체계와 연결 관계를 만들 수 있도록 하는 것이다.

사용자들은 체계화된 분류체계를 배우고 학습하여 체계를 분류하고 만드는 것이 아니라, 자발적으로 체계를 만들어가는 것이기 때문에 보다 편리하게 콘텐츠를 분류하고 체계화시킬 수 있다는 장점을 갖게 된다. 태깅과 폭소노미(folksonomy)를 이용한 기술에 대해서는 자동 태깅 기술과 효과적인 태깅 방법에 대한 연구를 비롯해 tag cloud 구성 기술, 다중 응용에서의 협업적 태깅 기술에 대한 연구, 폭소노미 기반의 관계 추출, 온톨로지와 연계한 폭소노미 기술 등에 대한 많은 연구들이 진행되고 있다[2]. 그러나 현재 진행 중인 연구들은 태깅을 위한 기술에 초점이 맞춰져 진행 중에 있으며, 사용자들이 태깅한 태그들을 이용한 지식체계 표현에 대한 연구가 미흡한 실정이다.

2.2 클러스터링 알고리즘

클러스터링은 정보 검색의 효율성과 유효성을 증대시키기 위한 목적으로 사용한다. 클러스터링 기법은 입력되는 정보의 순서에 따라 클러스터링 결과가 달라지는 단일처리 방법(single pass method)과 이의 단점을 보완한 재배치 방법(reallocation method) 그리고 대용량에 대한 탐색적인 기법으로 사전적인 정보 없이 의미 있는 자료구조를 얻으며 모든 형태의 데이터에 적용 가능한 K-Means 알고리즘이 있다. 그러나 이는 좌표평면 상에 좌표 값을 가지는 대상에 대한 클러스터링에만 적용되는 단점이 있다. 또한 단일링크 방법, 완전 링크 방법, 그룹 평균 연결 방법 등이 있으나, 이 기법은 클러스터링 오차가 적은 반면 대용량에 대한 처리 속도가 느리다는 단점을 가지고 있다[3].

이에 본 논문에서는 널리 잘 알려진 검증된 알고리즘인 Spectral Bisection Algorithm[4]을 적용하여 클러스터링 하였으며, 또한 더 이상의 클러스터링이 필요한지를 검증하기 위하여 클러스터의 질을 판단하는 기준으로 Modularity Function Q을 사용하였다[5].

2.3 토픽 맵

토픽 맵(topic map)은 지식구조를 묘사하고 지식구조와 정보자원을 연결하기 위해 만들어진 새로운 ISO 표준이다. 2000년 ISO/IEC 13250으로 발표되었고, 2001년 비정규기관인 Topicmaps.org에 의하여 XTM Topic Maps(XTM) 1.0 표준규격이 발표되었다. 2002년 5월에 두 번째 ISO/IEC 13250 개정판이 발표된 상태이다. 현재 XML을 주요 토픽맵 구문으로 사용하고 있으며 XTM 구문은 거의 모든 토픽맵 도구가 지원하고 있다[6].

토픽 맵은 대용량의 비구조화되고 비조직화된 정보를 효율적으로 검색하고 네비게이션하기 위한 해결책으로 제안되었다. 토픽 맵은 대용량의 정보를 분류하고 구조화하며, 의미론적인 연관관계를 설정할 수 있는 모델을 제시하고 있으며, 토픽 맵 모델은 기본요소로서 토픽(topic), 연관 관계(association), 어커런스(occurrence)로 원하는 지식을 쉽고 정확하게 찾을 수 있는 맵을 제시한다[6].

3. 제안 시스템

본 논문에서는 사용자가 요청한 검색 키워드와 유사한 태그들 간의 관계를 최적화된 표현 형식인 토픽 맵화하기 위한 방법을 제안한다.

제안하는 그림 1의 Tag 기반 토픽 맵 생성 시스템은 크게 웹상에 편재 되어있는 Resources 및 태그 정보들을 수집하여 DB화하기 위한 Tags Reader, 수집된 태그들 간의 빈도수 추출 및 팹핑을 수행하는 Tag Relation Mapping Module, 맵핑된 태그 및 빈도수를 기반으로 관련성 높은 태그 그룹으로 클러스터링 하는 Tag Clustering Module, 클러스터링된 태그들 간의 순위화 및 어소시에이션(Association) 부여를 위한 Ranking Module, 마지막으로 토픽 맵 생성을 위해 부여된 Association 정보를 이용한 Topic Map Generation Module로 구성된다.

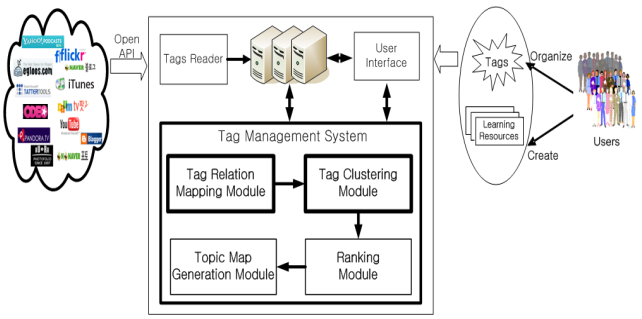

그림 1. Tag 기반 Topic Map생성 시스템 구조

이 중 본 논문에서는 수집된 태그 정보들을 토픽 맵화 하기위한 초기 단계로서 태그들 간의 맵핑을 통한 빈도수 추출 및 클러스터링을 위한 알고리즘 제시를 중심으로 소개한다.

3.1 연관 태그 맵핑

본 논문에서는 Flickr Open API를 이용하여 Resource 및 태그 정보를 수집하였다. 이중 컴퓨터와 관련된 상위 리소스 6개 및 그에 따른 태그 48개의 데이터는 표 1과 같다.

표 1. Flickr API를 통해 수집한 태그

Resource Image	Tags
Resource 1	work, computer, monitor, technology, screen
Resource 2	computer, dell, notebook, desktop, computer
Resource 3	notebook, screen, monitor, computer, date, desktop computer, laptop, keyboard, mouse
Resource 4	apple, computer, desktop computer, speaker, keyboard, mouse, ipod, monitor, screen
Resource 5	windows, screen, mouse, LCD, laptop, notebook, PC, computer, printer, technology, keyboard, monitor
Resource 6	mac, imac, desktop computer, setup, mouse, keyboard, macintosh, apple, computer, monitor
...	...

표 1의 수집된 태그 정보들을 이용하여 토픽 맵 생성을 위한 첫 번째 단계인 연관 태그들 간의 맵핑 및 빈도수를 추출하게 된다.

표 1에서 리소스 1에 대한 태그(work, computer, monitor, technology, screen)들은 서로 다른 사용자들이 리소스 1을 보고 느낀 것을 태그 정보로 표현한 것이다. 따라서 동일한 리소스에 대한 상이한 태그들은 서로 연관성이 있다고 정의할 수 있다.

이러한 연관성을 이용하여 표 1의 리소스 1, 2, 3, 4, 5, 6의 태그들을 서로 맵핑하게 된다. 여기에서 매핑이란 동일한 리소스에 대하여 서로 다른 사용자에 의해 부여된 태그들을 서로 관련이 있는 태그로 연관 관계를 부여하는 것이다. 또한 태그 맵핑 과정에서 태그들 간의 연관관계 및 동일 태그 출현 빈도를 이용하여 태그 빈도수를 추출하게 되며, 이는 식 (1), (2)와 같다.

$$A_G(i,j) = 0 \qquad \text{식 (1)}$$

: Tag i와 Tag j간에 관계가 없을 때.

$$A_G(i,j) = k \qquad \text{식 (2)}$$

: Tag i와 Tag j간에 k번 연관관계가 있을 때.

여기에서 $A_G(i,j)$는 태그 그래프 G의 인접행렬(adjacency matrix) A_G의 i열로서, Tag i와 Tag j가 동일한 리소스에서 동시에 출현하는 빈도수를 의미하며, 식 (1)에서 Tag i와 Tag j간에는 연관관계가 없음을 의미하며, 식 (2)$A_G(i,j) = k$는 Tag i와 Tag j간에 k번 연관관계가 있는 것을 의미한다.

이러한 과정에 의해 생성된 태그 그래프(tag graph) 및 빈도수는 그림 2와 같다.

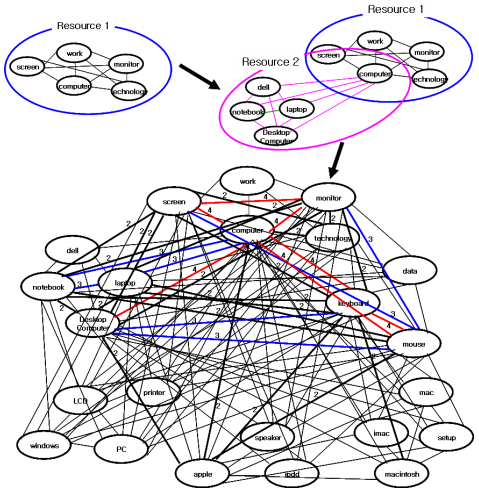

그림 2. 연관 태그 맵핑 및 빈도수 추출

3.2 클러스터링 알고리즘

3.1절에서 추출된 태그 그래프와 빈도수를 기반으로 두 번째 단계인 관련성 높은 태그 그룹으로 클러스터링(clustering) 하기 위해 다음 그림 3의 클러스터링 알고리즘을 적용한다.

```
//A_G = (a_ij) : Adjacency Matrix of Tag Graph G
//D_G = (d_ij) : Degree Matrix of Tag Graph G
//L_G = (D − A) : Laplacian Matrix of Tag Graph G
Find Laplacian Matrix L_G of the tag graph G by using
  adjacency matrix A_G.
Compute the value of modularity function Q_0 of the original
  graph.
While (Q_0 < Q_1) {
    Compute the eigenvector v_2 of L_G and eigenvalue
    λ_2(L_G)
    // λ_2(L_G) is the Second largest positive eigenvalue
    Bisect the vertices of the graph with v_2, and make
      partitioned graph G_1 with bisection.
    Compute the value of modularity function Q_1 of
      partitioned graph G_1
    if (Q_0 < Q_1) then {
        Q_0 ← Q_1
        G ← G_1
        }
}
```

그림 3. 클러스터링 알고리즘

위의 클러스터링 알고리즘의 진행과정을 살펴보

면, 크게 태그 그래프를 양분(bisection) 하는 과정과 양분된 그래프(partitioned graph)가 또다시 양분될 필요가 있는지 즉, 양분 과정을 마칠 것인지를 검증하는 과정으로 구성된다. 따라서 위의 알고리즘은 재귀적(recursive)으로 반복 진행되며, 태그 그래프의 태그들이 클러스터 단위로 충분히 클러스터링될 때까지 반복된다.

알고리즘에서 위와 같은 클러스터링 하는 방법을 Spectral Bisection[4] 방법이라 하며, 널리 잘 알려진 검증된 방법으로서 본 연구에 적용하였다. 또한 더 이상의 태그 그래프 분할이 필요한지를 검증하기 위하여 사용한 modularity function은 분할된 클러스터 의 질을 판단하는 기준으로 사용되며[5], 다음 식 (3)으로 정의 된다.

$$Q(P_k) = \sum_{c=1}^{k} [\frac{A(V_c, V_c)}{A(V, V)} - (\frac{A(V_c, V)}{A(V, V)})^2] \qquad \text{식 (3)}$$

여기서 P_k는 태그 그래프의 분할된 클러스터를, V_c는 분할된 클러스터 C에 속하는 태그들의 집합을 각각 의미한다.

위와 같은 과정에 의해 클러스터링된 결과의 예는 다음 그림 4와 같으며, 여기서 (A)는 클러스터링 알고리즘의 적용 전 리소스들에 태깅된 연관 태그들 간의 맵핑 결과의 예를 보여주고 있으며, (B)는 본 논문에서 Spectral Bisection 알고리즘 및 modularity function을 적용했을 때 연관도가 높은 그룹으로 클러스터링된 예를 보여주고 있다.

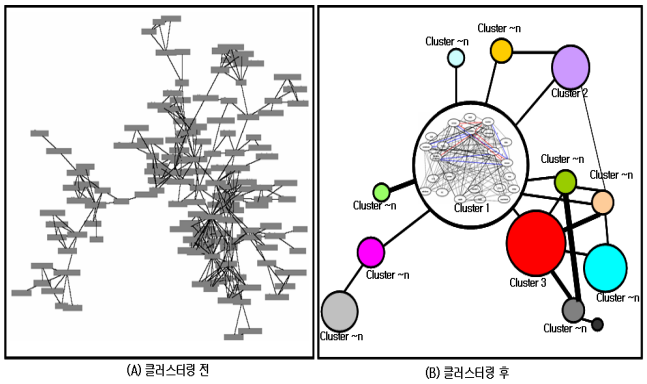

그림 4. 클러스터링된 결과의 예

이렇게 클러스터링된 클러스터 내의 태그들은 토픽 맵 생성에서의 토픽이 되며, 리소스는 어커런스가 된다. 또한 체계화된 지식표현을 위해 WordNet의 상하의 관계를 통해 태그들을 순위화 하며, 태그(Topic)들 간의 어소시에이션을 정의한다. 이러한 과정에 의해서 의미론적으로 연관된 정보들의 최적화된 표현 형식인 토픽 맵을 생성하게 되며, 이를 통해 사용자에게 효율적인 정보의 네비게이션 역할을 제공한다.

4. 결론 및 향후 연구방향

본 논문에서는 Web 2.0 환경에서 태그 정보를 기반으로 최적화된 표현 형식인 토픽 맵화 시키기 위한 시스템을 제안하였다. 이 중 토픽 맵 생성을 위한 초기 연구로서 수집된 연관 태그들 간의 맵핑 방법 및 클러스터링 알고리즘을 Spectral Bisection 알고리즘, modularity function을 적용하여 그 가능성을 보였다. 또한 클러스터링 된 정보를 기반으로 Tag들의 순위화 및 어소시에이션 부여를 위한 WordNet 적용 안을 제안하였다.

향후 연구 방향으로는 적용한 클러스터링 알고리즘을 기반으로 웹상에 산재되어 있는 태그정보들을 클러스터화 한다. 이를 WordNet에 적용하여 태그들 간의 순위화 및 어소시에이션 정의를 통한 토픽 맵 생성에 대한 연구를 수행할 계획이다.

참고문헌

[1] 강필구, 김남중, 이예슬, 채진석, “웹 2.0을 위한 효율적인 태그 관리 시스템의 설계 및 구축”, 정보과학회, Vol. 33, No. 2, 2006.

[2] 박영욱, “웹2.0 구현의 핵심, 태그”, 마이크로 소프트웨어, 5월호, 2006.

[3] M. Singh and P. Patel, and D. Khosla, "Segmentation of functional MRI by K-means clustering," IEEE Nuclear Science Symposium and Medical Imaging Conference, vol. 3, pp. 1732-1736, Oct 1995.

[4] U. Brandes, M. Gaertler, and D. Wagner. Experimentson graph clustering. In Proceedings of the 11th AnnualEuropean Symposium on Algorithms (ESA'03),of Lecture Notes in Computer Science, vol2832, p568-579. Springer-Verlag, 2003.

[5] S. White and P. Smyth. A spectral clustering approach to finding communities in graphs. In SIAM International Conference on Data Mining, 2005.

[6] Graham M.. "Topic Map technology - the state of the art," XML 2000 Conference & Exposition, Washington, USA, December 2000.