韓國IT서비스學會誌
第7卷 第4號
2008年 12月

연관 태그의 군집 알고리즘의 설계 및 구현*

박병재** · 우종우***

A Design and Development of A Related Tag Clustering Algorithm*

Byoung-Jae Park** · Chong-Woo Woo***

■ Abstract ■

Tagging represents one of the Web 2.0 technology, and has an appropriate mechanism for the classification of dynamically changing Web informations. This technique is capable of searching the Web informations using the user specified tags, but still it has a limitation of providing only the limited informations to the tags. Therefore, in order to search the related informations easily, we need to extend this technique further to search not only the desired informations through the designated tags and also the related informations. In this paper, we first have designed and developed an algorithm that can get a desired tag cluster, which is capable of collecting the searched tags along with the related tags. We first performed a test to compare the difference between the user collected tag data through RSS and the reduced data. The second test focused on the accuracy of extracted related tags that depends on the similarity functions, such as the Pearson Correlation and Euclidean. Finally, we showed the final results visually using the graph algorithm.

Keyword : Clustering, Tagging, Information Search, Visualization

논문투고일 : 2008년 10월 15일 논문수정완료일 : 2008년 12월 10일 논문게재확정일 : 2008년 12월 12일
* 본 논문은 2008학년도 국민대학교 교내연구비 지원을 받아 수행되었음.
** 국민대학교 컴퓨터공학부 대학원
*** 국민대학교 컴퓨터공학부 교수

1. 서 론

최근 웹 환경은 새로운 기술과 패러다임을 제공하면서 웹2.0으로 진화하였다. 웹 2.0으로의 진화는 정보 생산의 주체를 일반 사용자로 바꾸었으며, 그 결과 기존보다 다양하고 새로운 정보가 폭발적으로 증가하게 되었다. 정보의 증가로 인하여 분류 체계가 텍소노미(taxonomy)형태[7, 13]의 계층적 분류에서 폭소노미(folksonomy)형태[4, 9]의 태깅으로 변화하게 되었다. 텍소노미 형태의 계층적 분류는 미리 정해진 분류 체계를 가지고 있어 변화하는 웹의 정보를 처리하기에 한계가 있는 반면, 폭소노미 형태의 태깅은 사용자가 정보에 대하여 자유롭게 키워드를 작성하는 형태로서 계층적 분류 체계보다는 유연한 분류 체계를 가지고 있으며, 변화하는 정보를 분류하기에 적합한 구조를 가진다[2].

이러한 태깅은 유연하면서 역동적인 분류 체계를 제공해줄 뿐 아니라 작성된 정보를 검색하는데도 이용될 수 있다. 그러나, 검색한 태그에 대하여 한정된 정보만을 보여준다는 한계를 가지고 있다. 일반적으로 사용자들은 정보를 검색 할 때 관련 있는 정보를 더 찾기를 원하는 경우가 많기 때문에 태그의 검색 기능뿐만 아니라 검색한 정보의 관련 정보를 탐색할 수 있는 기능 또한 필요하다.

이러한 점에서, 태그를 이용하는 서비스에서는 정보 검색 시 태그 검색을 많이 사용하게 되며, 검색 태그와 관련이 있는 연관 태그를 보여줌으로써 관련 정보들을 탐색할 수 있는 기능을 제공할 수 있다. 연관 태그를 보여주는 것은 새로운 정보를 탐색할 수 있는 기능뿐만 아니라 연상되는 단어를 통해 기억나지 않은 검색 대상을 찾아주는 기능을 한다. 즉, 검색하려고 하는 검색 대상의 명확한 정보(검색 태그)를 모르는 상태에서 기억나는 단편적인 정보(연상 태그)를 통해서 검색 대상을 찾을 수가 있다. 예를 들면, 웹 2.0의 기술 중 하나인 "Ajax"의 정보를 찾고 싶은데 명확한 키워드가 생각나지 않고, "웹 2.0", "Javascript"등의 연상 단어만 생각 날 때는, 연상 단어로 검색하여 나오는 연관 태그들을 확인함으로써 자신이 찾고자 했던 대상 항목을 역으로 찾을 수 있을 것이다.

연관 태그를 보여주는 방식은 보통 나열식(listing)과 군집(clustering)하여 보여주는 방식이 있는데, 일반적으로 나열식으로 보여주기 보다는 비슷한 정보를 군집하여 보여주는 것이 관련 정보를 찾는데 있어 더 직관적일 수 있다. 본 논문에서는 검색한 태그의 연관 태그를 군집하여 보여주는 알고리즘을 제시하고, 알고리즘을 실행한 결과로 나타나는 데이터인 연관 태그의 정확도를 실험한다. 또한 군집 정보를 그래프로 시각화함으로서 보다 편리한 정보탐색의 기능을 제공한다.

본 논문의 구성은 다음과 같다. 제 2장에서는 연관 태그를 보여주는 사례 및 관련 알고리즘에 대해 알아보고, 제 3장에서는 제시한 태그 군집 알고리즘 및 태그의 시각화에 대해여 기술한다. 제 4장에서는 연관 태그 군집 알고리즘의 설계를 위한 실험 및 결과에 대해서 서술한다. 마지막으로 제 5장에서는 결론 및 향후 과제에 대해 기술한다.

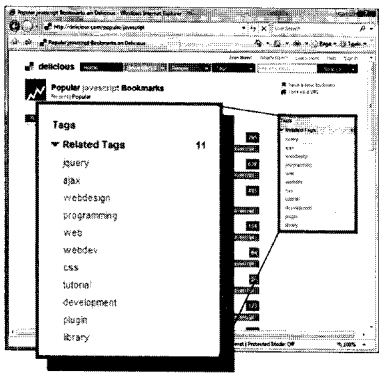
2. 관련 연구

연관 태그 탐색 기능은 주로 웹 2.0 사이트 대부분에서 지원하는 기능으로써, 이 기능을 지원하는 대표적인 사례로는 소셜 북마킹 사이트인 딜리셔스[6]와 사진 공유 사이트인 플리커[8]를 들 수 있다. 또 웹 2.0에서는 소셜 네트워킹[12]을 지원함으로써 많은 링크간의 탐색을 쉽게 할 수 있게 하는 시각화 역시 중요하게 되었다. 시각화의 대표적인 사례로는 Visual Thesaurus[16]를 들 수 있다.

2.1 딜리셔스의 연관 태그

딜리셔스에서는 북마크 정보를 찾기 위해 태그 검색을 사용하는데 [그림 1]과 같이 검색한 태그와 연관성이 높은 순으로 11개 정도를 추출하여 리스트로 보여준다. 추출된 연관 태그는 사용자가 몰

랐거나 잘 기억나지 않은 정보의 분류를 보여줌으로써 새로운 정보를 탐색할 수 있는 기능을 제공한다.



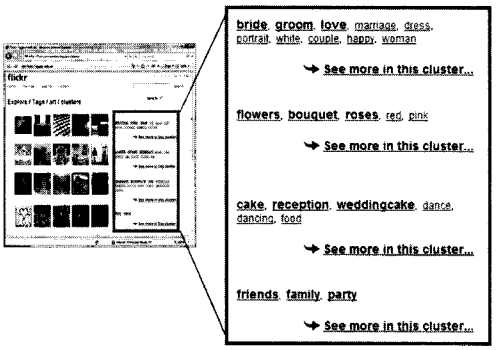
[그림 1] 딜리셔스의 연관 태그

2.2 플리커의 연관 태그 군집화

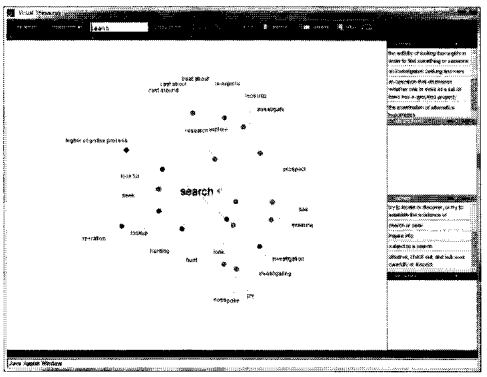
플리커는 태깅을 이용한 사진 공유 사이트로써 사진을 업로드할 때 사진에 대한 태그를 작성한다. 플리커는 딜리셔스보다 발전적인 탐색 기능을 제공하는데, [그림 2]와 같이 검색한 태그의 연관 태그를 단순 리스트가 아닌 몇 개의 유사 군집으로 보여준다. 예를 들어 "wedding"라고 검색했을 때 {bride, groom, love}, {flowers, bouquet, roses}, {cake, reception, weddingcake}, {friends, family, party}와 같이 "wedding"와 유사한 태그들을 몇 개의 군으로 묶어서 보여준다. 이렇게 연관 태그를 군집화하여 보여주는 것은 단순 리스트로 보여주는 것보다 보다 쉽게 정보를 찾고 탐색할 수 있게 한다.

2.3 어휘의 시각화

Visual Thesaurus[16]는 검색한 어휘와 관련된 어휘를 시각화하여 보여줌으로써 영어 교육에 많이 사용되는 프로그램이다. 예를 들면, 프로그램에서 "search"라는 어휘를 검색했을 경우 [그림 3]과 같이 관련 동의어, 복합어 그리고 단어와의 관계 등을 한 눈에 알아보기 쉽게 그래프로 시각화하여 보여준다. 그래프는 어휘나 태그의 관계처럼 네트워크 형태의 관계 구조[10]를 시각화하기에 적합하다.



[그림 2] 플리커의 군집화된 태그



[그림 3] Visual Thesaurus의 시각화

2.4 유사도 함수

유사도 함수는 군집화 알고리즘[5]에서 비슷한 개체를 찾기 위해 많이 사용되는 함수이다. 유사도 함수에는 다수가 있는데 그 중 많이 사용되는 함수로는 유클리디안(Euclidean)과 피어슨 상관계수(Pearson Correlation coefficient) 유사도 함수가 있다[11]. 유클리디안 함수는 벡터 $P=(p_1, p_2, \cdots, p_n)$와 $Q=(q_1, q_2, \cdots, q_n)$가 있을 때 두 벡터 사이의 거리 값을 이용하여 유사도를 계산하는 방식으로 수식은 다음과 같다.

$$\sqrt{(p_1-q_1)^2+(p_2-q_2)^2+\cdots+(p_n-q_n)^2}$$
$$=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2}.$$

상관계수 유사도 함수는 측정값 (x, y)에 대하여 n개의 측정값 (x1, y1) (x2, y2), …, (xń, yń)이 주어졌을 때 x, y사이의 유사도를 계산하는 방식으로 수식은 다음과 같다.

$$r=\frac{\frac{1}{n}\sum_{i=1}^{a}(X_i-\overline{X})(y_i-\overline{y})}{\sigma_x\sigma_y}$$

- $\overline{X}$, $\overline{y}$: x, y의 평균값,
- σx, σy : 각각 x, y의 표준편차

3. 군집 알고리즘의 설계 및 시각화

이 장에서는 검색 한 태그의 연관 태그를 군집하는 알고리즘과 이를 시각화하는 방법에 대해 서술 한다.

3.1 알고리즘 개요

웹 2.0 서비스 중 태그를 이용하는 서비스들의 기본 데이터 모델을 살펴보면 [그림 4]와 같이 나타나며, 일반적으로 이 모델을 확장하여 더 많은 기능을 제공한다. 본 논문에서는 기본 데이터 모델에서 사용하는 데이터 셋을 사용하여 알고리즘을 설계하였다. 알고리즘에 사용되는 데이터 셋은 다음과 같다.

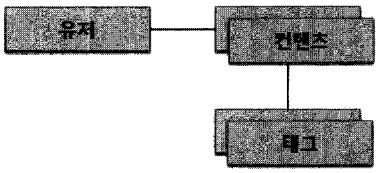
- 유저 : 컨텐츠를 등록한 사용자
- 컨텐츠 : 정보 자원 예) 북마크, 웹 문서
- 태그 : 컨텐츠에 작성한 키워드 정보

위 데이터 셋을 사용한 알고리즘의 간략한 절차는 다음과 같다.

① 연관 가중치를 이용하여 검색 태그의 연관 태그 30개를 추출한다.

② 유클리디안 유사도 함수를 이용하여 추출한 30개의 태그 중 검색 태그와 유사도가 높은 순으로 15개의 태그를 추출한다.

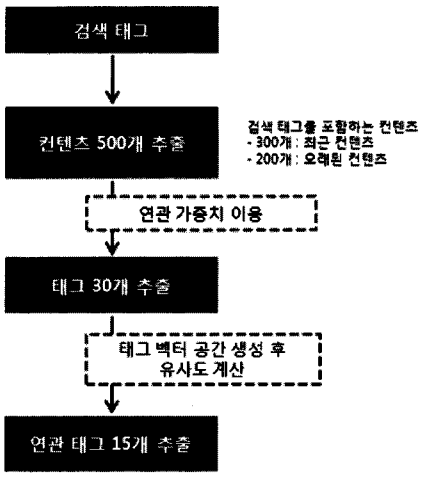
③ 추출된 태그 내에서 각각의 태그와 가장 유사도가 높은 태그를 찾아 군집을 형성 한다.



[그림 4] 태그 서비스의 데이터 모델

3.2 군집 알고리즘

3.2.1 태그 데이터 추출



[그림 5] 태그 데이터 추출 과정

군집화에 사용될 태그 데이터는 [그림 5]와 같은 과정을 통해 추출이 된다. 먼저 컨텐츠에 동시에 작성된 태그들의 연관 가중치를 이용하여 검색한 태그의 연관 태그를 30개를 추출한다. 제 4장의 실험 결과를 통해 연관 가중치 계산 시 사용되는 데이터는 검색 태그를 포함하고 있는 모든 컨텐츠의 태그 데이터가 아닌 컨텐츠 500개에 작성된 태그 데이터를 사용하였다. 컨텐츠 500개는 최신 컨텐

츠 300개(최근 날짜순 정렬)와 오래된 컨텐츠 200개(오래된 날짜순 정렬)로 구성하여 컨텐츠의 태그들이 작성 시간에 치우지지 않도록 하였다. 연관 가중치만을 이용하여 연관 태그를 추출하는 것은 일부 사용자들의 스팸 데이터들로 인해 유사하지 않은 태그들이 나타날 수 있기 때문에, 본 논문에서는 30개의 태그 중 유사도 계산을 통해 상위 15개를 추출하는 방식을 사용하였다.

검색한 태그와 추출된 태그와의 유사도 계산을 하기 위해서 <표 1>과 같은 벡터 공간 모델(Vector Space Model)을 만든다.

〈표 1〉 추출된 연관 태그의 벡터 공간

	Tag 1	Tag 2	Tag 3	Tag 4	…
Tag 1		30	20	10	
Tag 2	30		20	8	
Tag 3	20	20		42	
Tag 4	10	8	42		
…					

위 모델을 가지고 유클리디안 유사도 함수를 이용하여 각각의 태그와 유사도를 계산한 후 상위 15개를 추출하게 된다.

3.2.2 태그 군집

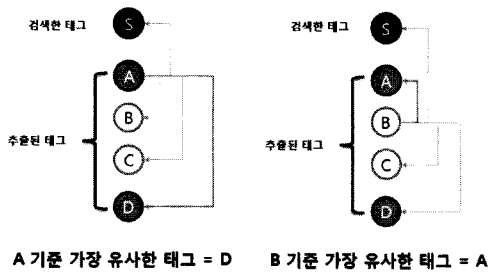
위와 같이 유클리디안 함수를 이용하여 유사도 계산을 통해 얻은 결과를 이용하여 다음과 같은 방식으로 태그군집을 하게 된다.

① [그림 6]에서 보듯이 각각의 태그와 가장 유사한 태그를 찾아서 연결한다. 예를 들어 S로 검색했을 때 추출된 태그가 A~E가 나왔다고 하자. 여기서 A와 가장 유사도가 높은 태그를 찾아 군집을 구성한다.

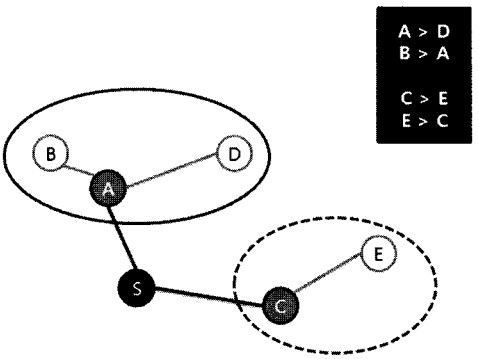
② 다음 B와 가장 유사도가 높은 태그를 찾아 군집을 구성한다. 이런 식으로 각각의 태그에 대해서 유사도가 높은 태그를 찾아 군집을 한다.

③ 유사도를 구한 결과가 A → D(A는 D와 가장 유사), B → A, C → E, E → C가 나왔을 경우 [그림 7]과 같이 2개로 군집이 구성됨을 알 수 있다.

실제 데이터에서는 군집이 2개일 수도 있고, 각각이 유사성이 없어서 독립적인 태그 군집으로 나올 경우 최대 15개의 군집으로 구성될 수 있다. 또한 이러한 군집을 구성할 때의 태그 간 유사 연결 정보는 군집 데이터를 시각화하는데 사용된다.



[그림 6] 가장 유사한 태그 추출

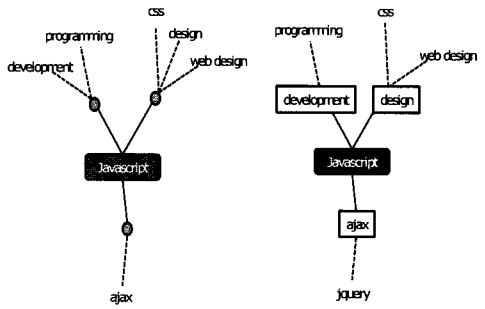


[그림 7] 연관 태그의 군집

3.2.3 시각화

정보의 시각화는 사용자의 UX(User Experience)[15]를 향상시킴으로써 정보에 대한 탐색을 보다 쉽게 할 수 있게 한다. 본 논문의 알고리즘을 통해 생성된 연관 태그의 결과를 시각화하기 가장 적합한 방법으로는 관련 연구에서 살펴본 Visual Thesaurus와 같이 그래프로 표현하는 방법이 있다. 검색한 태그의 연관 태그 데이터에 대한 시각

화는 간단히 보여주는 방법과 자세히 보여주는 방법으로 나눌 수 있는데, 간단히 보여주는 방법은 [그림 8] (a)와 같이 군집 내부의 태그 연관 관계를 제거하고 그래프로 나타내는 방법이다. 또한 자세히 보여주는 방법은 [그림 8] (b)와 같이 군집 내의 태그 연관 관계를 포함하여 보여주는 방법이다. 두 가지 방법 모두 장단점을 가지고 있기 때문에 사용자에게 두 가지 방법 모두 제공하고 사용자의 편의에 맞게 탐색 방법을 선택하여 사용할 수 있게 하였다.



[그림 8] 시각화 방법

4. 실험 및 결과

4.1 실험 개요

실험 데이터는 딜리셔스 RSS의 정보를 이용하여 수집하였으며, 실험 내용은 크게 두 부분으로 나뉜다. 첫 번째 실험은 태그 데이터 추출 시 사용하는 데이터의 크기를 축소하였을 때 원본 모델 데이터에서 추출한 결과와 얼마나 차이가 있는지를 알아보는 실험이다. 두 번째 실험은 유사도 함수에 따른 추출된 연관 태그가 얼마나 정확도를 가지고 있는지 알아보는 실험이다.

4.2 실험 데이터 수집

실험 데이터는 딜리셔스에서 제공하는 RSS로는 한 사용자의 모든 글을 수집할 수 없기 때문에 다음과 같은 절차를 통해 실험 데이터를 수집하였다.

① 1)의 RSS 정보를 이용하여 Popular 태그에서 사용자를 수집한다.

② 2)의 RSS 정보를 이용하여 사용자의 태그를 조사한 후 빈도수 상위 10개를 추출하여 3)의 RSS 정보를 이용하여 태그에 해당하는 사용자 북마크 정보(북마크에 작성된 태그 정보 포함) 4)를 수집한다. 위의 절차를 통해 수집된 실험 데이터는 <표 2>와 같다.

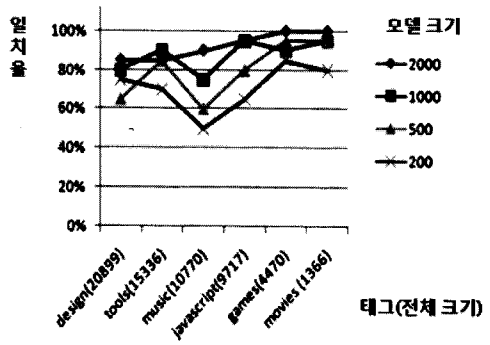
4.3 모델 크기에 따른 태그 추출 결과

많은 데이터에서 연관 태그를 추출하는 것은 시간이 오래 걸릴 수 있기 때문에 원본 모델 크기를 축소하여 추출했을 때 원하는 결과가 나오는지 확인하는 실험을 하였다. 실험은 [그림 9]와 같이 태그의 빈도별로 6개의 태그를 임의로 선정하여, 원본 모델 크기에서 추출한 결과와 비교해서 얼마나 일치하는지를 검사한다. 실험 결과, 모델 데이터 200개의 일치율은 낮았고, 모델 데이터 1000개와 2000개의 일치율은 높았으나, 모델 데이터 크기가 크다는 단점이 있었다. 따라서 본 논문에서는 비교적 높은 일치율을 보이면서 크기가 작은 500개의 모델 데이터를 통해 태그를 추출하였다.

〈표 2〉 수집된 실험 데이터

항 목	수
사용자 수	3,440명
북마크 수	351,866개
작성된 태그 총 수	1,265,096개
고유 태그 수	67,309개
북마크 당 평균 태그 수	약 3.5개
사용자 당 평균 고유 태그 수	약 20개

1) http://feeds.delicious.com/v2/{format}/popular/{tag}.
2) http://feeds.delicious.com/v2/{format}/tags/{user}.
3) http://feeds.delicious.com/v2/{format}/{user}/{tags}.
4) 딜리셔스의 RSS 피드는 피드 하나에 포함될 수 있는 컨텐츠가 15개로 제한되어 있다.

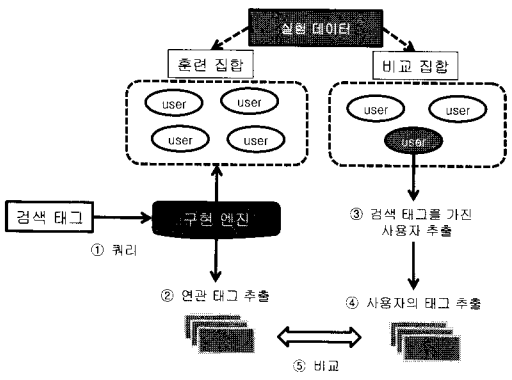


[그림 9] 모델 사이즈별 추출 태그 정확도

4.4 유사도 함수 적용 실험

이 실험은 연관 태그 추출 시 사용하는 유사도 함수에 따라 결과의 정확도가 얼마나 달라지는지 알아보는 실험이다. 실험은 [그림 10]과 같이 훈련 집합과 비교집합으로 나누어서 실험하였다. 먼저 훈련집합을 이용하여 검색 태그([그림 9]의 X축에 나타난 태그 사용)의 연관 태그를 추출한다. 그런 다음 비교집합에서 검색 태그를 가진 사용자의 태그와 훈련집합에서 추출된 연관 태그와 얼마나 일치하는지 검사한다.

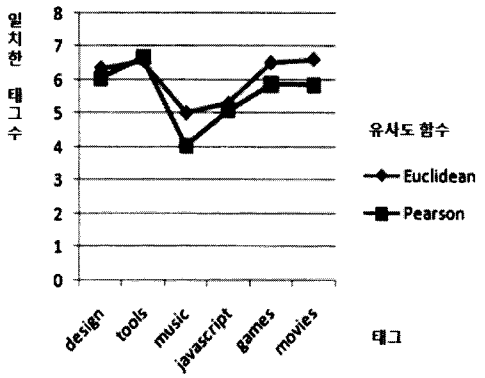
실험 결과는 [그림 11]에서 보듯이 유클리디안 함수를 사용하는 것이 피어슨 상관 계수(Pearson Correlation Coefficient) 함수를 사용하는 것보다 조금 더 좋은 정확률을 보였다.



[그림 10] 유사도 함수 적용 실험 과정

〈표 3〉 유사도 함수 적용 실험 결과

태그/유사도함수	Euclidean	Pearson
design	6.3684	6.0442
tools	6.5907	6.6667
music	5.0219	4.0647
javascript	5.314	5.1091
games	6.5323	5.8766
movies	6.6316	5.8588



[그림 11] 유사도 함수 적용 실험 결과 그래프

4.5 군집의 결과

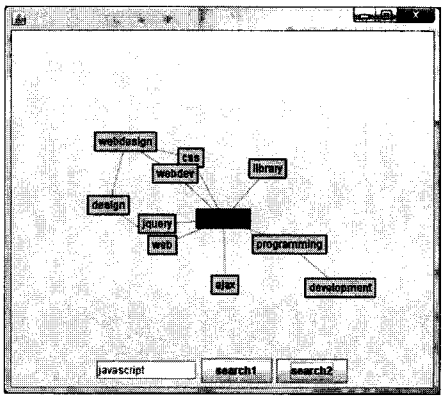
군집화의 결과는 <표 4>에서 보듯이 비교적 유사한 태그들끼리 묶여졌음을 알 수 있다. design 태그는 3개의 군집으로 묶이었고, javascript 태그는 5개, music 태그는 3개로 군집된 것을 볼 수 있다. javascript 태그의 경우 javascript의 api인 {mootools, jquery}, {yui, prototype}등으로 유사한 태그끼리 묶인 것을 확인할 수 있다. [그림 12]는 최종 결과를 시각화하기 위해 자바 언어로 구현된 어플리케이션의 모습이다. 시각화에는 그래프 알고리즘[14]을 사용하였다.

5. 결론 및 향후 과제

본 논문에서는 정보 탐색을 보다 쉽게 하기 위하여 검색한 태그의 연관 태그를 군집화하는 알고

〈표 4〉 군집의 결과

태 그	군 집
design	software, gallery, graphics
	webdesign, css, flash
	layout, html, inspiration, typography …
javascript	framework, reference, howto
	ajax. webdev
	yui, prototype
	dhtml, js. lightbox
	mootools, jquerys
	desigh, flash, html
music	mp3, audio, sound, radio
	downloads, artists, art
	streaming, quitar, reference



[그림 12] 군집화 알고리즘의 구현 결과

리즘을 제시하였다. 또한 유사도 계산을 하기 위한 태그 추출 시 소량의 모델 데이터에서 추출한 태그가 원본 모델 데이터에서 추출한 태그와 80% 정도 일치한다는 점을 실험을 통해 알아보았고, 이 실험 결과를 바탕으로 태그 당 500개의 모델 데이터를 사용하여 유사도 계산을 하였다. 또한 유사도 함수 선택을 위한 실험을 통해 더 정확도가 높은 유클리디안 함수를 이용하여 연관 태그를 군집화하였다. 군집의 결과, 비교적 유사한 태그들끼리 모이는 것을 확인하였다. 그리고 그 결과를 그래프 알고리즘을 이용하여 시각화함으로써 직관적인 정보 탐색의 기능을 제공하였다.

본 논문에서 제공하는 탐색 방법은 블로그 인텔리전스[1] 환경에서 컨텐츠를 탐색하거나, 시맨틱 웹에서의 검색을 구현하기[3] 위한 과도기적 단계에서 정보 탐색으로 활용할 수 있을 것이다.

향후 과제로는 유사도 계산 및 군집의 구성 시 사용된 연산의 복잡성을 줄이면서 보다 정확도가 높은 군집을 구성 할 수 있는 방안을 찾고 설계를 개선하고자 한다.

참 고 문 헌

[1] 김재경, 서유경, "블로그 인텔리전스", 「한국IT서비스학회」, 제7권, 제3호(2008), pp.71-83.

[2] 이강표, 김두남, 김형주, "웹 2.0 환경에서의 태깅 기술 동향", 「정보과학회지」, 제25권, 제10호(2007), pp.36-42.

[3] 한동일, 권혁인, 최호준, "시맨틱 검색 시스템의 구현과 평가에 관한 연구", 「한국IT서비스학회」, 제7권, 제3호(2008), pp.253-269.

[4] Adam Mathes, "Folksonomies-Cooperative Classification and Communication Through Shared Metadata", http://www.adammathes.com/academic/computer-mediatedcommunication/folksonomies.htm/.

[5] A. K. Jain, M. N. Murty, and P. J. Flynn, "Dat a Clustering : A Review", http://www.cs.rutgers.edu/~mlittman/courses/lightai03/jain99data.pdf.

[6] Delicious web site, http://del.icio.us.

[7] E. Arranga, T. Hubbel, A. Lorents, S. Shiflett, and J. Wessler, "Cobol Tools : Overview and Taxonomy", *IEEE Software*, Vol.17, No.2(2000), pp.59-75.

[8] Flickr web site, http://www.flickr.com.

[9] Folksonomy, http://en.wikipedia.org/wiki/Folksonomy.

[10] Scott A. Golder and Bernardo A. Huberman, "The Structure of Collaborative Tagging Systems", http://www.hpl.hp.com/research/idl/papers/tags/tags.pdf.

[11] Similarity Function, http://people.revoledu.com/kardi/tutorial/Similarity/index.htm.

[12] Social network, http://en.wikipedia.org/wiki/Social_network.

[13] Taxonomy, http://en.wikipedia.org/wiki/Tasonomy.

[14] Thomas M. J. Fruchterman and Edward M. Reingold, "Graph Drawing by Force-directed Placement", *Software-Practice and Experience*, Vol.21(1991), pp.1129-1164.

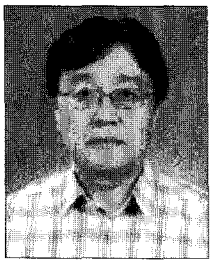
[15] UX(User Experience), http://en.wikipedia.org/wiki/User_experience.

[16] Visualthesaurus web site, http://www.visualthesaurus.com/.

◈ 저 자 소 개 ◈

박 병 재 (bejey79@gmail.com)

국민대학교에서 컴퓨터공학을 전공하고, 현재 국민대학교 컴퓨터공학부 석사 과정에 재학 중이다. 주요 관심분야는 웹 2.0, 시맨틱 웹, 인공 지능,데이터마이닝 등이다.



우 종 우 (cwwoo@kookmin.ac.kr)

Illinois Institute of Technology에서 전산학 박사학위를 취득하고, 1994년부터 현재까지 국민대학교 컴퓨터공학부 교수로 재직 중이다. 주요 관심분야는 웹 2.0, 인공 지능, 지능형 시스템 등이다.