태그 동시 출현의 동적인 특징을 이용한 개선된 태그 클라우드의 태그 선택 방법

(Improved Tag Selection for Tag-cloud using the Dynamic Characteristics of Tag Co-occurrence)

김 두 남 † 이 강 표 † 김 형 주 ††
(Dunam Kim) (Kangpyo Lee) (Hyoung-Joo Kim)

요 약 태깅 시스템은 인터넷 사용자로 하여금 태그라고 불리는 메타데이터를 글, 사진, 동영상 등에 부여하도록 하여서 컨텐츠의 검색 및 브라우징을 편리하게 하는 시스템이다. 컨텐츠의 브라우징을 위해서 태그 클라우드라는 시각적 인터페이스가 널리 쓰이고 있다. 태그 클라우드는 가장 빈도수가 높은 태그들을 알파벳 순으로 보여주고 폰트의 크기로 그 태그들의 빈도수를 반영한다. 하지만 기존의 태그 선택 방법은 몇 가지 단점들이 알려져 있다. 그래서 이 논문은 참신한 컨텐츠들을 찾을 수 있도록 Freshness라는 태그 클라우드를 위한 새로운 태그 선택 방법을 정의하였다. Freshness는 태그 동시 발생 확률 분포(tag co-occurrence probability distribution)가 동적으로 변화하는 것을 Kullback-Leibler divergence로 평균한 값이다. Allblog, Eolin, Technorati 등 세 개의 웹사이트로부터 실제 태그 데이터를 수집하여 우리의 태그 클라우드를 생성하는 시스템, 'Fresh Tag Cloud'를 구축하였다. 이 태그 클라우드를 Allblog에서 수집한 데이터에서 전통적인 태그 클라우드와 비교했을 때 중복평균이 87.5% 감소하여서 성능이 더 향상된 것을 확인할 수 있다.

키워드 : 태깅 시스템, 폭소노미, 태그 클라우드, 태그 동시 출현

Abstract Tagging system is the system that allows internet users to assign new meta-data which is called tag to article, photo, video and etc. for facilitating searching and browsing of web contents. Tag cloud, a visual interface is widely used for browsing tag space. Tag cloud selects the tags with the highest frequency and presents them alphabetically with font size reflecting their popularity. However the conventional tag selection method includes known weaknesses. So, we propose a novel tag selection method Freshness, which helps to find fresh web contents. Freshness is the mean value of Kullback-Leibler divergences between each consecutive change of tag co-occurrence probability distribution. We collected tag data from three web sites, Allblog, Eolin and Technorati and constructed the system, 'Fresh Tag Cloud' which collects tag data and creates our tag cloud. Comparing the experimental results between Fresh Tag Cloud and the conventional one with data from Allblog, our one shows 87.5% less overlapping average, which means Fresh Tag Cloud outperforms the conventional tag cloud.

Key words : tagging system, folksonomy, tag cloud, tag co-occurrence

· 본 연구는 국토해양부 첨단도시개발사업의 연구비지원(07첨단도시 A01), BK-21 정보기술 사업단, 지식 경제부 및 정보통신연구진흥원의 대학 IT연구센터 육성지원사업(IITA-2008-C1090-0801-0031)의 연구 결과로 수행되었음

† 비 회 원 : 서울대학교 컴퓨터공학부
dunam.kim@gmail.com
kplee@idb.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjkim@snu.ac.kr

논문접수 : 2008년 9월 30일
심사완료 : 2009년 4월 13일

Copyright©2009 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.

정보과학회논문지: 컴퓨팅의 실제 및 레터 제15권 제6호(2009.6)

1. 서 론

태깅은 사용자들이 블로그, 웹문서, 사진, 음악, 동영상 등의 웹 리소스에 직접적으로 관련 단어를 등록하여 검색 및 브라우징을 용이하게 하는 웹 2.0의 핵심기술 중의 하나이다. 태그라는 새로운 메타데이터(metadata)의 가장 큰 특징은 비전문가가 자유로이 입력할 수 있다는 점이다.

폭소노미(folksonomy)는 협력 태깅(collaborative tagging), 소셜 태깅(social tagging)과 동의어로 알려져있다. 위키피디아의 정의를 찾아보면 “폭소노미란 태그를 협력적으로 생성하고 관리하는 행위 및 방법”[1]을 의미한다. 협력 태깅을 사용하는 대표적인 서비스로 del.icio.us[2]가 있다. del.icio.us[2]는 소셜 북마킹 사이트인데 URL을 저장할 때 사용자가 관련 태그를 입력할 수 있다. 또한 다른 사람이 같은 URL을 자신의 북마크에 입력할 때 해당 URL에 부착된 태그들 중 가장 많이 사용된 것들을 참고해서 태깅을 할 수 있다.

폭소노미의 옹호론자들은 태깅 시스템의 여러 장점들을 제시한다. 첫째로 태깅 시스템은 분류 시스템(classification system)과 같은 고정된 트리형 계층구조를 취하지 않기 때문에 더욱 유연하게 컨텐츠의 속성을 표현할 수 있다. Golder[3]의 연구에서는 태깅 시스템과 분류 시스템의 차이점을 예시를 들어서 보여주고 있다. 또한 택소노미(taxonomy)와 같이 분류체계를 고려하면서 분류하지 않아도 되므로 사용하기가 훨씬 쉽다. 웹 2.0 사진공유 싸이트 Flickr[4]의 창업자 Butterfield는 “폭소노미는 올바른 택소노미의 90%의 가치를 가지고 있지만 10배는 단순하다.”라고 주장한다[5]. 태깅 시스템은 사용이 쉽기 때문에 소수의 전문가가 아닌 다수의 비전문가에 의해서 관리될 수 있다. 그러므로 인터넷에서 발생하는 대량의 컨텐츠에 적용할 수가 있다. 인터넷에서는 협업적 백과사전 프로젝트인 위키피디아(Wikipedia)[1]처럼 다수의 비전문가에 의해서 이루어지는 작업이 더 적절할 수 있다.

태그의 시각적인 브라우징을 위해서 많은 웹 2.0 서비스들은 태그 클라우드(Tag-Cloud)라 인터페이스를 제공한다. 태그들 중에 사용빈도가 높은 것을 뽑아서 알파벳순으로 정렬하고 그 중에서도 빈도가 높은 것은 큰 폰트로 표시한다. 태그 클라우드는 관련 컨텐츠가 많은 태그를 중심으로 브라우징할 때 유용하다. 최초의 태그 클라우드는 Flickr[4]에서 구현되었다(그림 1).

태그 클라우드는 널리 사용되며 간단하긴 하지만 단점 역시 존재한다. 태그 클라우드는 빈도수에 의해서만 태그를 선택한다. 그러나 문헌 식별값(term discrimination value)을 고려하면 빈도수가 높은 색인어는 문서들을 구별하기에 좋지 않다[6]. 그리고 벡터 스페이스 모델에서는 IDF를 통해서 많은 문서에 사용된 색인어에 대해서 가중치를 적게 부여하게 된다. Begelman[7]등의 연구에서는 매우 인기있는 태그들과 그에 관련된 태그들이 태그 클라우드를 대부분 차지하고 있다고 지적한다. 그리고 이 연구에서는 del.icio.us[2]의 태그 클라우드의 대부분이 웹디자인이나 기술에 관련이 있는 용어들이라고 예를 들고 있다. Xu[8]의 연구에서는 태그 클라우드의 태그 선택에 대해서 개선할 가능성을 제안하고 있다. Hassan-Montero[9]는 태그 클라우드의 새로운 태그 선택 방법 및 표현 방법을 제안하고 있다. 그리고 빈도수가 높은 태그들 중에는 상업적인 스팸 태그들이 다수 포함되어 있기 때문에, 빈도수로 태그를 선택한다면 스팸 태그들이 태그 클라우드에 나타날 수 있다.

비록 태깅 시스템이 본질적으로 동적인 특성을 갖지만, 태깅 시스템은 안정화되며[10] 전통적인 태그 클라우드 역시 안정화된다. 따라서 전통적인 태그 클라우드는 업데이트된 데이터를 기대하는 사용자들의 욕구를 만족시킬 수 없을 것이다. 첫번째로 전통적인 태그 클라우드의 태그들은 자주 바뀌지 않을 것이다. 그래서 Dubinko[11]는 “interestingness”를 정의함으로써 최근 며칠동안 다른 시기보다 많이 사용된 태그들을 선택할

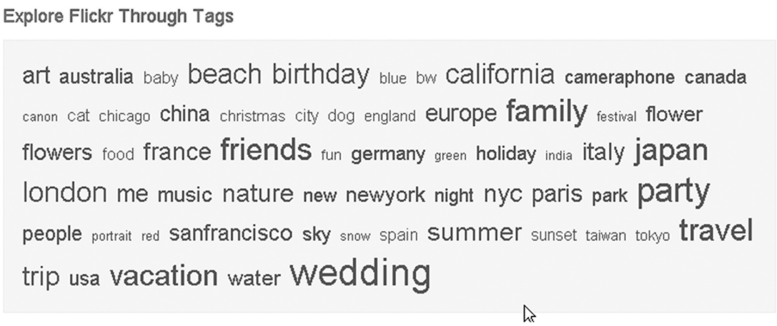

그림 1 Flickr의 태그 클라우드

수 있도록 했다. 두번째로 태그 동시 출현 네트워크가 안정화되기 때문에[10], 새로 만들어진 웹 리소스의 주제들 중 대부분은 과거의 웹 리소스의 주제들과 같을 것이다(3.2에서 태그와 주제의 관계에 대해서 좀 더 논의할 것이다). 이 문제를 해결하기 위해서, 이 논문에서는 태그 동시 출현 네트워크상에서 가장 동적인 부분을 발견함으로써 새로운 주제의 컨텐츠를 찾는 사용자들을 도울 것이다.

우리의 접근 방법은 태그 클라우드의 사용자들의 사용방식을 고려하고 있다. 태그 클라우드의 사용자들은 태그 클라우드의 태그들 중 하나를 선택하고 어떤 컨텐츠들이 나오는지 최근의 것부터 보기 시작할 것이다. 이 때 참신한 컨텐츠를 찾는 사용자들은 태그 클라우드의 태그들끼리 비슷한 컨텐츠를 갖는 것을 원하지 않을 것이다. 또한 태그를 선택한 뒤에 표시되는 컨텐츠들의 내용이 비슷한 것도 원하지 않을 것이다. 이 논문에서는 관련된 최근 컨텐츠의 주제가 급격하게 변하는 태그들을 우리의 태그 선택 방법인 Freshness를 이용하여 찾아서 태그 클라우드의 태그로 선택한다. 태그 동시 출현 확률 분포가 급격히 동적으로 변화하는 태그들이 높은 Freshness 값을 갖게 된다.

2장에서는 사용자 인터페이스와 태그 동시 출현 그리고 태깅 시스템의 동적인 특징에 대한 관련 연구를 살펴볼 것이다. 그리고 3장에서는 우리가 제시하는 태그 클라우드의 태그 선택 방법 Freshness에 대해서 설명한다. 4장은 Freshness를 사용한 태그 클라우드 시스템 Fresh Tag Cloud를 소개한다. 5장에서는 Freshness를 기존의 연구와 비교하여 성능평가를 수행하게 된다.

2. 관련연구

2.1 시각적 인터페이스

Begelman[7]은 태그들을 자동으로 클러스터링하여 다양한 주제들을 보여 줌으로써 태그 클라우드의 단점을 극복하려고 하였다. 태그 클라우드는 인기있는 한두 개의 주제들이 대부분의 태그들을 차지하기 때문에 다양한 주제를 대표하는 클러스터들을 만들어 낼 수 있다면 다양한 주제들을 표현할 수 있을 것이라고 주장한다.

Dubinko[11]의 연구는 “흥미로운” 태그들이 Flickr[4] 서비스에서 어떻게 변화하는 지 시각화하는 것이다. 이 연구에서 정의된 “흥미로움(Interestingness)”라는 태그 선택 방법은 특정한 시간 범위 내에서 특히나 많이 사용된 태그에게 높은 가치를 부여한다. 그리고 임의의 시간범위에 대응하는 태그들의 사용횟수를 빠르게 질의할 수 있도록 효과적인 인덱싱 알고리즘을 제안하고 있다. 공개된 서비스로 Yahoo의 tagline 서비스[12]가 있는데 각 날짜 별로 가장 “흥미로운” 태그들을 플래시 브라우저 플러그인을 이용해서 동적으로 시각화하고 있다.

Hassan-Montero[9]는 태그 클라우드를 개선하기 위한 새로운 태그 선택 방법과 표시 방법을 제안하고 있다. 이 연구에서 태그 유용성(tag usefulness)라는 개념을 통해서 동일한 리소스에 사용되는 다른 태그보다 그 리소스를 더 잘 표현할 수 있고, 다른 태그들보다 많은 리소스를 망라할 수 있는 태그를 찾으려고 했다. 또한 시각적으로는 Begelman[7]의 영향을 받아서 클러스터링된 태그 클라우드를 보여주고 있다.

2.2 폭소노미의 동적인 특징

Golder[3]의 논문에서는 협업적 태깅 시스템인 del.icio.us[2]에서 URL에 연결된 태그들의 상대적 비율의 변화에 대해서 다루고 있다. del.icio.us[2]는 협업적 태깅 시스템이기 때문에 기존에 시스템에 있는 URL을 자신의 북마크로 등록할 때 다른 사람의 태그를 참조해서 태깅을 할 수 있다. 본 논문에서는 여러 사람이 태깅을 해도 일정한 횟수의 북마킹 행위 뒤에는 URL에 관련된 태그들의 상대적인 비율이 일정해진다는 것을 보여주고 있다. 그리고 그 원인은 모방과 공유된 지식이라고 설명하고 있다.

Halpin[10]은 태깅 시스템에 대한 Golder[3]의 연구를 발전시키고 있다. 파워-로(power law)를 따르는 분포를 통해서 협업적 태깅 시스템의 안정성을 보이고 있다. del.icio.us[2]의 데이터를 이용하여 평가를 수행하며, 또한 태깅 시스템의 동적인 특성을 측정하기 위해서 Kullback-Liebler divergence를 사용하고 있다.

태깅 시스템의 동적인 특성(dynamics)에 대한 연구가 주로 협업적인 태깅 시스템에 집중되어 있었다. 비협업적 태깅 시스템에서는 모방을 통한 태깅은 할 수 없지만 공유된 지식에 의한 합의는 나타날 수 있을 것이다. 그리고 Flickr[4]나 Technorati[13], 혹은 블로그 사이트 등의 비협업적인(Non-collaborative) 태깅 시스템이 다수 존재하기 때문에 이런 연구를 비협업적 태깅 시스템에 적용하는 것이 의미가 있을 것이다.

2.3 태그의 동시 출현

Halpin[10]등의 연구에서는 협업적 태깅 시스템의 안정성과 태그 동시 출현 네트워크 간의 관계를 언급하고 있다. 태깅 시스템이 안정화되면 태그 동시 출현 네트워크도 안정화됨을 보여준다. Begelman[7]과 Hassan-Montero[9]는 태그의 동시 출현을 이용하여서 태그들을 클러스터링하였다. 태그들이 같이 쓰이는 횟수가 많을수록 더 관계가 친밀하다. 태그의 동시 출현 횟수를 관계의 척도로 사용하여 비슷한 태그들끼리 같은 클러스터 내에 속하게 된다. Brooks[14]는 태그 동시 출현을 이용하여서 태그의 계층구조를 생성하였다. 그렇게 만들어진 태그들의 계층구조는 사람이 직접 작업한 야후의 택소노

미와 유사점이 발견되었다. Mika[15]는 태그의 동시 출현으로 만들어진 관계와 자신의 새로운 접근방법을 비교하고 있다. 이는 태그 동시 출현으로 개념간의 관계를 추론해 낼 수 있다는 것을 보여준다. Schmitz[16]는 태그 동시 출현의 횟수를 태그들 사이에서 상대적으로 비교함으로써 태그들 간의 상하위 관계를 발견하였다.

3. 태그 선택 방법

3.1 태그 동시 출현

여러 단어가 하나의 문서에서 동시에 사용되는 현상을 동시 출현(co-occurrence)라고 한다. 그리고 자주 같이 사용되는 단어들 사이에는 관계가 있을 것이라고 가정할 수 있다. 예를 들면 Aloha라는 단어는 Hawaii라는 단어와 같이 쓰이는 횟수가 많으며 두 단어는 실제로도 많은 관계가 있다. 단어간의 동시 출현 현상은 문서 검색 시 추가 검색어를 생성하기 위해서 사용되며 기계 번역에서 의미상의 모호성을 해소하기 위해서도 사용된다.

태그의 동시 출현은 여러 개의 태그가 한 개의 사진, URL, 기사 등을 묘사하기 위해서 쓰일 때 일어난다. 단어의 경우와 마찬가지로 같이 사용된 태그들 사이에는 관계가 있을 것이라고 가정해 볼 수 있다.

태그 동시 출현 네트워크는 가중치가 있는 네트워크(weighted network)이다. 각 노드는 태그이고 링크는 두 개의 태그가 같은 웹 리소스에 사용될 때 그려지며 가중치는 두 개의 태그가 동시에 나타나는 웹 리소스의 수에 의해서 주어진다[17]. 태그 동시 출현 네트워크를 시각화해 보면 태그들의 관계에 대해서 조망해 볼 수 있다. 그림 2는 태그 동시 출현 네트워크를 이용하여 태그간의 관계를 시각화한 것이다[10]. 이 때 태그간의 관계는 식 (1)과 같이 정의된다. T_i, T_j는 거리를 측정할 2개의 태그이다. $N(T_i)$는 태그 T_i가 개별적으로 사용된 횟수이고 $N(T_i,T_j)$는 두 개의 태그가 같은 페이지에서 사용된 횟수이다.

$$Dist(T_i,T_j)=\frac{N(T_i,T_j)}{\sqrt{N(T_i)*N(T_j)}} \quad (1)$$

3.2 태그 동시 출현 확률 분포

이 논문에서는 태그의 동시 출현을 태그 별로 모델링하기 위해서 다음과 같은 수식을 정의 한다. 태그 T_i에 동시 출현이 일어날 때 같이 나타날 태그가 임의의 태그 T_j일 확률은 식 (2)와 같이 정의할 수 있다. n은 모든 태그의 수이다. $N(T_i,T_j)$는 태그 T_i와 태그 T_j가 같은 웹 문서에 사용된 횟수이다.

$$CoProb(T_j \mid T_i)=\frac{N(T_i,T_j)}{\sum_{k=1}^{n} N(T_i,T_k)} \quad (2)$$

태그 T_i에 대한 동시 출현 확률 분포는 식 (3)과 같이 정의된다. 이 확률 분포는 태그 T_i와 같이 사용된 태그가 x인 확률을 반환하는 함수이다. 확률 분포를 정의하는 이유는 확률분포들에 대한 Kullback-Leibler divergence를 계산하여 확률분포의 변동을 측정하기 위함이다.

$$CoProbDist(T_i)=(\lambda x, CoProb(x \mid T_i)) \quad (3)$$

그런데 이 확률 분포는 고정된 것이 아니고 태그 T_i에 동시 출현이 일어날 때마다 변화하게 되는데 이런

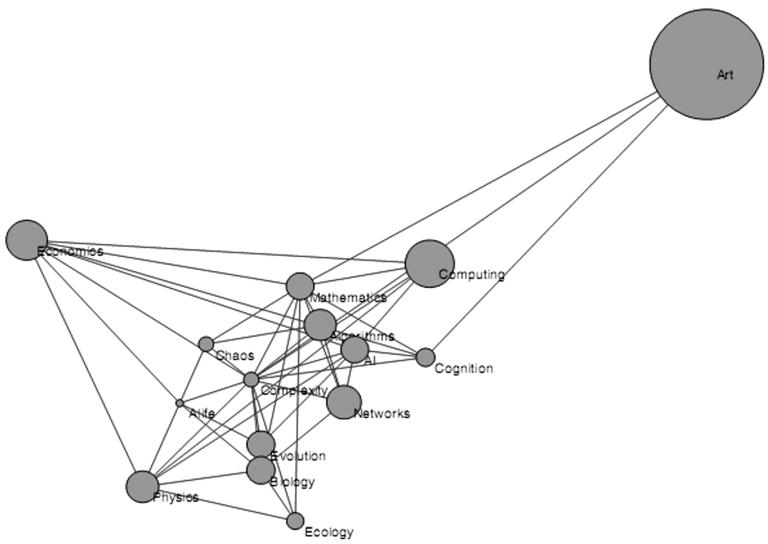

그림 2 태그간 동시 출현 관계의 시각화[10]

변화의 정도는 태그마다 다르다. 이 논문에서 웹 리소스의 주제는 관련된 태그들의 집합으로 대표될 수 있다고 가정하였다. 키워드들의 집합으로 문서를 표현하는 것은 정보검색 분야에서 일반적인 모델링 방법이다[18]. 그리고 Li[19]의 연구는 문서상의 '가장 중요한' 단어들이 태그에서 나타난다는 것을 보여준다. 태그들이 웹 리소스를 대표한다면 주제를 표현한다고 볼 수 있고 따라서 태그 동시 출현 확률분포가 매우 동적으로 변화한다면 해당 태그에 관련된 주제가 안정되어 있지 않고 계속 변화한다고 볼 수 있을 것이다. 우리의 접근방법은 태그 T_i에 대한 확률 분포 CoProbDist(T_i)가 동적으로 변화한다면 태그 T_i는 참신한 컨텐츠에 연결되어 있을 것이라고 추정하여 태그 클라우드에서 사용되도록 선택한다.

3.3 Kullback-Leibler Divergence

우리의 연구에서는 태그의 동시 출현 확률 분포가 얼마나 동적으로 변화하는 지 측정하기 위해서 Kullback-Leibler divergence를 사용한다. Kullback-Leibler divergence는 확률 분포간의 차이를 비교하기 위한 방법이다. Golder[3]가 처음으로 태깅 시스템의 안정성을 다루었지만 그래프의 형태를 통해서 설명했을 뿐 안정성의 측정방법을 제시하지는 못했다. 반면에 Halpin[10]은 협업적 태깅 시스템의 안정성을 이 방법으로 수치화해서 실험적인 증거를 보여준다. Kullback-Leibler divergence는 확률 분포 P, Q 사이의 거리를 측정하기 위한 방법인데 식 (4)와 같이 정의한다.

$$D_{KL}(P \parallel Q) = \sum_{x} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \quad (4)$$

Kullback-Leibler divergence는 항상 음수가 아닌 실수이다. 즉, $D_{KL}(P \parallel Q) \geq 0$이다. 어떤 확률 분포 P, Q에 대해서 P, Q가 동일한 확률분포일 때만 $D_{KL}(P \parallel Q)=0$이 된다. 또한 이 측정방법은 비교환적(non-commutative)이다. 즉, $D_{KL}(P \parallel Q) \neq D_{KL}(Q \parallel P)$

이 논문에서는 CoProbDist(T_i) 확률 분포에 Kullback-Liebler divergence를 적용하였다. 이 확률분포는 고정되어 있는 것이 아니라 태그 T_i가 새로운 웹 리소스에 사용될 때마다 달라진다. 이런 변화가 시간 순으로 일어날 때마다 달라지는 확률 분포의 차이를 구한다.

3.4 Freshness

이제 Freshness를 정의하려고 한다. 하지만 그전에 태그 동시 출현 확률 분포에 시간에 관련된 표현을 추가한다. N(T_i)는 태그 T_i가 사용된 횟수이고 1, 2, …, N(T_i)는 태그가 사용된 시점을 시간 순으로 정렬한 것이다. N(T_i,T_j,a,b)는 a번째 시점부터 b번째 시점까지 태그 T_i와 태그 T_j가 동시 출현한 횟수이다($a \leq b$). 그러므로 $0 \leq N(T_i,T_j,a,b) \leq (a-b+1)$. 태그 동시 출현 확률을 시점들을 고려해서 다시 정의해보면 식 (5)와 같다.

$$\mathrm{CoProb}(T_j \mid T_i, a, b) = \frac{N(T_i, T_j, a, b)}{\sum_{k=1}^{n} N(T_i, T_k, a, b)} \quad (5)$$

그리고 태그 동시 출현 확률 분포를 최근의 50개의 웹 리소스만을 사용하여 구하도록 식 (6)으로 재정의한다(C=50).

$$\mathrm{CoProbDist}_j(T_i) = (\lambda x, \mathrm{CoProb}(x \mid T_i, N(T_i) - C+1, N(T_i) - C+j)) \quad (6)$$

CoProbDist$_j$(T_i)는 태그 T_i가 사용된 최근 C개의 웹 리소스에서 j번째 웹 리소스까지의 태그를 추가했을 때의 태그 동시 출현 확률 분포이다. 최근의 웹 리소스 개수 C를 50으로 한정한 이유는 태그의 동시 출현이 많아질수록 매 시점 사이의 Kullback-Leibler divergence가 0에 가까워지기 때문이다. 그림 3은 Technorati[13]의 데이터를 이용하여 $D_{KL}(\mathrm{CoProbDist}_1(T_i) \| \mathrm{CoProbDist}_2(T_i))$부터 $D_{KL}(\mathrm{CoProbDist}_{49}(T_i) \| \mathrm{CoProbDist}_{50}(T_i))$까지의 값을 시각화한 것이다. 이 그림은 divergence값이 후반으로 갈수록 점차 0에 가까워짐을 보여준다.

Freshness는 특정 태그가 사용된 최근 웹 리소스에 대해서 각 시점 사이의 태그 동시 출현 확률 분포의 차이를 Kullback-Leibler divergence로 구하여 평균한 값이다. 이 논문에서는 Freshness를 정의함으로써 특정한 태그에 관련된 주제들이 얼마나 동적으로 변화하는지 측정한다. 그런데 초반에는 태그 데이터의 양이 적기 때문에 확률분포의 변동폭이 커서 Kullback-Leibler divergence값들이 후반에 비해서 크게 측정되는 경향이 있다(그림 3). 따라서 초반의 divergence 값 5개는 Freshness의 연산에서 제외시켰다. 그래서 Freshness는 식 (7)과 같이 정의된다.

$$\mathrm{Freshness}(T_i) = \frac{\sum_{j=6}^{C-1} D_{KL}(\mathrm{CoProbDist}_j(T_i) \parallel \mathrm{CoProbDist}_{j+1}(T_i))}{C-6} \quad (7)$$

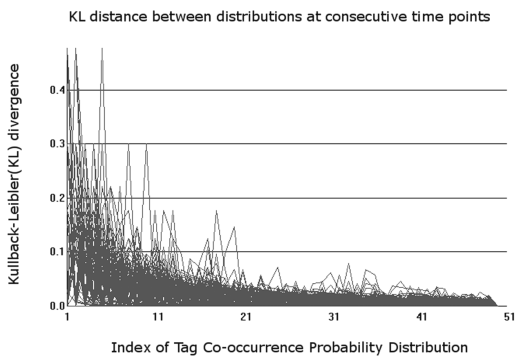

그림 3 연속된 태그 동시 출현 확률 분포 사이의 Kullback-Leibler divergence

이 논문에서는 변화의 정도가 큰 태그들은 참신한 컨텐츠를 가지고 있다고 추정하여 태그 클라우드에 사용되도록 선택한다. 반면에 전통적인 태그 클라우드는 빈도수가 높은 태그를 선택하여 전체 태그 공간을 대표하는 태그들을 선택하려고 하였다. 그런데 빈도수를 이용한 태그 선택 방법은 태그 공간의 모든 주제들을 담지 못하기 때문에[7] 대표성에 의문이 있다. 또한 사용자들이 생산하는 컨텐츠를 다루는 대부분의 서비스에서 최신의 컨텐츠를 강조하여 제공하고 있기 때문에 대표성이 태그 클라우드의 가장 중요한 요소가 아닐 수 있다. 사용자 생산 컨텐츠는 전통적인 뉴스 미디어와는 다르기 때문에 전통적으로 잘 다루어지지 않는 주제들이 나타날 경우가 많은데 Freshness를 사용하면 새로운 주제를 가진 참신한 컨텐츠들을 찾을 수 있다.

4. Fresh Tag Cloud

4.1 Fresh Tag Cloud의 설명

새로운 태그 선택 방법인 Freshness를 사용하여 다양한 태깅 시스템들의 태그 클라우드를 서비스하는 시스템 'Fresh Tag Cloud'를 구성한다. 그림 4는 Fresh Tag-Cloud 시스템의 전체적인 개략도이다. 첫 번째로 각종 웹2.0 서비스들로부터 태그들을 수집해온다. 각 서비스들은 HTML이나 RSS등의 형태로 최근 태그의 정보들을 알려준다. 여기서는 Allblog[20], Eolin[21], Technorati[13] 등의 웹 싸이트 3곳을 대상으로 하였다. 태깅 서비스에서 수집한 태그 데이터는 RDBMS에 단순한 형태로 저장된다. 그리고 50번 이상 사용된 태그들의 Freshness를 계산하여 RDBMS에 결과를 저장한다. 마지막으로 사용자들에게 시각화하기 위하여 Freshness가 높은 태그들을 선택하여 태그 클라우드를 만들어 낸다.

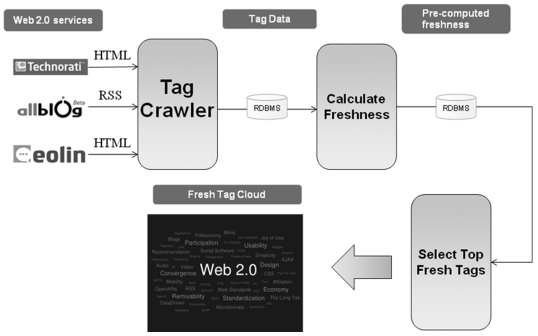

그림 4 Fresh Tag Cloud의 개관

4.2 예시

Fresh Tag Cloud는 Allblog[20], Eloin[21] 그리고 Technorati[13] 세 개의 웹사이트의 태그 클라우드를 보여준다. 그림 5, 6은 Eolin[21]에서 수집한 데이터를 사용한 태그 클라우드들이다. 그림 5는 빈도수만을 이용하여 태그를 선택한 전통적인 태그 클라우드이고 그림 6은 Freshness를 이용하여 선택한 우리의 태그 클라우드이다. 전통적인 태그 클라우드는 '포장이사', '대출', 'IS동영상'등의 홍보목적의 스팸 태그들이 있지만 우리의 태그 클라우드에서는 그러한 스팸 태그들이 나타나지 않는다. 이런 현상은 이런 스팸 태그들이 다수의 블로그 기사에서 사용되지만 같이 사용되는 태그들이 없거나 거의 고정되어 있기 때문이다. Fresh Tag Cloud에는 전통적인 태그 클라우드에 나타나지 않았던 '나', '생일', '다이어리', '육아' 등의 개인적인 기록에 대한 태그들이 다소 나타난다.

Popular Tag Cloud

CMA IS동영상 IT Video onemore photo UCC
강민경 게임 결혼 광고 구글 궁합 남대문 뉴스
다이어트 대출 동영상 레몬박스 리뷰 만화
발렌타인데이 블로거 블로그 사랑 사진 소녀시대
숭례문 실시간뉴스 애니 애니메이션 여행
연예뉴스 영어 영화 음악 이명박 이사
이사비용 이삿짐 인테리어 일본 일상 자동차 잡담 주식
책 티스토리 포장이사 프로그램

그림 5 전통적인 태그 클라우드

Fresh Tag Cloud

400D DVD OST 감기 강아지 건강 겨울 꿈
나 나무 네이버 노트북 눈 다이어리 대한민국 디자인
모바일 뮤직비디오 미국 발렌타인데이 보안 빅뱅
사랑 생각 생일 스트레스 시 신용카드 아토피 영국
영화 유머 유틸리티 육아 음악 이효리
인터뷰 일본어 일상 잡담 좋은글 지하철 추억
친구 커피 컴퓨터 테스트 하늘 행복 휴대폰

그림 6 우리의 Fresh Tag Cloud

5. 성능평가

5.1 실험 데이터

Allblog[20], Eolin[21], Technorati[13] 3개의 메타 블로그 사이트의 데이터를 수집하여 실험에 사용하였다. 메타 블로그 사이트란 개인이 운영하는 여러 블로그의 글들을 하나의 사이트에서 모아서 보여주는 서비스이다. 이 웹 사이트들은 크러울러(crawler)를 이용하여 등록된 블로그 사이트로부터 웹 문서를 수집한다. 국내에는 Allblog[20], Eolin[21], Blogkorea[22] 등의 서비스가 있으

표 1 실험 데이터

	allblog	eolin	technorati
수집기간	2008-02-18 ~ 2008-03-26	2008-02-08 ~ 2008-03-11	2007-10-29 ~ 2007-12-17
문서 수	383,789	161,257	43,975
태그 수	731,553	335,101	96,329

며 외국에는 Technorati[22]라는 서비스가 있다. 실험데이터의 수집기간 및 개수는 표 1에 기술된 바와 같다.

5.2 평가방법

태그 클라우드의 선택 방법을 평가하기 위해서 Hassan-Montero[9]의 연구에서 사용된 세가지 측정방법을 채택하였다.

- 범위(coverage) - 태그 클라우드로 연결되어 있는 웹 문서의 개수다. 태그 클라우드의 태그 중에 최소한 1개 이상의 태그를 포함한 문서의 개수이다.
- 중복 평균(overlapping average) - 태그 클라우드의 모든 태그 쌍이 평균적으로 가지는 중복된 문서의 개수
- 중복 표준편차(overlapping standard deviation) - 태그 클라우드의 모든 태그 쌍의 중복된 문서 수의 표준 편차

그리고 제약 조건으로 태그 클라우드의 각 태그가 사용된 최근의 문서만을 10에서 50까지 개수로 사용하였다. 그 이유는 태그를 클릭했을 때 기사들이 최근에 입력된 순으로 나타나기 때문이다. 또한 Fresh Tag Cloud의 태그들은 50개 이상의 기사가 연결되지 않을 수 있기 때문에 그 이상을 사용할 경우에 올바른 비교가 안될 수 있다.

5.3 태그 유용성(tag usefulness)

태그 유용성(tag usefulness)는 Hassan-Montero[9]의 연구에서 정의된 태그 클라우드의 개선된 태그 선택 방법이다. 이 논문에서는 이 방법을 Freshness와 비교하는 성능평가를 수행한다. 태그 유용성은 Salton[6]의 연구를 폭소노미에 적용한 것으로 벡터 공간 모델의 중심(centroid)을 찾는 수식을 변형한 것으로 다음 식 (8)과 같다.

$$F(T_j)=\sum_{i=1}^{i=n}\frac{\log(d_{ij})}{{m_i}^2} \quad (8)$$

n은 전체 문서의 수이고 m_i는 문서 D_i에 사용된 다른 태그들의 수이다. d_{ij}는 문서 D_i를 묘사하기 위해서 태그 T_j를 사용한 횟수이다. 본래의 수식은 협업적 태깅 시스템에만 적용이 가능하다. 따라서 우리의 태그 선택 방법과 비교하기 위해서 수식을 다음의 식 (9)로 변경하였다.

$$F(T_j)=\sum_{i=1}^{i=n}\frac{d_{ij}}{m_i^2} \quad (9)$$

5.4 실험결과

5.2에서 소개된 세가지 측정방법들을 이용하여 빈도수에 기반한 전통적인 태그 클라우드 선택 방법과 'freshness'를 사용하는 우리의 선택 방법, 그리고 Hassan-Montero[9]의 태그 클라우드 선택 방법인 태그 유용성을 비교하였다.

그림 7,8,9는 수집된 데이터가 가장 많은 Allblog[20]의 데이터를 사용한 결과이다. 표 2,3,4는 우리의 방법과 이전의 방법들을 비교한 결과인데, 10부터 50까지의 각 제약에 대한 평균값을 보여준다. 기존 방법들의 중복표준 편차가 높은 것은 스팸 태그들이 빈도수 높게 같이 사용되는 경향이 있기 때문인 것으로 보인다. 실험 결과는 Fresh Tag Cloud는 다른 태그 클라우드들에 비해서 태그들 사이에 중복된 문서가 훨씬 적다는 것을 보

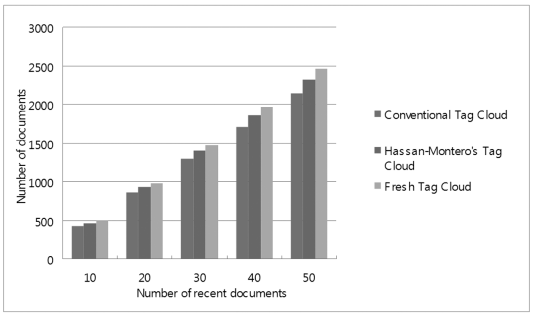

그림 7 Allblog의 데이터를 사용한 범위(coverage)

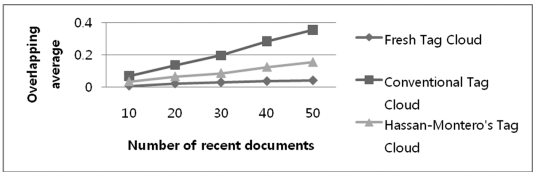

그림 8 Allblog의 데이터를 사용한 중복 평균 (overlapping average)

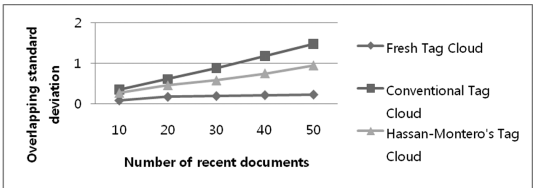

그림 9 Allblog의 데이터를 사용한 중복 표준편차 (overlapping standard deviation)

표 2 기존 방법 대비 평균적인 성능향상(Allblog)

Comparison	Coverage	Overlapping Average	Overlapping Standard Deviation
Conventional tag cloud	14.5% 향상	-87.5% 감소	-79.1% 감소
Hassan-Montero's tag cloud	5.5% 향상	-72.7% 감소	-69.7% 감소

표 3 기존 방법 대비 평균적인 성능향상(Technorati)

Comparison	Coverage	Overlapping Average	Overlapping Standard Deviation
Conventional tag cloud	6.9% 향상	-68.4% 감소	-72.0% 감소
Hassan-Montero's tag cloud	4.1% 향상	-56.7% 감소	-69.4% 감소

표 4 기본 방법 대비 평균적인 성능향상(Eolin)

Comparison	Coverage	Overlapping Average	Overlapping Standard Deviation
Conventional tag cloud	19.2% 향상	-91.5% 감소	-91.4% 감소
Hassan-Montero's tag cloud	8.2% 향상	-78.0% 감소	-84.4% 감소

여준다. 이것은 Fresh Tag Cloud의 태그들이 기존의 방법들을 이용해서 선택된 태그들 보다 더 나은 문헌 식별자(term discriminator)[6]임을 보여준다.

6. 결론 및 향후연구

우리의 연구는 빈도수만을 사용하여 태그를 선택하는 기존의 태그 클라우드 선택 방법을 개선하였다. Freshness라는 개념을 정의함으로써 안정화된 태그 동시 출현 네트워크에서 동시 출현 태그의 비율이 가장 동적으로 변하는 태그들을 선택하였다. 그리고 실험결과 Freshness가 가장 높은 태그들은, Allblog에서 수집한 데이터로 실험을 하였을 때, 전통적인 태그 클라우드보다 평균 중복이 87.5% 감소하였으며 Hassan-Montero의 개선된 태그 선택 방법인 태그 유용성으로 선택한 것보다 평균 중복이 72.7% 감소하였다(표 2). 그리고 스팸 태그들이 태그 클라우드에서 나타나지 않도록 하는 효과도 있었다.

이러한 우수한 결과는 문헌 식별값(discrimination value)[6]와 관계가 있을지도 모른다. 향후에는 이와 관련된 연구를 통해서 더 우수한 태그 클라우드 선택 방법을 연구할 것이다. 또한 태그 클라우드 이외에 태그 자동생성 분야에도 적용할 수 있을 것으로 기대하고 있다.

참 고 문 헌

[1] *Wikipedia*.
[2] *del.icio.us*.
[3] Golder, S.A. and B.A. Huberman, *Usage patterns of collaborative tagging systems*. Journal of Information Science. 32(2): p. 198, 2006.
[4] *Flickr*.
[5] Butterfield, S. (2004) *Sylloge*.
[6] Salton, G., A. Wong, and C.S. Yang, *A vector space model for automatic indexing*. Communications of the ACM. 18(11): pp. 613-620, 1975.
[7] Begelman, G., P. Keller, and F. Smadja, *Automated Tag Clustering: Improving search and exploration in the tag space*, in *WWW2006*. 2006: Edinburgh, UK.
[8] Xu, Z., et al., *Towards the semantic web: Collaborative tag suggestions*, in *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland, May*. 2006.
[9] Hassan-Montero, Y. and V. Herrero-Solana, *Improving Tag-Clouds as Visual Information Retrieval Interfaces*, in *Proceedings of International Conference on Multidisciplinary Information Sciences and Technologies.(Oct 2006)*. 2006.
[10] Halpin, H., V. Robu, and H. Shepherd. *The complex dynamics of collaborative tagging*. in *Proceedings of the 16th international conference on World Wide Web*. 2007.
[11] Dubinko, M., et al. *Visualizing tags over time*. in *Proceedings of the 15th international conference on World Wide Web*. 2006.
[12] *TagLines*.
[13] *Technorati*.
[14] Brooks, C.H. and N. Montanez. *Improved annotation of the blogosphere via autotagging and hierarchical clustering*. in *Proceedings of the 15th international conference on World Wide Web*. 2006.
[15] Mika, P., *Ontologies are us: A unified model of social networks and semantics*. Web Semantics: Science, Services and Agents on the World Wide Web. 5(1): pp. 5-15, 2007.
[16] Schmitz, P., *Inducing ontology from Flickr tags*, in *15th WWW Conference, Edinburgh*. 2006.
[17] Cattuto, C., et al., *Network properties of folksonomies*. AI Communications. 20(4): p. 245-262, 2007.

[18] Baeza-Yates, R. and B. Ribeiro-Neto, *Modern information retrieval*: Addison-Wesley Harlow, England. 1999.
[19] Li, X., L. Guo, and Y. Zhao, *Tag-based Social Interest Discovery*, in *Proceedings of the 17th International World Wide Web Conference*. 2008. pp. 675-684.
[20] *Allblog*.
[21] *Eolin*.
[22] *BlogKorea*.

김 두 남
2006년 서울대학교 언어학과 학사. 2008년 서울대학교 컴퓨터공학부 석사. 관심분야는 웹 2.0, 태깅, 집단지성, 데이터베이스, 정보검색

이 강 표
2004년 연세대학교 컴퓨터과학과(학사) 2006년 서울대학교 컴퓨터공학부(석사) 2006년~현재 서울대학교 컴퓨터공학부 박사과정 재학중. 관심분야는 데이타베이스, 웹 2.0, 시맨틱웹

김 형 주
1982년 서울대학교 전산학과(학사). 1985년 Univ. of Texas at Austin(석사) 1988년 Univ. of Texas at Austin(박사). 1988년~1988년 Univ. of Texas at Austin(Post-Doc). 1988~1990년 Georgia Institute of Technology(부교수). 1991년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이타베이스, XML, 생물정보학, 시멘틱웹, 웹 2.0