제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

TagPlus:Folksonomy에서 동의어 태그를 이용한 검색 기법

이선숙*, 용환승**
이화여자대학교 컴퓨터학과
e-mail:caffuchino@ewhain.net

TagPlus:A Retrieval Method using Synonym Tag in Folksonomy

Sun-Sook Lee*, Hwan-Seung Yong**
Dept. of Computer Science and Engineering, Ewha Womans University

요 약

Web 2.0은 사용자가 온라인에서 정보를 구축하고 공유하게 하는 World Wide Web 상의 2세대 서비스이다. Web 2.0 페이지에서는 사용자가 키워드인 태그를 이용하여 웹페이지나 인터넷상의 오브젝트를 분류하게 되는데 이를 Folksonomy 라 한다. Folksonomy 는 정보를 공동 작업으로 모으고 공유하는 사용자들에 의한 분류기법으로 Taxonomy 와는 다르며 장단점을 지니고 있다. 이 논문에서는 Folksonomy 의 단점중의 하나인 동의어 처리를 위하여 Wordnet 을 접목시킨 검색 기법을 제안한다.

1. 서론

Folksonomy[1]는 인터넷 사용자가 웹페이지, 온라인 사진, 웹링크 등의 콘텐트를 특별한 제한 없이 협력하여 분류하도록 하는 분류시스템이다. 사용자들에 의해 자유롭게 선택된 콘텐트의 식별 키워드, 카테고리 이름, 또는 메타 데이터를 태그[4]라 하고 태그를 이용하여 분류하는 작업을 태깅[4]이라 한다. 사용자는 자신만의 유일한 태그를 사용하여 웹페이지나 이미지등에 태깅한다. 이러한 웹페이지나 이미지는 자신을 식별하도록 하는 복수개의 태그를 가질 수 있다. 동일한 태그를 가진 이미지나 웹페이지는 함께 링크되며 사용자는 유사한 웹페이지나 이미지의 검색을 위하여 그 태그를 이용할 수 있다.

Folksonomy 에서 사용자는 누가 주어진 Folksonomy 태그를 만들었는지와 그 태그를 만든 사용자의 다른 태그들도 알 수 있다.

*본 연구는 한국과학재단 목적기초연구(R01-2006-000-10609-0) 지원으로 수행되었음

이렇게 Folksonomy 사용자들은 자신과 비슷하게 사고하는 다른 사용자의 태그집합을 알 수 있다. 결과적으로 사용자는 관련된 콘텐트를 용이하게 검색할 수 있다.

Folksonomy 는 태그를 이용하는 웹기반의 커뮤니티에서 나타난다. 이러한 커뮤니티에서는 사용자가 사진과 같은 사용자 제작 콘텐트를 분류하고 공유하도록 한다. 또한, 웹사이트, 책, 과학이나 학문 연구, 블로그 등의 기존의 콘텐트를 사용자들이 협력하여 분류하게도 한다.

전문적으로 발달된 통제된 어휘, Taxonomy[4] 와는 달리 Falksonomy 는 인터넷상의 오브젝트는 어떻게 분류되어져야 만 한다는 구체적인 정책이 없으므로 무질서하고 부정확하다. 어떠한 태그가 엉뚱한 웹사이트로 링크 될 수도 있다. 하지만 태그는 특정 권한을 가진 집단에 의한 것이 아니므로 태그에 의한 분류는 개인화되어 있다. 따라서 카테고리는 시간이 지남에 따라 많은 사람들에 의해 자연스럽게 구축되어진다. 이러한 이유로 사용자들은 자신의 관심분야의 태그들로 인터넷상의 오브젝트에 쉽고도

다양하게 접근할 수 있으며 자신과 비슷한 특성을 지닌 다른 사용자들과의 교류 또한 만들어진다. 이것은 Folksonomy 가 가진 가장 강한 강점이라고 할 수 있다.

통제적으로 관리 되지 못한 태그들로 인해 Folksonomy 에도 제한과 단점이 존재한다. 다양한 사용자들이 각기 다른 방식으로 태깅함으로써 Folksonomy 에서의 태그들은 자연적으로 애매모호성을 띄게 된다. 다의어가 그중 하나이다. 예를 들어, window 는 창문을 의미할 수도 있고 유리나 컴퓨터에서의 창을 의미할 수도 있다.

두문자어(Acronyms)도 애매모호성을 보여주는 또 다른 영역이다. 사회학 범위의 ANT(Actor network Theory)와 자바 컴퓨터 프로그래밍의 프로젝트 빌딩 툴인 ANT(Apache Ant)는 완전히 다른 영역이지만 같은 태그로 함께 나타날 수 있다.

많은 Folksonomy 사이트들이 한 단어의 메타 데이터만을 허용한다. 대표적인 Folksonomy 사이트인 Delicious(http://www.del.icio.us)는 태그에 스페이스를 허용하지 않는다. 이는 수많은 쓸모없는 복합단어들을 만들어낸다. 스페이스 없이 태그 하나에 여러 개의 단어를 한 번에 표현하고자 하기 때문이다.

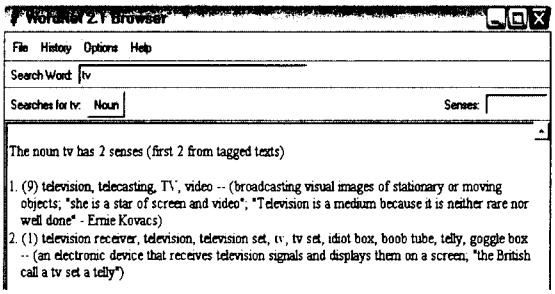
대다수의 태그들은 개인적인 용도로 이용되므로 친구들이나, 함께 일하는 그룹들 사이에만 통용되는 특수한 태그나 무의미한 태그들, 지극히 개인적이라서 다른 사용자들에게 공유되어질 확률이 없는 태그들이 존재하게 된다.

동의어 제어 또한 되고 있지 않다. "Apple macintosh computer" 와 관련된 정보에 "mac", "apple", "macintosh" 등 유사한 의미를 갖는 다른 태그들이 이용되어질 수 있다. 유사하지만 다른 형식을 취하는 단어들 또한 애매모호성을 증가시키는데 일조하는데 예를 들면 "tv" 와 "televesion", "Holland"와 "Dutch", 영어의 단수 복수 "flower" 와 "flowers" 등이 그것이다. 이들은 같은 의미를 지님에도 불구하고 서로 다른 태그로 인식되어 진다. 이렇듯 통제되지 않으면서 무질서한 태깅 용어들의 집합은 효과적인 검색을 보장하지 않는다.

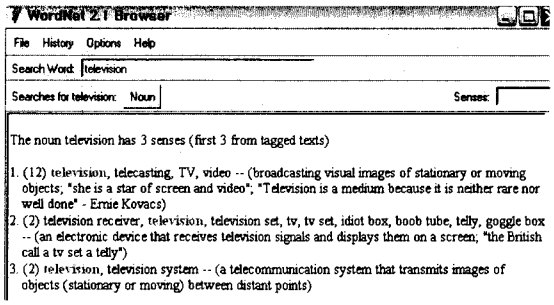
본 고에서는 2장에서 Folksonomy 단점중의 하나인 동의어 사용으로 인한 애매모호성을 감소시키기 위해 동의어 태그를 이용한 검색 모델(이하 TagPlus 라고 함)을 제안한다. 그리고 3장에서 제안한 모델을 기반으로 한 Flickr[7] 이미지 검색 시스템(이하 TagPlus 시스템이라고 함)을 소개하고 Flickr 사이트의 이미지 검색 결과와 비교하며, 마지막으로 4장에서 결론 및 추후 연구 계획을 제시한다.

2. 동의어 태그를 이용한 검색 모델

본장에서는 먼저 동의어의 처리를 위한 모델인 TagPlus에 대해서 설명한다. 형식상 다른 표현이지만 같은 의미를 가지는 동의어임을 우리가 알 고 있듯이 태깅 시스템에서도 이를 인지한다면 동의어가 가진 문제점은 해결될 수 있다. 이를 위하여 Wordnet[8] 의 동의어 그룹을 채택한다. Wordnet 은 영어 단어를 동의어 셋으로 그룹지어 단어의 의미뿐만 아니라 단어 사이의 관계를 기록해주는 사전이다. (그림 1)과 (그림 2)는 Wordnet 사전의 'tv'와 'television' 단어에 대한 사전 내용이다.



(그림 1) 'tv' 의 Wordnet 사전 데이터



(그림 2) 'television' 의 Wordnet 사전 데이터

(그림 1) 'tv' 의 두 번째와 (그림 2) 'television' 의 두 번째 내용은 텔레비전 신호를 받아 화면에 보여주는 전자 장치의 의미를 가지는 동의어이다. 이렇듯 'tv' 와 'television' 은 같은 동의어 그룹을 가짐을 Wordnet 데이터 상에서 확인할 수 있다.

TagPlus 의 검색 모델에서 사용자가 태깅시 오브젝트에 추가되는 태그는 기존의 키워드와 같은 하나의 단어가 아닌, <태그, 동의어 그룹> 과 같은 확장된 태그가 추가되어진다. 예를 들어, 사용자가 (그림

2)의 두 번째 의미를 갖는 'television'을 태그로 사용하게 되면 태깅은 다음과 같은 형식을 취하게 된다.

tagging(Object, TagPlus<tag, synonym group id>)

사전에 존재하지 않는 고유 명사나 개인적인 어휘를 태그로 이용할 경우도 고려해야 한다. 이를 위해선 사전을 이용한 태그인지 아닌지의 구분을 태깅에 포함할 수 있다. 'D' 는 사전을 통한 태깅, 'N' 은 사전을 이용하지 않은 태깅으로 구분하여 아래와 같은 형식을 취할 수 있다.

tagging(Object, 'D', TagPlus<'television', 301254>)
tagging(Object, 'N', 'television')

사전을 통하지 않는 태깅은 TagPlus 대신 기존의 태깅 방식을 이용한다.

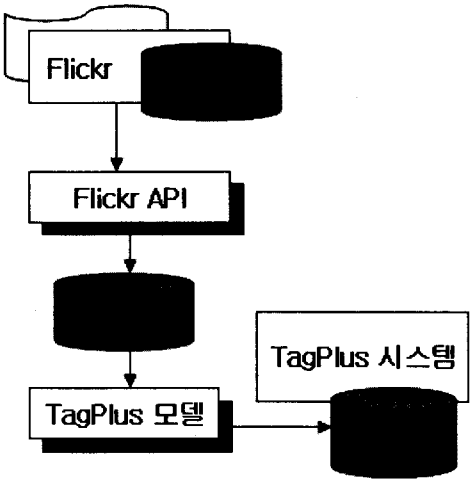
사용자가 검색을 할 경우에도 사전 데이터를 통해 선택된 어휘와 동의어 그룹 아이디로 TagPlus를 구성하여 검색한다. (그림 1)의 'tv' 사전 내용 중 사용자가 "2번" 의 의미를 선택하였을 경우 단순히 'tv' 가 아닌 <tv, 301254> 의 TagPlus로 오브젝트를 검색한다. '301254' 는 'tv'의 동의어 그룹 아이디이다. TagPlus에서는 'tv'와 'television' 은 같은 동의어 그룹 아이디 '301254'를 가지고 있으므로 두 단어 중 어느 한 개로 검색하더라도 두 단어 모두를 태그로 이용하여 따로 검색 결과를 얻었던 기존의 태깅 시스템에서와 같은 결과를 얻을 수 있다.

3. Tagplus 시스템:동의어 태그를 이용한 Flickr 이미지 검색 시스템

3.1 시스템 개요

TagPlus 시스템의 프로세스는 (그림 3)과 같다. 단계별 내용은 다음과 같다.

① Flickr 사이트의 이미지와 태그정보를 Flickr API를 이용하여 가져온다.
② Wordnet의 동의어 셋을 Flickr 태그에 적용하여 TagPlus 모델을 생성한다.
③ Flickr의 이미지 주소와 TagPlus를 TagPlus 시스템의 로컬 DB 에 저장한다.
④ TagPlus 시스템에서 TagPlus를 이용하여 검색한다.



(그림 3) TagPlus 시스템 프로세스

3.2 시스템 평가

Folksonomy 의 대표적인 사이트로 알려진 이미지 검색 사이트 Flickr 를 평가 대상으로 하여 TagPlus 모델을 적용한 TagPlus 시스템의 태그와 그 동의어에 의한 이미지 검색 결과를 평가한다.

<표 1> 실험에서 사용된 태그와 검색건수

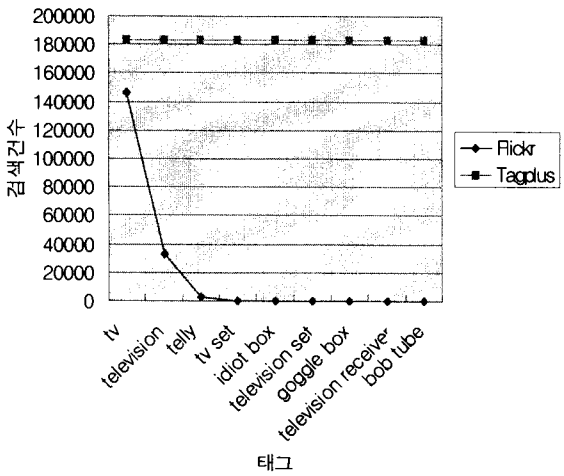
No.	태그명	검색건수
1	tv	146284
2	television	33442
3	telly	2456
4	tvset	211
5	idiot box	188
6	television set	26
7	goggle box	4
8	television receiver	0
9	bob tube	0

<표 1>은 Flickr 에서 'tv' 태그와 그의 동의어로 검색한 결과이다. Wordnet 에는 'tv'의 동의어가 <표 1> 에서 보여주듯이 9개로 정의 되어 있다.

(그림 4)는 TagPlus 시스템과 Flickr 사이트에서 'tv' 나 그의 8개 동의어로 이미지를 검색하였을 경우의 검색 결과를 보여준다. Flickr 사이트에서는 각 태그로 태깅되어진 이미지만을 검색 결과로 보여준다. 따라서 'tv'이외의 'tv'의 동의어로 태깅되어진 이미지는 따로 검색하는 수 밖에 없다. 게다가 동의

어를 잘 알지 못하는 경우에는 그러한 이미지를 볼 기회조차 없어지게 된다. 그러나 TagPlus 시스템에서는 'tv' 뿐만 아니라 'tv'의 동의어중 어느 단어를 태그로 사용하던지 182611건의 검색결과를 보여준다. 이는 Flickr 사이트에서 'tv'와 그 동의어로 각각 검색한 이미지들의 총 갯수이다. TagPlus를 이용하여 'tv'와 같은 동의어 그룹 아이디를 가지는 단어 전부를 태그로 하여 검색을 한 결과를 보여주기 때문이다. 따라서 사용자의 다양한 표현방식에 따라 다른 단어를 사용하여 태깅되었지만 같은 의미를 담은 태그를 가진 이미지의 검색도 가능하다.

Flickr 과 TagPlus 시스템의 검색 비교



4 결론

태그 사용자들은 지역적으로 문화적으로 다양하다. Folksonomy 식 접근의 강점은 개방성에 있다. 사용자는 원하는 대로 스스로의 기준에서 리소스를 묘사할 수 있다. 과연 태깅 시스템에서 다양한 사용자들 간에 태그 사용에 합의를 이끌어 내는 것이 바람직한 것인지, 합의에는 도달할 수는 있는지 등에 대한 답은 쉽지만은 않다. 태그들을 정리하거나 제약을 가함으로써 Folksonomy 의 본질과 매력을 잃을 수도 있다. 또한 개인적인 메타 데이터로부터 얻어지는 풍부한 데이터의 축소를 가져올 수도 있지만 메타데이터의 노이즈를 감소시킴으로써 검색의 정확도와 효율성을 향상시킬 수 있을 것이다.

본 고에서 제안한 동의어 처리 기법인 TagPlus 모델은 단순히 무질서한 태그를 정리하는 것이 목표가 아니다. TagPlus 모델은 현재의 Folksonomy 환경의 무질서한 태그에 효율적인 검색을 위하여 Wordnet 의 동의어 관계를 적용시켰다. TagPlus 시스템은 TagPlus 모델을 적용하여 동의어 그룹에 의한 태그들의 상호 관계를 바탕으로 검색하게 한다. 또한 Flcikr 의 이미지 검색 수준을 비교 시험하여 TagPlus 시스템의 검색결과의 효율성을 검증하였다.

본 고에서 제안한 TagPlus 모델을 기초로 다의어 등의 문제점도 보안할 수 있는 시스템의 확장은 향후 연구 과제이다. 동의어 태그에 의한 검색 결과와 마찬가지로 다의어 태그 또한 사용자의 검색결과 만족도에 크게 기여할 것으로 기대된다.

참고문헌

[1] Tim O'reilly, What Is Web 2.0, O'REILLY site, 30 Sep 2005 http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[2] Clay Shirky, "Folksonomy", Many to Many - A group weblog on social software, 24 Aug 2004

[3] Mariek Guy, Emma Tonkin, "Folksonomies : Tidying up Tags?", D-Lib Magazine Volumne12 Number, 1 Jan 2006

[4] Adam Mathes, "Folksonomies - Cooperative Classification and Communication Through Shared Metadata", Computer Mediated Communication - LIS590CMC, Dec 2004.

[5] Clay Shirky, " Ontology is Overrated : Categories, Links, and Tags", Economics & Culture, Media & Community, 2005

[6] Tom Gruber, "Ontology of Folksonomy : A Mash-up of Apples and Oranges", MTSR'05, Dec 2005

[7] Flickr http://www.flickr.com

[8] WordNet® http://wordnet.princeton.edu/