태깅 시스템의 태그 추천 알고리즘
(Tag Recommendation Algorithms in Tagging System)

김 현 우 † 이 강 표 † 김 형 주 ††
(Hyunwoo Kim) (Kangpyo Lee) (Hyoung-Joo Kim)

요 약 웹 2.0 시대에는 웹 상의 사용자들이 수많은 멀티미디어 컨텐츠를 생성함에 따라서 멀티미디어 검색이 더욱 중요하게 되었다. URL, 사진, 동영상과 같은 웹 컨텐츠를 설명하는 간단한 키워드인 태그는 웹 컨텐츠의 메타데이터 역할을 하고 있다. 태그가 달린 데이터의 양이 많아지면 훨씬 풍부한 메타데이터를 포함한 웹 컨텐츠를 대상으로 검색이 가능하기 때문에 태그를 이용한 검색으로 사용자가 원하는 결과를 찾을 수 있는 가능성이 높아지게 된다. 하지만 실제로 태그를 사용하는 사용자의 수는 많지 않다. 태그를 입력하는 과정이 번거롭기 때문이거나 어떠한 태그를 입력하는 것이 다른 사용자들로부터의 접근성을 높일 수 있는지 모르기 때문이다. 이러한 문제를 해결하기 위해서, 사용자의 태그 입력 과정을 도와주는 기법인 태그 추천이 연구되었다. 사용자가 어떠한 웹 컨텐츠를 게재하려고 할 때, 태그 추천 시스템이 해당 웹 컨텐츠에 적절한 태그를 추천하면, 사용자는 적절한 태그를 선택하는 것으로 태그 입력이 이루어진다. 본 연구에서는 이러한 태깅 시스템에서의 다양한 태그 추천 방법론을 분석하고, 분류하였다.

키워드 : 태그 추천, 태깅, 태그, 웹 2.0

Abstract In the era of Web 2.0, users create a number of their own Web contents. So, multimedia search becomes much more important than ever. A tag is a simple keyword which describes the Web contents including URL, pictures, and videos. Tags perform a role of descriptors of Web contents and Web metadata properly. If the number of tagged Web data increases, users are more likely to find the desired search result because the system includes the Web contents which have richer Web metadata. However, the number of users who use tags as Web metadata is relatively small. Because of the cumbersome process of adding tags, or users do not know what to add for the better accessibility from the public. Given situation, tag recommendation, which helps the process of adding tags, has been studied to solve these problems. When a user adds some Web contents, the tag recommendation system recommends relevant tags for the Web contents to the use, and the user selects recommended tags. We analyze and categorize various tag recommendation algorithms in tagging system.

Key words : Tag Recommendation, Tagging, Tag, Web 2.0

1. 서 론

웹(Web) 상에는 수많은 데이터(data)가 존재한다. 웹 상에서 사용자가 필요한 데이터를 찾기 위해서는 검색을 해야 한다. 지금의 검색 환경 이전에는 카테고리(category) 형태로 검색 서비스를 제공했다. 제공자(provider)가 데이터 혹은 정보(information)를 제공하면, 그 정보를 사용하고자 하는 사용자(consumer)가 제공자가 제공하는 카테고리를 직접 보면서 자신이 원하는 정보가 있는 카테고리에 가서 원하는 정보를 찾는다. 데이터의 양이 많지 않을 때는 카테고리 형태로 검색이 가능했지만, 웹 상의 데이터가 많아짐에 따라서 단순히 카테고리를 따라가서 데이터를 찾는 것은 어렵게 되었다. 그래서 문서에 존재하는 문자(text) 정보를 이용하는 검색 기술이 발전하기 시작했다. 사용자가 어떤 검색

· 본 연구는 BK-21 정보기술 사업단 및 지식경제부 및 정보통신산업진흥원의 대학 IT연구센터 육성·지원사업(NIPA-2010-C1090-1031-0002)의 연구결과로 수행되었음

† 비 회 원 : 서울대학교 컴퓨터공학부
hwkim@idb.snu.ac.kr
kplee@idb.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2010년 3월 2일
심사완료 : 2010년 7월 19일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제9호(2010.9)

키워드를 입력하면 그 키워드와 관련된 데이터를 나열하고 그 데이터 사이의 순위(ranking)를 매겨서 순위가 높은 순서대로 사용자에게 보여지게 된다[1]. 데이터가 웹 페이지인 경우, 웹 페이지 사이의 링크(link) 정보를 이용한 PageRank[2]와 같은 방법이 연구되었다. 하지만 사용자가 검색하고자 하는 데이터가 사진이나 동영상과 같은 멀티미디어(multimedia)인 경우, 각 데이터에 링크 정보가 존재하지 않고, 멀티미디어의 제목 이외에는 문자 정보가 존재하지 않기 때문이 이러한 방법을 그대로 적용할 수 없다.

최근, 수많은 사용자들이 사진과 동영상을 웹 상에 쏟아내고 있다. 기존의 방식대로 데이터 제공자들이 공급하는 데이터를 사용자들이 찾을 뿐만 아니라, 사용자들도 직접 데이터를 공급하는 제공자의 입장이 되었다. 사용자들의 참여, 공유, 개방의 웹 2.0 시대가 도래하였다. 이처럼 수많은 멀티미디어 데이터가 생성되고 있는 상황에서, 검색 시스템 성능의 중요성은 더욱 부각된다. 멀티미디어의 검색을 위해서 이미지 인식을 이용한 검색 기술 등이 연구되고 있다.

멀티미디어 검색에 도움을 줄 수 있는 것들 중 하나가 태그(tag)다. 태그란, 해당 데이터를 잘 설명할 수 있는 키워드를 의미한다. 이러한 태그는 해당 데이터를 웹 상에 게재한 사용자가 직접 입력하는 것이기 때문에 다른 어떤 웹 메타데이터(metadata)보다 그 데이터를 가장 잘 설명하는 기술어(descriptor)가 될 수 있다. 이미 많은 웹 사이트에서 태그 서비스를 제공하고 있다. 태그는 웹 상에 존재하는 멀티미디어 컨텐츠(contents)의 메타데이터 역할을 한다. 그렇기 때문에 웹 상에 태그가 달린 데이터가 많으면 많을수록, 태그를 이용해서 검색하는 사용자가 원하는 데이터를 찾을 수 있는 가능성이 높아지게 된다.

사용자들은 다른 사용자로부터의 접근성을 높이기 위해서 태그를 사용한다. 혹은 검색을 용이하게 하기 위해서 태그를 사용하기도 한다[3]. 하지만 태그를 사용하는 사용자들의 수는 그리 많지 않다. 태그를 입력하는 과정이 귀찮다고 생각할 수도 있고, 어떤 태그를 달아야 더 많은 사용자들에게 보여질 수 있는지 알기 어렵기 때문이다[4]. 이러한 사용자를 위해서 자신이 게재한 컨텐츠에 태그를 입력하는 과정을 좀 더 손쉽게 도와줄 수 있는 태그 추천이 필요하게 되었다. 본 논문에서는 이러한 태그 추천 방법론에 대해서 알아보고자 한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 태그와 태깅(tagging)에 대해서 설명하고, 3장에서는 태그가 검색에 어떻게 사용되고 있는지에 대해서 알아본다. 4장에서는 다양한 태그 추천 알고리즘을 분석 및 정리하고, 5장에서 결론을 맺는다.

2. 태그와 태깅

2.1 태그

태그란 문서, 사진, 동영상, 블로그, 웹 페이지(URL), 상품과 같은 웹 컨텐츠를 잘 설명할 수 있는 간단한 키워드 혹은 단어를 의미한다. 태그는 해당 웹 컨텐츠를 설명하는 웹 메타데이터 역할을 한다. 이러한 태그를 사용한 검색 시스템은 기존의 카테고리 시스템에 비해서 더 나은 검색 결과를 가져올 수 있다. 웹 상의 신문기사를 설명하는 키워드로써 태그를 사용하는 경우를 예로 들어보자. 기존의 카테고리 기반 시스템의 경우, 하나의 신문기사는 하나의 분류에 들어가야 한다. 김연아 선수와 박태환 선수에 관한 기사가 있다면, 이 기사는 피겨스케이팅 분류에 들어가거나 수영 분류에 들어가야 한다. 하지만 이 기사를 어느 한쪽 분류에만 저장한다면, 나머지 분류에 관한 정보를 잃게 된다. 이러한 문제를 해결하기 위해서 기사를 두 가지 분류에 모두 저장하는 방법을 사용할 수 있다. 하지만 이러한 방법의 경우 저장공간이 낭비된다는 단점이 있다. 태그를 이용한다면 기사의 메타데이터로서 **피겨스케이팅과 수영**을 태그로 사용하면 된다. 이렇듯 태그는 웹 메타데이터로서 유용하게 쓰일 수 있다.

태그는 문자 정보를 포함하고 있지 않은 사진, 동영상과 같은 웹 데이터의 경우에 훨씬 더 큰 역할을 할 수 있다. 예를 들면, 바다를 촬영한 사진에 태그가 전혀 달려있지 않으면 그 사진으로부터 얻을 수 있는 정보는 아무것도 것도 없다. 뿐만 아니라, 사진의 제목에 바다라는 단어가 들어가지 않는다면 바다라는 키워드로 검색했을 때 그 사진이 검색 결과에 나타나지 않는다. 하지만 이 사진에 **제주도, 푸른 바다, 여름**과 같은 태그가 달려 있는 경우에는 이 태그들로부터 사진에 관한 정보를 얻을 수 있게 된다. 제주도와 관련된 사진을 검색하거나, 여름과 관련된 사진을 검색하는 사용자에게 검색 결과로서 제공될 수 있다.

2.2 태깅

태깅이란, 태그를 웹 컨텐츠에 추가하는 작업을 의미한다. 태깅에는 일반적인 태깅과 협력태깅(collaborative tagging)이 있다. 일반적인 태깅의 경우에는 하나의 웹 컨텐츠에 대해서 한 명의 사용자만 태그를 입력할 수 있는 환경을 의미한다. 협력 태깅은 하나의 웹 컨텐츠에 대해서 여러 명의 사용자들이 태그를 입력할 수 있는 환경을 의미한다.

웹 브라우저의 즐겨찾기를 웹 상에 옮겨놓은 Deli-

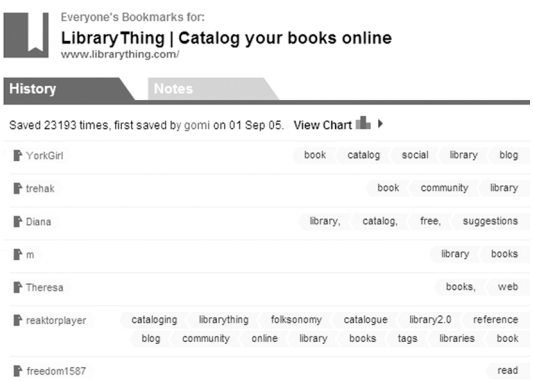

그림 1 협력태깅 예제

cious[1]는 협력 태깅 환경을 제공한다. 그림 1의 예제에서 볼 수 있듯이, 여러 명의 사용자들이 www.librarything.com이라는 하나의 URL에 대해서 태그를 입력하였다. 이러한 협력 태깅 환경에서 사용자들이 입력한 태그가 이루는 분류 체계를 폭소노미(folksonomy)라고 한다. 폭소노미는 대중을 뜻하는 folk와 기존의 분류 체계를 의미하는 taxonomy의 합성어이다.

협력 태깅 환경에서는 집단 지성(collective intelligence)를 이끌어 낼 수 있다. 특정 웹 컨텐츠에 대해서 단 한 명의 사용자가 입력한 태그는 한쪽으로 치우쳐 있을 수 있기 때문에, 그러한 태그 데이터를 웹 메타데이터로서 사용하는 것은 잘못된 결과를 초래할 수 있다. 하지만 한 명의 사용자가 아닌 다수의 사용자가 입력한 태그 데이터에서는 집단 지성을 이용하여 해당 웹 컨텐츠를 가장 적절하게 설명할 수 있는 웹 메타데이터로서의 태그들을 추출할 수 있다.

2.3 태깅 시스템

현재 Delicious, CiteULike[2], BibSonomy[3], 네이버 블로그[4], 싸이월드 미니홈피[5], 티스토리[6], 아마존[7] 등 다양한 웹 사이트에서 태그 시스템을 제공하고 있다. Gmail[8]에서는 이메일에 다양한 라벨을 붙여서 보관할 수 있게 되어있다. Gmail에서는 라벨이라는 이름으로 서비스를 제공하고 있지만, 결국 이메일에 태그를 붙이는 것으로 볼 수 있기 때문에 Gmail 역시 태그 시스템을 제공하고 있다고 볼 수 있다. Delicious는 사용자 URL을 저장하고 해당 URL에 대해서 태그를 입력할 수 있는 태

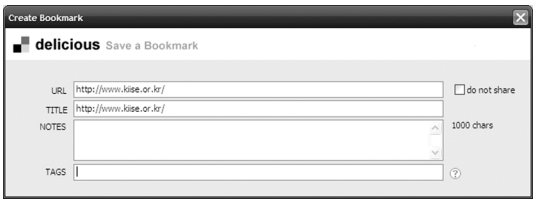

그림 2 Delicious 북마크 저장 예제

깅 시스템을 제공하고 있다.

예를 들어, 한국정보과학회 홈페이지를 저장하려고 하면 그림 2와 같은 화면을 볼 수 있다. 그림 2의 'TAGS'에 사용자가 입력하고자 하는 태그를 입력할 수 있다. 태깅 시스템을 제공하는 대부분의 웹 사이트들도 이와 비슷한 사용자 인터페이스를 제공하고 있다. 아마존의 상품들에도 해당 상품을 설명하는 키워드 혹은 카테고리 라벨(category label)로서의 태그가 사용되고 있다.

태깅은 다음과 같이 정의할 수 있다[5].

사용자(user) 집합 : U

웹 컨텐츠(resource) 집합 : R

태그(tag) 집합 : T

사용자들이 입력한 포스트(post) 집합 : $P \subseteq (U \times R \times T)$

즉, 사용자들의 태깅으로 이루어진 폭소노미는 어떤 사용자가 특정 웹 컨텐츠에 태그를 입력한 트리플(triple)로 이루어진 포스트의 집합으로 정의할 수 있다.

태깅 시스템은 태깅을 하는 사용자의 권리에 따라서 자가 태깅(self-tagging), 허가기반 태깅(permission-based), 자유 태깅(free-for-all)의 3가지로 분류할 수 있다[6]. Technorati[9], YouTube[10] 등에서 사용하는 자가 태깅은 해당 웹 컨텐츠를 저장한 사용자만 태그를 달 수 있는 시스템이다. Flickr[11] 등에서 사용하는 허가기반 태깅은 자신을 포함해서 허가된 사용자만이 태그를 입력할 수 있는 권한이 존재하는 시스템이다. Delicious, Yahoo! MyWeb[12] 등에서 사용하는 자유 태깅은 웹 컨텐츠에 대해서 어떤 사용자든지 태그를 달 수 있는 시스템을 말한다. 자유 태깅은 앞서 언급한 협력 태깅과 같은 형태의 태깅 시스템이라고 볼 수 있다.

하나의 웹 컨텐츠에 대해서 태그를 중복해서 달 수 있는지 없는지에 따라서 백모델(bag-model)과 집합모델(set-model)로 구분할 수 있다. 백모델에서는 태그가 중복해서 나타날 수 있지만 집합모델에서는 하나의 태그는 딱 한번만 나타난다. 백모델에서는 사용자들이 특정 태그를 얼마나 사용했는지 빈도수를 측정해서 알 수

1) http://www.delicious.com
2) http://www.citeulike.org
3) http://www.bibsonomy.org
4) http://blog.naver.com
5) http://www.cyworld.com
6) http://www.tistory.com
7) http://www.amazon.com
8) http://www.gmail.com
9) http://www.technorati.com
10) http://www.youtube.com
11) http://www.flickr.com
12) http://www.yahoo.com

있지만 집합모델에서는 하나의 태그는 한번만 나타나기 때문에 빈도수를 측정할 수 없다.

3. 태그와 검색

태그를 사용하는 여러 가지 목적 중 하나는 나중에 해당 웹 컨텐츠를 검색하기 위해서다. 예를 들어 **정보과학회**라는 태그로 검색을 하면, 수많은 웹 컨텐츠 중에서 **정보과학회**라는 태그가 달려 있는 웹 컨텐츠가 결과에 나타날 것이다. 앞서 언급한 바와 같이 멀티미디어의 경우에는 이러한 태그가 검색에서 유용하게 사용될 수 있다. 태그를 이용한 검색 시스템이 기존의 검색 시스템보다 더 나은 결과를 나타낼 수 있고[7], 태그와 검색 기술과 합쳐지면 흥미있는 웹 컨텐츠를 찾는데 강력한 도구가 될 수 있다[8].

특정 컨텐츠를 찾기 위한 태그 검색 이외에도, 사용자들이 많이 사용한 태그를 보여주기 위한 태그 클라우드(tag cloud)라는 시각화(visualization) 기법이 존재한다.

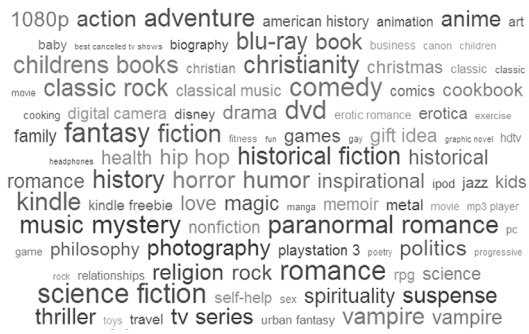

그림 3 Amazon의 태그 클라우드[13]

태그 클라우드는 태그가 사용된 빈도수를 기준으로 많이 사용된 태그는 글자 크기를 크게 하고 적게 사용된 태그는 글자 크기를 작게 하는 방식을 이용한 태그 시각화 기법이다. 사용자는 태그 클라우드에 나타난 태그를 클릭하는 것만으로 해당 태그가 사용된 정보들을 검색할 수 있다. 또한, 협력 태깅 환경의 폭소노미를 태그의 상하위 관계를 추출하여 시각화함으로써 사용자에게 정보를 제공할 수 있다[9].

4. 태그 추천 알고리즘

4.1 태그 추천

태그 추천이라는 단어는 두 가지 의미로 사용될 수 있다. 첫 번째로, 사용자가 어떤 웹 컨텐츠에 태그를 입력하고자 할 때, 시스템이 해당 웹 컨텐츠와 관련 있는 태그를 사용자에게 제시함으로써 사용자의 태그 입력 과정을 도와주는 기법을 태그 추천이라고 한다. 또 다른 태그 추천의 의미로는 태그를 이용해서 사용자에게 아이템을 추천하는 방법을 의미한다[6,10,11]. 아마존과 같은 온라인 쇼핑 사이트에서 기존의 사용자들이 상품에 대해서 입력한 태그들을 이용해서 새로운 사용자에게 상품을 추천하는데 태그를 이용할 수 있고, 태그를 이용해서 음악을 추천할 수도 있다. 본 논문에서는 사용자가 웹 컨텐츠에 태그를 입력하는 과정을 돕는 첫 번째 의미로서의 태그 추천에 대해서 다룬다.

사용자들이 태그를 사용하는 이유 중 하나는 나중에 웹 컨텐츠의 검색의 편리성을 위해서이다. 또한, 자신의 블로그 글이나 사진과 같은 웹 컨텐츠가 좀 더 많은 대중들에게 노출될 수 있도록 하기 위해서 태그를 사용한다[3]. 하지만 기존의 카테고리 시스템에 비해서 태그가 가지는 장점에도 불구하고 태그를 이용하는 사람들이 많지 않다. 그 이유는 태그를 다는 작업이 귀찮거나 어떠한 태그를 달아야 다른 사람들에게 노출이 많이 되는지 알 수 없기 때문이다[12]. 이러한 상황에서 사용자들의 태그 입력 과정을 돕기 위해서 태그 추천이 연구되었다. 태그 추천 시스템으로 인해서 웹 컨텐츠가 태그를 메타데이터로 가질 가능성이 높아지고, 사용자에게 웹 컨텐츠가 무엇에 관한 것인지 상기시킬 수 있고, 시스템이 추천한 태그를 사용자가 선택함으로써 태깅 작업에 도움을 줄 수 있다[4].

태그 추천 과정은 다음과 같이 정의할 수 있다[4].

사용자 집합 : U

웹 컨텐츠 집합 : R

태그 집합 : T

함수 s : $U \times R \rightarrow \widetilde{T}$, $\widetilde{T}$ 는 $\widetilde{T} \subseteq T$를 만족하는 태그 집합

함수 s의 도메인 : dom_s, $dom_s \subseteq U \times R$

위와 같은 정의에서, $u \in U$인 사용자 u와 $r \in R$인 웹 컨텐츠 r이 입력으로 들어 왔을 때 그에 알맞은 태그 집합 $\widetilde{T}(u,r) \subseteq T$를 출력하는 것이 태그 추천 시스템에서 추천이 이루어지는 과정이다.

이미 몇몇 웹 사이트에서 태그 추천을 제공하고 있다. Delicious가 URL에 대한 태그 추천을 제공하고 있고, 티스토리도 사용자가 입력한 블로그 글에 대한 태그 추천을 제공하고 있다.

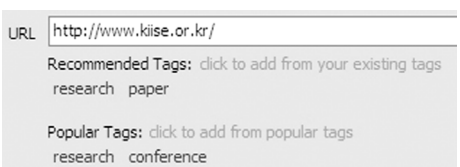

그림 4 Delicious의 태그 추천

13) http://www.amazon.com/gp/tagging/cloud

그림 5 티스토리의 태그 추천

Delicious의 경우, 그림 4에 나타난 것과 같이 추천태그(recommended tags)와 인기태그(popular tags)를 제공하고 있다. 추천태그란 현재 URL을 입력하려고 하는 사용자가 이전에 입력한 태그들 중에서 이 URL에 알맞은 태그를 추천하는 것이다. 인기태그는 다른 사용자들이 이 URL에 대해서 어떤 태그들을 가장 많이 사용하였는가를 빈도순으로 나열한 것이다. 티스토리의 경우, 그림 5에 나타난 것과 같이 추천태그와 태그뽑기를 제공하고 있다. 추천태그는 Delicious와 마찬가지로 사용자가 이전에 입력한 태그 중에서 입력한 블로그 글에 알맞은 태그를 추천하는 것이다. 태그뽑기의 경우에는 현재 사용자가 입력하고 있는 글에서 태그가 될 가능성이 높은 단어들을 추출하여 사용자에게 추천하는 것이다. 시스템에서 추천된 태그들을 단순한 클릭만으로 사용자가 추가할 수 있게 되어 있다. 이와 같은 방법으로 사용자가 편리하게 태그를 추가할 수 있는 방법을 제공함으로써 사용자의 태그 입력 과정을 돕고, 시스템이 양질의 태그를 추천함으로써 태그 데이터의 질을 높이고, 태그가 달려있는 웹 컨텐츠의 양을 늘릴 수 있다.

4.2 태그 추천 알고리즘

최근, 많은 태그 추천 알고리즘들이 연구되고 있다. 이러한 태그 추천 알고리즘들은 다음과 같은 다양한 기법을 이용해서 사용자의 태그 입력 과정에 도움을 주고 있다.

- 웹 컨텐츠의 내용 분석
- 태그 동시출현(co-occurrence) 빈도 이용
- 기계학습(machine learning) 기법 이용
- 자연언어처리(natural language processing) 기법 이용
- 소셜 네트워크(social network)를 이용한 협력 필터링(collaborative filtering)
- 개인 태그 입력 기록(personomy) 이용

위의 방법을 이용한 태그 추천 알고리즘에 대해서 다음에서 자세히 알아본다.

4.2.1 웹 컨텐츠의 내용 분석

태그를 추천하고자 하는 웹 컨텐츠에 문자 정보가 존재하는 경우, 해당 텍스트 정보를 분석하여 태그 추천에 이용할 수 있다. 웹 컨텐츠가 블로그 글이라면, 블로그 글의 문자 정보에 특화된 태그 추천 방법을 사용할 수 있다[13]. 기존의 다른 사용자들이 입력한 블로그 글과 태그 정보를 저장한다. 사용자가 새로운 블로그 글을 입력했을 때, 이미 저장해 둔 블로그 글 중에서 사용자가 입력한 글과 비슷한 블로그 글을 찾아서 그 블로그 글에 달려 있는 태그들을 추천한다. 단순히 글에 달려 있는 모든 태그를 추천하는 것이 아니라 태그들을 여과(filtering)하고 태그 간의 순위(ranking)를 매겨서 태그를 추천하고자 하는 블로그 글과 관련 있는 태그를 추천한다.

Sood[14]는 Mishne[13]의 방법을 향상 시킨 방법을 제안하였다. Mishne의 방법에서는 블로그 글에 달려있는 태그들을 그대로 사용하였지만 Sood의 방법은 태그 데이터를 그대로 사용하는 것이 아니라 처리 속도를 빠르게 하기 위해서 태그 데이터를 압축(compression) 및 선별하는 과정과 선별된 태그들의 타당성을 확인(validation)하는 과정을 추가하였다. 태그 데이터를 압축해서 선별된 태그들만 사용하지만, 원래 태그 데이터가 가지고 있는 정보는 잃어버리지 않는 수준에서의 압축이고, 압축된 데이터를 다시 검열하는 과정이 있기 때문에 Mishne의 방법에 비해서 더 좋은 결과를 나타내었다.

기존의 사용자들이 입력한 태그 데이터만 이용하는 경우, 태그 데이터에 존재하지 않는 태그는 추천될 수 없다. 하지만 위와 같은 방법은 웹 컨텐츠의 내용 분석을 통해서 다른 사용자들이 입력하지 않은 새로운 태그를 추천할 수 있는 장점을 가지고 있다[15]. 그러나 텍스트 정보를 가지고 있는 블로그 글에는 좋은 성능을 보일 수 있지만, 사진이나 동영상과 같이 문자 정보가 전혀 없는 웹 컨텐츠에 적용할 수 없다는 단점이 있다.

4.2.2 태그 동시 출현 빈도 이용

태그 추천 알고리즘에서 가장 많이 쓰이는 기법 중 하나는 태그 동시 출현 빈도(tag co-occurrence frequency)를 이용한 방법이다[8]. 태그 동시 출현 빈도라는 것은 두 태그가 함께 얼마나 자주 쓰였는지 횟수를 측정한 값을 의미한다. 수많은 태그들 사이의 동시 출현 빈도를 측정해서 분석한 데이터를 태그 추천에 이용하는 것이다. 예를 들어, **정보과학회**와 **논문**이라는 태그는 서로 관련이 있기 때문에, **정보과학회**와 **논문**이라는 태그가 함께 자주 쓰였을 확률이 높다. 하지만 정보과학회와 호랑이는 서로 관련도가 낮기 때문에 **정보과학회**와 **호랑이**라는 태그가 함께 자주 쓰였을 확률은 낮을 것이다. 이처럼 사용자들이 함께 자주 사용한 태그는 서로 연관도가 높다. 기존의 태그 데이터를 분석할 때, 이러한 정보를 이용해서 사용자들에게 연관도가 높은 태그를 추천하고자 하는 것이 동시 출현 빈도를 이용한 태그 추천 방법이다. 하지만 단순히 두 태그의 동시 출현 빈도만을 측정하는 것은 의미가 없다. 각 태그가 얼마나 많이 사용되었는지도 고려해야 한다[16].

Sigurbjornsson[16]은 사용자들이 태그를 어떻게 입력하는지, 사용자들이 입력한 태그에는 어떤 정보들이 있는지를 분석하고, 태그 동시 출현 빈도를 이용한 태그

추천 방법을 제안하였다. 사용자가 몇 개의 태그를 입력하면, 해당 태그와 동시 출현한 태그들을 태그 추천을 위한 후보 태그로 선별한다. 선별된 태그 중에서 가장 많이 중복된 태그를 추천하는 방법과 각 태그의 동시 출현 빈도 값의 합이 높은 태그를 추천하는 방법을 제안하였다.

동시 출현 빈도는 태그의 범위를 어떻게 보냐에 따라서 두 가지로 나뉠 수 있다. 주어진 웹 컨텐츠가 있을 때, 기존의 사용자들이 해당 웹 컨텐츠에 입력한 태그만 보고 동시 출현 빈도를 측정할 수 있다. 반면에 주어진 웹 컨텐츠와 상관 없이 전체 태그 데이터를 보고 동시 출현 빈도를 측정할 수 있다. 전자는 해당 웹 컨텐츠에 특화된 폭소노미를 보는 것이고, 후자는 해당 웹 컨텐츠 뿐 아니라 전체 사용자들이 구성한 폭소노미를 보는 것이다. 웹 컨텐츠에 특화된 폭소노미를 보는 경우에는 주어진 데이터의 양이 적어서 동시 출현 빈도가 외곡될 수 있고, 전체 폭소노미를 보는 경우에는 해당 태그가 가지는 다양한 의미가 희석되고 일반적인 정보 한 가지만 얻을 수 있다.

Wu[17]는 기존의 태그 동시 출현 빈도에 기반한 태그 추천 알고리즘의 단점을 지적하고 이를 극복하기 위한 방법을 제안하였다. 이 방법은 태그 동시 출현 빈도는 물론, 태그 컨텐츠 연관도(tag contents correlation), 이미지 기반 태그 연관도(image based tag correlation)과 같은 추가적인 요소를 보고 사진에 대한 태그 추천을 할 수 있는, 사진에 특화된 태그 추천 방법이다. 태그 동시 출현 빈도만을 이용하는 경우 soccer와 football이 서로 연관이 있다는 것은 찾아낼 수 없다. 왜냐하면 축구와 관련된 사진에 태그를 붙인다고 했을 때, soccer라고 쓰는 사용자가 있고, football이라고 쓰는 사용자도 있지만 soccer와 football을 동시에 사용하는 사용자는 없기 때문이다. 이러한 태그 동시 출현 빈도의 단점을 극복하기 위해서 비슷한 사진들을 추출해 낼 수 있는 VLM(visual language model)을 제안하고 이 모델을 태그 추천에 이용하였다.

4.2.3 기계학습 기법 이용

기계학습 분야의 기법들을 태그 추천에 적용할 수 있다. Katakis[18]는 URL의 제목과 같이 웹 컨텐츠 자체에 존재하는 문자 정보를 분석하고, multilabel text classification을 사용하여 태그를 추천하는 방법을 제안하였다. 이 태그 추천 방법은 어떤 URL과 사용자가 주어졌을 때, 기존의 데이터에 존재하는 사용자, URL, 태그 정보를 이용해서 기계학습을 수행한다. 기존의 학습 집단(training set)에 주어진 URL이 존재한다면 그 URL에서 가장 많이 쓰인 태그를 추천한다. 만약 학습 집단에 주어진 URL이 존재하지 않는 경우, 주어진 사용자가 학습 집단에 존재한다면 그 사용자가 쓴 태그 중에서 가장 많이 쓰인 태그를 추천한다. 기존의 학습 집단에 주어진 URL와 사용자가 모두 존재하지 않는다면, multilabel text classifier를 이용해서 태그를 추천하는 방법을 사용한다.

대부분의 태그 추천 논문들은 추천된 결과가 얼마나 정확한지, 사용자들에게 얼마나 도움이 되는가에 대한 연구를 진행하였다. 하지만 태그 추천은 사용자가 웹 컨텐츠를 입력하는 단계에서 수행되기 실시간으로 때문에 태그 추천의 수행 속도도 중요한 요소 중 하나이다. Song[19]은 태그 추천의 정확도와 자동화 뿐만 아니라 효율성을 중점적으로 태그 추천 방법을 제안하였다. Poisson mixture model을 사용하여, 태그 추천을 기계학습의 관점으로 바라보았다. 이를 통해 자동적이고 실시간 태그 추천이 가능한 태그 추천 방법을 제안하였다.

4.2.4 자연언어처리 기법 이용

웹 페이지의 문서(document), 사용자 정보(user profile)을 분석해서, 새로운 문서가 입력으로 들어왔을 때 분석된 사용자 정보와 이전의 문서에 달려 있는 태그를 기준으로 새로운 태그를 추천할 수 있다[20]. Naïve Bayes Classifier, Word Sense Disambiguation과 같이 자연언어처리 분야에서 사용되는 방법들을 태그 추천에 이용하였다.

사람이 태그를 다는 과정은 사람이 생각을 전개해 나가는 과정으로 볼 수 있다. 사람의 사고의 과정은 어느 한 생각에서 다른 생각으로 갑자기 전이되는 것이 아니라 연관성에 기반해서 전개된다. 그렇기 때문에 사용자가 연속으로 입력한 태그들은 서로 연관이 있을 확률이 높다. 모든 태그들의 동시 출현 빈도를 분석하는 것이 아니라 서로 연속되어 있는 태그들의 동시 출현 빈도만을 사용하면 적은 태그 데이터의 저장만으로 빠른 처리 속도와 좋은 추천 결과를 얻을 수 있다. 이를 위해, 자연언어처리 분야에서 사용되는 bigram 방법과 연관 규칙(association rule)을 태그 추천에 이용할 수 있다[12]. 연관 규칙은 데이터 마이닝(data mining) 분야에서 많이 사용되고 있는 기술이다. 기존의 태그 데이터에서 연관 규칙을 추출하여 계산해야 하는 태그의 양을 줄임으로써 추천 속도를 높이고, 전체 데이터를 저장하는 것이 아니라 bigram 태그 데이터만 저장함으로써 데이터 저장 용량을 줄일 수 있다.

4.2.5 소셜 네트워크를 이용한 협력 필터링

협력 필터링 방법은 기법의 간단함과 좋은 결과 때문에 일반적인 추천 시스템에서 가장 많이 쓰이는 방법론 중 하나이다[5]. 태그 추천을 하는데 있어서 사용자 정보를 사용하는 것이 그렇지 않는 방법에 비해서 더 좋은 결과를 나타낼 수 있다. 기존의 사용자 프로필, 태깅

패턴과 같은 정보를 기반으로 유사한 사용자를 찾아서 이를 태그 추천에 이용할 수 있다.

사용자에게 태그 추천을 함에 있어서 태그 정보뿐만 아니라 사용자 정보도 함께 사용하는 경우, 더 나은 결과를 얻을 수 있다[4]. 주어진 사용자의 가장 가까운 이웃(nearest neighbor)를 찾아서 그 이웃이 사용한 태그를 추천한다. 가장 가까운 이웃을 찾는 데에는 사용자-웹 컨텐츠 사이의 관계를 볼 수 있고 사용자-태그 사이의 관계를 볼 수도 있다. 사용자-웹 컨텐츠 사이의 관계를 보는 경우, 같은 웹 컨텐츠에 태그를 단 횟수가 많으면 많을수록 가까운 이웃이 된다. 사용자-태그 사이의 관계를 보는 경우, 같은 태그를 사용한 횟수가 많으면 많을수록 가까운 이웃이 된다. 이러한 방식으로 소셜 네트워크를 구성한 후, 같은 네트워크 상의 사용자가 입력한 태그를 추천받게 된다. 두 가지 방법을 실험적으로 비교해 본 결과, 사용자-태그 사이의 관계를 보는 경우가 더 좋은 결과를 나타내었다.

태그 데이터는 기본적으로 사용자, 웹 컨텐츠, 태그의 삼중관계(ternary relation)을 이루고 있다. 어떤 사용자가 어떤 웹 컨텐츠에 어떤 태그를 달았는지 저장한다. 하지만 이러한 삼중관계를 태그 추천에 바로 이용하는 것이 어렵기 때문에 다른 분야에서도 많이 사용되고 있는 협력 필터링 방법을 사용하거나 기존의 삼중관계를 이중관계(binary relation)으로 나누어서 사용한다. 대부분의 태그 추천 알고리즘은 웹 컨텐츠-사용자의 관계를 이용하거나 사용자-태그의 관계를 이용한다[5]. 하지만 웹 컨텐츠, 사용자, 태그의 삼중관계를 이중관계로 나누어서 태그 추천에 이용하는 경우, 원래의 태그 데이터가 가지고 있는 정보가 사라질 수 있다. 그렇기 때문에 Symeonidis[21]는 Tensor Dimensionality Reduction을 사용하여, 삼중관계를 이중관계로 분리하지 않고 삼중관계 그대로 데이터를 분석하여 태그 추천에 이용하였다.

4.2.6 개인 태그 입력 기록 이용

사용자 정보를 사용하지 않은 방법들은 웹 컨텐츠의 정보와 다른 사용자들이 입력한 태그를 분석하여 태그 추천에 사용하였다. 하지만 Lipczak[22]은 외부의 데이터만을 이용한 것에 비해서 사용자의 개인 태그 입력 기록(personomy)을 이용하게 되면 더 좋은 결과를 나타낼 수 있다는 것을 입증하였다. 웹 컨텐츠가 웹 페이지 혹은 URL인 경우, 사용자의 데스크톱의 정보를 이용할 수 있다[23]. 사용자의 데스크톱에 저장되어 있는 문서를 검색하여 관련된 키워드를 추출하여 그 키워드를 해당 URL을 위한 태그로 추천하는 방법이다.

Garg[24,25]는 전체 문자 정보가 존재하지 않는 태깅 환경에서 개인 태그 입력 기록을 이용하여 동적 갱신(dynamic update)이 가능한 태그 추천 방법을 제안하였다. 기존의 정적(static)인 태그 추천 알고리즘은 사용자와 웹 컨텐츠가 주어지면 그 상황에서 태그를 추천하고 더 이상의 태그 추천은 진행되지 않았다. 하지만 동적 갱신이 가능한 태그 추천 방법에서는, 시스템이 태그를 추천하여 사용자가 추천된 태그들 중에서 하나를 선택하면, 사용자가 선택한 태그를 기반으로 다시 태그 추천이 이루어지는 방식이다. 뿐만 아니라, 사용자가 이전에 입력한 태그 데이터인 개인의 태그 기록(tag history) 정보를 이용함으로써 개인화된 태그 추천이 가능하다는 장점이 있다. 즉, 사용자가 다르면, 같은 웹 컨텐츠에 대해서 서로 다른 태그 추천 결과가 가능하다.

지금까지 알아본 다양한 태그 추천 방법론에 대해서 해당 알고리즘에서 사용한 기법을 표 1에 정리하였다.

표 1 태그 추천 방법론 분류

사용 기법	내용 분석	동시 출현 빈도	기계 학습	자연 언어 처리	협력 필터링	개인 태그 입력
Basile, P. *et al.*	√			√	√	
Chirita, P. A. *et al.*	√					
Garg, N. *et al.*		√	√			√
Katakis, I. *et al.*			√			
Kim, H. *et al.*		√		√		
Lipczak, M.	√					√
Marinho, L. B. *et al.*					√	
Mishne, G.	√					√
Sigurbjornsson, B. *et al.*		√				
Song, Y. *et al.*			√			
Sood, S. C. *et al.*	√					
Symeonidis, P. *et al.*					√	
Tatu, M. *et al.*	√					
Wu, L. *et al.*	√	√				
Xu, Z. *et al.*		√				

4.3 태그 추천의 의의

태그 추천은 사용자의 태그 입력 과정에 도움을 주고, 폭소노미의 질을 높일 수 있다. Suchanek[26]은 사용자들이 입력한 태그들이 과연 얼마나 의미가 있는지, 태그 추천이 사용자들의 태깅 과정에 있어서 어떠한 영향을 끼치는지 분석하였다. 다른 사용자들이 가장 많이 사용한 태그를 추천하는 단순한 태그 추천 모델과 FMTS (Frequence Move-to-Set) 모델, MST(Move-to-Set) 모델을 만들어서 태그 추천이 사용자들의 태깅 과정에 어떠한 영향을 끼치는지 알아보기 위한 실험을 진행하였다. 실험 결과, 다른 사용자들이 이미 사용한 태그를 추천하는 것이 그렇지 않는 태그를 섞어서 추천하는 것보다 더 나은 결과를 나타내었다. 또한 사용자의 태깅 패턴은 태그 추천 시스템이 존재하는 경우에 영향을 많

이 받는 다는 것을 알 수 있었다. 즉, 태그 추천이 사용자들에게 도움을 주고 있다는 결과를 얻어낼 수 있었다.

5. 결론 및 향후 연구

본 논문에서는 현재 연구되고 있는 다양한 태그 추천 방법론에 대해서 분석 및 분류하였다. 대부분의 방법론이 태그를 추천하고자 하는 웹 컨텐츠의 내용을 분석하거나, 기존의 사용자들이 입력한 태그 데이터에서 태그 동시출현 빈도를 측정하여 태그 추천에 이용하였다. 추가적으로 기계학습, 자연언어처리, 협력 필터링 등의 기법을 이용하여 추천 시스템의 성능을 높이고자 노력하였다.

추천 시스템이란, 추천하고자 하는 아이템의 평점이나 사용자의 과거 행동을 기반으로 알맞은 아이템을 사용자에게 추천하는 시스템을 의미한다. 태그 추천 시스템도 이와 같이 태그 데이터를 기반으로 사용자가 웹 컨텐츠에 태그를 입력하려고 할 때, 알맞은 태그를 사용자에게 추천하는 시스템을 의미한다. 다른 추천 시스템과 마찬가지로 태그 추천 시스템은 태그를 입력하는 사용자의 의도를 파악해야 한다. 사용자가 웹 컨텐츠를 잘 묘사하는 태그를 원할 수도 있고, 웹 컨텐츠가 속하는 분류를 위한 태그를 원할 수도 있다. 세부적인 내용을 담고 있는 태그를 최대한 많이 추천 받고 싶을 수도 있고, 웹 컨텐츠를 대표할 수 있는 큰 개념의 태그를 추천 받고 싶을 수도 있다. 다양한 사용자의 의도를 파악하여 그에 알맞은 태그를 추천해야 하는 것이 태그 추천 시스템의 어려움 중 하나이다. 뿐만 아니라, 기존의 정보검색 시스템이 가지고 있는 문제도 태그 추천 시스템에서 발생한다[14]. 태깅 시스템은 입력하는 태그에 대한 전체적인 분류 체계가 존재하지 않기 때문에 사용자는 자신이 원하는 어떤 단어든지 태그로 사용할 수 있다. 그리고 같은 의미를 가지는 다양한 단어가 존재하고, 하나의 단어도 다양한 의미를 가질 수 있기 때문에 다의어(polysemy) 문제와 유의어 문제를 가지게 된다.

태그는 주관적이고, 구조화되어 있지 않은 본질을 가지고 있다. 사용자들은 객관적이고, 형식에 맞는 태그를 입력하고자 하는 것이 아니라, 단순히 자신이 원하는 태그를 입력하고자 한다. 그렇기 때문에 다른 추천 분야에 비해서 사용자의 선호도에 맞는 태그를 추천하는 일은 쉽지 않다. 앞서 언급한 연구들은 기존의 태그 데이터를 분석해서 태그를 추천하는 것에 주안점을 주고 문제에 접근하였다. 하지만 위에서 언급한 문제들을 해결하기 위해서는 데이터를 분석하는 것은 물론이고, 사용자의 의도를 파악하고 태그가 가지는 자체적인 문제를 해결하는 것에 주안점을 두고 연구를 진행해 나가야 할 것이다.

참 고 문 헌

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval:* Addison-Wesley Reading, MA, 1999.

[2] S. Brin and L. Page, "The Anatomy of a Large-scale Hypertextual Web Search Engine," *Computer Networks and ISDN Systems*, vol.30, pp.107-117, 1998.

[3] M. Ames and M. Naaman, "Why We Tag: Motivations for Annotation in Mobile and Online Media," in *SIGCHI Conference on Human Factors in Computing Systems*, 2007.

[4] L. B. Marinho and L. Schmidt-Thieme, "Collaborative Tag Recommendations," presented at the 31st Annual Conference of the Gesellschaft für Klassifikation, 2007.

[5] R. Jaschke, *et al.*, "Tag Recommendations in Folksonomies," presented at the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, 2007.

[6] A. T. Ji, *et al.*, "Collaborative Tagging in Recommender Systems," presented at the Advances in Artificial Intelligence, 2007.

[7] J. Morrison, "Tagging and Searching: Search Retrieval Effectiveness of Folksonomies on the World Wide Web," *Information Processing & Management*, vol.44, pp.1562-1579, 2008.

[8] Z. Xu, *et al.*, "Towards The Semantic Web: Collaborative Tag Suggestions," in *Proceedings of the Collaborative Web Tagging Workshop in 15th International Conference on the World Wide Web*, 2006.

[9] K. Lee, *et al.*, "Folksoviz: A Subsumption-based Folksonomy Visualization using Wikipedia Texts," presented at the 17th International World Wide Web Conference, 2008.

[10] K. H. L. Tso-Sutter, *et al.*, "Tag-aware Recommender Systems by Fusion of Collaborative Filtering Algorithms," presented at the ACM Symposium on Applied Computing, 2008.

[11] M. Vojnovic, *et al.*, "Ranking and Suggesting Popular Items," *IEEE Transactions on Knowledge and Data Engineering*, vol.21, pp.1133-1146, 2009.

[12] H. Kim, *et al.*, "Tag Suggestion Method based on Association Pattern and Bigram Approach," presented at the International Workshop on e-Activity, 2009.

[13] G. Mishne, "AutoTag: A Collaborative Approach to Automated Tag Assignment for Weblog Posts," presented at the 15th International World Wide Web Conference, Edinburgh, Scotland, 2006.

[14] S. C. Sood, *et al.*, "TagAssist: Automatic Tag Suggestion for Blog Posts," presented at the International Conference on Weblogs and Social Media 2007.

[15] M. Tatu, *et al.*, "Tag Recommendations using Bookmark Content," presented at the The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2008.
[16] B. Sigurbjornsson and R. v. Zwol, "Flickr Tag Recommendation based on Collective Knowledge," presented at the 17th International World Wide Web Conference, 2008.
[17] L. Wu, *et al.*, "Learning to Tag," presented at the 18th International World Wide Web Conference, 2009.
[18] I. Katakis, *et al.*, "Multilabel Text Classification for Automated Tag Suggestion," presented at the Discovery Challenge of The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2008.
[19] Y. Song, *et al.*, "Real-time Automatic Tag Recommendation," presented at the Annual ACM Conference on Research and Development in Information Retrieval, 2008.
[20] P. Basile, *et al.*, "Recommending Smart Tags in a Social Bookmarking System," presented at the The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2008.
[21] P. Symeonidis, *et al.*, "Tag Recommendations based on Tensor Dimensionality Reduction," presented at the ACM Conference on Recommender Systems, 2008.
[22] M. Lipczak, "Tag Recommendation for Folksonomies Oriented towards Individual Users," presented at the The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, 2008.
[23] P. A. Chirita, *et al.*, "P-TAG: Large Scale Automatic Generation of Personalized Annotation TAGs for the Web," presented at the 16th International World Wide Web Conference, 2007.
[24] N. Garg and I. Weber, "Personalized Tag Suggestion for Flickr," presented at the 17th International World Wide Web Conference, 2008.
[25] N. Garg and I. Weber, "Personalized, Intercative Tag Recommendation for Flickr," presented at the ACM Conference on Recommender Systems 2008.
[26] F. M. Suchanek, *et al.*, "Social Tags: Meaning and Suggestions," presented at the Conference on Information and Knowledge Management, 2008.

김 현 우
정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 16 권 제 2 호 참조

이 강 표
정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 16 권 제 3 호 참조

김 형 주
정보과학회논문지 : 컴퓨팅의 실제 및 레터 제 16 권 제 3 호 참조