Journal of Korea Multimedia Society Vol. 13, No. 5, May 2010(pp. 730-739)

태그 기반 토픽맵 생성 시스템의 설계 및 구현

이시화[†], 이만형[††], 황대훈[†††]

요 약

웹2.0환경에서의 핵심적인 기술은 태깅이며, 현재 블로그와 같은 웹 문서에서부터 이미지, 동영상 등과 같은 멀티미디어 데이터에 이르기까지 폭넓게 적용되고 있다. 그러나 태깅에 사용된 태그가 정보 검색에 재사용되어 검색의 효율성을 극대화 시킬 것이라는 기대와는 달리 실제로는 태그가 가지는 근본적인 한계들로 인해 만족스럽지 못한 검색결과가 나타나고 있다. 이에 본 연구에서는 태그 클러스터링을 통한 이미지 검색에 대한 선행연구를 기반으로 의미론적 지식체계인 토픽맵 생성 시스템을 설계 및 구현하였다. 구현 결과 클러스터 내의 태그 정보들은 토픽맵에서의 토픽으로 자동 생성되었으며, 생성된 토픽맵의 토픽들 간에는 WordNet을 적용하여 의미연관관계를 부여하였다. 또한 토픽 쌍에 적합한 어커런스 정보들을 추출하여 토픽들에 부여함으로서 의미론적 지식체계인 토픽맵을 생성하였다. 이와 같이 생성된 토픽맵은 사용자의 정보검색 요구에 대한 시맨틱 내비게이션의 제공을 가능하게 할 뿐만 아니라 풍부한 정보제공이 가능하다.

Design and Implementation of Topic Map Generation System based Tag

Si-Hwa Lee[†], Man-Hyoung Lee[††], Dae-Hoon Hwang[†††]

ABSTRACT

One of core technology in Web 2.0 is tagging, which is applied to multimedia data such as web document of blog, image and video etc widely. But unlike expectation that the tags will be reused in information retrieval and then maximize the retrieval efficiency, unacceptable retrieval results appear owing to toot limitation of tag. In this paper, in the base of preceding research about image retrieval through tag clustering, we design and implement a topic map generation system which is a semantic knowledge system. Finally, tag information in cluster were generated automatically with topics of topic map. The generated topics of topic map are endowed with mean relationship by use of WordNet. Also the topics are endowed with occurrence information suitable for topic pair, and then a topic map with semantic knowledge system can be generated. As the result, the topic map preposed in this paper can be used in not only user's information retrieval demand with semantic navigation but alse convenient and abundant information service.

Key words: Web2.0(웹2.0), Tag(태그), Clustering(클러스터링), Image Retrieval(이미지 검색), Topic Map(토픽맵), WordNet(워드넷)

※ 교신저자(Corresponding Author) : 황대훈, 주소 : 경기도 성남시 수정구 복정동 산 65번지 경원대학교 새롬관 5-14호(461-701), 전화 : 031)750-5327, FAX : 031)757-6715, E- mail : hwangdh@kyungwon.ac.kr
접수일 : 2009년 11월 30일, 수정일 : 2010년 1월 14일
완료일 : 2010년 2월 2일

[†] 준회원, 경원대학교 전자계산학과 박사과정
(E-mail : leesihwaman@gmail.com)
[††] 준회원, 현대전문학교 교수
(E-mail : atomv@nate.com)
[†††] 종신회원, 경원대학교 교수
※ 이 연구는 2010년도 경원대학교 지원에 의한 결과임

1. 서 론

웹 서비스는 점차 동적이고 능동적으로 변화하고 있으며, 이러한 웹 서비스 변화의 흐름을 잘 반영하는 것이 웹2.0이다[1].

웹2.0에서 대부분의 정보는 사용자에 의해 생산되고, 사용자가 붙인 태그에 의해 정보들을 체계화 시킨다[2]. 그러나 현재 태그기반 시스템은 사용자들의 부정확한 태깅과 비구조화된 태그 구조로 인해 낮은 정보 검색 결과 및 비효율적인 정보 내비게이션을 제공하고 있다[3]. 이와 같은 태그가 가지는 한계로 인해 현재 태그에 관한 서비스들은 효율적인 태깅이나, 태그 구름(tag cloud)에 초점이 맞춰져 서비스되어지고 있다. 이를 개선하려는 노력의 일환으로 많은 연구들이 진행되어 오고 있으나 태그의 근본적인 한계는 극복하지 못하고 있는 실정이다.

이에 본 연구에서는 웹상에 산재되어 있는 리소스 및 그에 따른 태그정보들을 수집하여 태그가 가지는 첫 번째 문제점인 부정확한 태그로 인한 낮은 검색의 문제점을 해결하기 위해 부정확한 태그들은 제거하고 연관성이 높은 태그 그룹으로 클러스터링하기 위한 연구 [4]를 선행연구로 진행하였으며, 본 논문에서는 생성된 클러스터를 기반으로 의미론적 지식체계인 토픽맵을 반자동으로 생성하기 위한 시스템을 설계 및 구현하였다.

구현결과 생성된 토픽맵은 기존 태그 기반 시스템의 비구조화된 태그를 통한 정보 내비게이션이 아닌 토픽들 간의 의미연관관계를 활용한 시맨틱 내비게이션이 가능할 뿐만 아니라, 실험 데이터의 부정확한 태그들을 클러스터링 과정을 통해 또 다른 토픽맵으로 생성하는 향상된 결과를 도출하였다

2. 관련 연구

2.1 웹2.0환경에서의 태그

현재 웹2.0 환경에서 핵심적인 기술 중 하나인 태그는 검색에 재사용되어 검색의 효율성을 극대화 시킬 것이라는 기대와는 달리 실제로는 태그가 가지는 한계로 인해 만족스럽지 못한 검색결과를 도출하고 있다[5,6].

이와 같은 한계로는 첫째, 태그 기반 검색 시스템은 리소스에 태깅된 태그들 중 부정확한 태그들로 인해 검색에 있어서 정확도가 떨어진다. 둘째, 태깅된 태그는 비구조화된 데이터로서, 태그들 간의 어떠한 관계가 있는지를 판단할 수 없다. 이는 태깅의 주목적중 하나인 태그를 통한 정보 내비게이션을 방해하는 원인이 되며, 정보 내비게이션에 대한 검색 결과 또한 첫 번째 문제에 따라 낮은 검색 결과를 제공할 수밖에 없다.

이와 같은 태그가 가지는 문제점을 해결하고자 Christopher 등[7]은 블로그 상에 존재하는 뉴스 문서들을 수집하여 TF와 IDF의 평가 방법론을 이용하여 유사문서 추출 및 중요도가 높은 태그부터 낮은 태그 순으로 구조화된 클러스터로 생성하는 방법론을 제안하였다. 그러나 이 연구에서 제안한 방법론은 특정 콘텐츠들에 대해서만 효율적인 것으로 분석되었으며, 또한 클러스터링 과정 중 어느 시점이 높은 태그이고 어느 시점부터가 낮은 태그인지에 대한 명확한 제시는 못하고 있다.

강필구 등[8]은 태그 기반 시스템에 적합한 데이터베이스를 제시하고 연관 태그 및 대표태그를 추출하는 방법 및 그를 통해 트리화된 구조로 검색결과를 제시하는 연구를 진행하였다. 그러나 미리 구조화된 데이터베이스 및 연관 태그와 대표태그를 사용자가 선택하는 방식을 사용한다는 단점을 가지고 있다.

Wiener 등[9]는 시맨틱 웹(semantic web) 기술을 사용하여 디지털 이미지에 주석을 달고 관리하는 도메인 독립적인 이미지 어노테이션 툴을 제안하였다. 사용자는 이미지 또는 이미지의 특정 영역에 태그가 아닌 온톨로지의 개념으로 어노테이션하며, 자동 생성된 이미지의 메타데이터를 웹에 게시한다. 또한 단일 온톨로지만으로 이미지를 표현하기에는 부족하기 때문에 사용자가 여러 개의 도메인 온톨로지를 불러와 이를 이용하여 이미지를 설명할 수 있게 한다. 그러나 미리 정의된 여러 분야에 적용 가능한 도메인 온톨로지가 필요하며, 실세계에 존재하는 수만은 도메인에 적합한 온톨로지를 생성하여 적용하기에는 어려운 문제점을 가지고 있다.

2.2 토픽맵 생성 연구

토픽맵은 2001년 "ISO/IEC 13250 Topic Map"이라는 명칭으로 표준화 되었다. 제정된 XTM(XML Topic Map)의 핵심 엘리먼트는 "TAO of Topic Map"라고 불리는 토픽(Topic), 연관관계(Association), 어커런

스(Occurrence)로써, 지식층은 상위계층으로 토픽과 토픽간의 어소시에이션으로 구성된다[10].

현재 토픽맵 구축 연구로는 Librelotto [11]은 XML 문서에서 사용자가 지정하는 엘리먼트를 추출하여 XTM 문서로 변환하여 준다. 이 방법은 사용자가 해당 XML 문서 집합의 스키마와 맵핑되는 XTM 엘리먼트들을 XSTM 언어로 직접 명세하므로, 특정 도메인에 대한 토픽맵을 비교적 정확하게 생성할 수 있다. 그러나 이러한 방법은 아주 잘 구조화된 XML 데이터가 있어야만 하며, 또한 토픽 간의 연관성은 예측하기 어려우므로 사용자가 미리 정의된 스키마를 기반으로 함에 따라서 한정된 토픽맵이 생성될 수밖에 없다는 단점을 가지고 있다.

함화진 [12]은 XML로 구조화된 뉴스 데이터로부터, 연관규칙 마이닝 결과를 분석함으로써 토픽 간의 연관관계를 사용자가 인지하기 위한 연구를 진행하였다. 이 방법은 토픽맵 수동 구축 시 사용자의 인지에 도움을 줄 수 있다. 그러나 이 방법 또한 뉴스 데이터와 같은 완전 구조화된 자료를 필요로 하며, 연관규칙 마이닝의 결과 자체가 연관관계가 아니므로, 인간에 의해서 재해석되어야 한다.

Kohler 등[13]은 문서 색인화기법과 정보검색 기법을 이용하여 토픽맵의 반자동 생성을 위한 연구를 진행하였으며, 크게 토픽들의 추출과 연관관계 추출로 나눈다. 토픽의 추출은 문서 인덱싱 기법을 사용한 전형적인 인덱싱 기법을 제시하며, 1단계 문자요소 분석, 2단계 불 용어 제거, 3단계 스테밍의 단계를 거친다. 그러나 추출된 토픽 리스트를 바탕으로 인간이 개입해서 하나하나 분류해 나가야 하는 어려움을 가지고 있다. 또한 문서에 한정된다는 단점을 가지고 있다.

3. 시스템 설계

본 논문에서는 웹상에 산재되어있는 태그 정보들을 수집하여 의미론적 지식체계인 토픽맵을 생성하기 위한 그림 1의 시스템을 제안 및 구현한다.

선행 연구로는 연관 태그들 간의 맵핑 및 상호 연관성이 높은 태그 그룹으로 클러스터링하기 위한 연구 및 연관성이 높은 태그들을 클러스터에 포함할지 유무를 결정하는 Threshold 값 선정에 대한 연구를 [14]에서 진행하였다. 또한 생성된 클러스터를 이미지 검색에 적용하기 위한 알고리즘 및 검색결과를 [15]에서 진행하였다.

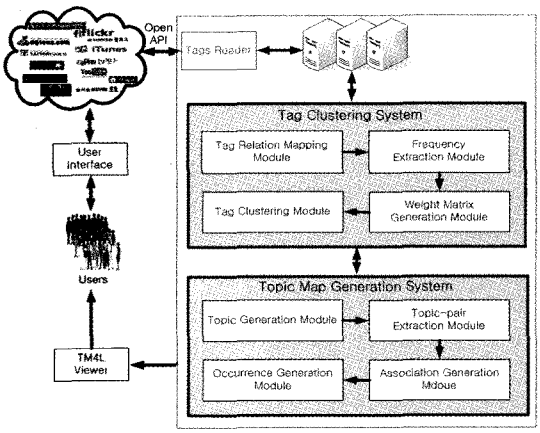


그림 1. 토픽맵 생성 시스템

본 논문에서는 선행연구의 결과를 기반으로 의미론적 지식체계인 토픽맵 생성을 위한 그림 1의 토픽맵 생성 시스템(Topic Map Generation System)을 중심으로 연구하였다.

3.1 토픽 생성 모듈(Topic Generation Module)

토픽맵 생성을 위한 첫 번째 과정은 생성하고자 하는 토픽맵의 공유 시 사용자들 간의 토픽맵을 설명할 수 있는 명세가 필요로 하며, 이를 위한 UI(User Interface) 및 XTM(XML Topic Map : 토픽맵 표현 구문으로서 2001년 ISO 표준으로 제정) 구문은 그림 2와 같다.

UI의 구성은 생성하고자 하는 토픽맵의 주제 및 간단한 설명을 위한 Title, Description과 생성자, 배포자 및 생성 날짜를 명시하기 위한 Creator, Publisher, Creation Date, 마지막으로 생성된 토픽맵의 언어 및 생성된 토픽맵의 포맷을 정의하기 위한 Language, Format으로 구성되어져 있다.

명세서 생성 화면에서의 Title, Description, Creator,

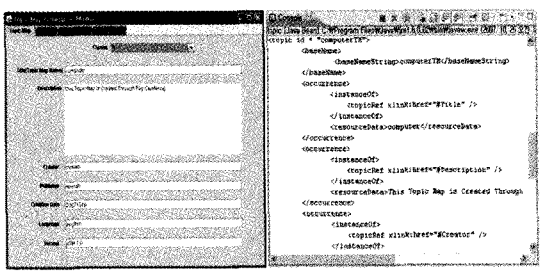


그림 2. Topic Map에 대한 명세 및 XTM 구문

Publisher, Creation Date, Language, Copyright는 생성된 XTM 문법에서 각각의 <topic id=""> 및 <baseName>을 가지는 토픽으로 생성된다. 그리고 명세를 대표하는 <topic id = "computerTM">라는 토픽 타입의 <occurrence>요소에 <instenceOf>요소의 <topicRef>요소를 이용하여 토픽 Title을 지시하고 있다. 또한 <resourceData>요소를 통해서 데이터 리소스는 computer라는 것을 표현한다. 이를 서술적으로 표현하면, computerTM라는 토픽 타입은 Title라는 토픽을 가지고 있으며, 리소스 형태는 데이터형인 computer를 가지고 있다와 같이 표현 가능하다.

토픽 생성 모듈은 선행연구에서 생성된 클러스터 내의 태그 정보들을 토픽맵에서의 토픽으로 자동 생성하기 위한 모듈이며, 이를 위한 토픽 생성 UI는 다음 그림 3과 같다.

토픽 생성 UI는 토픽들을 리스트화하기 위해 Topic List(①)와 리스트 화된 각각의 토픽들에 고유한 식별자를 부여하기 위한 Topic ID(②), 마지막으로 생성된 토픽맵에서의 토픽들을 사람이 이해할 수 있도록 도움을 주는 BaseName(③)으로 구성된다.

토픽맵 상에서의 Topic ID는 토픽과 토픽을 구분하는 고유한 식별자(ID : identifier)로서, 본 연구에서는 연관 태그 맵핑[4]을 통해 토픽들 간의 유일성을 보장하였으며, 이를 통해 태그 이름을 Topic ID로 자동 부여하였다. 또한 토픽의 기본이름(base name)은 사람이 여러 토픽들 중 각각의 토픽을 이해 할 수 있게 하기 위한 목적으로 사용된다. 복수개의 지정이 가능하나 본 연구에서는 클러스터링을 통해 추출된 토픽들의 경우, 태그 기반 사이트에서 빈번하게 사용되어지는 태그들이 추출되어진 토픽으로서, 이 또한 토픽 명으로 자동 부여 하였다.

그림 3의 토픽 생성에 따른 XTM 구분은 클러스터 1[4]내의 태그 computer, apple, laptop, desktop, keyboard, monitor, macintosh, mac이 토픽맵에서의 토픽으로 생성된 결과를 보여주고 있다. 이 중 computer 토픽에 대해 고유한 식별자 <topic id = "computer">와 <baseName>으로 <baseNameString> computer</baseNameString>으로 생성된 결과를 보여주고 있다.

3.2 토픽 쌍 추출 모듈(Topic-pair Generation Module)

연관 토픽 쌍 추출 모듈은 3.1절에서 생성된 토픽들 간에 의미 관계 부여를 위해 연관관계로 구성된 태그 쌍(tag-pair)을 추출하기 위한 모듈이다.

이를 위해 [4]의 클러스터링 기반 탐색 알고리즘 중 태그 쌍을 추출하는 방법론을 이용하였으며, 토픽 쌍 추출을 위한 알고리즘은 다음 그림 4와 같다.

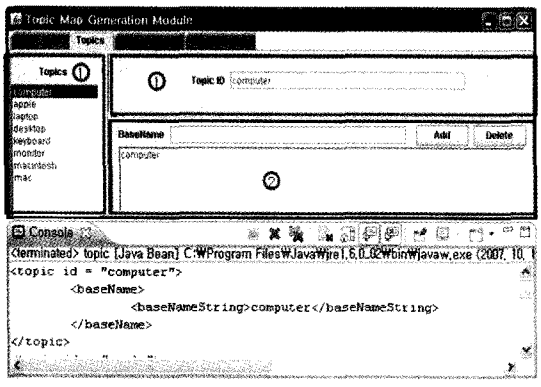


그림 3. 토픽 속성 생성 UI

```
// C(i) : 사용자에 의해 선택된 클러스터
//A(i,j) : C(i)에 포함된 tag들의 가중치 행렬
//Max(A(i,j)) : 가중치 행렬 A(i,j)의 원소 중 최대값을 가지는 원소로서,
                tag i와 tag j의 가중치
//T(k) : Max(A(i,j))의 tag c와 tag j에 incident한 모든 tag들의 집합
//B(l,m) : T(k) 에 포함된 태그들의 가중치 행렬
//Max(B(l,m)) : 가중치 행렬 B(l,m)의 원소 중 최대값을 가지는 원소로서,
                tag l과 tag m의 가중치

    //A(i,j)가 empty가 될 때까지 반복
    Repeat {

        //가중치 행렬 A(i,j)의 원소 중 최대값을 가지는 Max(A(i,j)) 추출
        Extract Max(A(i,j)) in A(i,j)

        //Max(A(i,j))의 tag i와 tag j에 incident한 모든 tag들을 C(num)
        에서 탐색하여 T(k)를 구성
        Find T(k)

        //B(l,m)이 empty가 될 때까지 반복
        Repeat {

            //가중치 행렬 B(l,m)의 원소 중 최대값을 가지는 Max(B(l,m)) 추출
            Extract Max(B(l,m))

            //B(l,m)에서 Max(B(l,m))는 삭제
            Remove Max(B(l,m)) from B(l,m)

        } until (B(l,m)==empty)

            //A(i,j)에서 B(l,m)는 삭제
            Remove B(l,m) from A(i,j)

    } until (A(i,j)==empty)
```

그림 4. 연관 토픽쌍 추출 알고리즘

알고리즘의 진행과정은 크게 3단계로 진행되어지며, 첫 단계로, 생성된 클러스터들 중 사용자가 선택한 클러스터 *C(i)*의 가중치 행렬 *A(i,j)*(①)의 원소 중 최대 가중치를 가지는 *Max(A(i,j))*(②)를 선택하여 선택된 *tag i*와 *tag j*를 연관 태그로 추출한다. 두 번째, 선택된 *Max(A(i,j))*의 *tag i*와 *tag j*에 incident 한 모든 tag를 *C(i)*에서 탐색하여 *T(k)*(③)를 구성한다. 마지막으로 *T(k)*에 포함된 가중치 행렬인 *B(l,m)*의 원소 중 최대값을 가지는 원소 *Max(B(l,m))*(④)를 선택하여 tag i와 tag m을 연관 태그로 추출한다. 그 후, *B(l,m)*에서 *Max(B(l,m))*는 삭제되며, *B(i,m)*가 empty가 될 때까지 반복한다.

이러한 과정은 *A(i,j)*내에 empty가 될 때까지 반복 진행된다.

선행 연구에 적용한 결과 클러스터 1은 26쌍의 연관 태그 쌍과 클러스터 2는 6쌍의 연관 태그 쌍을 추출하였으며, 클러스터 1에 해당하는 26쌍의 연관 태그 쌍은 다음 그림 5와 같다.

Relation Topic List

Topic1	Topic2	Weight
computer	apple	27
computer	desktop	23
computer	laptop	20
computer	mac	19
computer	macintosh	18
apple	desktop	18
apple	mac	17
computer	monitor	13
computer	keyboard	12
apple	laptop	9
apple	keyboard	8
apple	monitor	6
apple	desktop	5

Relation Topic List

Topic1	Topic2	Weight
mac	macintosh	12
mac	desktop	6
macintosh	keyboard	6
mac	keyboard	6
macintosh	laptop	5
mac	laptop	5
mac	monitor	5
macintosh	monitor	4
macintosh	desktop	3
desktop	monitor	5
desktop	laptop	4
monitor	keyboard	5
keyboard	laptop	4

그림 5. 클러스터 1의 연관 토픽쌍

3.3 연관관계 생성 모듈(Association Generation Module)

연관관계 생성 모듈은 추출된 토픽 쌍 에 의미관계를 부여함으로서 의미론적인 지식체계를 구축 가능하게 한다.

본 논문에서는 추출된 연관 토픽쌍에 의미연관관계 부여를 위해 영어를 기반으로 한 어휘적 지식 모델인 워드넷(WordNet)을 적용하였다. 워드넷은 단어 상의 의미론적 또는 사용 패턴에 관련된 정보로, 단어 간의 연관성을 구축한 데이터베이스라고 할 수 있다[16]. 워드넷는 두 단어 간의 연관관계, 단어의 상위어, 하위어, 동의어 등의 관계를 추출해 낼 수 있는 자바 기반의 워드넷 라이브러리(JWNL : Java Word Net Library)를 통해 공개 배포하고 있다. 따라서 워드넷은 토픽들 간의 연관성을 알아내기 위해 사용될 수 있다.

그림 6은 워드넷을 이용해 연관관계를 생성하기 위한 UI를 보여주고 있다. UI는 사용자 참여에 의한 의미관계를 부여 시 두 토픽 간의 연관관계 및 각각의 토픽에 상위어, 하위어, 동위어를 사용자가 선택적으로 참조할 수 있도록 관계를 제공한다. 이를 위해 그림 6의 ①과 워드넷 공개 라이브러리를 직접 호출할 수 있는 ③, 선택된 관계를 브라우징 하기 위한 ④로 구성되어져 있다.

이 중 본 시스템은 자동화된 연관관계 부여를 위해 ①의 선택된 두 토픽들 간의 연관관계 추출기능(②)을 이용하여 3.2절에서 추출된 두 토픽들 간의 연관관계를 자동으로 부여하였다. 또한 ②를 통해 추출하지 못하는 토픽쌍들의 관계에 대해서는 사용자들의 참여를 통해 ①과 ③의 기능들을 참조하여 관계를 반자동으로 부여한다.

그림 7은 연관관계 생성 UI를 보여주고 있으며, 추출된 토픽쌍들을 리스트로 보여주기 위한 Relation Topic List(①)와 리스트된 토픽들 중 선택된 토픽들의 정보를 보여 주기위한 ②, 선택된 두 토픽들 간에 의미적인 관계의 타입을 정의하기 위한 Relationship Type(③)으로 구성되어져 있다. 또한 정의된 관계 타입에 따라 A토픽 관점에서 B토픽과의 관계와 B토

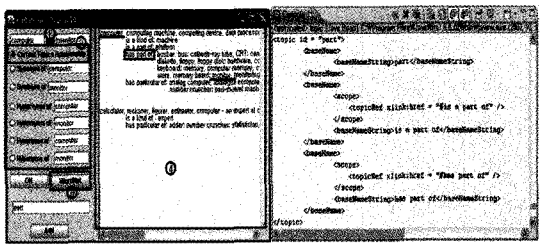


그림 6. WordNet을 이용한 연관관계 추출

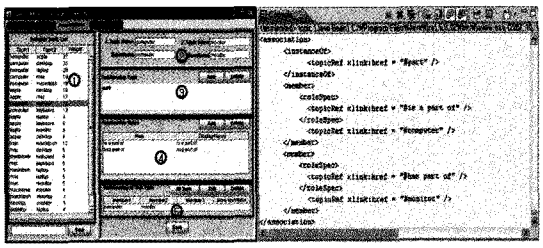


그림 7. 추출된 연관관계 및 XTM 구문

픽 관점에서 A토픽과의 관계를 역할로 정의하기 위한 Relationship Roles(④)와 연관 타입과 멤버로 구성된 결과를 보여주기 위한 Relationship of This Type(⑤)으로 구성된다.

생성된 XTM 구문에서의 연관관계는 "part"가 연관관계 타입으로 정의되고 <member>요소의 <roleSpec> 요소를 통해 part에는 "is a part of"와 "has part of"라는 역할을 하는 연관관계가 있음을 <topicRef>요소를 통해 정의되며, computer 관점에서 monitor와의 관계는 "has part of", monitor 관점에서 computer와의 관계는 "is a part of"라는 연관관계로 정의된다. 이 결과를 통해 "monitor는 computer에 일부이다"와 같은 서술적인 의미가 표현 가능하다. 그러나 computer와 apple의 경우 WordNet 상에서 apple는 과일의 사과로 정의되어 짐에 따라 연관관계 추출이 불가능하며, 이와 같은 토픽 쌍들에 대해서는 사용자의 참여가 필요한 부분이다. 또한 design과 graphic와 같은 관계는 WordNet 상에서도 정의 되어있지 않으며, 사용자에 의해 정의되기에도 애매하여 정의되지 않은 토픽 쌍들에 대해서는 "related to"로 자동 정의하도록 하였다.

3.4 어커런스 생성 모듈(Occurrence Generation Module)

연관관계로 생성된 토픽들은 하나 이상의 정보 리소스와 연결되는데, 이와 같은 역할을 하는 것이 어커런스이다. 어커런스는 토픽과 리소를 연결하는 리소스 레퍼런스 혹은 리소스 그 자체인 문자열 값 중 하나의 형태를 가진다. 본 논문에서의 리소는 이미지(image)를 대상으로 하고 있으며, 토픽들에 적합한 리소스를 추출하기 위해 [4]의 클러스터링 기반 탐색 알고리즘을 적용하였다.

다음 그림 8은 어커런스 생성 결과 및 이를 관리하기위한 UI를 보여주고 있으며, 선택된 토픽쌍(①)들에 부여된 리소스 정보들을 브라우징하기 위한 ②와 잘못된 어커런스들을 수정 및 삭제하기 위한 ③으로 구성되어져 있다. 또한 ③의 View 버튼을 통해 토픽에 부여된 리소스들을 확인할 수 있다. 그림 7은 computer 토픽에 대해 문자열 리소스로 "Image"를, 외부 연결 리소스는 "http://www.flickr.com/photos/kros/259803598"로 자동 생성된 어커런스의 일부분을 보여주고 있다. 이와 같은 결과를 통해 "computer는 Image http://www.flickr.com/photos/kros/259803598을 가지고 있다"와 같은 의미를 가지게 된다.

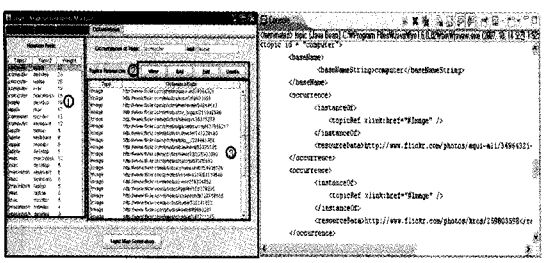


그림 8. 추출된 어커런스 및 XTM 구문

4. 시스템 구현 및 검증

4.1 구현환경

제안한 시스템의 구현 및 실험 환경은 Windows 2000 Server 환경에서 구현되었고, 하드웨어는 Pentium 4 CPU 3.0 GHz, 1GB RAM을 사용하였으며, 실험 데이터 수집을 위해 Flickr Open API 1.0 및 이를 저장하기 위한 데이터베이스 MySQL Server 5.0을 사용하였다. 또한 알고리즘 구현을 위한 언어는 J2SDK 1.5와 도구로 Eclipse 3.3.1 및 JWNL을 이용하였으며, 생성된 온톨로지를 검증 및 가시화하기 위한 도구로서 TM4L Viewer 0.1을 사용하였다[17].

4.2 토픽맵 검증

그림 9는 토픽맵에 대한 명세를 생성한 결과를 보여주고 있으며, 토픽 맵에 대한 주제, 정의, 생성날짜, 사용 언어, 토픽맵 포맷, 생성 자, 배포 자와 같은 정보들을 통해 토픽맵 공유 시 토픽맵에 대한 명세를 제공 가능하다.

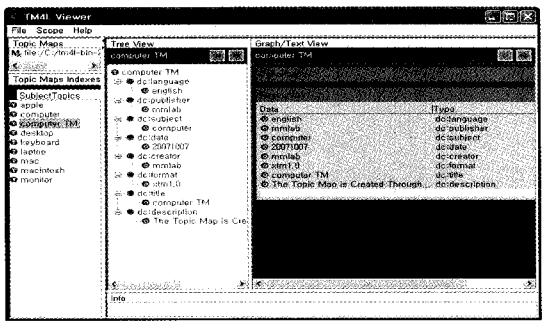


그림 9. 토팩맵에 대한 명세

그림 10과 그림 11는 제안한 방법론들에 의해 생성된 클러스터 1과 2[4]에 해당하는 토픽맵을 보여 주고 있으며, 클러스터 1내의 태그들을 토픽맵에서의 토픽들로 생성된 결과는 그림 10의 ①과 같다. 또한 3.2의 연관 토픽쌍 추출 및 3.3의 의미연관관계 부여를 통해 생성된 토픽들 간의 의미연관관계는 트리 형태로 표현된 ②와 그래프로 표현된 ④와 같이 생성되었으며, 4.3에서 생성된 각각의 토픽들에 부여된 이미지 리소스들은 ③과 같다.

이와 같이 생성된 토픽맵을 통해 computer 토픽의 part로는 토픽 keyboard와 monitor가, computer의 kind로는 laptop, mac, macintosh가 있으며, computer에 company로는 토픽 apple이 있다와 같은 서술적 표현이 가능하다.

의미론적 지식체계인 토픽맵은 의미연관관계를 통해 시맨틱 내비게이션을 제공 가능하다. 그 결과는 그림 12와 같으며, 사용자는 토픽 apple을 클릭하여 그와 연관된 토픽들 중 mac을 의미연관관계를 통해 kind라는 관계임을 알고 정보를 내비게이션 가능하다. 또한 토픽 mac과 연관관계로 구성된 토픽들

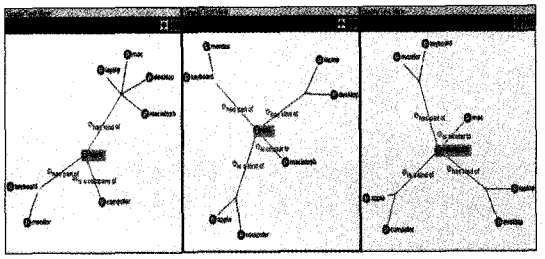


그림 12. 토픽맵 생성결과를 통한 시맨틱 내비게이션

중 similar라는 의미연관관계를 통해 mac과 macintosh는 동일한 의미를 가지는 토픽임을 인지하고 검색해 나갈 수 있다.

그림 13과 그림 14는 향상된 검색결과를 제공하기 위해 3.4에서 생성된 각각의 토픽들에 부여된 이미지 리소스들 중 본 논문의 실험 데이터(키워드 computer를 통해 수집된 1~5page의 이미지)로 사용된 그림 10(1 page)에 해당하는 리소스들[4]을 추출한 결과이다.

이 결과를 통해 기존 Flickr 사이트와 같은 태그 기반 사이트들은 단순히 키워드인 computer를 포함하는 태그들을 가진 이미지를 보여줌에 따라서 사용

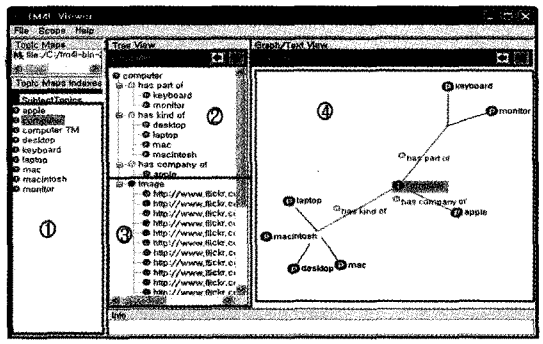


그림 10. 생성된 클러스터 1의 토픽맵

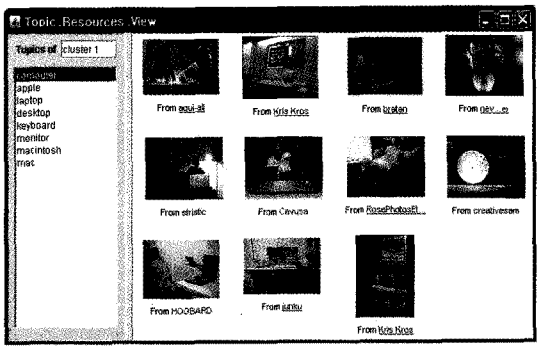


그림 13. 클러스터 1 토픽맵의 computer 토픽 리소스

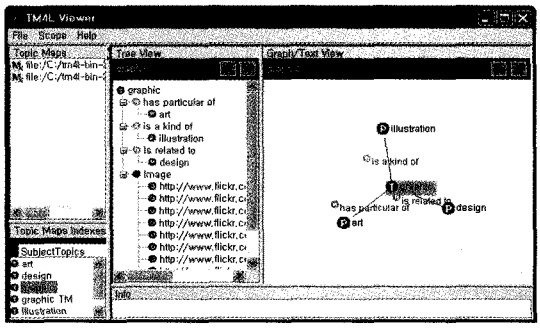


그림 11. 생성된 클러스터 2의 토픽맵

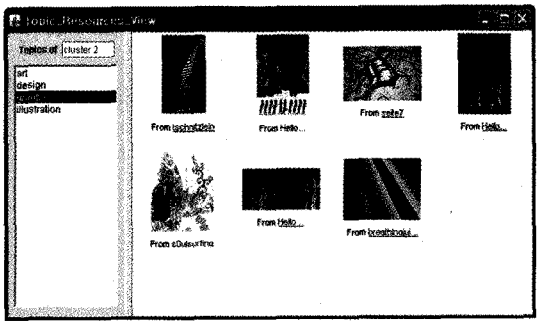


그림 14. 클러스터 2 토픽맵의 graphic 토픽 리소스

자가 의도한 키워드와 다른 이미지들을 많이 내포한 결과를 보여준다.

그에 반해 본 논문에서 제안한 시스템에 의해 추출된 결과는 사용자가 요청한 computer와 관련성이 높은 이미지들로 분류 가능할 뿐만 아니라 연관관계로 생성된 각각의 토픽들에 적합한 이미지 리소스가 부여된 결과를 도출하였다. 또한 Flickr 사이트의 잘못된 결과 또한 연관도가 높은 이미지들로 분류 시키는 향상된 결과를 도출하였으며, 그 결과는 그림 14와 같다.

4.3 실험 및 평가

본 논문에서는 제안한 시스템을 통해 기존 태그들이 가지는 문제점 중 두 번째 문제점인 비효율적인 정보 내비게이션의 문제점을 4.2의 의미론적 지식체계인 토픽맵 생성 및 검증을 통해 해결하였다.

본 절에서는 첫 번째 문제점을 해결하기 위한 방안으로 제시한 토픽맵 내에 토픽들에 부여된 정보 리소스들과 태그 기반 대표 사이트인 Flickr 검색결과와의 비교평가를 통해 향상된 검색결과를 제시한다.

비교평가 방법으로는, Flickr Open API를 이용하여 수집한 표 2에 해당하는 이미지들과 표 2의 데이터를 제안한 시스템에 적용하여 생성된 표 3의 computer, apple, jaguar관련 토픽맵의 computer, apple, jaguar 토픽에 각각 부여된 이미지들의 검색결과를 비교 평가하였다. 비교평가 기법으로는 정보검색 시스템의 성능 측정 지표로 활용되고 있는 정확성(precision)과 재현율(recall)평가 기법을 적용하였다[18].

여기에서 정확한 이미지란, 추출된 이미지들 중 이미지 내에 키워드(computer, apple, jaguar)와 관련된 이미지를 포함하고 있으면, "정확", 그렇지 않으면, "부정확"으로 정의하였다.

아래의 표 3은 각각 관련 토픽맵에서 토픽에 부여된 이미지들의 정확성 및 재현율을 실험한 결과다. 실험 결과 Flickr에서 수집한 1~5page에 해당하는 computer 관련 리소스(120개) 중 computer 토픽에 부여된 정확한 이미지는 59개, 부정확한 이미지는 61개로 정확성과 재현율은 각각 49%이다. 이에 반해 제안한 시스템에서는 리소스 중 추출된 이미지는 62개로 정확한 이미지는 54개, 부정확한 이미지는 8개로 정확성과 재현율은 각각 87%와 45%로 나타난다. 부정확한 이미지가 8개 추출된 이유는 사용자가 잘못 태깅하여 추출된 경우로, 제안한 토픽쌍을 이용한 검색 결과의 측면에서는 정확한 결과이다. 그러나 computer를 포함하지만 추출하지 못한 경우는 computer 하나만 포함하여 연관 태그쌍으로 추출되지 못한 경우로서, 이러한 제안 시스템의 문제점은 향후 해결해야하는 문제점이다.

또한 Apple과 Jaguar의 검색은 두 가지 관점으로 나눠 정확성과 재현율을 측정하였다. Flickr에서 fruit 관점의 Apple을 검색할 경우 각각 34%의 정확성과 재현율을 보이며 company 관점에서는 각각 53%를 보여준다. 그에 반해 제안 시스템의 경우 fruit 관점

표 1. 평가 항목

키워드	상위 1~5Page	태그 수
computer	120개의 이미지	836
apple	120개의 이미지	1017
jaguar	120개의 이미지	996

표 2. 표 1의 평가 항목을 제안 시스템에 적용한 결과

키워드	threshold 값	클러스터 수	생성된 토픽맵	토픽맵 내의 토픽
computer	9	2	computer 관련	computer, apple, laptop, desktop, keyboard, mac, monitor, macintosh
			design 관련	design, graphic, art, illustration
apple	7	2	apple (fruit 관점)	apple, fruit, green, red
			apple (company 관점)	apple, mac, macintosh, powerbook, ipod, imac
jaguar	9	2	jaguar (animal 관점)	jaguar, animal, zoo, cat, bigcat, black, specialanimal
			jaguar (car 관점)	jaguar, car, auto, e-type

표 3. computer, apple, jaguar토픽의 평가 결과

키워드	평가 항목	Flickr					제안 시스템						
		리소스 수	정확	부정확	정확성 precision	재현율 Recall	리소스 수	추출된 이미지 수	정확	부정확	추출하지 못한 이미지 수	정확성 precision	재현율 Recall
Computer	Computer 관련	120	59	61	49%	49%	120	62	54	8	5	87%	45%
Apple	Apple (fruit 관점)	120	41	79	34%	34%	120	43	38	5	3	88%	32%
	Apple (company 관점)	120	64	56	53%	53%	120	60	58	2	6	97%	48%
Jaguar	jaguar (animal관점)	120	66	54	55%	55%	120	65	62	3	4	95%	52%
	jaguar (car 관점)	120	45	75	38%	38%	120	52	44	8	1	85%	37%

에서는 88%의 정확성과 32% 재현율, company관점에서 97%의 정확성과 48%의 재현율을 보여준다. Jaguar의 경우 animal과 car의 관점으로 볼 수 있다. 이에 Flickr에서의 정확성과 재현율은 animal관점에서 각각 55%, car의 관점에서 각각 37%로 나타난다. 제안 시스템의 결과 animal 관점에서는 95%의 정확성과 52%의 재현율을 보여주며, car의 관점에서는 85%의 정확성과 37%의 재현율을 보여준다.

이 결과를 통해 Flickr 사이트의 정확성 및 재현율은 평균 45.8%로 평가되었다. 그에 반해 제안 시스템은 평균 90.4%의 정확성과 42.8%의 재현율로 평가됨에 따라, 재현율 면에서는 평균 3% 떨어지는 결과를 도출하였지만, 정확성 면에서는 평균 44.6%의 향상된 정확성을 도출하였다.

5. 결 론

본 논문에서는 선행연구 결과를 기반으로 의미론적 지식체계인 토픽맵 생성을 위한 시스템을 설계 및 구현하였다.

구현결과 기존 태그 기반 시스템의 비구조화된 태그를 통한 정보 내비게이션이 아닌, 토픽들 간의 의미연관관계를 활용한 시맨틱 정보 내비게이션이 가능하였다. 또한 실험 데이터들을 제안 시스템에 적용한 결과 부정확한 이미지에 태깅된 태그들을 클러스터링 과정을 통해 또 다른 토픽맵으로 생성하는 향상된 결과를 도출할 수 있었다.

제안한 시스템의 성능평가를 위해 태그 기반 대표 사이트 중 하나인 Flickr와의 이미지 검색결과의 정확성과 재현율을 비교평가 하였으며, 그 결과 재현율 면에서는 평균 3% 떨어지는 결과를 도출하였지만, 정확성 면에서는 평균 44.6%의 매우 향상된 검색결과를 도출할 수 있었다.

향후 연구 과제로는 워드넷을 통해 추출할 수 없는 토픽 쌍들의 연관관계에 대한 자동화생성기법 및 성능 평가에서 도출된 부정확한 이미지와 추출하지 못한 이미지를 해결하기 위한 연구를 진행할 예정이다. 또한 웹2.0이 정보의 생산과 공유의 초점을 맞췄다면, 다가올 웹3.0의 중심은 개인화인 만큼 제안한 시스템을 개인화하기 위한 연구가 필요하다.

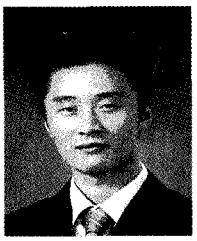
참 고 문 헌

[1] 정부연, "2006년 인터넷 화두 웹2.0(Web2.0)," 정보통신정책, 제18호, 제387호, 2006.

[2] 홍성태, 임일, "웹2.0 환경에서 정보 분류와 필터링, 그리고 협업을 위한 기술의 동향 및 발전 방향," *Telecommunications Review*, 제17권, 제4호, 2007.

[3] Time O'Reilly, "What is Web2.0," http://www.oreilly.net.com/pub/a/oreilly/time/news/20-05/09/30/what-is-web-20.html, 2005.

[4] 이시화, 무효려, 이만형, 황대훈, "web2.0 환경에서의 Tag Clustering 시스템 설계 및 구현," 한국멀티미디어학회 춘계 학술대회, 제10권, 제1호, 2007.
[5] 이강표, 김두남, 김형주, "웹2.0 환경에서의 태깅 기술 동향," *한국정보과학회지*, 제25권, 10호, 2007.
[6] 박영진, 송길영, 김경서, 송성환, "웹2.0과 정보 검색," *ITFIND 주간기술동향*, 제12권, 제5호, 2006.
[7] C. H. Brooks and Nancy Montanez, "Improved Annotation of the Blogosphere Via Autotagging and Hierarchical Clustering," International Conference on World Wide Web, 2006.
[8] 강필구, 채진석, "웹2.0을 위한 효율적인 태그 관리 시스템의 설계 및 구축," 한국정보과학회, 제33권, 제2호, 2006.
[9] C. H. Wiener, J. Golbeck, A. Schain, and J. A. Hendler, "Annotation and Provenance Tracking in Semantic Web Photo Libraries," International Provenance and Annotation Workshop, 2006
[10] Ontopia.net, "The TAO of Topic Maps," http://www.ontopia.net/topicmaps/materials/tao.html/
[11] G. R. Ramalho, J. C. Ramalho, and P. R. Henriques, "TM Bulder : An Ontology Builder based on XML Topic Maps," *Clei Electronic Journal*, Vol. 7, No. 2, 2004.
[12] 함화진, "토픽맵 반자동 구축도구의 설계 및 구현," 이화여자대학교 석사학위 논문, 2005.
[13] C. Kohler, A. Korthaus, and M. Schader, "Automatic Topic map Generation from a Conventional Document Index," IASTED International Conference, 2004.
[14] 이시화, 이만형, 김용수, 황대훈, "web2.0에서 효율적인 Tag Clustering을 위한 Threshold 선정에 관한 연구," 한국멀티미디어학회 춘계 학술대회, 제10권, 제2호, 2007.
[15] 이시화, 이만형, 황대훈, "Web2.0 환경에서의 효율적인 이미지 검색을 위한 태그 클러스터링 시스템의 설계 및 구현," 한국멀티미디어학회 논문지, 제11권, 제8호, 2008
[16] JWNL, http://www-nlp.stanford.edu/nlp/javadoc/jwnl-docs/net/didion/jwnl/JWNL.ht_ml, Princeton University.
[17] TM4L, http://compsci.wssu.edu/iis/nsdl/index.html, Winston Salem State University.
[18] S. M. Shafi and R. A. Rather "Precision and Recall of Five Search Engines for Retrieval of Scholarly Information in the Field of Biotechnology," Webology, Vol.2, No.2, 2005



이 시 화

2005년 서울보건대학 컴퓨터정보과 졸업
2005년 블루M 개발실 연구원
2007년 경원대학교 전자계산학과 석사과정 졸업
2008년~현재 경원대학교 전자계산학과 박사과정
관심분야 : e-Learning, Context- Aware, Semantic Web, Web2.0

이 만 형

1997년 건양대학교 정보통신학과 (학사)
1999년 경원대학교 전자계산학과 (석사)
2007년 경원대학교 전자계산학과 (박사수료)
1999년~현재 현대전문학교 교수
관심분야 : e-러닝, Semantic Web, 보안

황 대 훈

1997년 동국대학교 수학과(학사)
1983년 중앙대학교 전자계산학과 (석사)
1991년 중앙대학교 전자계산학과 (박사)
1983년~1985년 한국산업경제기술연구원(KIET) 연구원
1987년~현재 경원대학교 교수
2009년 한국멀티미디어학회 회장
관심분야 : e-러닝, Semantic Web, 유비쿼터스 컴퓨팅