블로그 검색에서의 태그 계층구조를 이용한 포스트 군집화

Post Clustering Method using Tag Hierarchy for Blog Search

이기준(Ki Jun Lee)[*], 김경민(Kyungmin Kim)[**], 이명진(Myungjin Lee)[***], 김우주(Wooju Kim)[****], 홍준석(June S. Hong)[*****]

초 록

웹 3.0으로 진화중인 웹 환경 하에서 블로그는 사용자 주도적인 웹의 특성을 가장 잘 표현하는 집합체 중 하나로, 기존의 웹 정보자원과 구분되는 새로운 형태의 지식베이스로써의 역할을 담당하고 있다. 기존의 웹 정보자원들이 사이트 단위로 광범위한 주제를 다루었던 것에 반해, 블로그의 정보자원은 사용자의 관심사에 따라 특정 정보들이 블로그 단위로 밀집되어 있으며 또한 사용자 태깅에 의해 게시된 정보자원에 대한 분류기준을 가지고 있다. 본 연구에서는 이러한 블로그의 특징들을 이용하여 보다 좀 더 효과적인 정보검색에 활용하기 위하여 블로그의 제목 키워드나 태그를 활용하여 태그 계층구조를 만들고 그 계층구조를 적용한 포스트군집화 방법론을 개발하여 기존의 블로그 검색과는 다른 특성을 가진 검색결과를 제시하였다. 이를 위하여 블로그 태그간의 관계성이 반영된 태그 계층구조를 생성하고, 태그 유사도에 따른 태그군집화 방법을 개발하였다. 본 논문은 제안된 방법론을 구현한 프로토타입 시스템을 통해 실제사례에서의 연구의 적용 가능성을 판단하였으며, 군집 유사도 평가기준인 CSIM(Cluster SIMilarity)을 사용하여 골든 스탠다드의 유사도 비교를 통해 개발된 방법론과 시스템의 성과를 평가하였다.

ABSTRACT

Blog plays an important role as new type of knowledge base distinguishing from traditional web resource. While information resources in their existing website dealt with a wide range of topics, information resources of the blog are concentrated in specific units of information depending on the user's interests and have the criteria of classification forresources published by tagging. In this research, we build a tag hierarchy utilizing title keywords and tags of the blog, and propose apost clustering methodology applying the tag hierarchy. We then generate the tag hierarchy reflected the relationship between tags

---

본 연구는 지식경제부와 한국산업기술진흥원의 전략기술인력양성사업으로 수행된 결과임.
이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 기초연구사업 지원을 받아 수행된 것임(2011-0027007).

* 팬택 중앙연구소
** 연세대학교 공과대학 정보산업공학과 박사과정
*** 연세대학교 공과대학 정보산업공학과 Post-Doc
**** 연세대학교 공과대학 정보산업공학과 교수
***** 교신저자, 경기대학교경상대학경영정보학과 부교수

2011년 09월 01일 접수, 2011년 10월 17일 심사완료 후 2011년 11월 18일 게재확정.

and develop the tag clustering methodology according to tag similarity. In this paper, we analyze the possibility of applying the proposed methodology with real-world examples and evaluate its performances through developed prototype system.

키워드 : 블로그 검색, 태그 계층구조, 포스트 군집화
Blog Search, Tag Hierarchy, Post Clustering

1. 서 론

웹 3.0으로 진화하고 있는 현재의 웹은 인터넷 사용자가 직접 만들어내는 게시물들로 구성되는 웹 사이트인 블로그의 숫자가 크게 증가하면서[26], 블로그는 인터넷상에서의 새로운 정보의 원천으로, 그리고 새로운 연구주제로 주목받고 있다[16]. 블로그의 포스트는 블로그의 소유자인 블로거에 의해 작성되며, 작성된 포스트들은 작성자의 주 관심사를 나타내고, 또한 그 분류는 작성자의 주관적인 판단에 따른다는 특징이 있다. 이와 같이 블로그 정보자원들은 주관적인 분류체계를 따르는 특징들 때문에 기존의 검색 엔진들이 제공하는 검색 알고리즘으로는 효율적인 블로그 검색에 한계가 있다. 이에 따라 현재 데이터 마이닝, 검색 엔진 알고리즘, 인공 지능 등의 여러 가지 기법을 사용한 블로그 검색 연구들이 진행되고 있으며 여러 분야의 연구주제로 새롭게 제시되고 있다.

블로그 검색에 관한 기존 연구들은 대부분 블로그들끼리의 관계, 블로거의 평판, 태그, 댓글, 디렉토리 등을 이용하여 검색을 수행하는 방법론을 채택하고 있다[31-34]. 이 중 블로그 소유자의 주관적인 분류체계가 반영된 태그를 이용한 검색 방법은 블로그 포스트의 핵심 주제어가 태그로 표현될 수 있다는 점에 착안한 검색 방법으로, 블로그의 특성을 제대로 검색에 반영할 수 있는 방법론 중 하나이다. 하지만 선행연구들 중에서 태그를 이용한 블로그 검색 방법론은 태그로 사용된 키워드들 간의 중복여부만을 이용하고 있을 뿐, 태그 키워드들 간의 유사성을 검색에 전혀 활용하지 못하고 있다[7]. 이에 본 논문에서는 태그 키워드 간의 유사성이 반영하기 위하여 태그 유사도에 따라 작성된 태그계층구조를 이용하여 블로그 검색결과의 군집화 작업을 수행하고자 한다. 태그의 계층구조는 블로그 포스트에서의 태그 간 동시출현 빈도를 반영하여 동적으로 생성되며, 이는 중복된 태그가 존재하지 않는 두 개의 포스트들일지라도 연관도가 높은 태그로 태깅되어 있다면, 그 두 개의 포스트는 유사한 내용을 포함한다고 판단할 수 있다.

본 논문은 태그 계층구조를 이용한 블로그 검색 방법론의 제시를 위해 제 2장에서 관련 연구들을 소개하고 그 속에서 본 연구의 발전 방향을 기술한다. 제 3장에서는 본 연구의 검색 알고리즘과 프로토타입 시스템을 제시하며, 제 4장에서 프로토타입 시스템을 이용하여 수행한 블로그 검색 실험을 분석하여 제시된 알고리즘의 타당성을 평가한다. 마지막으로 제 5장에서는 연구를 종합하여 결론을 제시하고 향후 연구 주제들을 소

개하고자 한다.

2. 관련 연구

2.1 문헌조사

블로그 검색에 관련된 연구는 현재 다양하게 진행 중이며, 각각의 연구의 특징은 <표 1>에 요약된 바와 같다. <표 1>에서 나타난 것과 같이 국외 연구로는 크게 블로그 검색을 위한 효율적인 쿼리 생성방법론에 관한 연구[20], BLOGRANGER를 통해 생성된 다양한 지표를 활용하는 블로그 검색연구[13], 포스트나 블로거의 피드백 정보를 이용한 검색 연구[28], 태그 클러스터링을 통한 의미적으로 연관된 태그추출 연구[8] 등이 있으며, 국내 연구로는 태그 온톨로지를 이용한 표준화된 태그 추천[3], 태그 유사도를 통한 블로그를 추천[5], 태그 네트워크를 이용한 개인화 북마크 추천[6], 사용자의 태그를 분석한 스패머 판별 연구[4], 태그 및 트랙백을 이용한 블로그 검색 실험[2], 블로그의 사용자 태그를 이용한 블로그 자동분류 성능 향상 연구[1] 등이 있다.

본 연구는 '태그는 사용자의 의도가 반영된 키워드로서 태그 유사도를 통해 산출된 태그 계층구조는 의미기반 검색에 활용될 수 있다'는 전제에서 시작한다. 이미 여러 분야에서 태그의 이러한 특성을 이용한 다양한 연구는 존재한다. 예를 들어, 사용자가 사용한 태그를 분석함으로써 사용자가 고의적으로 시스템을 악용하는 스패머인지를 그 판별하는 연구[4]가 있으며, 사용자 태그를 이용한 개인화된 북마크 추천에 관한 연구[5]가 있다. 이러한 연구들은 태그를 통해 사용자의 의도 혹은 사용자의 취향을 분석하여 적용한 연구이다. 이와 마찬가지로 본 연구에서는 동일한 포스트에 붙이는 태그에 대한 사용자 의도가 포스트와 관련된 연관성이 있는 키워드들의 나열이라고 가정하고, 결과적으로 포스트에 붙어있는 사용자 태그를 통해 도출된 태그 계층구조는 의미적으로 연관성이 있는 키워드들의 계층 구조라고 가정한다.

본 연구의 목적은 의미기반 블로그 검색을 가능하게 하기 위하여 태그 계층구조를 이용한 블로그 검색결과에 대한분류이다. 이는 사용자 키워드에 따른 블로그 검색 이후, 결과로 나열된 블로그 포스트들을 유사한 결과들끼리 묶어서 제시함으로써, 결과적으로 의미기반 블로그 검색을 가능하게 한다. 앞서 조사한 블로그 관련 연구들은 블로그 검색 자체에 초점을 맞추거나[13, 20, 28], 태그 클라우드 형성 혹은 연관 태그 추천[8, 3], 계층구조를 지원하지 않는 태그 네트워크를 활용한 검색 및 분류[5, 6, 2, 1]에 초점을 맞추고 있지만, 본 연구는 검색엔진을 통한 '키워드 기반 블로그 검색결과'를 '의미기반 블로그 검색결과'로 나타내기 위하여 태그 유사도를 이용하여 작성된 태그 계층구조를 활용한 블로그 검색결과의 그룹화 및 재순위화에 주안점을 둔다. 물론 재순위화를 통해 검색 능력을 향상시키는 관련 연구들도 존재하지만[6, 2, 1], 기존의 연구는 계층구조가 없는 태그 네트워크를 사용하는데 반해, 본 연구는 태그 계층구조를 이용한 블로그 군집화를 이용함으로써 의미기반 블로그 검색을 가능하게 하고자 한다.

〈표 1〉 국내외 관련연구 동향

종류	관련연구		
국외 연구	Mishne and Rijke[20]	목적	블로그 검색을 위한 효율적인 쿼리 생성 방법론
		특징	일반 웹 페이지와 블로그의 차이를 세부적으로 제시하여 그 차이로부터 효율적인 검색을 위한 쿼리 생성 방법을 제시
	Fujiki et al.[13]	목적	BLOGRANGER를 통해 생성된 다양한 지표를 활용하는 블로그 검색연구
		특징	블로그의 제목과 블로거, 블로거의 평판을 수치화하여 좀 더 실용성 있는 블로그 정보를 제공
	Takama et al.[28]	목적	포스트나 블로거의 피드백 정보를 이용한 검색 연구
		특징	블로거의 블로그 정보가 반영된 키워드 맵을 통해 검색을 수행하여 사용자 중심의 검색 결과를 얻을 수 있음
	Begelman et al.[8]	목적	태그 군집화를 통한 태깅 서비스의 향상
		특징	태그 공간의 태그들을 군집화 함으로써, 질의어 태그에 대한 의미적으로 관련이 있는 태그들을 제공. 결과적으로, 사용자의 태그 공간에서의 검색 혹은 탐색의 효율성을 향상
국내 연구	김재승 외[3]	목적	태그 온톨로지를 이용한 표준화된 태그 추천
		특징	이미 존재하는 문서에 대한 자동 태깅과, 신규 추가되는 문서에 대한 태그 추천 방법을 제시 함.
	심학준 외[5]	목적	태그 유사도 측정을 통해 사용자가 이미 구독중인 블로그와 유사한 블로그를 추천
		특징	각 블로그에 달려 있는 태그들에 대해 유사도를 계산하는 식을 제안하고, 블로그의 유사성을 평가하고 추천
	엄태영 외[6]	목적	태그 네트워크를 이용한 개인화 북마크 추천
		특징	본 논문은 북마크 검색에 대해 개인화된 검색결과를 추천하기 위해 사용자 태그를 기반으로 딜리셔스가 제공하는 북마크들의 순위를 재순위화 하는 방법론을 제안
	김찬주, 황규백[4]	목적	사용자의 태그를 분석하여 소셜북마킹을 이용하는 사용자가 스패머인지를 판별
		특징	사용자 태그에 대해 기계학습의 다양한 방법을 적용하여 사용자가 스패머인지 아닌지를 예측하는 모델을 생성하고 각기 다른 기계학습법에 대한 성능 비교
	김은희, 정영미[2]	목적	태그 및 트랙백을 이용한 블로그 검색 실험
		특징	사용자가 블로그 부여한 태그 및 트랙백을 이용하여 블로그 페이지의 검색 실험을 수행하여 가장 향상된 검색 결과를 나타내는 방법론을 제시
	김기현, 정영미[1]	목적	블로그의 사용자 태그를 이용한 블로그 자동분류 성능 향상
		특징	사용자의 태그나 태그 확장을 적용하여 블로그 자동분류 성능을 향상 시킬 수 있음을 제시

2.2 정보검색 기법

본 절에서는 블로그 포스트들의 어휘빈도 비교를 통해 태그 계층구조를 생성하며, 태그 계층구조를 통한 블로그 군집화를 수행하기 위하여 정보검색이론 연구분야에서 적용되는 어휘가중치 산출기법들과 군집화 알고리즘을 활용하고 있다.

2.2.1 가중치 산출기법

어휘의 가중치를 산출하는 방법은 그 접근방법에 따라 크게 어휘가 가지는 시소러스나 텍소노미의 계층구조를 이용하는 '경로기준(Path-Based)', '정보량기준(Information content-Based)' 방법과 어휘를 표현한 벡터 간의 비교를 통해 어휘가중치를 산출하는 '용어정의기준(Gross Based)', '벡터기준(Vector Based)' 방법으로 나누어진다[30]. 각 접근방법을 개략적으로 설명하면 다음과 같다.

- 경로기준 방법 : 사전에 정의된 계층구조(시소러스 혹은 텍소노미)로부터 어휘간 거리를 이용하여 가중치를 산출한다[17, 22, 29].
- 정보량기준 방법 : 사전에 정의된 계층구조로부터 어휘가 가지는 정보량을 계산하여 가중치를 산출한다[24].
- 용어정의기준 방법 : 시소러스에 포함된 용어정의를 이용, 용어정의에서 중복되는 어휘의 빈도를 계산하여 가중치를 산출한다[18].
- 벡터기준 방법 : 벡터로 표현된 어휘들의 코사인 내적을 통해 유사도 및 가중치 산출한다[33].

일반적으로 시소러스는 소수의 도메인 전문가들에 의해 설계되므로 개념간의 관계 정의(예를 들어 상위어, 하위어, 유사어)에 있어 신뢰성과 무결성을 보장한다. 또한 이미 정의된 계층구조를 사용하므로 어휘가중치 및 유사도 계산이 간단하다는 장점이 있다. 이에 많은 '경로기준 방법'과 '정보량기준 방법'의 가중치 산출방법을 적용한 선행연구들이 'WordNet'과 같은 시소러스를 이용하여 수행되었다. 그러나 미리 정의된 시소러스를 이용하여 어휘 가중치를 산출하는 방법은 컨텍스트 내에서의 어휘간의 함의적인 관계를 파악하기에는 한계가 있고, 사전에 구조가 정의되어 있지 않으면 적용할 수 없다는 단점이 있다.

어휘 가중치 산출을 위해 또 다른 접근방법으로 '용어정의기준 방법'와 '벡터기준 방법'의 가중치 산출 방법이 있다. 이 방법들은 어휘를 벡터구조로 표현하기 때문에 사전에 시소러스나 텍소노미를 생성하지 않아도 벡터 비교를 통하여 어휘에 가중치를 할당할 수 있다는 장점이 있으나, 벡터 비교를 통하여 얻어진 어휘간의 의미론적 관계가 정확히 무엇인지는 파악하기 힘들다는 단점이 있다. 예를 들어, 벡터 비교를 통하여 관계가 있다고 파악된 두 어휘가 정확히 유사어관계인지 또는 상·하위어 관계인지를 식별하는데 어려움이 있다.

본 연구에서는 위에 설명한 각 접근방법의 장단점을 반영하여 태그 가중치를 산출하였다. 기본적으로 본 연구에서의 태그 가중치는 태그의 동시출현 빈도를 이용하여 계층구조를 생성한 후, 태그 계층구조에 '경로기준 방법'을 적용하여 가중치를 계산하는 방법을 사

용하였다.

2.2.2 군집화 알고리즘

본 연구에서는 블로그 포스트의 분류를 위하여 각 포스트의 제목을 벡터 형태로 표현한 후 벡터 형태로 표현된 포스트의 클러스터링을 수행한다. 클러스터링이란 텍스트 마이닝과 정보검색이론 분야의 연구에서 많이 적용되는 대표적인 비지도학습 방법이며, 군집화 방법에 따라 계층구조 클러스터링(Hierarchical Clustering), 분할구조 클러스터링(Partitional Clustering), 그리고 스펙트럼구조 클러스터링(Spectral Clustering)으로 나누어진다[36].

가장 일반적으로 사용되는 클러스터링 알고리즘은 상향식 계층군집화(Agglomerative hierarchical clustering) 알고리즘과 K-means 알고리즘이 있는데[26], K-means 알고리즘은 상향식 계층 클러스터링보다 성능은 떨어지지만 효율적인 군집화 알고리즘으로 평가되고 있다[25, 12]. 본 연구의 블로그 검색은 군집화 성능보다 효율성에 더 초점을 맞추어 K-means 알고리즘을 본 연구의 군집화 알고리즘으로 적용하였다.

K-means 알고리즘은 벡터 공간내의 K개의 중심점은 K개의 클러스터로 표현할 수 있다는 전제를 가지고 수행되며, 일반적으로 각 벡터의 코사인 유사도 산출을 통해 클러스터를 형성한다. K-means 알고리즘은 클러스터를 생성하는 방법에 따라 bisecting, antipole, multipole 등의 다양한 확장 알고리즘이 존재하지만[26, 10], 본 연구에서는 기본적인 K-means 알고리즘을 적용하였다.

3. 블로그 검색 시스템

블로그 검색의 결과인 포스트들을 제시하는데 있어 현재의 검색엔진들은 검색대상 리소스 측면과 블로그 검색목적 측면에서 블로그 검색의 한계를 보여준다. 이는 현재의 검색엔진들이 검색대상과 검색목적 측면에서 범용성을 중심으로 개발되어 있기 때문이다.

검색대상 측면에서 보면, 현재의 검색엔진들은 웹 상의 다양한 정보를 모두 다루기 위한 방향으로 설계되어 있어, 블로그 포스트와 같이 특성이 있는 자료들을 효과적으로 분류하는 것에 적합하지 않다. 예를 들어, 블로그는 블로거에 따라 개성적인 주제와 분야가 탄생하고 신조어도 끊임없이 창출되어 일정한 기준으로 분류하기 어려운 특징을 가지고 있을 뿐만 아니라, 블로그는 개인 블로거들에 의해 제한 없이 다양한 형태로 관리되기 때문에 범용성 중심으로 개발된 현재의 디렉터리 분류구조를 이용한 검색엔진들은 효율적인 블로그 검색을 수행하는데 한계가 있다.

또한 블로그 검색목적 측면에서 보면, 사용자들이 블로그 검색을 하는 경우는 일반적인 웹 검색으로 얻을 수 있는 전문적이고 체계적인 뉴스나 저널, 스폰서 링크 얻기 보다는 주관적인 관점으로 작성된 정보를 얻기 위한 경우가 많다. 이렇게 검색목적 측면에도 대체적으로 블로그 검색은 일반 웹 검색과는 다른 목적을 갖기 때문에 계층적 분류 시스템을 이용한 일반 웹 검색엔진들이 효과적인 블로그 검색 결과를 제시하기에는 한계가 있다.

본 연구에서는 이러한 한계점을 극복하기 위해 정적인 디렉토리 구조가 아닌 동적인 태그 계층구조를 활용하여 블로그 검색 결과

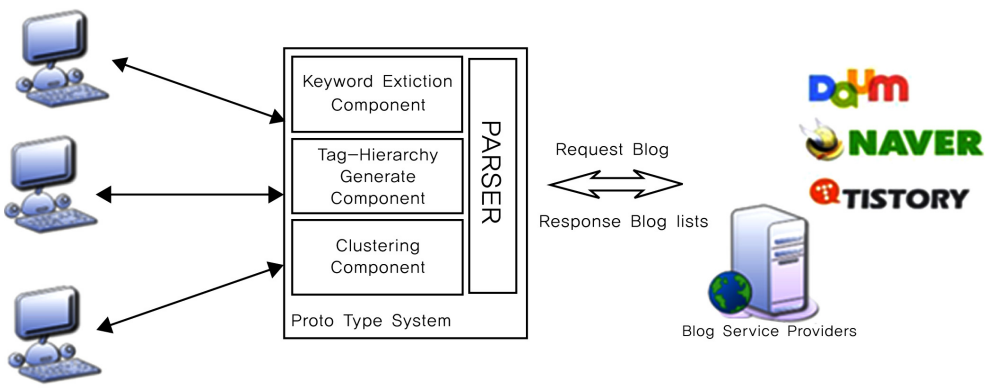
인 포스트 군집화를 수행하고자 한다. 즉, 어떤 키워드에 의해 검색된 포스트들이 일률적으로 이미지, 동영상 등으로 분류되기 보다는 사용자들이 올린 블로그 포스트들을 분석하여 가장 많이 반복된 주제어들을 추출하여 카테고리를 구성하고, 그것들을 그 검색어에 대한 카테고리로 삼아 포스트를 클러스터링 한다. 따라서 정해진 카테고리에 맞춰 분류하는 작업없이 사용자는 그 검색어가 가진 모든 화제어(카테고리 표제어)들을 확인할 수 있을 것이며, 그 화제어에 맞추어진 포스트 클러스터들을 제공받을 수 있을 것이다. 추가로 클러스터링 된 각각의 그룹을 분석하여 일종의 순위와 통계적 수치를 제공함으로써 '어떤 화제어가 해당 검색어에 대해가장 많이 포스팅 되었는가'와 같은 추가적인 정보도 한눈에 확인할 수 있게 된다. 위에 제시한 검색 목표를 달성하기 위해 본 논문의 알고리즘은 아래와 같이 3가지 전략을 가지고 블로그 검색을 수행한다.

- 검색엔진이 제공하는 정적인 카테고리를 통해 클러스터링을 하기 보다는 검색어의 특징을 반영하는 동적인 카테고리들로 포스트군집화를 수행하여 검색어 자체의 분포를 반영한다.
- 현재의 날짜순으로 보여주는 결과 제시가 아닌 해당 검색어에 관하여 가장 많이 포스팅되어 있고 가장 중요하고 가장 인기있는 자료를 우선적으로 볼 수 있는 클러스터링 체계를 마련한다.
- 검색어를 입력하였을 때 제공되는 결과로부터 사용자는 한 눈에 해당 검색어에 대한 추세와 분포를 파악할 수 있어야 한다. 따라서 중복되는 결과를 줄여 제시하는 것이 필요하다.

3.1 시스템 구성도

본 논문의 프로토타입 시스템은 포탈사이트로부터 블로그 정보를 요청한 후 검색을 수행한다.

프로토타입 시스템은 다음 검색엔진의 결과를 그대로 사용하며, 다음과 같은 절차를



〈그림 1〉 블로그 검색시스템 구성도

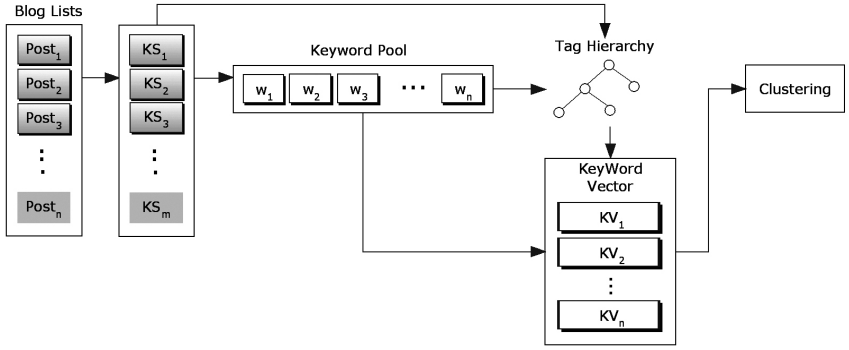
수행한다. 우선 사용자의 질의어에 따라 시스템은 블로그 서비스 제공자로부터 블로그 검색 결과인 포스트 리스트를 파서로 받아온다. 파서(Parse)에서는 검색된 결과들 중 프로토타입 컴포넌트에서 필요로 하는 정보만(블로그 포스트 제목, 태그)을 추출하여 키워드 추출 컴포넌트(Keyword Extraction Component)에 전달한다. 키워드 추출 컴포넌트는 토크나이징, 불용어 제거, 중복어휘제거 작업등을 수행하고, 이를 통해 블로그의 포스트 제목과 태그를 키워드 벡터로 변환하여 태그 계층구조생성을 위한 선처리 과정을 수행한다. 태그 계층구조 생성 컴포넌트에서는 키워드 추출 컴포넌트에서 생성된 키워드 벡터를 이용하여 키워드간의 동시출현 빈도를 산출함으로써 태그 계층구조를 생성한다. 태그 계층구조가 생성되고 나면클러스터링 컴포넌트에서 생성된 계층구조를 반영하여 포스트의 군집화를 수행한다.

3.2 블로그 검색 알고리즘

본 연구에서 K-means 알고리즘은 태그 계층구조의 생성과 키워드 벡터의 군집화에 사용된다. K-means를 적용하여 관련 있는 정보들을 군집화하기 위해서는, 우선 키워드에 키워드간의 관련성이 반영된 수학적 수치를 할당하는 작업이 선행되어야 하는데[19], 본 연구에서 이 과정은 벡터를 형성하고 벡터 간의 거리공식을 적용함으로써 달성된다.

<그림 2>는 알고리즘의 수행 절차를 나타낸다. 먼저 사용자 쿼리에 따른 블로그 검색 리스트를 검색엔진으로부터 파싱하면서 알고리즘은 수행된다. 알고리즘은 파싱된 검색결과로부터 블로그 포스트의 제목과 태그들을 추출하여 형태소 분석을 실시하며, 그것을 바탕으로 전체 키워드 벡터 풀을 만든다. 키워드 풀이 생성되면 키워드 집합과 비교하여 태그 계층구조를 생성하고, 다시 각각의 제목을 벡터 풀과 대조하여 키워드 벡터를 만든다. 키워드 벡터를 만드는 작업까지 완료되면 그 벡터를 중심으로 K-means 클러스터링을 수행하며, 그 결과들을 정렬하여 태그 계층구조를 반영한 검색결과를 생성한다.

알고리즘에서 사용될 용어를 정의하면 다음과 같다.



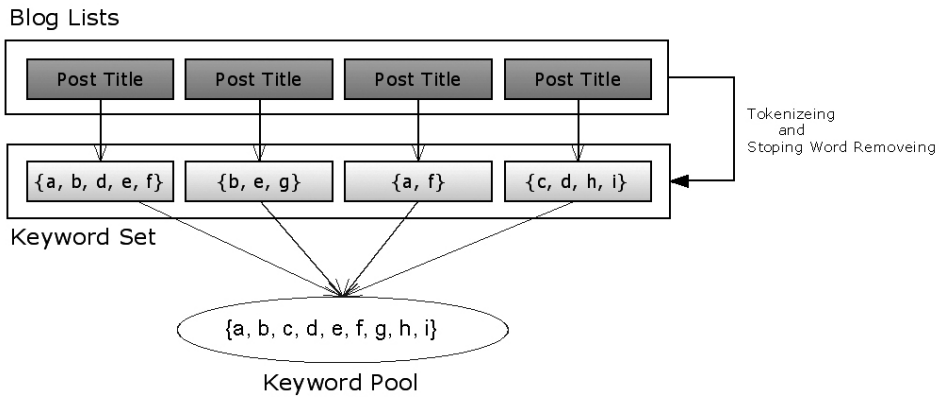
〈그림 2〉 포스트 군집화 알고리즘의 수행 단계

- 정의 1 : $T = \{t_1, t_2, \cdots, t_n\}$: 검색 결과로 주어진 블로그 포스트를 가리킨다. 블로그 검색엔진으로부터 검색결과를 받으면 그들의 포스트를 각각 순서대로 $t_1, t_2, \cdots, t_n$으로 정의한다. 만약에 블로그 검색엔진의 검색 결과가 100개라면 n = 100이 된다.
- 정의 2 : 키워드 집합(Keyword Set) : 블로그 포스트의 제목과 태그들로부터 벡터의 요소를 추출하고 불필요한 부분들을 제거하기 위해 형태소 분석을 시행한다. 이 때 각 블로그의 제목과 태그들을 형태소 단위로 분리하여 나열함으로써 하나의 블로그는 일정한 형태소 그룹을 이루게 된다. 이 그룹을 키워드 집합으로 정의하며, $KS = \{ks_1, ks_2, \cdots, ks_n\}$로 정의한다. 숫자 n은 각각의 블로그로부터 추출되므로 정의 1에 정의된 n의 값과 일치한다. 그리고 이러한 키워드집합은 모두 형태소, 즉 단어들로 구성되어 있으므로 각각의 키워드 집합은 $KS = \{w_1, w_2, \cdots, w_i\}$로 표현된다.
- 정의 3 : 키워드 풀(Keyword Pool) : 키워드 집합에 포함된 모든 형태소들을 중복없이 나열하여 만든 키워드 풀은 KP로 정의하고 $KP = \{w_1, w_2, \cdots, w_m\}$로 나타낸다. 키워드 풀은 현재 검색결과의 제목과 태그들에 포함된 모든 형태소들을 하나씩 가지고 있으며 KS와의 비교가 가능한 형태가 된다. 숫자 m은 검색 결과가 가지고 있는 모든 형태소의 개수를 의미한다.
- 정의 4 : 워드 벡터(Word Vector) : 키워드 풀에 속한 각각의 키워드들이 블로그 포스트 제목과 태그들에 동시 출현된 빈도를 산출하여 워드벡터 WV = {가중치 1, 가중치 2, …, 가중치$_m$}의 가중치를 할당한다. 워드벡터는 클러스터링을 수행하여 태그 계층구조를 생성하는데 이용한다.
- 정의 5 : 키워드 벡터(Keyword Vector) : 키워드 벡터는 각 블로그에 포함되어 있는 형태소의 벡터를 나타낸다. KP에 매칭되는 형태소소유여부에 따라 0(포함되지 않은 경우) 이나 1(포함된 경우)의 값을 가지게 된다. 예를 들어 KS와 KP를 비교할 때, KS_i란 키워드 집합이 키워드 풀의 KP_j에 해당하는 형태소를 가지고 있다면 KV_{ij}값은 1이 된다.

태그 계층구조의 목적은 태그들 사이의 연관성 관계를 기반으로 계층구조를 작성하고, 직접적으로 연관이 없어 보이는 태그들도 관련도를 가지게 함으로써 좀 더 진보된 클러스터링 결과를 제공하는 것이다. 이러한 목적 달성을 위해 태그 계층구조는 블로그 포스트의 제목과 태그들에 대하여 형태소 분석 및 불용어 제거, 키워드 집합 생성, 키워드 풀 생성의 절차를 통해 키워드 간의 관계를 반영하여 생성되며 태그 계층구조는 다음의 절차를 통해 생성된다.

• Step 1 : 형태소 분석 및 불용어 제거

검색 엔진에서 제공받은 블로그 검색결과를 가지고 K-means를 수행하기 위해 블로그 자료들 은클러스터링 수행에 적합한 형태로 변환하여야 한다. 이를 위해 포스트의 제목과 태그들을 벡터로 변환하는 수치화 작업을 진



〈그림 3〉 키워드 풀 생성 예시

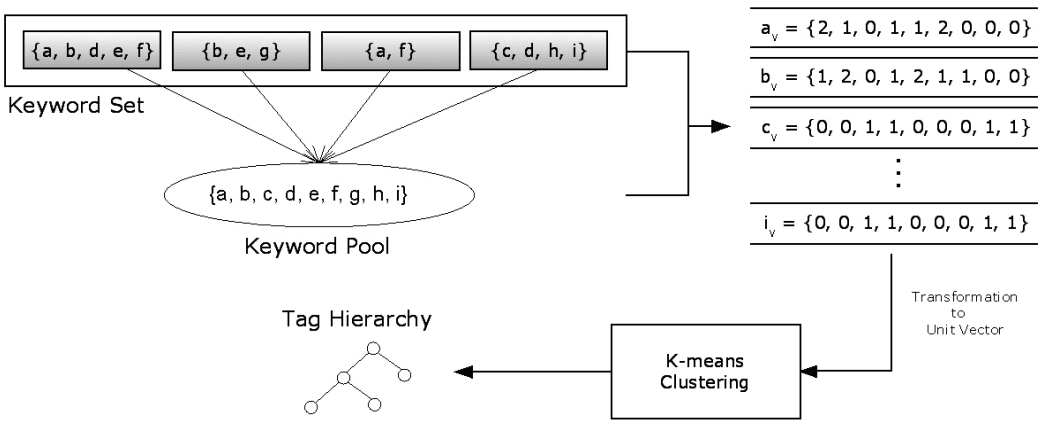
행한다. 수치화 작업을 위하여 포스트 제목과 태그들에 대한 형태소 분석을 실시하며 형태소 하나하나가 한 벡터의 차원을 이루게 된다. 해당 형태소의 포함 여부에 따라 각각의 벡터 요소들은 1 또는 0의 값을 가지게 된다. 본 연구는 형태소 분석기를 사용하여 제목에서 형태소를 분석하게 되는데 이러한 과정에서 블로그의 특성상 자주 등장하게 되는 이모티콘을 없애고, 노이즈를 일으켜 효율성을 떨어뜨리는 한 글자짜리 형태소나 조사와 관사 같은 단어들도 제거된다.

• Step 2 : 키워드 풀 생성

포스트들 간의 거리측정 및 비교 가능한 기준은 벡터 곱을 통해 달성된다. 이를 위해 모든 형태소에 대한 정보를 포함하고 있는 표준벡터에 대한 정의가 필요하며, 이러한 표준벡터로 모든 포스트로부터 추출된 키워드 풀(KP)을 사용한다. 키워드 풀은 해당 검색어에 대한 블로그 검색을 통해 나온 키워드들을 중복 없이 모두 담고 있는 풀을 의미하며, <그림 3>은 키워드 풀의 생성 예시를 보여준다. 그림으로 나타난 바와 같이 키워드 풀은 포탈 검색엔진으로부터 얻은 블로그 리스트를 전 처리하여 포스트별로KS을 얻은 후, KS에 들어 있는 모든 키워드의 형태소들을 순서대로 KP에 입력하면서 중복된 형태소를 제거함으로써 얻는다.

• Step 3 : 태그 계층구조 제작

태그 계층구조는 앞서 제작한 키워드 풀을 기반으로 제작된다. 태그 계층구조 생성을 위한 벡터는 키워드 풀의 요소를 서로 비교하면서 임의의 태그들이 동일한 KS에 포함된 경우가 몇 번 있는가에 따라 벡터 값을 부여한다. 즉, 동일한 블로거가 하나의 포스트 제목과 태그들에서 임의의 두 키워드를 동시에 발생한 정도에 따라 벡터 값이 부여된다. 이러한 방식을 따라 산출된 각각의 키워드벡터들은 단위벡터로 변환 후 K-means를 통해 클러스터링 되고, 군집 분석을 통해 유사도가 높은 벡터들끼리 그룹화하여 몇 개의 태그 그룹으로 분류한다. 그리고 다시 각각의 그룹



〈그림 4〉 태그 계층구조 생성 예시

에 대해 위와 같은 클러스터링 작업을 반복적으로 적용함으로써 전체 태그 집합에 대한 태그 계층구조를 생성한다.

<그림 4>는 태그 계층구조의 생성 예시를 나타낸다. 태그 계층구조의 생성을 위한 워드 벡터는 KS에서의 각 태그의 동시출현 빈도를 통해 산출된다. 예를 들어, 태그 a는 전체 KS에서 두 번 출현하므로 a는 2가 되고, {b, d, e, f}는 a와 동시에 한 번씩 출현하므로 1의 값을 가지게 된다. 즉, 키워드 풀 KP의 모든 태그들에 대하여 설명된 방법으로 WV를 산출하여 보면, <그림 4>와 같이 av = {2, 1, 0, 1, 1, 2, 0, 0, 0}, bv = {1, 2, 0, 1, 2, 1, 1, 0, 0}, cv = {0, 0, 1, 1, 0, 0, 0, 1, 1}, …., iv = {0, 0, 1, 1, 0, 0, 0, 1, 1}의 벡터들을 얻을 수 있으며, 이 벡터들을 단계적으로 반복하여 클러스터링하여 계층적인분류 작업이 수행되고, 이러한 작업의 결과로 태그 계층구조를 구축한다.

• Step 4 : 키워드 벡터의 계산

마지막으로 키워드 벡터를 만드는 과정을 통해 군집화 알고리즘을 수행하기 위한 모든 준비 과정이 끝나게 된다. 이 과정을 통해 각각의 포스트가 서로 얼마나 닮은 정보를 가지고 있는지 유추해 볼 수 있는 기준을 마련하게 된다.

예를 들어 예제와 같은 키워드 풀 KP = {a, b, c, d, e, f, g, h, i}가 있는 경우에 키워드 벡터를 만들기 위한 KS_1 = {a, b, d, e, f}과 KS_2 = {b, c, f, g}는 키워드 풀과의 비교를 통해 KS_1과 KS_2의 KV는 각각[1, 1, 0, 1, 1, 1, 0, 0, 0]과 [0, 1, 1, 0, 1, 1, 0, 0]이 된다. 이에 더하여 본 연구에서는 태그 계층구조를 키워드 벡터에 반영하기 위해 계층구조를 바탕으로 계산된 값을 사용한다. 이 때 부여되는 벡터 값은 Leacock-Chodorow Measure[17]를 사용하였다. Leacock-Chodorow Measure는 0부터 무한대까지의 값으로 어휘 유사도를 표현하기 때문에, 본 연구에서처럼 키워드벡터의 가중치를 업데이트하는 용도로 적용하기 위해서는 Measure를 수정할 필요가 있다. 이에 본 연구에서는 Hyperbolic Tangent를 이용하여 정규화를 수행하며, 범

위조정을 위해 조정가중치인 α(0≤α≤1)를 곱하여 태그간의 정규화 된 유사도 값을 산출한다. 변환된 유사도 산출 공식은 다음과 같다.

Distance (tag1, tag2) = (length(tag1, tag2) +1)/((2×d)+1)
where,
length(tag1, tag2) : 계층구조에서 tag1과 tag2의 가장 짧은 경로의 길이
D : 계층구조의 최대깊이

Similarity(tag1, tag2) = α×tanh(1/Distance (tag1, tag2)-1)
where,
α : 조정가중치 (상수 0.2를 사용)
tanh(x) : Hyperbolic-Tangent

Max Similarity : 유사도 값 중 최대 유사도 값

유사도 공식을 통해 산출된 각 가중치들은 포스트의제목과 태그에는 포함되지 않았지만, 대상 포스트와 관련성이 존재한다고 추정되는 태그의 가중치이다. 산출된 태그 가중치 값이 클수록 포스트와 관련성이 높은 키워드가 된다.

<그림 5>와 <그림 6>은 태그 계층구조를 적용한 키워드벡터 KV'_1의 생성 예를 나타낸다. <그림 5>와 같이 키워드 풀이 KP = {a, b, c, d, e, f, g, h, i}이고 키워드 집합 KS_1 = {a, b, d, e, f} 이면, 키워드 벡터는 KV_1 = {1, 1, 0, 1, 1, 1, 0, 0, 0}으로 표현된다. 그리고 키워드벡터 KV_1에 태그 계층구조가 반영된 KV'_1을 산출하기 위해 <그림 6>에서 제시한 유사도 공식을 이용하여 c_{weight}, g_{weight}, h_{weight}, i_{weight}을 산출한다. c_{weight}의 산출 예를 살펴보면, a-c(태그a에서 태그c까지의 길이), b-c, d-c, f-c는 모두 같은 Diatance를 가지고, a-c는 e-c보다 짧은 Diatance를 가지므로c_{weight}의 Max Similarity는 similarity(a, c)가 된다. 즉 c_{weight} = (0.2)×tanh(1/((2+1)/(2×2+1))-1)로 0.116이 된다. 같은 방법으로 g_{weight}, h_{weight}, i_{weight}를 산출하여 KV_1에 반영하여 태그 계층구조가 적용된 KV'_1을 산출한다.

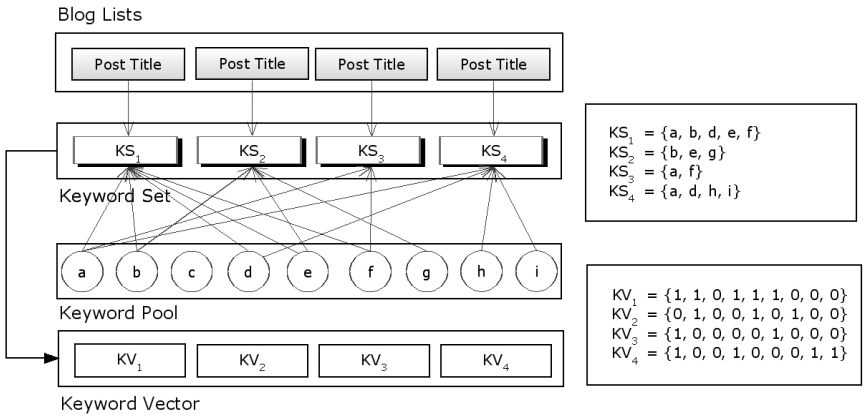
모든 KV_n은 태그 계층구조에 의해 KV'_n으로 갱신한 후, 단위벡터로 표준화시켜 군집화 알고리즘을 적용하기 위한 준비과정을 완료한다.

• Step 5 : 포스트 군집화

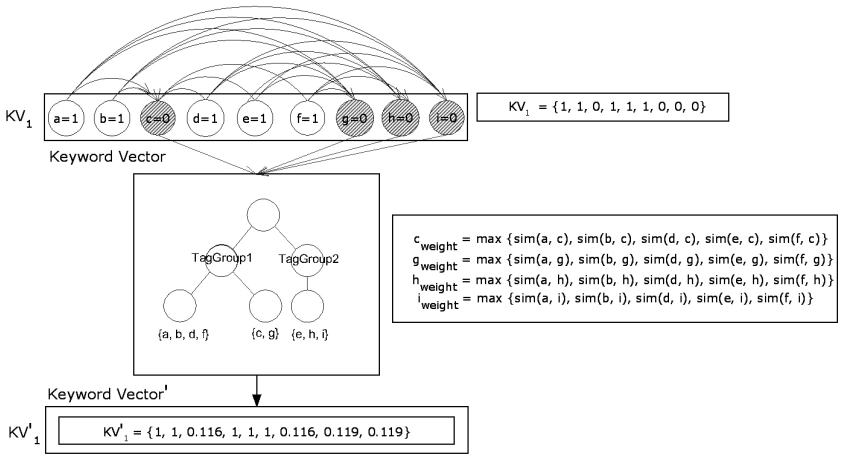
■ 첫 클러스터링 K값의 결정
K-means 클러스터링 적용할 때, 군집화 결과 그룹의 개수(K)는 군집화 성능에 중요한 영향을 미친다. 대부분의 경우 K는 정적인 변수로 설정되어 그 개수에 맞추어 클러스터링 결과를 도출하도록 하고 있지만 매 번 다른 개수의 블로그와 다른 방향의 주제어들을 가지는 검색어에 맞추어 일정한 K값을 정하는 것은 매우 어려운 일이다. 따라서 본 연구에서는 검색어에 따라 각기 다른 K값을 두고자 하였으며 그 이상적인 K값을 본 연구에서는 “Gold Standard”를 통해 산출하였다. -실험에 수행하기 이전에 실험대상이 되는 블로그들을 미리 선정하여 하나하나 방문하여 보며

그룹화를 진행하였으며 이 때 선정된 K값을 각각 실험군의 K값이 되도록 설정하였다.

- 중심 값 결정과 클러스터링의 반복
  K-means는 최적의 클러스터가 생성될 때까지 반복하여 군집을 재생성한다. 즉, K-means에 의해 일단 클러스터가 생성되면 클러스터된 모든 그룹들에 대하여 중심 값을 다시 계산하고 모든 KS를 대상으로 K개의 클러스터에 순서대로 벡터 곱을 수행한다. 그 후 수행결과 벡터 곱이 가장 컸던 그룹에 다시 그 KS를 배정하며 마지막 KS까지 벡터 곱과 배정 과정을 수행한다. 이로 인해 새로운 클러스터링 결과가 산출되면 다시 중심 값을 구하는 과정을 반복 수행하여 최적의 클러스터링 상태를 이루도록 한다.



〈그림 5〉 키워드 벡터생성 예시



〈그림 6〉 태그 계층구조가 반영된 키워드 벡터생성 예시

■ 종료시점의 결정

K-means가 중단되는 임계값을 구하는 것도 K-means를 다루는데 있어 중요한 요소 중 하나이다. 최적의 클러스터링는 새롭게 형성된 결과들이 직전의 클러스터링 결과와 비교하였을 때 아무런 변화가 일어나지 않았을 때이다. 그러므로 클러스터링을 반복적으로 진행할 때 그룹의 포스트의 개수를 세는 변수를 마련하여 각 그룹의 개수정보를 유지함으로써 최적의 종료시점을 결정하였다.

• Step 6 : 군집화된 검색결과의 제시

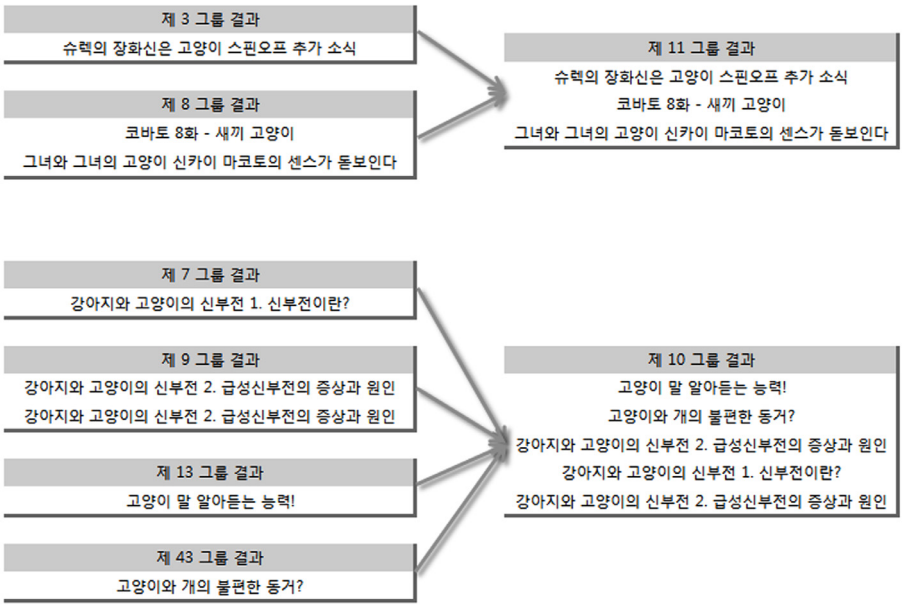
클러스터링의 결과제시 방법은 블로그 검색목적이 반영되어야 하는 핵심요소 중에 하나이다. 본 연구에서는 클러스터링된 결과를 그룹의 크기에 따라 순위를 부여하는 방법을 선택하였는데, 이를 통해 사용자들은 해당 검색어에 따라 군집화된 그룹별로 포스트의 수가 어떻게 분포하는지 좀 더 빠르고 쉽게 판단할 수 있게 된다. 클러스터의 순위화는 가장 크기가 크게 클러스터링된 그룹을 먼저 제시함으로써 가장 필수적인 포스트들를먼저 제시하도록 하고, 또한 모든 그룹의 옆에는 전체 검색 결과 중에 이 그룹이 차지하는 비율을 표시하여 각 그룹의 중요도에 대한 추가적인 정보를 제공한다. 결과적으로 최종 검색 결과를 제시받은 사용자는 어떠한 화제가 얼마만큼 중요한지, 그 검색어에 대해서는 중요하게 다루어지는 문제들이 몇 가지나 있는지를 그 키워드의 카테고리에 따라 빠르게 파악할 수 있다.

4. 시스템 성능평가

제 3장에서 제시한 알고리즘과 시스템에 대한 성능 및 실효성을 평가하기 위해 본 연구에서는 프로토타입을 구현하였다. 그리고 작성된 알고리즘의 효과를 평가하기 위한 방법으로 Cluster SIMilarity(CSIM)[13]가 사용되었다. CSIM은 Rand's method[23]에서 고안된 것으로 군집화 된 그룹들의 유사도를 판단할 때 사용된다.

본 연구에서는 정답 클러스터와 실험군의 CSIM을 비교하여 논문에서 제시하는 태그계층 구조를 이용한 블로그 검색 알고리즘의 성능을 평가하였다. 첫 번째 실험군은 사용자가 자신의 블로그의 특징을 표현하고 분류하기 위해 달아놓은 제목과 태그들의 리스트만을 사용하였다. 두 번째는 태그를 단순히 반영하기보다는 본 연구의 목적인 태그들 사이의 관계를 기반으로 계층 구조를 작성하여 직접적으로 연관이 없어 보이는 태그들도 관련도를 가지게 하여 좀 더 진보된 클러스터링 결과를 제공하기 위한 실험군으로 하였다.

블로그 검색 대상은 블로그 전문 사이트인 티스토리[35]와 다음블로그를 대상으로 하였으며, 검색어는 블로그의 다양한 주제를 고려하여 여러 가지 분야에 걸쳐 포스팅이 되어 있고 블로그들이 개성을 살려 다양한 포스팅을 하고 있는 '고양이'라는 단어를 선정하였다. 실험 대상이 되는 블로그는 수작업 강도를 고려해 200개로 한정하였고 티스토리와 다음의 블로그에서 무작위로 선정하였다. 본 연구의 평가 알고리즘으로 사용한 CSIM방법은 우선 정답이 되는 클러스터링 결과와 비교대상으로써 프로토타입을 거쳐 나오는 결



〈그림 7〉 태그 계층구조를 통한 성능 향상의 예

과가 필요하다[19]. 이 두 결과를 비교하여 얼마나 두 그룹이 일치하는지를 수치로 알려주는데 1에 가까운 숫자를 가질수록 두 개의 일치도가 높다는 것을 나타낸다. 정답이 되는 클러스터링 결과는 직접 하나하나의 블로그를 방문해가며 제목이나 태그에 연연하지 않고 모든 면을 판단하여 작성되었으며 똑같은 200개의 자료를 클러스터링 하였다.

	실험군 1	실험군 2
CSIM	0.715	0.724

실험결과, 실험군 1의 경우(태그 리스트만을 사용) 0.715, 실험군 2는(태그 계층구조를 적용) 0.724의 CSIM 값을 나타냈다. 태그 계층구조를 사용한 실험군 2의 수치는 그냥 태그만 적용한 실험군 1보다 향상된 군집화 성능을 보여주었다. 성능 향상의 정도가 미약한 이유는 사용자 태그의 개수 부족을 보완하기 위하여 포스트 제목에서 추출한 키워드들을 동시에 사용함으로써 분류체계에 대한 정확하지 않은 정보가 추가되었기 때문으로 판단된다. 그러나 구체적인 성능 향상의 예로 <그림 7>의 그룹 3과 그룹 8은 실험군 2에서는 제목이 다르고 태그 그룹도 달라 다른 그룹으로 분류되었던 포스트들이 실질적으로 애니메이션을 다루고 있다는 점에서 한 그룹으로 묶일만한 포스트들이었고 태그 계층 구조를 이용한 실험군 2에서는 위와 같은 그룹들이 하나의 그룹으로 묶이게 되었다는 점을 들 수 있다. 실험 결과는 본 연구에서 제안하는 방법론이 텍스트 기반만으로는 잡아줄 수 없는(즉, 관련 주제는 비슷하지만 포스트의 태그나 제목의 키워드가 다른) 블로그들을

하나의 그룹으로 군집화 할 수 있는 방법을 제공하고 있는 것이다.

5. 결 론

본 논문은 블로그 검색 결과를 그룹화하여 제시함으로써 여러 사용자들에게 좀 더 유용한 정보를 줄 수 있을 것이라는 가정하에 블로그 검색 알고리즘을 제시하였다. 블로그 검색 알고리즘은 블로그 태그간의 관계성이 반영된 태그 계층구조의 생성과 그 활용을 위해 키워드 유사도 산출방법을 기반으로 키워드유사도에 따른 키워드 군집화 방법을 제시하였다. 제시된 방법론의 평가를 위해본 논문에서는 프로토타입 시스템의 구현하고, 실제 사례에서의 실효성을 평가 하였으며, 또한 CSIM(Cluster SIMilarity)을 사용하여 본 연구결과와 골든스탠다드의 유사도 비교를 통해 제시한 알고리즘의 성능을 평가하였다.

본 연구에서 제시한 방법론은 블로거들이 블로그 포스트에 적절한 제목과 태그를 달지 않으면 성능을 보장할 수 없다는 단점이 존재하지만, 제 4장에서 보여주는 것처럼 블로그 포스트에 포함되지 않은 태그들도 관련성 정도에 따라 가중치를 할당하여 잠재적으로 유사성이 있는 포스트들의 군집화를 가능하게 한다. 현실적으로 본 논문에서 제시하는 알고리즘이 잠재적 유사성이 있는 모든 포스트의 태그들을 검색할 수는 없지만 알고리즘의 결과로 제시된 포스트로부터 연관 검색어를 역산출 할 수 있는 수단으로 사용될 수 있을 것이다. 또한 각 기업들도 자신들의 이미지나 새로 출시한 상품에 대한 대중의 의견이 궁금한 경우에, 위의 알고리즘을 이용한 검색을 시도한다면 다른 웹 사이트가 제공해 주지 못한 여러 주관적인 정보(사용자의 관점이 반영된 정보)들을 손쉽게 확인할 수 있을 것이다.

참 고 문 헌

[1] 김기현, 정영미, “이용자 태그 확장을 통한 블로그 자동분류 성능 향상에 관한 연구”, 제16회 한국정보관리학회 학술대회 논문집, pp. 43-48, 2009.

[2] 김은희, 정영미, “사용자 태그와 중심성 지수를 이용한 블로그 검색 성능 향상에 관한 연구”, 정보관리학회지, 제27권, 제1호, pp. 61-77, 2010.

[3] 김재승, 문현정, 우용태, “태그 온톨로지를 이용한 자동 태깅 및 태그 추천 기법”, 한국전자거래학회지, 제14권, 제4호, pp. 167-179, 2009.

[4] 김찬주, 황규백, “소셜 북마킹 시스템의 스패머 탐지를 위한 기계학습 기술의 성능 비교”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 제15권, 제5호, pp. 345-349, 2009.

[5] 심학준, 윤태복, 이지형, “메타정보를 활용한 블로그 추천방법”, 한국지능시스템학회 2010 년도 춘계학술대회 학술발표 논문집, 제20권, 제1호, pp. 96-97, 2010.

[6] 엄태영, 김우주, 박상언, “태그 네트워크를 이용한 개인화 북마크 추천시스템”, 한국전자거래학회지, 제15권, 제4호, pp.

181-195, 2010.

[7] 이기준, 이명진, 김우주, "주제 유사성 기반 클러스터링을 이용한 블로그 검색 기법 연구", 한국지능정보시스템학회, 제 15권, pp. 61-74, 2009.

[8] Begelman, G., Keller, P., and Smadja, F., "Automated tag clustering : Improving search and exploration in the tag space," Citeseer, 2006.

[9] Broder, A., "A taxonomy of web search," pp. 3-10, 2002.

[10] Cantone, D., Ferro, A., Pulvirenti, A., Recupero, D. R., and Shasha, D., "Antipole tree indexing to support range search and k-nearest neighbor search in metric spaces," Knowledge and Data Engineering, IEEE Transactions on, Vol. 17, pp. 535-550, 2005.

[11] Chung, Y. M. and Lee, J. Y., "A corpus based approach to comparative evaluation of statistical term association measures," Journal of the American Society for Information Science and Technology, Vol. 52, pp. 283-296, 2001.

[12] Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W., "Scatter/gather: A cluster-based approach to browsing large document collections," pp. 318-329, 1992.

[13] Fujiki, T., Nanno, T., Suzuki, Y., and Okumura, M., "Identification of bursts in a document stream," pp. 55-64, 2004.

[14] Grahl, M., Hotho, A., and Stumme, G., "Conceptual clustering of social bookmarking sites," pp. 356-364, 2007.

[15] Gurevych, I., "Using the structure of a conceptual network in computing semantic relatedness," Natural Language Processing-IJCNLP 2005, pp. 767-778, 2005.

[16] Kumar, R., Novak, J., Raghavan, P., and Tomkins, A., "On the bursty evolution of blogspace," World Wide Web, Vol. 8, pp. 159-178, 2005.

[17] Leacock, C. and Chodorow, M., "Combining local context and WordNet similarity for word sense identification," WordNet : An electronic lexical database, Vol. 49, pp. 265-283, 1998.

[18] Lesk, M., "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," pp. 24-26, 1986.

[19] MacQueen, J., "Some methods for classification and analysis of multivariate observations," p. 14, 1967.

[20] Mishne, G. and Rijke, M. de, "A study of blog search," Advances in Information Retrieval, pp. 289-301, 2006.

[21] Patwardhan, S. and Pedersen, T., "Using WordNet-based context vectors to estimate the semantic relatedness of concepts," Making Sense of Sense : Bringing Psycholinguistics and Computational Linguistics Together, p. 1, 2006.

[22] Rada, R., Mili, H., Bicknell, E., and Blettner, M., "Development and application of a metric on semantic nets," Sys-

tems, Man and Cybernetics, IEEE Transactions on, Vol. 19, pp. 17-30, 1989.

[23] Rand, W. M., "Objective criteria for the evaluation of clustering methods," Journal of the American Statistical association, Vol. 66, pp. 846-850, 1971.

[24] Resnik, P., "Using information content to evaluate semantic similarity in a taxonomy," Arxiv preprint cmp-lg/9511007, 1995.

[25] Sarle, W. S., "Algorithms for Clustering Data," Vol. 32, ed: JSTOR, pp. 227-229, 1990.

[26] Steinbach, M., Karypis, G., and Kumar, V., "A comparison of document clustering techniques," pp. 525-526, 2000.

[27] Sun, A., Suryanto, M. A., and Liu, Y., "Blog classification using tags : An empirical study," pp. 307-316, 2007.

[28] Takama, Y., Kajinami, T., and Matsumura, A., "Application of keyword map-based relevance feedback to interactive blog search," pp. 112-115, 2005.

[29] Wu, Z. and Palmer, M., "Verbs semantics and lexical selection," pp. 133-138, 1994.

[30] Zesch, T. and Gurevych, I., "Wisdom of crowds versus wisdom of linguists-measuring the semantic relatedness of words," Natural Language Engineering, Vol. 16, pp. 25-59, 2010.

[31] (2011/04/26), Bloglines. Available : http://www.bloglines.com/.

[32] (2011/04/26), Blogpulse. Available : http://www.blogpulse.com/.

[33] (2011/04/26), BLOGRANGER. Available : http://ranger.labs.goo.ne.jp.

[34] (2011/04/26), BlogWatcher. Available : http://blogwatcher.pi.titech.ac.jp.

[35] (2011/04/26), Tistory. Available : http://www.tistory.com/.

[36] (2011/04/26), Wikipedia. Available : http://en.wikipedia.org/wiki/Cluster_analysis.

저 자 소 개

이기준 (E-mail : backdale@hotmail.com)
2008년 연세대학교 컴퓨터산업공학과 (학사)
2010년 연세대학교 정보산업공학과 (석사)
현재 팬택 중앙연구소
관심분문야 웹서비스, 유비쿼터스 웹서비스, 검색엔진

김경민 (E-mail : milren78@gmail.com)
2008년 UTS(University of Technology Sydney) Information Technology(e-business management) (석사)
현재 연세대학교 정보산업공학과 박사과정
관심분문야 시맨틱웹 환경의 의사결정 지원시스템, 시맨틱웹마이닝

이명진 (E-mail : xml@yonsei.ac.kr)
2010년 연세대학교 정보산업공학과 (박사)
현재 연세대학교 정보산업공학과 Post-Doc
관심분문야 시맨틱웹과 의미기반검색, 시맨틱웹 포털

김우주 (E-mail : wkim@yonsei.ac.kr)
1987년 연세대학교 BBA과정 (학사)
1994년 KAIST 경영과학 (박사)
현재 연세대학교 정보산업공학과 교수
관심분문야 시맨틱웹, 시맨틱웹 환경의 의사결정지원시스템, 시맨틱웹 마이닝, 지식관리 및 인공지능웹서비스

홍준석 (E-mail : junehong@kyonggi.ac.kr)
1989년 서울대학교 경영 (학사)
1991년 KAIST 경영과학 (석사)
1997년 KAIST 경영과학 (박사)
현재 경기대학교 경영정보학과 부교수
관심분문야 시맨틱 검색, 온톨로지 추론, 지능형에이전트, 자동협상시스템