신뢰할 수 있는 온라인 커뮤니티 기반의 효율적인 웹 검색 시스템 설계

정회원 김 영 안*, 박 상 관**° 종신회원 이 상 훈*

Design of Efficient Web Search System Based on Reliable Online Community

Young-an Kim*, Sang-kwan Park**° *Regular Members*, Sang-hoon Lee* *Lifelong Member*

요 약

웹상에 존재하는 정보의 량이 방대해 질수록 사용자가 원하는 정보를 찾는데 더 많은 노력이 필요하게 되었다. 따라서, 원하는 정보를 보다 정확하고 효율적으로 검색하기 위한 많은 검색 방법들이 연구되고 있다. 하지만 기존 검색방법들은 사용자의 검색 의도를 충분히 반영하지 못하기 때문에 검색 의도에 부합되는 적합한 정보를 찾기 어려운 단점을 가지고 있다. 따라서 사용자들의 검색 의도에 맞는 정보를 효율적으로 검색하기에 부족하다.

본 논문에서는 웹 사용자들의 관심사와 검색 의도를 파악하여 이를 검색에 반영함으로써 정보를 효율적으로 검색하기위한 방안을 제안한다. 이를 위해 형성된 커뮤니티들 중 신뢰할 수 있는 온라인 커뮤니티를 식별하고, 식별된 커뮤니티를 기반으로 검색을 수행한다. 실험결과 제안하는 방법이 기존 방법보다 검색의 정확성 측면에서 향상된 것을 확인할 수 있다.

Key Words : Web users, Interest, Information Search, Interest, Reliable Online Community, accuracy

ABSTRACT

As the amount of information on the web augments, more efforts become subsequently necessary to find the information that users want. Researchers have studied and tried a number of methods to find the wanted information more accurately and effectively, It is hard to find the information that suits to what users really want to find on the web, because previous search methods have not reflected users' aim in searches.

This paper suggests an effective method in information searching by reflecting web users' interests and aim in searches. This process involves distinguishing the reliable online communities and searching based on the distinguished communities. Newly suggested method shows an improvement in accuracy in comparison to the previous methods.

Ⅰ. 서 론

최근 웹 검색 시스템에서 사용자의 질의(Query)에 대한 검색 결과의 정확도를 향상시키기 위한 연구들이 활발하게 진행되고 있다. 대표적인 연구로는 제한 검색(Limit Search), 포커스 크롤러(Focused Crawler), 웹 문서 클러스터링(Web Document Clustering) 등이 있다. 제한 검색은 검색 범위를 특정 사이트 또는 도메인으로 한정시켜 검색 결과를 제공하는 방법이며, 포커스 크롤러는 질의가 주어진

* 국방대학교 국방과학학과 교수(roundsun@kndu.ac.kr),(hoony@kndu.ac.kr), ** 공군본부(parksangkwan@naver.com), (° : 교신저자)
논문번호 : KICS2011-09-404, 접수일자 : 2011년 9월 19일, 최종논문접수일자 : 2012년 1월 13일

시점에 질의와 관련 있는 웹 페이지들만을 수집하여 결과로 반환하는 방법이다[1-2]. 웹 문서 클러스터링은 많은 양의 웹 페이지(또는 사이트)를 서로 관련 있는 페이지 그룹을 형성하는 방법이다[3]. 그러나 위에서 설명한 방법들은 다음과 같은 단점을 가지고 있다. 먼저 제한 검색은 검색의 범위를 URL에 의해 명시되는 사이트 또는 도메인들로만 제한할 수 있을 뿐이며, 의미적으로 관련된 사이트들로 제한할 수 없다. 또한 포커스 크롤러는 질의 시점에 웹 페이지들을 수집하기 때문에 질의에 대한 결과를 획득하기까지 처리 시간이 오래 걸린다. 웹 문서 클러스터링은 많은 양의 웹 페이지 또는 사이트를 대상으로 복잡한 처리를 수행하므로 공간적·시간적 비용이 크다.

일반적으로 인터넷 상에서의 온라인 커뮤니티의 정의는 "네티즌들이 직접 정보를 생산, 공유하고 이들이 모여 활동할 수 있는 인터넷 상의 공간"이다[4]. 본 논문에서는 이와 같은 온라인 커뮤니티를 대상으로 웹 서비스 기술을 적용함으로써 의미적으로 관련된 사이트를 찾지 못하는 제한 검색의 문제점과 포커스 크롤러 및 웹 문서 클러스터링이 가지는 질의 처리 시간에 대한 문제를 해결함으로서 검색의 효율성을 향상시키는 방안을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들에 대하여 설명하고, 3장에서는 제안방안인 신뢰할 수 있는 온라인 커뮤니티 기반의 효율적인 검색 방법에 대해 기술한다. 4장에서는 제안방안을 기존의 일반적인 검색 방법과 비교 평가하고, 마지막 5장에서는 논문의 결론을 맺는다.

Ⅱ. 관련연구

2.1. 커뮤니티

커뮤니티라고 하는 것은 개인 사용자들 간의 상호 교류를 뜻하며, 웹 2.0시대의 도래와 함께 다양한 형태의 온라인 커뮤니티들이 형성되고 발달하게 되었다. 이러한 온라인상의 커뮤니티를 통해 커뮤니티에 속한 사용자들은 서로 자신들의 관심 정보를 손쉽게 공유할 수 있게 되었다. 커뮤니티를 통한 사용자 간의 교류는 사용자의 능동적 참여가 가능한 양방향성의 특징을 가지고 있다. 커뮤니티는 개방적인 특성을 가지고 있기 때문에 해당 커뮤니티에 부합하는 목적을 지닌 어떠한 사용자들도 참여 할 수 있으며, 사용자들이 공유하는 데이터 또한 커뮤니티의 목적이나 취지에 주로 적합한 것들이다.

또한 커뮤니티는 추천이나 태깅(Tagging)과 같은 기술들이 접목 되면서 개인 포털화 시대를 이끌어 내었으며, 이러한 포털화된 커뮤니티는 사용자들의 다양한 의견과 그들이 제공하는 뉴스 기사나 구독자들의 피드백(Feedback)이 되어 정보의 빠른 공유를 가능하게 하고 있다.

2.2. 제한 검색

제한 검색은 그림 1과 같이 중앙 데이터베이스에서의 기본 동작 방식은 서비스 기술, 서비스 등록 및 발견, 서비스 간의 통신 관점에서 정의 된다.

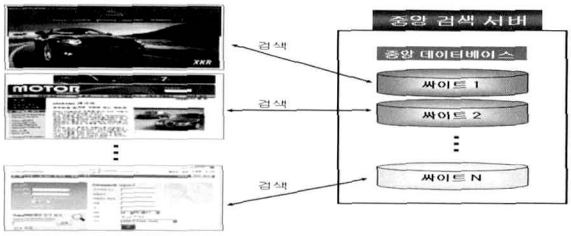


그림 1. 사이트 제한 검색의 개념
Fig. 1. Concept of site limit search

제한 검색을 사용하기 위하여 웹 사이트 관리자가 자신의 사이트를 검색 시스템에 등록하면, 웹 로봇[5]이 등록된 사이트에 포함되어 있는 웹 페이지를 수집하여 중앙 데이터베이스에 저장한다. 이때 어떤 사이트로부터 수집된 웹 페이지인지 나타내는 정보를 함께 저장하며, 사이트 제한 검색 요청이 들어오면 이 정보를 사용하여 해당 사이트로부터 수집된 웹 페이지에 대해서만 검색을 수행한다. 사이트 제한 검색 기능을 사용하면 웹 사이트에 검색 엔진을 설치하지 않고도, 마치 웹 사이트에 검색 엔진을 설치하여 운영하는 효과를 볼 수 있다.

2.3. 웹 문서 클러스터링

웹 문서 클러스터링은 웹 문서를 대상으로 기존 명사 위주의 키워드 뿐 아니라 인명, 지명, 회사명, 물품명 등을 자동으로 인식하는 개체명 인식 결과를 클러스터링 자질로 활용하는 방법이다[6].

MST(Minimum Spanning Tree) 클러스터링은 MST를 서브트리들로 나누어 클러스터를 구하는 방법이다. MST 클러스터링은 클러스터 개수를 사전에 정하지 않아도 클러스터를 구할 수 있으며 여러 가지 데이터 분포에서도 잘 동작하는 장점이 있다.

2.4. 웹 서비스

웹 서비스는 웹상에서 정의된 모듈화 된 소프트웨어 컴포넌트로서, 개방형 표준 데이터 표현 기법

인 XML과 인터넷 프로토콜을 결합시킨 분산 컴퓨팅 기술이다. 웹 서비스는 SOAP[1], WSDL[2], UDDI[3]를 통해 SOA의 주요요소인 메시지, 서비스 인터페이스, 서비스 공개 및 발견 체계를 구현한다. 따라서 웹 서비스는 SOA 구축에 필요 한 표준 기술들을 제공한다. 그림 2에서 볼 수 있듯이 웹 서비스의 기본적인 아키텍처는 SOA를 채택하고 있다.

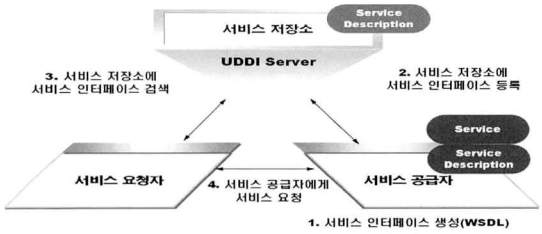


그림 2. 웹 서비스 아키텍처
Fig. 2. Web service architecture

Ⅲ. 커뮤니티 기반 웹 검색 체계

제안하는 검색 시스템은 'RSS Address Manager Tool'과 'Client View'의 두 개 프로그램으로 구성되어 있다[7]. 'RSS Address Manager Tool'은 검색에 필요한 커뮤니티를 등록하고 커뮤니티 내 정보를 읽어 들여 분석 한 후 정제된 단어 형태로 UDDI라 불리는 레지스터에 저장하기 위한 프로그램이다. 즉 웹 사용자가 검색을 수행하기 이전까지 정보를 정제하여 준비시키는 단계로써 모든 절차는 자동으로 수행된다. 'Client View'는 야후, 구글, 네이버 등과 같이 사용자가 검색을 수행하기위해 사용하는 인터페이스이다. 즉, 웹 사용자는 'Client View'를 이용하여 검색을 수행하게 된다.

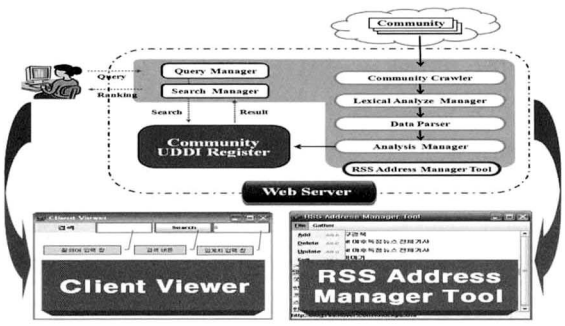


그림 3. 커뮤니티 기반 웹 검색 시스템 구조
Fig. 3. Community based web search system structure

1) SOAP(Simple Objective Access Protocol): 웹 서비스에서 메시지 전달 기반
2) WSDL(Web Service Description Language): 웹 서비스 기술 언어
3) UDDI(Universal Description Discovery Integration): 웹 서비스 관련 정보의 공개와 탐색을 위한 표준

3.1. RSS 크롤러

RSS[4] 크롤러는 웹 서비스의 제공자(Provider) 역할을 담당한다. RSS 크롤러는 사전에 관리자에 의해 수동으로 등록된 RSS 주소들을 통해 각 커뮤니티에서 문서들을 읽어 들여 각 커뮤니티 특징을 추출하는 기초 자료들로 사용된다. RSS 크롤러는 수집한 각 문서들을 색인하기 위해 형태소 분석 관리자(Lexical Analysis Manager)와 상호작용 하며, 주기적, 순차적, 병렬적으로 동작한다.

3.2. 형태소 분석 관리자

형태소 분석 관리자는 RSS 크롤러로부터 전달된 각 문서들을 형태소 분석하고 분석 결과를 데이터 파서에게 전달하는 역할을 한다. 형태소 분석은 한글 형태소 분석기(HAM: Hangul Analysis Module)를 활용하며, 기본적인 어휘 분석과 함께 불용어 제거(Stop-word), 어근 추출(Stemming) 과정을 거친다. 분석된 형태소들은 각 커뮤니티의 특징 벡터를 만들기 위해 데이터 파서(Parser)에게 전달된다.

가중치 값을 통해 선택된 각 단어들은 다시 벡터 모델로 표현되어 검색을 위해 사용될 수 있도록 UDDI에 저장되는데 그 형태는 다음과 같다.

$$<t0 : tf0 : df0, \cdots; ti : tfi : dfi, \cdots; tn : tfn : dfn> \quad (1)$$

*ti*는 *i*번째 단어(Term)을 의미하고, *tfi*는 *ti*가 발생하는 빈도수를 나타낸다. 위와 같은 과정을 통해 최종적으로 뽑힌 최상위 가중치의 값을 가진 단어를 통해 커뮤니티에 있는 정보가 어떤 정보를 담고 있는지를 알 수 있다.

3.3. 데이터 파서

데이터 파서는 RSS 크롤러로부터 읽어 들인 각 커뮤니티의 각 문서에 대해 형태소 분석기를 통해 얻어진 각 문서의 형태소 집합을 벡터 모델로 표현한다. 각 문서들을 벡터 모델로 표현함으로써 해당 커뮤니티의 특징을 결정하기 위한 연산을 비교적 쉽게 할 수 있다. 예를 들면 각 커뮤니티의 특징을 결정하기 위해 각 커뮤니티에서 수집된 문서들에 공통적으로 많이 나타나는 키워드나 단어들을 비교적 쉽게 찾을 수 있다. 데이터 파서는 이러한 연산들을 비교적 쉽게 할 수 있도록 각 문서에 대한 벡터 모델을 다음과 같이 구성한다.

4) RSS(Really Simple Syndication): 뉴스나 블러그 사이트에서 주로 사용하는 콘텐츠 표현 방식

<t0 : tf0, ···; ti : tfi, ···; tn : tfn> (2)

*ti*는 *i*번째 단어를 의미하고, *tfi*는 *ti*가 발생하는 빈도수를 나타낸다. 벡터모델로 표현된 문서들은 해당 커뮤니티의 특징을 추출하기 위해 분석 관리자에게 전달된다.

3.4. 분석 관리자

분석 관리자는 데이터 파서로부터 전달된 각 문서에 대한 벡터들에 근거해서 해당 커뮤니티의 특징을 결정하는 단어들의 집합을 추출하고 이를 다시 새로운 벡터 모델로 표현하고 검색에 사용할 수 있도록 UDDI에 저장하는 기능을 한다. 분석 관리자는 각 커뮤니티의 특징을 결정짓기 위해 해당 커뮤니티의 모든 문서들에서 가장 일반적으로 발생하는 단어들을 추출한다. 이를 위해 DF(Document Frequency)[5]를 이용하며 이는 각 문서들이 벡터 형태로 표현 되어 있어 비교적 쉽게 연산된다.

시스템에서는 DF값이 일정 임계값 이상이 되는 단어에 대해서만 특징값으로 사용한다. 또한 단어들의 가중치 *w*를 계산(TF-DF : Term Frequency, Document Frequency)해서 최상위 가중치 값을 가진 단어들을 선택한다. 이때, 가중치 값에 대한 임계값 이상인 단어들만 선택한다. 특징 선택에 사용되는 가중치의 계산은 다음의 계산식을 통해 이루어진다.

$$w_{d,t} = tf_{d,t} \times df_{t,d} \quad (3)$$

$$w_{q,t} = tf_{q,t} \times df_{t,d} \quad (4)$$

$tf_{d,t}$는 문서 *d*에서 단어 *t*가 발생하는 빈도수를 나타내고, $tf_{q,t}$는 질의어 *q*에서 단어 *t*가 발생하는 빈도수를 나타내며, $df_{t,d}$는 전체 문서집합에서 단어 *t*가 발생하는 문서의 수를 의미한다. 이렇게 추출된 단어들은 기존에 UDDI 레지스터가 갖고 있는 정보와 비교해서 중복성 검사를 통해 해당 정보가 없을 경우에는 UDDI 레지스터에 등록을 한다.

웹 서비스의 요청자(Requester)인 사용자가 웹 서버를 통해 질의어를 입력하면 질의매니저(Query Manager)에 의해 전처리 과정을 거쳐, 검색매니저(Search Manager)가 UDDI 레지스터에 있는 정보를 검색하게 된다. 이때 UDDI 레지스터에 있는 정보와 사용자의 질의어간 유사도 평가를 하며, 유사도 순서에 따른 검색 결과들을 웹 서버를 통해 사용자에게 제공한다.

$$Sim(R_d, Q) = \frac{\vec{R_d} \cdot \vec{Q}}{|\vec{R_d}| \times |\vec{Q}|} = \frac{\sum_{t=1}^{t} w_{d,t} \times w_{q,t}}{\sqrt{\sum_{t=1}^{t} w_{d,t}^2} \times \sqrt{\sum_{t=1}^{t} w_{q,t}^2}} \quad (5)$$

유사도 측정은 식(5)을 사용해서 이루어지며, 이는 질의어와 커뮤니티 간 유사도를 측정하는 것으로 벡터 공간 모델의 공식이다. 식(6)은 식(5)의 즉 단어의 가중치를 계산하는 공식인데 제안하는 방식은 *idf*의 의미와 상반되기 때문에 1에서 *idf*값을 뺀 후 *df* 값을 계산하여 가중치 값을 얻었다.

$$w_{i,j} = f_{i,j} \times (1 - \log \frac{N}{n_i}) \quad (6)$$

식(7)은 식(6)의 $f_{i,j}$ 값을 구하는 공식이다.

$$f_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \quad (7)$$

최상위 가중치 값을 가진 단어를 통해 커뮤니티가 어떤 정보를 담고 있는지를 알 수 있다.

3.5. UDDI 레지스터

UDDI 레지스터는 분석 관리자, 질의어 관리자, 검색 관리자와 상호작용하며, 각 커뮤니티의 특징 벡터를 저장하고 있다. UDDI 레지스터는 각종 정보들을 생성, 저장, 검색할 수 있는 XML 기반의 자료 저장장치를 의미 한다.

데이터 파서에서 분석된 단어들은 UDDI에 저장되어 사용자가 질의할 때 검색에 이용된다. UDDI를 통한 검색 시 가장 큰 문제점은 순위화를 지원하지 않는다는 것이다. 이 문제를 해결하기 위해 본 논문에서 제안하는 시스템은 정보검색 시스템에서 사용하는 코사인 유사도 기법을 사용해 사용자 질의어와 UDDI내의 정보 간 유사도 측정을 해 이에 대한 정렬을 수행함으로써 결과의 순위화 문제를 해결하였다. 사용자가 자신이 원하는 정보를 찾기 위해 질의를 했을 때 검색 매니저는 사용자의 질의와 UDDI에 있는 정보와의 유사도를 평가해서 유사도가 가장 높은 커뮤니티 정보를 제공해 주기 때문에 검색결과가 기존 시스템에 비해 보다 효과적이

5) IDF(Inverse Document Frequency)는 전체문서 집합에서 일부문서에 집중적으로 나타나는 단어에 대해 높은 가중치를 할당하는 방법이고, DF는 전체문서 집합에서 해당단어가 나타나는 빈도수를 의미한다. 즉 DF는 보다 많은 문서에서 해당 단어가 많이 나타날수록 높은 가중치를 할당하는 방법이다. 제안 시스템에서 커뮤니티를 특징지을 수 있는 단어들을 추출하기 위해서는 각 단어가 커뮤니티 내의 여러 문서에서 두루 나타날수록 보다 효과적일 것이기 때문에 DF를 적합한 가중치 계산을 위한 요소로 선정하여 특징값으로 사용한다.

고 정확하다. 또한 기존의 RSS 기반으로 수집된 정보를 일일이 확인하는 문제점을 UDDI 랭킹을 통해 보다 확실한 정보를 사용자의 질의에 적합한 선별된 정보를 제공하도록 한다.

3.6. 질의어 관리자

웹 서비스 요청자인 사용자와 직접 상호작용하며, 사용자로부터 웹 서버를 통해 질의어를 입력 받아 질의어에 대한 전처리 과정을 수행 한다. 전처리 과정은 형태소 분석과정과 분석된 질의어의 벡터 모델 표현 과정으로 이루어진다. 첫 번째 과정은 형태소 분석 관리자와 상호작용을 통해 입력된 질의어를 형태소 단위로 분석하는 과정을 의미한다. 두 번째 과정은 데이터 파서와 상호작용으로 이루어지며 분석된 질의어를 벡터 모델로 표현하는 과정이다. 전처리 과정이 끝난 질의어는 검색 관리자에게 전달되어 커뮤니티 내 문서를 검색하는데 사용된다.

3.7. 검색 관리자

UDDI 레지스트리에 있는 정보를 검색하고 질의어와의 유사도에 따라 검색된 결과를 웹 서버를 통해 사용자에게 반환한다. UDDI 레지스트리에서 문서들의 검색은 두 단계로 이루어진다. 먼저 유사한 커뮤니티를 찾고 해당 커뮤니티에서 문서를 검색하는 과정으로 이루어진다. 질의어와 커뮤니티 간 유사도 측정은 입력된 질의어와 커뮤니티의 특징 벡터 간 유사도를 측정함으로써 이루어진다.

3.8. 시스템 구조에서 RSS의 역할

월드 와이드 웹(World Wide Web)의 보편화로 인하여 하루에도 100만 개가 넘는 웹 페이지가 생성되고 있으며, 또한 그 만큼의 웹 페이지가 사라지고 있다. 이렇게 급속하게 증가하고 변화하는 웹 문서로 인하여 커뮤니티의 특징 또한 빠르게 변화한다. 이러한 변화에 동적으로 대응하기 위해 제안 시스템은 RSS를 적용하였다.

등록된 커뮤니티는 커뮤니티의 회원이나 커뮤니티 운영자에 의해 수시로 정보가 업데이트 되는데 이렇게 업데이트 되는 정보는 RSS에 의해 자동으로 시스템에 알려지게 되고 크롤러와 데이터 파서는 위의 과정을 다시 반복하게 된다. 다시 추출된 단어들은 기존에 UDDI 레지스터가 갖고 있는 정보와 비교해서 중복성 검사를 통해 해당 정보가 없을 경우에는 UDDI 레지스터에 등록을 한다. UDDI에 등록된 커뮤니티의 특징들은 커뮤니티에 업데이트 되는 내용을 RSS를 통하여 분석하고 분석된 내용에 맞게 커뮤니티의 특징을 변화시켜 UDDI에 자동으로 저장한다. 이 시스템을 통해 사용자는 정보를 손쉽게 제공 받을 수 있다.

Ⅳ. 실험 및 평가

이 장에서는 제안 시스템의 신뢰성과 효율성을 증명하기 위해 수행한 실험환경 구성과 결과에 대해 기술한다. 사용하는 UDDI 레지스터는 기존 웹에서 제공되었던 서비스 환경에서 로컬에 맞게 구축하였다.

4.1. 커뮤니티 등록

실험에 사용되는 커뮤니티는 RSS 서비스에서 제공되는 웹 페이지들을 그림 4와 같이 자체 개발 프로그램인 “RSS Address Manager Tool”을 이용하여 등록한다. 즉 평소 관심이 있거나 검색에 유용하게 활용 될 수 있는 커뮤니티들의 URL 주소를 입력창에 간단하게 등록 할 수 있으며 등록된 커뮤니티들의 목록은 “RSS Address Manager Tool”의 화면을 통해 웹 사용자들에게 제공된다.

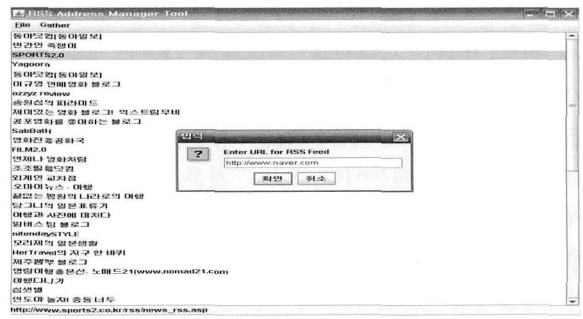


그림 4. 커뮤니티 등록
Fig. 4. Community registration

4.2. 등록된 커뮤니티 내 정보 크롤링

그림 5는 등록된 커뮤니티 내 정보를 크롤링하기 위한 과정으로 ‘RSS Address Manager Tool’의 ‘Crawl’ 메뉴를 통해 간단하게 이루어진다. 툴(Tool)을 통해 각 커뮤니티에 사용자들이 올린 게시글이나 자료에 대해 크롤링 가능하며, 분류 방식은 게시글 제목을 주제로 하여 분류하였고, 각 주제에 따른 내용을 UDDI 레지스터에 저장했다. 이와 같이 주제에 따라 분류된 커뮤니티로부터 추출된 정보는 그림 6과 같이 UDDI에 등록되게 된다.

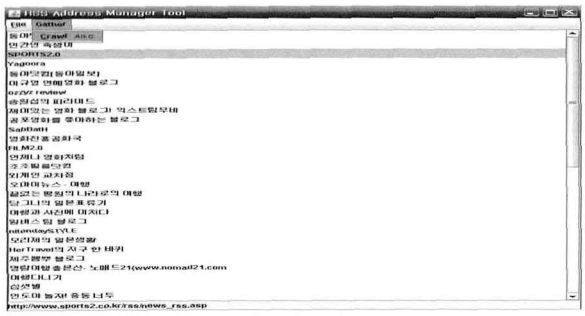


그림 5. 커뮤니티 내 정보 크롤링
Fig. 5. Information Crawling within community

Pick	Details	Object Type	Name	Description	Version	VersionComment	Pin
	Details	Organization	freebXMLRegistry	freebXML Registry	1.1		
	Details	Organization	레몬이나 자몽 껍질을 버리지 말고 반병에 넣어 냉장고에 보관해 놓다. 가스대	레몬이나 자몽 껍질을 버리지 말고 반병에 넣어 ...	1.1		
	Details	Organization	감자 삶기, 다섯가지 방법	감자 삶기, 다섯가지 방법 가계부프로그램 가계부...	1.1		
	Details	Organization	[다이어트]작은 얼굴, 동안얼굴 만들기	[다이어트]작은 얼굴, 동안얼굴 만들기 가계부프...	1.1		
	Details	Organization	재래식 된장으로 끓인 구수한 된장국	재래식 된장으로 끓인 구수한 된장국 가계부 프로그...	1.1		
	Details	Organization	와이드 하게 세상을		1.1		
	Details	Organization	[다이어트]작은 얼굴, 동안얼굴 만들기	[다이어트]작은 얼굴, 동안얼굴 만들기 가계부프...	1.1		
	Details	Organization	효율환자를 위한 6가지 허리강화 운동	1. 허리 비틀기, 플어올리기 운동 이 운동을 첫...	1.1		
	Details	Organization	성인병 예방하는 바다 영양제 다시마	이채를 먹지 않는 아이들의 경우 다시마가 풍풍한...	1.1		
	Details	Organization	다이어트의 기본은 장이 불끗해야한다	합격된 생활습관 등으로 인해 확연과는 점점 멀어...	1.1		

그림 6. UDDI 레지스터 저장 형태
Fig. 6. UDDI register storing form

4.3. 등록된 커뮤니티 검증

커뮤니티를 등록하고 등록된 커뮤니티 내의 정보를 크롤링하여 UDDI 레지스트리에 저장 한 이후에는 커뮤니티의 신뢰성을 검증하게 된다. 웹상에서 커뮤니티로 등록되어 있다 할지라도 실제 많은 정보들이 커뮤니티와 관련 없는 정보들로만 구성되어 있을 수 있기 때문이다. 이와 같은 커뮤니티 검증 과정을 통해 차후 검색에 커뮤니티를 반영 할 경우 웹 사용자는 검색 과정에서 불필요한 노력을 감소와 사용자의 검색 의도에 더 부합하는 검색결과를 획득 할 수 있다.

Name	Slot Type	Value(s)
다이어트칼로리	Term	[1]
트랙백	Term	[1]
체질별다이어트	Term	[1]
웰빙음식과건강	Term	[1]
마켓	Term	[1]
다이어트음식단	Term	[1]
의상	Term	[1]
다이어트음식	Term	[1]
엠에인	Term	[1]
체어스칼정보	Term	[1]
다이어트요리	Term	[1]
정보	Term	[1]
스트레칭	Term	[1]
클릭	Term	[1]
http	Term	[1]
주소	Term	[1]
by	Term	[1]
분류	Term	[1]
파리잡	Term	[1]
출처	Term	[2]
전체보기	Term	[1]
Posted	Term	[1]
비키니	Term	[1]
복위별다이어트	Term	[1]
slender.tistory.com/trackback/2410	Term	[1]

그림 7. 한글 형태소 분석기를 이용해 추출한 단어(TF)
Fig. 7. Extracted words form HMA

UDDI 레지스트리에 저장된 정보들은 데이터 파서를 통해 커뮤니티별 수집된 자료간 공통된 특징을 추출하기 위해 한글 형태소 분석기(Hangul Morpheme Analyzer: HMA)을 이용하여 Stop-word와 Stemming을 통해 자료들을 벡터 모델로 표현한다.

그림 7은 커뮤니티에 있는 문서의 내용을 형태소 분석하여 추출한 단어를 나타낸 것이다. 각 커뮤니티 내에는 수많은 문서들이 있기 때문에 전체 문서를 대상으로 형태소 분석을 실시하여 벡터상에 표현한다. 추출된 단어들은 각각의 가중치를 계산해서 커뮤니티를 위한 특징 벡터를 형성한다. 다양한 커뮤니티에서 질의와 관련된 커뮤니티를 찾기 위해 검색을 한다.

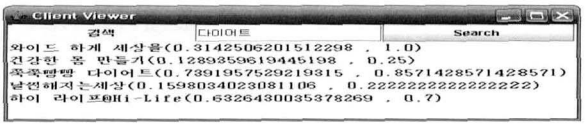


그림 8. 웹 서비스를 이용한 커뮤니티 검색
Fig. 8. Web service using community search

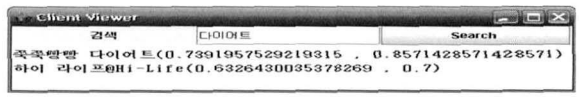


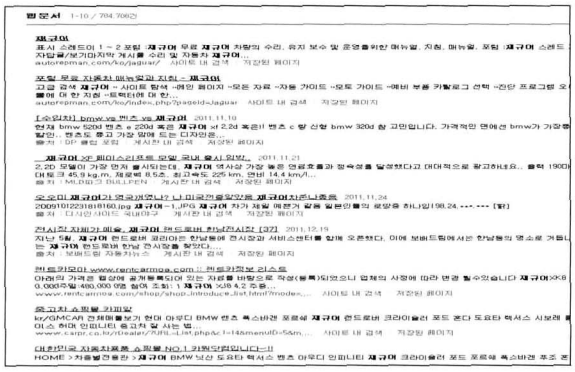
그림 9. 특정 임계치 이상의 값을 갖는 커뮤니티 필터링
Fig. 9. Community filtering surpassing particular critical value

그림 8은 '다이어트'라는 질의어를 예로 커뮤니티를 검색한 결과이며 질의어에 대한 각 커뮤니티별 코사인 유사도에 따른 유사도 값과 포커스 크롤링 값을 제공한다. 이때 유사도와 포커스 크롤링 값이 특정 임계치보다 큰 커뮤니티는 검증된 것으로 인정한다. 이와 같이 검증된 커뮤니티는 별도의 목록에 저장되며 웹 사용자가 검색을 할 경우 검증된 커뮤니티를 통해 검색 결과를 제공한다. 즉 커뮤니티는 특정 임계치(유사도/포커스 크롤링 값)가 모두 0.5 이상의 값을 갖는 커뮤니티에 대해서만 필터링을 하여 검색되어 지며, 이와 같은 커뮤니티들은 각각의 유사도와 포커스 크롤링 값을 기준으로 재순위화되어 최종적으로 웹 사용자에게 제공된다.

그림 9는 특정 임계치 이상의 값을 갖는 커뮤니티들이 재순위화 되어 제공되는 결과를 나타낸다.

4.4. 커뮤니티 검증 및 검색

그림 10은 검색자가 야생동물 재규어에 대한 정보를 검색하기 위한 의도를 가지고 질의어 "재규어"를 입력 한 후 각각 검색 한 결과이다. 그림 10(a), 그림 10(b) 및 그림 10(c)는 각각 네이버, 다음 및 구글 검색엔진에 의한 검색 결과로 사용자의 검색 의도와는 무관하게 자동차 브랜드에 대한 검색 결과가 상위에 순위화 되어 있음을 확인 할 수 있다.



(a) 네이버(http://www.naver.com) 검색 결과

(b) 다음(http://www.daum.net) 검색 결과

(c) 구글(http://www.google.co.kr) 검색 결과

이와 같은 검색결과는 야생동물 “재규어”에 대한 정보를 검색하고자 하는 웹 사용자와는 무관한 정보로 웹 사용자는 검색을 반복하거나 스크롤을 통해 원하는 정보가 나올 때 까지 다른 페이지를 계속 클릭해야 하는 불필요한 노력을 해야 한다.

그림 10(d)는 논문에서 제안하는 신뢰할 수 있는 온라인 커뮤니티 기반의 검색결과, 즉 클라이언트 뷰(Client View)에 의한 검색결과로, 검색자의 의도와 부합되는 야생동물에 관련된 정보가 첫페이지에 등장하는 것을 확인 할 수 있다.

(d) 클라이언트 뷰(Client View) 검색 결과

그림 10. 질의어 “재규어”에 대한 웹문서 검색 결과('11.12.26)
Fig. 10. Web document search results from the query "jaguar"

즉 일반 검색 방법보다 제안 검색 방법이 검색자의 의도가 반영된 보다 정합한 정보를 제공하는 것을 확인 할 수 있다. 이 같은 결과는 평소 관심 있는 커뮤니티를 웹 사용자가 등록하고 등록된 커뮤니티 내의 정보가 커뮤니티를 대표 할 수 있는 관련성 있는 정보를 얼마나 많이 보유하고 있는지에 대한 검증과정을 거쳐 검색에 반영하기 때문이다.

V. 결 론

본 논문에서 제안한 신뢰된 커뮤니티를 기반으로 한 웹 검색을 통해 포커스 크롤링, 웹 문서 클러스터링, 제한 검색 등의 기존 검색방법의 문제점을 해결하고자 하였다. 제안한 시스템은 커뮤니티 내의 대부분의 정보가 크롤러에 의해 특정 어휘나 주제에 의해 하나로 분류 될 수 있기 때문에 포커스 크롤링이 가지는 질의 처리 시간보다 상대적으로 빠르고 제한 검색이 가지는 정보 범위의 한계와 신뢰된 데이터 획득을 어느 정도 더 보장 받을 수 있다.

웹 문서 클러스터링이 가지는 문제점은 문서의 분류에 대해 색인 등의 전처리 과정을 거치기 때문에 복잡하고 유지 보수가 어렵다는 것이다. 이러한 문제점은 UDDI에 등록된 커뮤니티의 특징들은 커뮤니티에 업데이트 되는 내용을 RSS를 통하여 분석하고 분석된 내용에 맞게 커뮤니티의 특징을 변

화시켜 저장함으로써 상대적으로 유지보수가 쉽다.

또한 특정 질의 즉 주제에 대해 크롤링 되어있는 상태를 확인하여 커뮤니티가 예상치와 동일하게 웹 사용자의 특정 관심사를 반영 할 경우 해당 커뮤니티를 검색에 반영함으로써 기존의 방법 보다 개인이 찾고자 하는 검색 결과를 보다 정확하고 근접하게 찾아 줄 수 있다.

향후 연구로 제안한 시스템의 성능에 대한 정확한 검증을 위해 정량적 분석을 실시할 계획이다.

참 고 문 헌

[1] S. Sizov, J. Graupmann, M. Theobald. "From Focused Crawling to Expert Information: an Application Framework for Web Exploration and Portal Generation." *VLDB*, 2003

[2] S. Chakrabarti, M. van den Berg and B.Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." *WWW-8*. 1998

[3] Zamir, O. and Etzioni, O., "Web Document Clustering: a Feasibility Demonstration," *In Proc. 19 Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval*, pp. 46-54, Melbourne, Australia, June 1998.

[4] 네이버 용어 (http://terms.naver.com/item.nhn?dirId=200&docId=11311)

[5] Shkapenyuk, V. and Suel, T., "Design and implementation of a High Performance Distributed Web Crawler," *In Proc, of the 18thInt'l Conf. on Data Engineering*, San Jose. California, Feb. 2002

[6] Oren Zamir and Oren WEtzioni, " Grouper: A Dynamic Clustering Interface to Web Search Results," *Proc of WWW8*, 2009

[7] H, Lee. and J, Kwon., "Personalized RSS Search Service Using RSS Characteristics and User Context," *Proceedings of the International MultiConference of Engineers and Computer Scientists 2008 Vol I*, pp.19-21

[8] Barry Smyth., "A Community-Based Aproach to Personalizing Web Search," *IEEE Computer Society Press Los Alamitos, Volume 40 Issue 8*, August 2007

[9] Laurence A. F. Park Kotagiri Ramamohanarao., "Mining Web Multi-resolution Community-based Popularity for Information Retrieval," *CIKM 2007*, November 6-8

[10] Said Kashoob James Caverlee Krishna Kamath., "Community-Based Ranking of the Social Web," *HT'10*, June 13-16, 2010

김 영 안 (Young-an Kim) 정회원

1988년 2월 금오공과대학 전산공학과 졸업

1996년 3월 Keio University Dept. of Information and Computer Science (공학석사)

2008년 2월 경희대학교 컴퓨터공학박사)

2009년 2월~현재 국방대학교 국방과학학과 교수

<관심분야> Ad-hoc Network, Routing Protocol, DTN, VANET, WMN, 정보검색

박 상 관 (Sang-Kwan Park) 정회원

1996년 2월 공군사관학교 졸업

2009년 2월 국방대학교 국방과학학과 (공학석사)

2009년 3월~현재 공군본부

<관심분야> 정보검색, 소셜 네트워크

이 상 훈 (Sang-Hoon Lee) 종신회원

1978년 2월 성균관대학교 정보통신공학과 졸업

1989년 2월 연세대학교 산업대학원 전산학과 (공학석사)

1997년 3월 일본 교토대학교 정보공학 (공학박사)

1998년~2000년 서일대학 겸임교수, 충남산업대학교 교수

2000년~현재 국방대학교 국방과학학과 교수

<관심분야> 정보검색, 멀티미디어 데이터베이스, HCI