웹 서비스를 이용한 커뮤니티 검색

정찬백[O] 김태환 전호철 최중민
한양대학교 컴퓨터공학과
{cbjeong[O], kimth, hcjeon, jmchoi}@cse.hanyang.ac.kr

Community Retrieval Using the Web Services

ChanBack Jeong[O], Taehwan Kim, Hochul Jeon, Joongmin Choi
Department of Computer Science & Engineering
Hanyang University

요 약

가공해서 사용하는 정보량이 많아질수록 원하는 정보를 찾는데 더 많은 노력이 필요하게 되었다. 따라서 사람들은 인터넷상에서 원하는 정보를 검색하는 여러 방법들을 고안해 왔으며, 이렇게 구현된 검색 알고리즘은 검색 질의와 유사한 문서가 대중에게 얼마나 관심을 받고 있는지 그 정도에 따라 검색순위 상위에 링크된다. 하지만 웹 문서의 폭발적인 증가로 해당 질의에 대한 검색 결과 문서의 수가 급격히 늘어나면서 사용자를 만족시키기가 점점 어렵게 되었다. 이러한 문제를 해결하기 위한 방안으로 네티즌들이 직접 정보를 생산, 공유하고 이들이 모여 활동할 수 있는 커뮤니티를 형성하기 시작했다.

이 논문에서는 정보의 공유를 목적으로 하는 커뮤니티를 인터넷상의 표준화된 웹 서비스(Web Services) 기술인 UDDI에 저장하고, SOAP 프로토콜을 이용하여 플랫폼에 상관없이 사용자 검색 질의와 가장 유사한 커뮤니티를 검색하여 제공하는 방법을 제안 한다.

1. 서 론

웹 검색 시스템은 웹에서 원하는 정보를 보다 쉽게 찾기 위한 도구로서 그 중요성이 점차 부각되고 있다. 최근 들어 웹 검색 시스템에서 사용자의 질의에 대해 정확도가 높은 충분한 양의 검색 결과를 제공하고자 하는 연구들이 활발하게 진행되고 있다. 그 대표적인 연구로는 제한 검색(limit search), 포커스 크롤러(focused crawler), 웹 문서 클러스터링(web document clustering) 등이 있다. 제한 검색은 현재 입력한 검색어의 검색 결과를 줄이고자 할 때 이용하는 검색 방식으로 검색 범위를 특정 사이트 또는 도메인으로 한정시켜 검색 결과를 제공하는 방법이다[1]. 포커스 크롤러는 웹 문서가 가지는 정보들은 URL을 이용하여 문서간에 연결되어 있다는 특성을 이용하여 질의가 주어진 시점에 질의와 관련 있는 웹 페이지들만을 수집하여 결과로 반환하는 방법이다[2][3][4]. 웹 문서 클러스터링은 클러스터를 구하기 위해 많은 양의 사이트들 또는 웹 페이지들을 서로 관련 있는 웹 페이지들끼리 클러스터링하는 방법이다[5].

그러나 위에서 설명한 연구들은 다음과 같은 단점을 가지고 있다. 제한 검색은 검색의 범위를 URL에 의해 명시되는 사이트 또는 도메인들로만 제한할 수 있을 뿐이며, 의미적으로 관련된 사이트들로 제한할 수 없다. 포커스 크롤러는 질의 시점에 웹 페이지들을 수집하기 때문에 질의 처리 시간이 오래 걸린다. 웹 문서 클러스터링은 클러스터를 구하기 위해 많은 양의 사이트들 또는 웹 페이지를 대상으로 복잡한 처리를 수행하므로 공간적 시간적 비용이 크다.

이러한 문제점을 해결하기 위해서 본 논문에서는 정보의 공유를 목적으로 하는 커뮤니티를 대상으로 웹 서비스 기술을 적용 함으로써 제한 검색이 가지는 의미적으로 관련된 사이트를 찾지 못하는 문제점과 포커스 크롤러 및 웹 문서 클러스터링이 가지는 질의 처리 시간에 대한 문제를 해결하려 한다. 일반적으로 인터넷 상에서의 온라인 커뮤니티의 정의는 " 네티즌들이 직접 정보를 생산, 공유하고 이들이 모여 활동할 수 있는 인터넷 상의 공간" 이다[6].

본 논문의 구성은 다음과 같다. 먼저 2장에서는 본 논문에서 언급된 연구들의 이론적 배경에 대하여 설명한다. 3장에서는 본 논문에서 제시하고 있는 시스템의 구조와 웹 서비스를 이용한 커뮤니티 검색 방법에 대해 설명한다. 4장에서는 실험 결과를 통하여 기존의 방법과 본 논문에서 제시한 방법을 비교 평가하고 5장에서 향후 연구를 기술하고 결론을 내린다.

2. 관 련 연 구

본 장에서는 관련 연구로서 웹 검색 시스템과 관련된 주요 기술들과 웹 서비스에 대해 설명한다.

2.1 제한 검색(limit search)

사이트 제한 검색은 그림 1과 같이 중앙 데이터베이스에

서의 기본 동작 방식은 서비스 기술, 서비스 등록 및 발견, 서비스 간의 통신 관점에서 정의된다.

사이트 제한 검색은 저장된 전체 데이터 중에서 지정된 사이트의 데이터만을 검색하는 기능이다. 제한 검색을 사용하기 위하여 웹 사이트 관리자가 자신의 사이트를 검색 시스템에 등록하면, 웹 로봇[7]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[1].

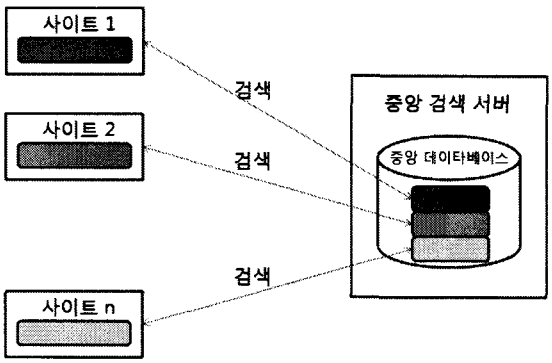

그림 1. 사이트 제한 검색의 개념.

2.2 웹 문서 클러스터링 (web document clustering)

웹 페이지의 클러스터링은 단어 또는 링크 등을 이용하여 웹 페이지 간의 유사도를 측정하는 과정과 측정된 유사도를 바탕으로 기존의 데이터 클러스터링 알고리즘을 적용하는 과정으로 이루어 진다.

Minimum spanning tree(MST)클러스터링은 클러스터링 알고리즘 중 하나로 MST를 subtree들로 나누어 클러스터를 구하는 방법이다[8]. MST클러스터링은 클러스터의 개수를 사전에 정하지 않아도 클러스터를 구할 수 있으며 여러 가지 데이터 분포에서도 잘 동작하는 장점이 있다.

2.3 웹 서비스(web services)

웹 서비스는 웹 상에서 정의된 모듈화 된 소프트웨어 컴포넌트로서, 개방형 표준 데이터 표현 기법인 XML과 인터넷 프로토콜을 결합시킨 새로운 패러다임에 의해 탄생된 분산 컴퓨팅 기술이다. W3C는 웹 서비스를 URI에 의해 인식되는 소프트웨어 애플리케이션으로서, 인터넷 기반의 프로토콜을 통하여 교환되는 XML기반 메시지를 사용하여 다른 소프트웨어 에이전트들과의 직접적인 상호작용을 지원한다고 정의한다[9].

웹 서비스는 SOAP, WSDL, UDDI를 통해 SOA의 주요 요소인 메시지, 서비스 인터페이스, 서비스 공개 및 발견 체계를 구현한다. 따라서 웹 서비스는 SOA 구축에 필요한 표준 기술들을 제공한다. 그림 2에서 볼 수 있듯이 웹 서비스의 기본적인 아키텍처는 SOA를 채택하고 있다. 웹 서비스의 기본 동작 방식은 서비스 기술, 서비스 등록 및 서비스 간의 통신 관점에서 정의 된다.

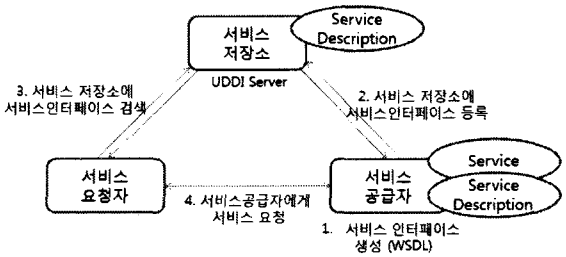

그림 2. 웹 서비스 아키텍처

3. 웹 서비스를 이용한 커뮤니티 검색 시스템 구조

3.1 시스템 구조

본 논문에서 사용하고 있는 시스템의 구조는 그림 3과 같은 구조로 이루어져 있다.

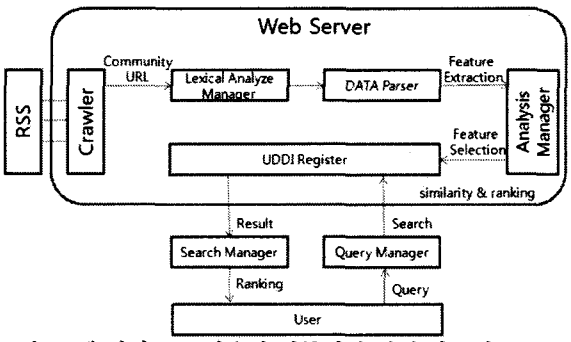

그림 3. 웹 서비스를 이용한 커뮤니티 검색 시스템 구조

그림 3에서 웹 서비스의 제공자(Provider) 역할을 담당하는 RSS Crawler는 사전에 등록된 RSS 주소들을 통해 문서들을 읽어 들이는 기능을 하며, 이렇게 읽어 들인 문서들은 각 커뮤니티의 특징을 추출하는 기초 자료들로 사용된다. 데이터 파서(DATA Parser)는 각 커뮤니티별 수집된 자료간 공통된 특징을 추출하기 위해 한글 형태소 분석기(Lexical Analyze Manager:HAM)을 이용해서 Stop-word와 Stemming작업을 통해 읽어온 자료들을 벡터 모델로 표현한다. 이때 단어들의 가중치를 계산(TF-DF : Term Frequency, Document Frequency)해서 최상위 가중치 값을 가진 단어들을 추출(Feature Extraction)한다.

$tf_{d,t}$는 문서 d에서 term t가 발생하는 빈도수를 나타내고, $tf_{q,t}$는 질의어 q에서 term t가 발생하는 빈도수를 나타내며, $df_{t,d}$는 전체 문서집합에서 term t가 발생하는 문서의 수를 의미한다.

본 논문에서는 IDF 대신 DF를 사용한다. 그 이유는 다음과 같다. IDF는 전체 문서 집합에서 일부 문서에 집중적으로 나타나는 단어에 대해 높은 가중치를 할당하는 방법이다. 반면에, DF는 전체 문서 집합에서 해당 단어가 나타나는 빈도수를 의미한다. 즉, 보다 많은 문서에서 단어가 나타날수록 높은 가중치를 할당하는 방법이다.

제안 하는 시스템에서 커뮤니티를 특징 지을 수 있는 단어들을 추출하기 위해서는 각 단어가 커뮤니티 내의 여러 문서에서 두루 나타날수록 보다 효과적일 것이다. 때문에 IDF가 아닌 DF가 적합한 가중치 계산을 위한 한 요소가 된다.

$$w_{d,t} = tf_{d,t} \times df_{t,d} \quad (1)$$

$$w_{q,t} = tf_{q,t} \times df_{t,d} \quad (2)$$

위와 같은 과정을 통해 최종적으로 뽑힌 최상위 가중치의 값을 가진 단어를 통해 커뮤니티에 있는 정보가 어떤 정보를 담고 있는지를 알 수 있다.

이렇게 추출된 단어들은 기존에 UDDI 레지스터가 갖고 있는 정보와 비교해서 중복성 검사를 통해 정보가 있는지 없는지 확인하고 해당 정보가 없을 경우에는 UDDI 레지스터에 등록을 한다.

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

$$Sim(R_d, Q) = \frac{\vec{R_d} \cdot \vec{Q}}{|\vec{R_d}| \times |\vec{Q}|} = \frac{\sum_{t=1}^{t} w_{d,t} \times w_{q,t}}{\sqrt{\sum_{t=1}^{t} w_{d,t}^2} \times \sqrt{\sum_{t=1}^{t} w_{q,t}^2}} \quad (3)$$

3.2 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 본 시스템은 RSS[10]를 적용한다. UDDI에 등록된 커뮤니티의 특징들은 커뮤니티에 업데이트 되는 내용을 RSS를 통하여 분석하고 분석된 내용에 맞게 커뮤니티의 특징을 변화시켜 UDDI에 저장한다. 이 시스템을 통해 사용자는 급속하게 변화하는 정보를 손쉽게 제공 받을 수 있다.

4. 실험 및 평가

4.1 시스템 구성 환경

이 장에서는 제안한 시스템의 효율성을 확인하기 위해 수행한 실험의 환경과 결과에 대해 기술한다.

UDDI 레지스터는 기존 웹에서 제공되었던 서비스 환경을 로컬에 맞게 바꿔서 사용했다.

4.2 실험

실험을 위해 정보검색의 기본척도인 정확률과 재현율을 다음과 같이 계산한다.

$$\text{정확률(P)} = \frac{\text{질의에적합한커뮤니티수}}{\text{검색된커뮤니티수}} \quad (4)$$

$$\text{재현율(R)} = \frac{\text{질의에적합한커뮤니티수}}{\text{전체커뮤니티수}} \quad (5)$$

정확률과 재현율이 높다는 전제하에 다음과 같은 실험을 통해 포커스 크롤링의 질의 처리 시간과 비교 평가한다. 본 논문에서 정의하고 있는 온라인 커뮤니티는 " 특정 주제에 대하여 사용자가 정보를 직접 생산, 공유하고 이들이 활동할 수 있는 인터넷 상의 공간" 이기 때문에 커뮤니티 자체가 이미 특정 질의 즉 주제에 대해서 크롤링되어 있는 상태라 볼 수 있다. 이를 증명하기 위해 다음과 같은 실험을 한다. 커뮤니티 내의 정보들이 크롤러를 통해 분류 했을 때 자료 대부분이 특정 어휘나 주제에 의해 하나로 분류 된다면 이는 해당 질의에 대해 이미 크롤링 된 상태라고 할 수 있다. 각각의 커뮤니티가 가지는 tf-df값이 임계값(threshold) 보다 큰 커뮤니티에 대해 크롤러가 얼마나 많은 커뮤니티 내의 자료를 특정 어휘나 주제에 대해 하나로 분류할 수 있는지 임계값을 변화시켜 실험한다.

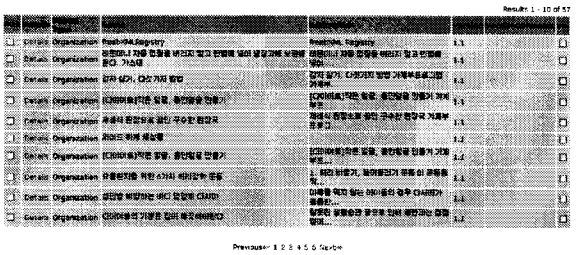

그림 4. 커뮤니티 검색을 위한 RSS 크롤링

실험을 위해 RSS를 이용한 임의의 커뮤니티를 대상으로 각 커뮤니티에 사용자들이 올린 게시글이나 자료를 크롤링 하였다. 분류 방식은 게시글의 제목을 주제로 하여 분류했고, 각 주제에 따른 내용을 UDDI 레지스터에 저장했다. 이렇게 주제에 따라 분류된 여러 개의 커뮤니티에서 추출된 정보는 그림 4와 같이 UDDI에 등록되게 된다. 이 정보들은 다시 데이터 파서(DATA Parser)통해 각 커뮤니티별 수집된 자료간 공통된 특징을 추출하기 위해 한글 형태소 분석기(Lexical Analyze Manager:HAM)

을 이용해서 Stop-word와 Stemming작업을 통해 읽어온 자료들을 벡터 모델로 표현한다.

다이어트칼로리	Term	[1]
트랙백	Term	[1]
체질별다이어트	Term	[1]
황빙음식과건강	Term	[1]
아래	Term	[1]
다이어트식단	Term	[1]
의상	Term	[1]
다이어트음식	Term	[1]
연예인	Term	[1]
헤어스타일정보	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]
과격적	Term	[1]
출처	Term	[2]
전체보기	Term	[1]
Posted	Term	[1]
비키니	Term	[1]
부위별다이어트	Term	[1]
slender.tistory.com/trackback/2410	Term	[1]

그림 5. 한글 형태소 분석기를 이용해 추출한 단어(TF)

그림 5는 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 계산(TF-DF : Term Frequency, Document Frequency)해서 커뮤니티를 위한 특징 벡터를 형성한다.

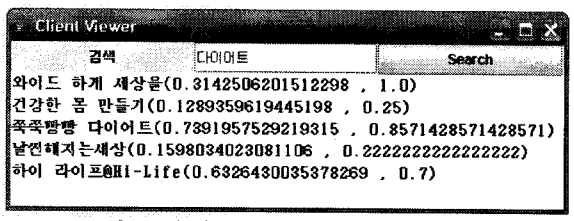

그림 6. 웹 서비스를 이용한 커뮤니티 검색

다양한 커뮤니티에서 질의와 관련된 커뮤니티를 찾기 위해 검색을 한다. 그림 6은 ' 다이어트' 라는 질의어를 예로 커뮤니티를 검색했다. 검색 결과에서 질의에 대한 각 커뮤니티별 코사인 유사도에 따른 유사도 값을 계산하며, 이와 함께 질의어에 대한 각 커뮤니티의 포커스 크롤링 값을 계산 했다.

포커스 크롤링값은 다음과 같이 정의 한다.

$$FC = \frac{n}{N} \tag{6}$$

이때, n은 각 커뮤니티에 있는 문서 중에서 질의어가 나타나는 문서의 수, N은 각 커뮤니티에 있는 전체 문서의 수를 의미한다.
이것은 질의와 관련해서 단순히 TF-DF값으로 커뮤니티를 알려주는 것 이외에 포커스 크롤링된 값을 보여줌으로써 해당 커뮤니티가 질의에 대해서 관련성이 더 높다는 것을 알 수 있다. 하지만, 각 커뮤니티 내에 관련된 문서 이외에 다른 주제에 관련된 문서가 함께 포함되어 있을 경우, 포커스 크롤링 값은 작게 나오게 된다.

5. 결론 및 향후 연구

5.1 결론

본 논문에서 제안한 웹 서비스를 이용한 커뮤니티 검색은 관련 연구에서 언급한 포커스 크롤링, 웹 문서 클러스터링, 제한 검색의 각 문제점을 해결하고자 하였다.
위 실험에서 보여지는 것과 같이 커뮤니티 내의 대부분의 정보가 크롤러에 의해 특정 어휘나 주제에 의해 하나로 분류된 상태이기 때문에 포커스 크롤링이 가지는 질의 처리 시간보다 상대적으로 빠르고 제한 검색이 가지는 검색 결과를 줄이고자 하는 목적을 달성했다.
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 변화시켜 저장함으로써 상대적으로 유지 보수가 쉽다.

5.2 향후 연구

커뮤니티의 특징 집합과 질의간 유사도를 활용해 의미적 접근을 시도했지만 만족할 만한 성과를 얻지는 못했다. 이는 의미적 접근이 아닌 형태적 유사성에 근거했기 때문으로 풀이 된다. 이를 해결하기 위해 온톨로지를 구축하고 시맨틱 정보를 활용한 연구가 필요하다. 또한 사용자 마다 같은 주제에 대해서도 기준이 다를 수 있기 때문에 개인의 성향에 따른 커뮤니티 검색에 대한 연구도 필요하다.

후 기

이 논문은 " 국가 IT 온톨로지 인프라 기술개발" 정보통신부 선도과제 성과의 일부입니다.

참 고 문 헌

[1] 이재길, 이민재, 김민수, 황규영, " 오디세우스 객체 관계형 DBMS를 사용한 사이트 제한 검색의 구현,"

한국정보과학회 춘계학술발표회 논문집, pp. 755-757, 2003년 4월.

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

[3] De Bra, G. Houben, Y. Koranatzky and R. Post. " Information Retrieval in Distributed Hypertexts." *Proceedings of the 4th RIAO Conference*, 481-491, New York, 1994.

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

[5] 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.

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

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

[8] Zahn, C., " Graph Theoretical Methods for Detecting and Describing Gestalt Clusters," *IEEE Trans. On Computers*, Vol. C-20, No. 1, pp. 68-86, Jan. 1971

[9] W3C Web Services WG. "Web Services Architecture", http://www.w3.org/TR/ws-arch/ W3C Working Group Note 11 February 2004.

[10] W3C, RSS, http://web.resource.org/rss/1.0/

[11]UDDI. " The UDDI Technical White Paper" , http://uddi.org/pubs/Iru_UDDI_Technical_White_Paper.pdf, UDDI.org, September 2000.