Open API를 활용한 다국어 정보검색 시스템 모델링에 관한 연구

황세찬 * 김흥철* 김선진** 정주석** 강신재***
(Se-Chan Hwang *, Heung-Cheol Kim *, Seon-Jin Kim**, Ju-Seok Jeong**, Sin-Jae Kang***)

요 약 본 논문은 오픈 API를 이용하여 다국어 정보검색 시스템을 모델링하는 방법론을 제시한다. 웹 2.0이 대두되면서 웹 2.0의 개념을 활용한 기술들이 발달하고 있는데, 그 중 한 기술이 오픈 API이다. 기업에서 개발한 새로운 서비스나 기능, 데이터 등을 API로 공개함으로써 사용자들이 공개된 API를 이용하여 새로운 서비스를 쉽게 개발할 수 있게 되었다. 본 연구에서는 구글, 플리커, 유튜브, 네이버, 다음 등의 사이트에서 제공하는 오픈 API를 이용하여, 다국어 정보 검색 시스템을 구현하였다. 구글 번역 API를 이용하여 한국어 질의어를 검색 대상 언어(영어, 일본어, 중국어 등)로 번역한 후, 소셜 웹 사이트(플리커, 유튜브, 다음, 네이버 등)의 정보를 검색하고, 검색된 결과 내 텍스트를 다시 한국어로 번역한 후, 통합된 검색 결과를 사용자에게 보여준다.

핵심주제어 : 교차언어 정보검색, 오픈 API, 매쉬업

Key Words : cross-language information retrieval, open API, mashup

Ⅰ. 서 론

최근 웹 2.0에서는 웹 개방성을 기본으로 하는 오픈 서비스가 이슈가 되고 있다. 오픈 서비스는 키워드검색, 상품 정보 등 각 회사의 정보를 공개하는 것으로 사용자는 공개된 서비스를 이용하여 새로운 서비스를 개발할 수 있게 되었다. 오픈 서비스는 웹 서비스 기반 오픈 API로 제공되며 이러한 서비스의 등장으로 매쉬업(mashup) 서비스가 각광 받고 있는 추세이다. 오픈 API란 자사의 API를 웹 서비스를 통해 외부로 공개한 것을 말하고, 매쉬업은 두 가지 이상의 서로 다른 웹 서비스를 결합하여 새로운 서비스를 만드는 것을 의미한다. 즉, 사용자가 원하는 서비스를 제공하기 위해서 공개된 서비스를 이용하여 새로운 서비스를 구현하는 것이다. 이러한 방식은 기본에 존재하고 있는 데이터들을 가지고 섞어서 새로운 데이터를 만들어 사용자가 원하는 시스템에 맞게 가공하는 것이기 때문에 빠르고 쉽게 개발할 수 있다는 장점이 있어 큰 관심을 받고 있다[1].

오픈 API는 플랫폼으로서의 웹을 실현하기 위해 데이터 또는 서비스를 널리 분산시키는 목적을 가지며, 이들 오픈 API들의 매쉬업은 실제 프로세스를 가지지 않은 새로운 서비스를 재창조해 낸다. 오픈 API는 단순함을 유지하기 위해 자신을 설명할 수 있는 표준적인 문서를 가지지 않으며 REST (REpresentational State Transfer), XML-RPC (Remote Procedure Call), SOAP(Simple Object Access Protocol) 등의 다양한 통신 프로토콜로 구현된다. 매쉬업 애플리케이션은 구조적으로 API 제공자, 호스팅 사이트, 소비자로 나누어진다[2].

국제화 시대에 언어 장벽으로 인해 인터넷 상의 방대한 정보를 제대로 활용하지 못하는 것은 큰 문제이기 때문에, 본 연구에서는 사용자의 모국어로 질의어를 입력하여 외국어로 되어 있는 정보를 검색할 수 있는 다국어 정보검색 시스템을 모델링하고자 한다.

* 대구대학교 정보통신대학 컴퓨터IT공학부 석사과정
** 대구대학교 정보통신대학 컴퓨터IT공학부 학사과정
*** 대구대학교 정보통신대학 컴퓨터IT공학부 교수

다국어 정보검색(Multilingual Information Retrieval)의 정의는 서로 다른 언어로 이루어진 정보들로부터 원하는 정보를 검색하는 것을 말한다. 사용자가 언어에 구애받지 않고 여러 언어의 문서를 검색해서 원하는 정보를 얻게 해 주는 것이다. 예를 들어 한글로 검색하면 한국어 문서뿐만 아니라 일본어, 중국어, 영어 문서를 모두 사용자에게 제시해 줄 수 있도록 하는 것이다. 교차 언어 정보검색(Cross-Language Information Retrieval)은 질의어에 사용한 언어와 다른 언어로 이루어진 문서를 검색하는 것을 말한다[3].

교차 언어 정보검색을 위한 접근 방법은 통계적 기법과 번역 기법으로 나눌 수 있다. 통계적 기법은 언어 번역을 하지 않고 교차 언어 간의 연관관계 정보를 만드는데, 이 방법에서는 대량의 양국어 병렬 말뭉치(bilingual parallel corpus)가 필요하다. 번역 기법은 질의어의 언어를 문서의 언어로 번역하는 질의번역 방법이나, 문서의 언어를 질의어로 번역하는 문서번역 방법을 통해서 질의어와 문서의 언어를 같은 언어로 만들고 검색을 수행한다. 고품질의 기계번역 시스템을 사용 가능한 경우에는 문서번역 기법에 의한 교차언어 정보검색이 가능하지만, 대량의 문서집합에 대해 검색하는 경우에는 모든 문서를 번역해야 하기 때문에 실용적이지 않다. 질의어 번역방법은 대역어 사전을 이용해서 번역을 하는데, 대부분의 연구에서는 양국어 사전이나 다국어 사전이 이용 가능한 경우, 단순하고 실용적이기 때문에 사전을 이용한 질의번역방법(Dictionary-based query translation method)을 채택해 오고 있다[4].

본 논문에서는 오픈 API로 다국어 사전 정보와 번역 기능을 제공하고 있는 서비스를 이용하여 번역기능을 구현하고자 한다. 본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 살펴보고, 3장에서는 본 논문에서 제안한 시스템을 구현하고 인터페이스를 제시한다. 마지막으로 4장에서는 결론 및 향후 과제에 대해서 기술한다.

II. 관련연구

현재 오픈 API는 수백 개 이상 존재하며, 음악, 검색, 지도와 같이 지속적으로 API가 추가되는 분야가 있는 반면, 파일 공유, 지불 수단 등 초기에 등록된 이후 더 이상 변화가 없는 분야와 보안, E-mail, 데이터베이스 등 최근에 API가 제공되기 시작한 분야도 있다. 대표적인 오픈 API로는 구글 번역(Google AJAX Language), 플리커(Flickr), 유튜브(Youtube) 등이 있다[5]. 국내에서는 네이버(Naver)와 다음(Daum)에서 제공하는 오픈 API가 많이 사용된다.

오픈 API의 기본적인 개념은 그림 1과 같다. 오픈 API 제공자는 오픈 API 사용자들에게 각각 인증키를 분배하고 사용자들은 그 인증키를 소유함으로써 제공자의 오픈 API를 사용할 수 있게 되는 것이다[4].

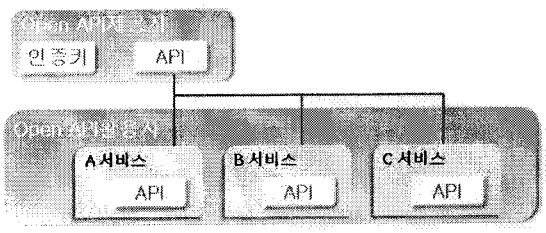

그림 1. 오픈 API의 개념

플리커(Flickr)는 온라인 사진 공유 커뮤니티 사이트로 소셜 웹의 대표적인 사이트이다. 이 서비스는 개인 사진을 교환하는 목적 이외에도 사진을 올려 저장하는 용도로 쓰이기도 한다. 사용자는 태그를 이용해서 사진들을 분류하는 것이 가능한데, 이것은 나중에 검색자가 장소 이름이나 주제 같은 것을 가지고 검색하는 일을 용이하게 해준다.

유투브(YouTube)는 온라인 동영상 공유 커뮤니티 사이트로 플리커와 같이 소셜 웹의 대표적인 사이트이다. 사용자가 영상 클립을 업로드하거나, 보거나, 공유할 수 있다. 구글 번역 API(Google AJAX Language API)는 구글에서 오픈한 번역 API이다. 이것은 41개국 간의 간단한 문장을 번역할 수 있다.

III. 연구 설계

이 장에서는 본 논문에서 제안한 시스템을 구현하고 인터페이스를 제시한다.

아래의 그림 2는 본 시스템의 구조도이다. 사용자가 질의어를 시스템에 보내게 되면 번역 API나 사전 API를 이용하여 질의어를 번역한 후 번역된 질의어를 이용하여 소셜 웹 사이트에서 질의어에 해당하는 검색 결과를 추출하여 사용자한테 보여준다.

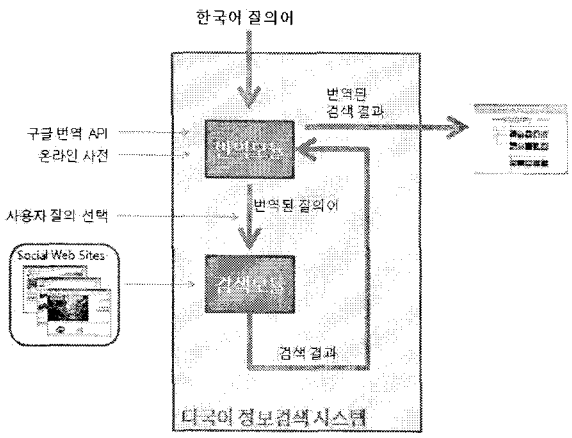

그림 2. 시스템 구조도

3-1. 번역 모듈

본 논문에서는 Google AJAX Language(구글 번역)를 이용하여 사용자의 모국어를 다른 언어로 번역하여 검색을 수행한다. 하지만 Google AJAX Language는 간단한 문장까지 번역이 가능하지만 변역 결과가 하나만 제공되기 때문에 일부의 사람들에겐 만족할 만한 결과를 주기 어렵다. 이러한 단점을 보완하기 위하여 온라인 사전을 이용한다. 현재 대부분의 사이트에서는 사전 기능을 저작권 때문에 오픈 API로 공개하지는 않는다. 따라서 본 논문에서는 웹 사이트에서 제공하는 온라인 사전을 HTML Parser를 이용하여 정보를 추출, 가공하여 사용한다. Google AJAX Language과 온라인 사전의 내용을 사용자에게 제시하고, 선택하게 함으로써 사용자마다 같은 질의어라도 다른 번역이 가능하기 때문에 다른 검색 결과가 나타나게 된다.

3-2. 소셜 웹 API 검색 모듈

본 논문에서는 대표적인 소셜 웹 사이트인 플리커와 유튜브, 그리고 국내의 다음(Daum)과 네이버(Naver)의 오픈 API를 이용하여 검색 모듈을 구성하였다. 플리커로부터는 반환값을 JSON 형식으로 받아서 파싱하고 URL, TITLE, LINK 등을 추출한다. 유튜브와 다음, 네이버로부터는 반환값을 XML 형식으로 받아서 파싱하여 URL, TITLE, LINK 등을 추출한다. 이렇게 추출된 정보에 번역기능을 한 번 더 적용하게 되는데, 추출된 TITLE을 한국어로 번역하고, 외국어로 된 원래의 TITLE과 함께 보여 줌으로써 사용자의 이해를 돕는다.

3-3. 결과 화면 및 인터페이스

3-2절에서 추출한 정보를 사용자에게 편리한 형태의 인터페이스를 구성하여 보여준다. 각각의 오픈 API에서 검색된 결과 중 상위 5개만을 보여주고 각각의 데이터마다 그 데이터의 타이틀과 링크, 번역 타이틀을 보여준다.

그림 3은 사용자가 입력한 질의어를 번역한 화면이다. 아래의 화면은 한글 질의어 "눈"을 입력하고 영어로 번역 하였을 때의 결과화면이다. 왼쪽에는 한글로 뜻이 나오고 오른쪽에 영어로 번역된 단어들이 나온다. 사용자는 이 중 자신이 원하는 뜻을 선택하게 된다.

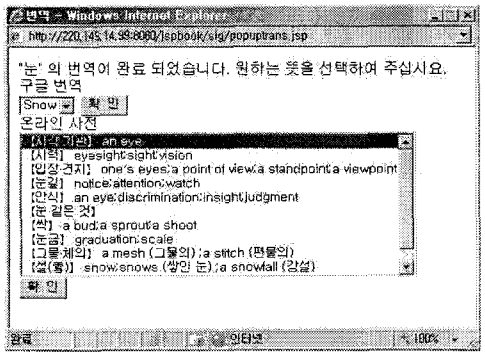

그림 3. 대역어 선택 화면

다음은 번역 기능을 사용하지 않고 검색을 한 경우의 결과 페이지이다. 왼쪽에는 질의어에 대한 추천 검색어가 나타난다. 번역을 하지 않았기 때문에 검색 결과와 타이틀, 링크만 보여준다.

그림 4. 검색결과(번역안함)

다음은 번역 기능을 사용하여 영어로 검색을 한

경우의 결과 페이지이다. 번역 기능을 사용하였기 때문에 검색된 결과가 모국어가 아닐 경우 타이틀을 모국어로 번역하여 사용자에게 보여준다. 이때 각 나라의 언어별로 고유의 색을 지정하여 번역된 타이틀을 각 나라의 지정된 고유색으로 보여준다. 나머지는 번역을 하지 않은 경우와 동일하다.

그림 5. 검색 결과(영어 번역)

다음은 질의어(コミック :코믹)를 일본어로 번역을 하여 검색한 경우의 결과 페이지이다.

그림 6. 검색 결과 (일본어 번역)

Ⅳ. 연구의 의의

본 논문에서는 오픈 API를 이용하여 다국어 정보검색 시스템을 모델링하였다. 구글의 오픈 API인 Google AJAX Language API를 이용하여 질의어를 번역하여 검색을 수행한다. 하지만 번역 결과가 하나만 제시되기 때문에 완전한 질의어 번역을 원하는 사용자에게는 부족한 면이 있다. 따라서 본 연구에서는 온라인 사전을 추가로 이용하여 이 부분을 보완한다. 구글의 번역결과와 온라인 사전의 검색결과를 사용자에게 나열하여 보여준 후, 사용자가 원하는 뜻을 선택하게 하고, 선택된 번역 질의를 가지고 소셜 웹 사이트의 오픈 API를 이용하여 검색을 수행하게 된다.

검색된 결과 내의 텍스트도 번역 기능을 이용하여 번역된 형태로 사용자에게 제시한다.

향후 연구로는 사용자 프로파일과 검색 이력 등에 근거하여 자동으로 대역어를 선택/추천하는 기능과, 더 나아가 개인의 성향에 특화된 검색과 추천을 할 수 있는 지능형 서비스에 관한 연구를 할 계획이다.

참 고 문 헌

[1] 고윤미, 오기남, 권경희, "GPS와 오픈 API를 이용한 모바일 일정관리 매쉬업 서비스 구현", 한국컴퓨터종합학술대회 논문집, Vol.35, No.1(D), 2008, pp.281-284

[2] 김진한, 이병정, "웹 서비스와 오픈 API를 사용한 SOA 기반 동적 서비스 합성 프레임 워크", 정보과학회논문지:소프트웨어및 응용, 제 36권, 제 3호, 2009, pp.187-199

[3] 최용석, 최기선, "과도한 지식을 요구하지 않는 공통기반축에 의한 용어 번역과 한영 교차정보검색에의 응용", 한국 인지과학회 논문지, 제 14권, 제 1호, 2003, pp.29-40

[4] 이경순, "한국어-영어/일본어-영어 교차언어정보검색에서 클러스터분석을 통한 성능 향상", 정보처리학회논문지, 제11-B권, 제 2호, 2004, pp.233-240

[5] 천동석, 차승준, 김경옥, 이규철, "키워드를 이용한 효율적인 웹서비스 및 오픈 API 검색 엔진 개발", 한국컴퓨터종합학술대회 논문집, Vol.35, No.1(C), 2008, pp.159-164