교차언어 정보검색을 통한 DL연합
Integrating Digital Libraries with Cross-Language IR

맹성현, 충남대학교 컴퓨터과학과
장명길, 전자통신연구원
Sung Hyon Myaeng, Chungnam National University
Myung-Gil Jang, ETRI

디지털도서관의 특징으로 가장 중요한 것은 상호운용성의 보장이다. 이질언어가 사용되는 데이터베이스를 검색할 수 있는 기능을 제공하는 것은 국가간 디지털도서관의 상호운용성 제공에 필수적이며 국가 경계를 초월하는 디지털도서관 연합의 초석이 된다. 본 논문에서는 질의에 사용되는 언어와 검색되는 문서를 기술하는 언어가 다른 경우를 일컫는 교차언어검색(Cross-Language Information Retrieval)연구 및 관련 기술 개발 현황을 소개하고 본 연구팀이 개발한 효율적이고 효과적인 교차언어검색 방법과 그 성능을 기술한다. 이 방법은 한영 교차언어검색 상황에 적응되었는데, 교차언어검색에 필요한 언어 자원을 최소하면서도 단일언어검색의 신뢰도 수준에 매우 가깝게 접근한다는 특징을 가지고 있다.

1.개요

인터넷과 웹의 발전으로 전세계에 분산된 디지털 도서관에 더욱 쉽게 접근하여 정보를 획득할 수 있는 환경이 마련되었다. 또한 미국을 비롯한 영어권 디지털 도서관 뿐 아니라 아시아, 유럽에서는 영어가 아닌 자국어 언어로 작성된 디지털 도서관 시스템의 구축이 매우 활발히 이루어지고 있다. 하지만 이러한 디지털 도서관의 대부분은 자국어 언어를 사용하여 접근할 수 있는 단일언어 환경의 디지털 도서관 시스템으로 구축되고 있다.

디지털도서관의 특징으로 가장 중요한 것은 상호운용성의 보장이다. 이질언어가 사용되는 데이터베이스를 검색할 수 있는 기능을 제공하는 것은 국가간 디지털도서관의 상호운용성 제공에 필수적이며 국가경계를 초월하는 디지털도서관 연합의 초석이 된다. 뿐만 아니라, 국내에서 제공하는 디지털정보에 타 언어 문서가 포함되어있는 경우도 흔히 존재한다.

이렇게 질의에 사용되는 언어와 검색되는 문서를 기술하는 언어가 다른 경우를 교차언어검색이라 하는데, 이용자는 일반적으로 자국어를 사용하여 자국어가 아닌 영어를 비롯한 외국어로 쓰여진 원하는 정보를 검색하는 방법을 이용할 수 있다. 매년 영어 외의 언어로 된 웹 문서의 급격한 증가로 교차 언어 검색을 포함한 다국어검색에 대한 요구가 증가하고 있고, 이러한 요구에 발맞추어 최근에 국제 학술대회와 프로젝트를 통한 활발한 연구가 이루어지고 있다.

본 논문은 다국어 전자 도서관 시스템 환경에서 다국어 접근을 위한 인터페이스로 교차언어 정보검색을 활용하는 문제에 대하여 설명한다. 다음 2 장에서는 디지털도서관에서의 다국어 정보 접근의 문제에 대하여 살펴보고 3 장에서는 교차언어 정보검색의 기술 현황을 소개한다. 4장에서는 본 연구팀의 접근 방법을 소개하고 마지막으로 결론을 맺는다.

2. 디지털 도서관에서의 다국어 정보접근

디지털 도서관은 다매체(multi-media), 다언어(multi-lingual) 디지털 정보의 색인과 검색의 문제를 다루어야 한다(Borgman, 1997). 여기서는 다매체 디지털정보의 색인과 같은 문제는 다루지 않고 다국어 정보의 디지털 도서관에서의 정보접근의 문제만을 한정적으로 다룬다. 일반적으로 다국어 정보 접근(multilingual information access)의 목적은 사용자의 정보 요구가 표현된 언어를 포함하여 다수의언어로 표현된 정보를 얻는 것을 일컫는다. 이것은 다국어 정보 검색과 거의 유사한 의미로 사용되기도 하나, 다국어 정보의 디스플레이 및 입력 방법 등도 포함하는 포괄적인 개념으로 분산된 디지털 도서관에서 중요한 의미를 갖는다. 특히 분산 환경에서 다양한 검색서비스로부터 정보를 획득하는 경우, 다국어 정보 접근은 매우중요한 문제가 된다.

디지털 도서관에서 다국어 정보 접근의 문제는 사실상 매우 복잡한데, Peters와 Picchi(1997)는 이를 두 가지 기본적인 이슈로 정리하였다.

- 다국어 인식과 표현
- 다국어 혹은 교차언어 검색

2.1 다국어 인식 및 표현 문제

첫 번째 이슈는 사용자가 위치한 장소나 저장된 정보에 쓰여진 언어에 상관없이 디지털 도서관 정보에 접근할 수 있게 하는 것이다. 다양한 언어로 작성된 문서를 자유롭게 디스플레이하고 질의를 손쉽게 입력할 수 있도록 하는 문제와 다양한 문자의 코딩 문제 등이 여기에 속한다. 또한 임의의 문서가 주어졌을 때, 이 문서에 사용된 언어를 인식하는 문제도 여기에 속한다.

언어의 코드 문제는 최근에 국제 표준으로 채택된 ISO 10646의 Unicode를 사용하는 경우 16 비트 코드를 따르는 아시아, 유럽의 여러 언어를 호환성 있게 표현하고 해석하여 이러한 언어를 사용하는 문서를 디지털도서관을 통해 교환하는데 근본적인 해결책을 제시해 주고 있다. 임의의 문서에 사용되는 언어의 인식 문제도 연구가 되어 언어 및 학습 데이터의 크기에 따라 혼돈되는 정도가 다르기는 하지만 대부분언어의 경우 학습 데이터의 크기가 큰 경우 99%이상의 인식율을 가진 방법이 개발 되었다(Ludovik, 1999).

입력 방법의 경우 중국어나 일본어와 같이 한자로 인한 기본 글자수가 큰 경우오래 전부터 연구되어 왔는데, 일본어의 경우, 히라가나 혹은 카다가나를 사용하여 소리에 의해 입력한 후 시스템이 제공하는 한자 후보 중에 하나를 선택하는 방법을 주고 사용하고 있고, 중국어의 경우 pinyin이라고 불리는 로마자화 시킨 방법을 사용하고 있다. 그러나 이러한 방법들은 모두 각 언어에 대한 상당한 수준의 지식을 요구하고 있고 소프트웨어가 제공하는 입력 방법에 대한 숙련도를 요구한다. 이러한 문제점 들은 아래에서 언급하는 교차언어 검색을 통해 어느 정도 해소될 수 있다.

디지털도서관을 사용하여 다국어 문서에 접근하는 경우, 임의의 문서가 주어졌을 때 여기에 사용된 코딩 기법과 언어를 인식한 후 이를 디스플레이하기 위해서는 이 언어에 대한 폰트가 설치되어 있어야한다. 웹 브라우저에서 디스플레이 언어를 선택하여 해결되는 경우도 있으나 아직 포함되지 않은 언어 코드가 있을 뿐만 아니라 부분적으로 다른 언어를 가진 문서를 사용자의 디스플레이 하거나 자동적으로 언어를 인식하여 디스플레이 하는 기능은 아직 제공되지 않는다. 이러한 문제에 대한 부분적인 해결 방법으로 최근에 발표된 방법은 필요한 폰트를 해당 소프트웨어에 설치하지 않은 상황에서도 요구된 문서와이를 디스플레이 하는데 필요한 폰트에 해당하는 최소한의 문자형(glyph)을 같이 전송해 준다(Maeda et a1., 1998).

2.2. 분산 다국어 DB 검색 문제

첫 번째 이슈가 주로 언어를 구성하는문자 형태의 다양성에서 오는 표층적인 문제라면, 다국어 검색 문제는 언어간의 의미적 호환성을 위한 것이라 할 수 있다. 특정 지역 혹은 기관의 디지털도서관에서도 다양한 언어로 쓰여진 문서가 있는 경우 다국어 정보검색의 문제가 대두된다. 그러나 이 문제는 분산 디지털도서관 환경에서 문서를 통합적으로 검색하는 연합탐색(Federated Search)을 수행하는데 있어 반드시 해결되어야 하는 심각한 문제이다.
연합탐색기는 사용자가 작성한 하나의 질의를 사용하여 분산되어 있는 다양한 디지털도서관으로부터 필요한 정보를 검색한 후 그 결과를 융합한 후 최종 결과를 제시함으로써 분산 투명성을 제공한다. 이 과정에서 사용자 질의는 다양한 언어로 변환되어 해당 디지털도서관으로 전송되어야하는데, 변환 및 분배를 맡는 소프트웨어가 지역적으로 분산되어 있는 디지털도서관이 가지고 있는 데이터베이스의 언어 및 특성을 파악하는 것이 중요하다. 이 문제를 체계적으로 해결하는 하나의 방편으로 연합탐색에 필요한 다양한 정보의 교환을 정의하는 STARTS프로토콜(Gravano *et al.*, 1997)을 기반으로 언어 정보를 교환하는 방법이 제시되고 있다(Hayashi *et al.*.1999).

3. 교차언어 정보검색

교차언어 정보검색은 다국어 정보검색의 부분에 해당하는 개념으로, 하나의 언어를 사용하여 이와 상이한 언어로 된 문서를 검색할 수 있게 하는 정보검색을 말한다. 교차언어 정보검색은 질의의 언어와 문서의 언어와의 차이를 극복하기 위하여 질의변환을 수행해야 하는데 질의 언어를 문서언어로 변환하는 질의 변환이나 혹은 문서를 질의 언어로 변환하는 문서 변환이 있을 수 있다. 최근에는 이들 방법을 혼합한 방법(McCarley, 1999)이나 직접적인 언어변환을 수행하지 않는 LSI(Latent Semantic Indexing)를 사용한 교차언어 검색 방법(Dumais et al., 1997) 등 많은 연구가 이루어지고 있다.
Oard & Hackett(1997)의 문서 변환에서는 문서들을 질의 언어로 변환하는데 고품질 기계번역 시스템을 활용하고 있지만, 이 방법은 현재의 기술 수준으로 대규모로 적용되기에는 비실용적이다(Carbonell *et al.,* 1997). 반면에 질의를 문서 언어로 변환하는 질의 변환은 문서 변환과 비교하여 훨씬 간단하고 더욱 경제적이기 때문에 실용적인 방법으로 부각되어 왔다. 질의 변환은 크게 사전 기반 접근 방법, 시소러스기반 접근 방법, 그리고 코퍼스 기반 접근방법으로 구분할 수 있는데, 이들 방법 중하나 혹은 둘 이상이 함께 적용될 수 있다.

이러한 교차언어 정보검색의 접근 방법들에 대한 대표적인 기술을 먼저 살펴보고(맹성현, 1999) 본 연구팀이 개발한 기술에 대한 개요 및 접근 방법을 4장에 소개한다.

3.1사전 기반 방법

사전 기반 질의 변환 방법은 대역 사전을 사용하는 비교적 단순한 질의 변환 방법 이다. 대역 기계판독형사전(MRD: Machine Readable Dictionary)을 이용하여 질의 단어 혹은 구를 문서 언어로 변환하는데 이러한 사전은 다른 언어 자원들보다 손쉽게 구할 수 있어 가장 쉽게 구현이 가능하여 많이 채택되고 있다. 하지만 사전의 단순한 사용에 의한 질의 변환은 정확율과 재현율로 측정한 검색 성능이 단일언어 검색의 40%-60% 정도에 불과한 것으로 나타났다(Ballesteros, 1997). 검색 효과하락의 원인으로는 근본적으로 대역 사전의 엔트리가 하나 이상의 의미를 가지는 변환 모호성이 발생하기 때문이고, 그 밖에 전문 용어나 외래어와 같은 미등록 단어의 변환 실패, 그리고 구절과 같은 여러 단어가 하나의 질의 용어를 구성하는 경우의 변환 실패나 부분적인 변환으로 인한 원인들이 있다.

최근에는 기본적으로는 사전 기반 질의 변환을 채택하지만 질의 변환의 모호성 문제에 대처하여 검색 성능을 향상시키기 위하여 다른 추가적인 자원들을 함께 이용하는 연구들이 있다. Yamabana(1996)가 제안한 DMAX(Double MAXimize) 방법은 대역 사전으로 변환된 가능한 후보들로부터 모호성을 해소하기 위하여 원시 단어와 목적 단어 사이의 공기 빈도를 동시에 최대로 하는 단어 쌍을 선택하는 통계적인 단어 선택 방법을 제시하였다. Twenty-One 시스템(Kraaij, 1997)에서 구현된 질의 변환 방법은 네덜란드어-(독어, 불어, 영어, 스페인어) 대역 사전 뿐 아니라 표준 자연어처리 도구들을 사용하는데, 대역 문서 코퍼스로부터 추출된 명사구 후보들에 근거하여 변환 명사구들의 모호성을 해소한다. Hull(1997)은 가중치를 사용한 Boolean 모델에 기반한 질의 변환 방법을 통하여 질의 단어들의 Boolean 조합을 형성하면서 단어들에 대한 가중치를 계산한다.

3.2 시소러스 기반 방법

교차언어 검색에서 다국어 시소러스를 사용하는 초기의 방법으로 통제 어휘를 사용한 방법은 문서들이 미리 정해진 어휘로 수작업으로 색인되고 사용자는 같은 어휘를 사용하여 질의를 표현하도록 하였다. 이 방법은 좁은 도메인에서는 단일어 검색의 100%까지 검색 효과를 나타내기도 하나 확장성에 있어 제한된다는 단점을 가지고 있다.

다국어 시소러스 기반 방법은 각 언어로부터 선택된 용어들을 언어 독립적인 개념 표시자(concept descriptor)의 공통 집합에 매핑하여 검색 문서의 선택이 이들 개념 표시자의 정확한 매칭에 의하여 이루어지도록 하는 방법이다. 영어의 경우 Roget 시소러스나 WordNet을 이용한 다국어 시소러스 기반 교차언어 검색 방법이 소개되고 있다. 여기서 사용하는 다국어 시소러스는 기존의 시소러스를 번역하거나 기존시소러스들의 통합을 통하여 구축될 수 있으나 몇 가지 한계점을 가진다. 즉 시소러스에 사용된 용어의 문화적 경계를 초월하는 지식구조를 생성하는 문제, 도메인 유지를 위한 전문가에 의한 수작업 색인의 부담, 새로운 도메인으로 확장의 어려움, 그리고 사용자의 개념 표시자 선택 및 지식 구조 탐색 등의 어려움이 있다.

하지만 최근에는 다국어 시소러스의 구축과 관련하여 이러한 한계점을 극복하기 위한 여러 가지 노력들이 시도되고 있다. 대규모 다국어 코퍼스를 이용한 자동 시소러스 생성 방법의 연구, 자동 시소러스 병합 도구 개발, 기계의 도움을 받는 다국어색인 도구, 지식 구조의 시각화를 위한 그래픽 인터페이스 개발이 이루어지고 있다.

3.3 코퍼스(corpus) 기반 방법

다국어로 이루어진 코퍼스는 교차언어검색에 매우 중요한 자원으로 사용되는데, 코퍼스간의 정렬정도에 따라 병행(parallel)코퍼스, 비교(comparable)코퍼스, 비정렬(unaligned) 코퍼스로 구분된다. 병행 코퍼스는 각 언어로 된 두 코퍼스가 동일 내용의 문서나 문장 혹은 용어 단위로 정렬된 코퍼스이고, 비교 코퍼스는 두 코퍼스의 상응하는 문서가 각각의 내용을 번역한 형태가 아니라, 단지 동일 주제의 내용으로 된 구성된 경우이다. 비정렬 코퍼스는 동일 영역의 문서로 이루어 졌을 뿐 문서 간 내용 상의 대응성도 가지지 못한 코퍼스이다.
코퍼스를 이용한 질의 변환 연구를 몇 가지 살펴본다. Davis(1998)는 UN 병행 코퍼스를 사용하였는데, 사전을 이용하여 영어 질의를 스페인어로 변환한 후, 다양한 후보 중 최적의 스페인어 질의를 선택하기 위하여 병행 코퍼스를 사용한다. 영어 질의로 검색된 문서집합과 각 스페인어 질의로 검색된 문서집합 간의 유사성 정도를 계산한 후, 영어문서 집합과 가장 유사한 문서집합을 검색한 스페인어 질의를 최종질의로 선택하고 실제 스페인어 문서 검색에 사용한다. Carbonell이 이끄는 CMU팀(1998)에서는 일반화된 벡터공간모델(Generalized Vector Space Model) 방법을 사용하는데, 원시 언어와 대역 언어의 코퍼스를 두 개의 행렬로 표현하고 이들의 열을 기준으로 문서 쌍이 매치되도록 하여 질의 변환과 문서 변환에서는 각 행렬의 열을 기준으로 유사도를 계산하여 교차언어 검색을 수행한다. 이와 유사한 방법으로 LSI를 이용한 교차언어 문서검색 방법(Dumais *et al.*, 1997)은 기본적으로 문서와 질의의 용어들의 행렬을 이용하는데, 이들을 하나의 동일한 공간에 매핑시켜 검색을 수행한다. 이 방법은 특별히 변환을 수행하지 않고도 교차언어 검색 효과를 얻을 수 있어 현재 교차언어 검색의 새로운 이론으로 연구되고 있다.

4. 실용성 위주의 한영교차언어 검색

본 팀의 연구에서는 코퍼스나 시소러스기반 질의 변환 방법보다는 단순성과 실용성에 초점을 두어 사전 기반 질의 변환 방법을 채택하였다. 사전을 기반으로 질의변환을 한 후 여기서 발생하는 모호성 해소를 하기 위해 목적 언어 코퍼스, 즉 검색 대상 문서 집합을 이용하는 접근 방법을 사용하였다.
일반적으로 교차언어 문서검색에서는 해결해야 하는 세 가지 문제가 있다(Grefenstette, 1998). 첫 번째 문제는 한 언어로 쓰여진 질의를 다른 언어로의 변환을 수행하는 방법에 관한 것이고, 두 번째 문제는 변환된 후보 단어들 중에서 어떤 불필요한 단어들을 제거하는가를 결정하는 것이다. 이는 단어 모호성 해소의 문제에 해당된다. 세 번째 문제는 하나 이상의 후보 단어들의 상대적인 중요성도에 따라 적절히 가중치를 주는 방법을 결정하는 것이다. 본 접근 방법에서는 마지막 두 가지 문제들에 초점을 맞추어 대역 문서 집합으로 부터 추출한 상호 정보(Church,1990)통계치를 이용하여 변환 모호성 해소와 가중치 부여의 문제에 대처하는 비교적 단순하면서도 효과적인 방법을 채택하였다.

4.1 변환 모호성의 분석

질의 변환을 수행하는 가장 간단한 방법은 대역 사전을 사용하는 것이지만, 대역 사전에 존재하는 일 대 다 매핑 때문에 변환 모호성의 문제를 가지게 된다. 예를 들어, "자동차 공기 오염"이라는 세 단어로 구성된 한국어 질의의 변환에 한국어-영어대역 사전이 직접 사용될 때 각 단어는 다수의 영어 단어들로 변환될 수 있다. 질의의 첫 번째 단어 "자동차"는 motocar, automobile, car와 같은 의미적으로는 비슷하지만 다른 영어 단어들로 변환된다. 두번째 단어 "공기"는 다른 의미를 가지는 영어 단어들인 air, atmosphere, empty vessel, bowl들로 변환된다. 그리고 마지막단어 "오염 "은 pollution과 contamination으로 변환된다. 다수의 후보 단어들을 질의로 그대로 사용하는 것은 단일언어 문서검색에서 재현율을 증가시키는 데 유용할 수 있는 반면에, 단어들의 의미 모호성 해소에서 실패하는 경우에는 검색 효과에 나쁜 영향을 미칠 수 있다고 이전의 연구들은 지적한다. 예를 들어, empty vessel과 같은 구절은 질의의 의미를 전체적으로 변경시킬 수도 있으며, 심지어 pollution과 동의어인 contamination 같은 단어는 의미에서의 약간의 차이로 인하여 관련이 없는 문서들을 검색하게 될 수도 있다.

표1. 모호성의 정도

	Words			Word Pairs		
	# in S. Lan.	# in T. Lang.	Average Ambiguity	# in S. Lan.	# in T. Lang.	Average Ambiguity
Title	48	158	3.29	24	212	8.83
Short	112	447	3.99	91	1459	16.03
Long	462	1835	3.97	423	6196	14.65

표 1은 한국어-영어 대역 사전이 형태소 분석과 태깅 후에 단순히 사용되는 경우에 질의어 변환에서 모호성이 어느 정도 나타나는 지를 보여준다 (Jang *et* al.,1999). 세개의 행 title, short, long은 TREC 문서집합에서 질의를 구성하는 세 가지 다른 방법들을 가리킨다. 표의 왼쪽은 각 질의에 대하여 한국어 단어에 대한 대역 영어단어들의 평균 수를 나타내고 반면에 오른쪽은 한국어 단어 쌍으로부터 만들어질 수 있는 대역 영어 단어 쌍들의 평균 수를 나타낸다. 이것은 질의 변환의 모호성 해소과정에서 평균적으로 9 개 이상의 가능한 쌍들로부터 하나를 선택해야 한다는 것을 의미한다.

4.2 상호 정보를 이용한 한국어-영어 질의 변환

본 연구에서는 교차언어 문서검색을 위하여 병렬 코퍼스, 비교 코퍼스, 혹은 다국어 시소러스 등의 구축하기 힘든 자원들에 의존하지 않고 사전에 기반한 실용적인 질의변환 방법을 사용한다. 마찬가지로 교차언어 문서검색 환경에서 항상 얻을 수 있는 대역 언어로 된 문서집합에서 추출한 상호 정보를 이용하여 질의 변환의 변환 모호성을 해결하는 접근 방법을 채택하였다. 본 논문의 한국어-영어 질의어 변환 방법은 그림 1과 같이 네 단계로 수행된다.

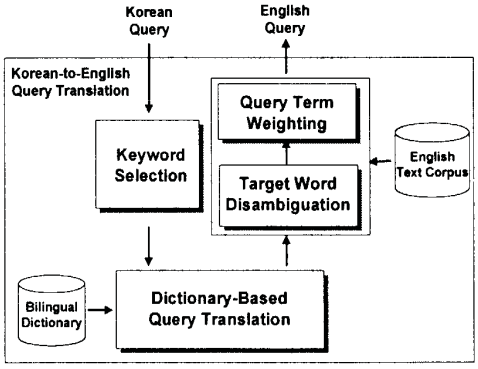

그림 1. 한국어-영어 질의어 변환 단계

즉 키워드 추출, 사전 기반 질의어 변환, 대역 단어 의미 모호성 해소, 그리고 질의 텀 가중치 부여이다. 특히 세 번째와 네 번째 단계에서 사용하는 목적 언어코퍼스는 교차언어 검색 환경의 대부분언어 쌍에서도 상대적으로 쉽게 구할 있는 자원을 활용하는 실용성에 초점을 주었다. 본 연구에서는 TREC-6 교차언어 문서검색 환경의 한국어-영어 질의 변환방법의 실험에서는 1988~1990 AP 뉴스퍼스로부터 추출한 상호 정보를 사용하다.

본 질의 변환에서 사용하는 상호 정보는 단어가 하나의 코퍼스에 동시에 나타는 정도인 공기 통계치에 근거하여 계산되는데, 단어들 사이의 상관성을 나타내 측정 장치로 사용된다. 상호 정보 MI(x, y),는 다음 공식으로 정의된다(Church & Hanks, 1990).

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)} = \log_2 \frac{N \cdot f_w(x,y)}{f(x)f(y)}$$

여기서 x 와 y 는 ω 단어들의 범위 내에서 함께 나타나는 단어들이다. 본 연구에서는 숙어나 구절 같은 표현들을 구성하는단어들의 문맥 관계를 포함하도록 하기 위하여 w=6으로 정하였다.

상호 정보는 질의 변환에서 한국어 질의 단어가 하나 이상의 영어 단어들로 변환될 때 가장 가능성 있는 변환들을 선택하는 모호성 해소와 단어들의 중요도를 나타내는 가중치 부여의 기준으로 사용된다. 변환 모호성을 해결하는 본질의 변환 방법의 모호성 해소 및 가중치 부여 방법은 가장 큰 상호 정보 값을 가진 단어 쌍을 중심으로 그 단어 쌍의 전 후로 모호성 해소와 가중치 부여를 적용하여 확장해 나가는 방법이다(Jang *et* al., 1999, 장명길 외,1999). 이 방법은 상호 정보 값들의 상대적이면서 절대적인 중요도를 감안하여 가장 높은 상호 정보 값을 가진 단어 쌍을 중심으로 먼저 후보들을 선택하고 진행한다는 것이 특징이다. 즉 통계적으로 볼 때 동시에 출현할 가능성이 많은 단어쌍을 선택한 후 이들과 동시에 출현할 단어들을 찾아나가는 방식이다. 여기서 질의 용어 가중치 부여는 개념적으로 관련이 없는 용어들을 잘라내는 모호성 해소 과정에 덧붙여 사용하는데, 기본 아이디어는 가장 좋은 후보에게 큰 가중치를 주고 나머지 후보들에는 나머지 가중치를 똑같이 나누어 부여하는 것이다.

TREC-6 교차언어 테스트 문서집합을 사용한 실험에서 본 한-영 질의 변환 방법의 검색 효과는 단일언어 문서검색 경우의 85% 까지 도달하는 성능을 보여 사전만을 이용하는 타 연구 결과에 비해 매우 높은 신뢰도를 보여 주었다. 현재 다양한 모호성 해소 방법에 대한 종합적인 연구를 통해 성능 향상을 도모하고 있으며 제시하는 접근 방법이 제공할 수 있는 최대치까지 도달할 수 있는 방법에 대한 연구를 수행하고 있다.

5.결론

본 논문에서는 효과적인 분산 디지털도서관의 구축을 위해 필수적으로 해결되어야 할 다국어 정보접근 문제에 대한 최근의 연구 동향을 소개하였다. 언어를 구성하는 문자의 특성에 의해 발생하는 표층적인 문제와 언어간 존재하는 의미상의 차이점을 극복하는데 있어 가장 시급히 해결되어야 할 교차언어검색 문제로 구분하여 해결되어야 할 문제점들을 설명하였고 현재진행 중이거나 해결된 방법들을 설명하였다.
교차언어검색의 경우 국내외적으로 현재 매우 활발한 연구가 진행되고 있는데, 본 연구팀에서는 다국어 디지털도서관 시스템을 위한 실용적인 교차언어 방법을 개발하였다. 이를 본 연구팀에서 수년간 개발해 온 MIRAGE-II(Myaeng *et al.*, 1999)에 적용할 예정이며, 그 신뢰도 향상을 위한 연구를 지속할 예정이다.
한영 교차언어검색 기술이 그 효용성면에서 가장 중요하긴 하지만 다국어 디지털 도서관에서의 교차언어 검색용 인터페이스의 개발 관점에서는 기타 언어에 대한 자유로운 교차언어 검색 기능을 제공하는 것이 필요하다. 그러나 많은 언어의 경우 한국어와 그 언어간의 기계판독형 사전이 존재하지 않거나 사용할 만한 품질이 못되는 경우가 많다. 다행히 대부분의 목적 언어와 세계공용어인 영어간에는 이용할 만한 사전이 존재할 것으로 기대되므로 한국어-영어-목적언어 순으로 질의 변환하는 방법의 개발이 필요하다. 즉 영어를 피봇언어로 사용한 다른 두 언어 간의 질의 변환 문제에 대한 연구를 수행할 예정이다.

참고 문헌

[Ballesteros, 1997] Lisa Ballesteros and W. Bruce Croft, "Phrasal Translation and Query Expansion Techniques for Cross-lingual Information Retrieval", *SIGIR'97*, 1997.
[Borgman, 1997] Christine L. Borgman, "Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries: Or How Do We Exchange Data In 400 Languages?", *D-Lib Magazine*, June 1997.
http://www.dlib.org/dlib/june97/06borgman.html
[Carbonell *et al.*, 1998] J.G. Carbonell, Y. Yang, R.E. Frederking, R.D. Brown, Yibing Geng and Danny Lee, "Translingual Information Retrieval: A Comparative Evaluation". In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1998.
[Church, 1990]Kenneth W. Church and Patrick Hanks, Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics*, Vol. 16, No.1, pp. 22-29, 1990.

[Davis, 1998] Mark Davis, "On the Effective Use of Large Parallel Corpora in Cross-Language Text Retrieval", In Cross-*Language Information Retrieval* (ed: Gregory Grefenstette,), Kluwer Academic Publishers, 1998.
[Dumais *et*. al., 1997] S. T. Dumais, T. A. Letsche, M. L. Littman and T. K. Landauer, "Automatic Cross-Language Retrieval using Latent Semantic Indexing", In *Proceedings of AAAI Spring Symposium* on *Cross-Language Text and Speech Retrieval*, 1997.
[Grefenstette, 1998] Gregory Grefenstette, Cross-Language Information Retrieval, Kluwer Academic Publishers, 1998.
[Hayashi *et*. al., 1999] Yoshihiko Hayashi, Genichiro Kikui and Toshiaki Iwadera, "A Scalable Cross-Language Metasearch Architecture for Multilingual Information Access on the Web" In *Proceedings of Machine Translation Summit VII '99,* Singapore, September 1999.
[Hull, 1996] David Hull, "A Weighted Boolean Model for Cross-Language Text Retrieval". In *Proceedings of the 19th Annual ACM SIGIR Conference on Information Retrieval,* Zurich, Switzland, 1996.
[Kraaij, 1997] Wessel Kraaij and Djoerd Hiemstra, "Cross Language Retrieval with the Twenty-One System". In Proceedings of the Sixth Text Retrieval Conference *(TREC-6)*, NIST, 1997.
[Jang *et al*., 1999] Myung-Gil Jang, Sung Hyon Myaeng and Se Young Park, "Using Mutual Information to Resolve Query Translation Ambiguities and Query Term Weighting". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, June 1999.
[Ludovik, 1999] Ludovik, Y. & Zacharski, R., "Multilingual Document Langauge Recognition for Creating Corpora." In *Proceedings of Machine Translation Summit VII '99*, Singapore, September, 1999.
[Maeda *et al*., 1998] Maeda, A., Dartois, M., Fujita, T., Sakaguchi, T., Sugimoti, S., & Tabata, K., "Viewing Multilingual Documents on Your Local Web Browser," *Communications of ACM*, 41 (4).
[McCarley, 1999] J. Scott McCarley, "Should we Translate the Documents or the Queries in Cross-Language Information Retrieval?". In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, June 1999.
[Myaeng *et* al., 1999] Myaeng, S. H., Lee, M.-H., & Kang, J.-H, "Virtual Documents: a New Architecture for Knowledge Management in Digital Libraries" To appear in *Proc. of the 2nd Asian Digital Libraries, to be held in* Taipei, Nov.
[Oard, 1997] Douglas W. Oard, "Serving Users in Many Languages: Cross-Language Information Retrieval for Digital Libraries", *D-Lib Magazine*, December 1997.
http://www.dlib.org/dlib/december97/oard/12oard.html

[Oard. 1997] Douglas W. Oard and Paul Hackett, Document Translation for the Cross-Language Text Retrieval at the University of Maryland, In *Proceedings of the Sixth Text Retrieual Conference (TREG-6)*, NIST, 1997.
[Peters, 1997] Carol Peters & Eugenio Picchi, "Across Languages, Across Cultures: Issues in Multilinguality and Digital Libraries", *D-Lib Magazine*, May 1997.
http://www.d1ib.org/dlib/may97/peters/05peters.html
[Yamabana, 1996] Kiyoshi Yamabana, Kazunori Muraki, Shinichi Doi and Shin-ichiro Kamei, "A Language Conversion Front-End for Cross-Language Information Retrieval". In *Proceedings of the 19th Annual ACM SIGIR Conference on Information Retrieval*, Zurich, Switzland, 1996.
[맹성현, 1999] 맹성현, "교차언어 정보검색", *한글 및 한국어정보처리 자연언어처리 튜토리얼*, 1999년8월.
[장명길 외,1999] 장명길, 맹성현, 박세영,"한-영 교차언어 정보검색에서 상호정보를 이용한 질의 변환 모호성 해소 및 가중치부여 방법", *제11회 한글 및 한국어 정보처리 학술대회*, 1999년10월.