94-b-1

특정어 파일에 대한 연구

A Study on Special Matching Term File

김경주 중앙일보사 데이터뱅크국

KYOUNG JU, KIM
JOONGANG DAILY NEWS DATABANK BUREAU

자동색인 시스템의 색인어 선정 능력을 향상시키기 위한 특정어파일 구축을 제안한다. 특정어파일을 도입함으로써 색인어 선정시 좀더 포괄적인(또는 세부적인) 주제어선정을 돕고 또한 전조합색인의 문제점으로 야기되기 쉬운 검색누수현상을 막을 수 있다. 특정어파일은 시소러스 기반 자동시스템의 성능을 강화하거나 시소러스파일의 대용으로 이용할 수 있을 것이다.

I. 서론

현대의 대량 정보화사회에서는 많은 정보들이 컴퓨터에 저장 관리되고 필요에 의해 이용자에게 서비스되는 정보관리 시스템이 널리 이용되고 있다. 이러한 정보관리 시스템의 수준은 그 시스템이 제공하는 정보서비스의 정확성과 정보추출의 속도에 의해 결정되며, 이때의 정확성과 속도는 그 시스템이 유지, 관리하는 색인어의 정확성에 의존한다.

정보서비스 시스템에서 정보의 내용을 표현하는 색인을 추출하는 방법으로 이전에는 주로 수작업 색인이 사용되었다. 그러나 정보량의 방대함, 색인어의 불일치성 같은 문제가 제기됨에 따라 컴퓨터에 의한 자동색인이 부각되기 시작했다.

자동색인이란 데이터베이스에 저장된 많은 정보중에서 탐색어에 따라 필요한 정보를 즉시 찾아낼 수 있도록 각각의 정보에다 그 정보를 대표할 수 있는 색인어들을 자동으로 컴퓨터가 부여하는 시스템을 말한다.

본 논문에서는 형태소분석과 시소러스를 이용한 자동색인 시스템에 특정어파일을 도입함으로써 색인어 선정시 좀더 포괄적인(또는 세부적인) 주제어 선정을 돕고 또한 전조합색인의 문제점으로 야기되기 쉬운 검색누수현상 등을 막는 방안을 제안한다.

본 연구에서 사용한 색인시스템은 중앙일보사에서 신문기사DB용으로 구축한 시스템이다.

II. 관련 연구

1. 자동색인 시스템

자동색인 방식은 크게 두 종류로 구분된다. 첫째는 어구의 출현빈도를 고려하는 통계적인 방식으로 여기에는 단순빈도에 의한 추출법(Luhn의 모델), 확률을 이용하는 방법(2 Poisson 모델), 분산을 이용하는 방법(Dennis-Salton 모델), 문서를 n차원(n개의 색인어)의 벡터로 표현하는 벡터공간모델(Vector Spacd Model) 등이 있다.

이 방식들은 주로 식별력만을 고려하며 부적절한 어구를 색인어로 선택할 수 있고, 적절한 구단위의 색인 후보어를 찾아내기 어려우므로 사용자의 질의에 적합한 문서를 검색하기에 다소 미흡하다.

둘째는 언어정보를 이용하여 문서의 의미를 바탕으로 하여 색인어를 추출하는 방식으로 형태소처리만을 하는 방식, 특정어구에 관련된 명사를 추출하는 방식, 명사구를 처리해 추출하는 방식, 구문분석이나 의미분석을 통해 명사의 역할을 규명해 추출하는 방식 등이 있다.

언어정보를 이용하는 방식의 대부분은 구문분석과 의미분석을 통해 각 명사구의 색인어로서의 자격 여부를 판정하는 방식을 채택하고 있다. 이때 자격판정 기준으로 주로 사용되는 것은 의존문법등에 의해 명사구들의 의존관계나 표현유형을 분석하여 색인과 검색에 이용하는 방식이나 격문법을 이용해 각 명사구의 역할을 규명하여 그 중요도를 결정하는 방식 등이 있다. 또한 이같은 방법과 더불어 시소러스를 이용하여 개념의 상하위관계를 고려하는 방식들도 개발되어 있다.

2. 시소러스

시소러스란 후조합을 위해 설계된 색인어의 통제어휘집으로, 용어간의 동등관계,계층관계, 관련관계를 상호 대응시켜 표준화된 관계지시기호로 명확하게 표시하고 식별할 수 있도록 배열하고 구조화한 것이다.

그러나 이상과 같은 전통적인 개념의 시소러스는 어휘간의 관계규정이 모호한 경우가 많으며, 상황과 주제에 따라 달리 구성될 수 있는 관계식이 일방적으로 결정되는 제한점을 갖고 있다. 또한 복합적인 관계용어집이 되기 보다는 세분된 주제내에서만 제한적으로 구축되는 경향이 있고, 용도에 있어서도 색인어의 통제와 전거 이상의 기능을 수행하기 어렵다.

따라서 최근에는 전통적인 시소러스가 갖는 제한점과 모순을 제거하기 위해 의미구조를 갖는 시소러스, 동적시소러스 등에 대한 연구들이 수행되고 있다.

본 연구에서는 기본적으로 전통적인 개념의 시소러스에 근거한 자동색인시스템을 가정하되, 시소러스의 자동색인 기능을 향상시킬 수 있는 특정어파일에 대한 개념을 제안하기로 한다. 특히 특정어파일은 색인시스템에서 이미 선정된 주제어의 어(語)특정에 의한 파생 주제어 선정기능을 하므로 장차 동적시소러스 구축시 유용한 도구가 될 것으로 보인다.

III. 시소러스 기반 색인시스템

1. JOINS 색인시스템

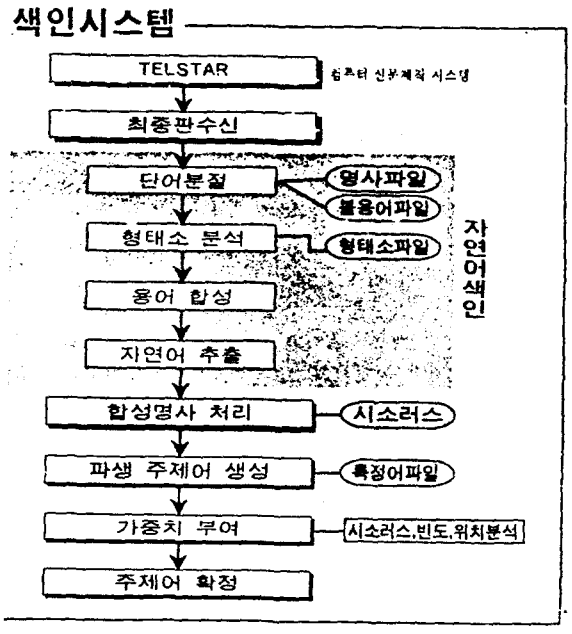
국내에 시소러스의 개념이 도입된 것은 20년 전의 일이다. 그간 학계에서는 시소러스에 대해 많은 연구를 수행하였지만 예석하게도 실용화된 예를 찾아보기는 어렵다. 단지 일부 정보처리기관에서 자체 정보관리를 위해 제한된 분야를 대상으로 한 책자형 시소러스를 개발한 예가 있을 뿐이다.

그러나 최근 정보관리분야에 본격적으로 컴퓨터가 도입되고 데이터베이스가 확산되면서 정보검색시스템의 효율을 향상시키기 위한 도구로써 시소러스에 대한 관심도 고조되었다. 최근 1~2년 사이 각계의 정보관련기관들은 시소러스가 정보처리의 모든 문제를 해결해줄 것이라는 기대감에 부풀어있는 것처럼 보인다.

본 난에서는 현재 국내에서 실용화된 유일한 시소러스 기반 자동색인 시스템인 중앙일보사의 JOINS 색인시스템에 대해 살펴본다.

JOINS 색인시스템의 특징은 합성명사 처리 및 가중치 부여시 시소러스를 이용한다는 것이다. 즉 자연이색인 시스템에서 추출된 모든 명사를 시소러스파일과 대조하여 동의어, 유사동의어 등의 비디스크립터를 디스크립터로 변환하는 한편, 시소러스파일에 매칭이 되는지의 여부와 매칭이 될 경우 시소러스파일 내에서의 어특징에 따라 중요도를 부여받게 된다.

또한 JOINS 시스템은 특정어파일에 의한 파생주제어 생성 기능을 도입하였다.



그러나 상기 시스템을 제외하면 현재 국내에서 시소러스개발을 추진하고 기관들중 특정어파일 구축을 고려하고 있는 기관은 찾아보기 어렵다. 이는 다음과 같은 시소러스 기반 색인 시스템의 한계를 고려하지 못한 결과이다.

2. 시소러스 기반 색인 시스템의 한계

시소러스 기반 색인 시스템은 기본적으로 문장 내에 발생한 용어가 어떤 형태로든 시소러스파일의 관련용어와 매칭된다는 것을 전제로 한다. 문장 내에 동의어나 유사동의어로도 등장하지 않아 자동색인에 의한 주제어 추출이 될수 없는 경우는 시소러스로도 해결할 수 없다.

바로 이런 이유에서 문맥을 이해할 수 있는 완벽한 지식베이스를 구축하지 않는 한 사람에 의한 후통제 작업이 필요하게 된다. 이런 예는 다음과 같은 경우에 해당한다.

- 눈을 질렀다 ..... 방한
  대상문헌내용 기대되는 색인어
- 눈이 났다 ..... 화재
  대상문헌내용 기대되는 색인어

IV. 특정어파일 구축

1. 특정어파일의 필요성

특정어파일이란 색인시 색인 대상 용어의 어(語)특징을 이용하여 자동으로 파생주제어를 발생시킴으로써 좀더 포괄적인(또는 세부적인) 주제어 선정을 돕기 위한 용어파일이다. 또한 복합개념의 용어를 개념어 단위로 분절 색인하므로써 전조합색인의 문제점으로 지적되기 쉬운 검색 누수현상을 막을 수도 있다.

가장 중요한 것은 앞장에서 살펴본 것 처

럼 시소러스 기반 색인시스템에서도 해결할 수 없는 문제점으로 지적된, 대상문헌에 발생하지 않은 용어도 개념관계상 필요하다면 특정어파일에 정의된 특정어관계를 이용하여 색인어로 선정할 수 있다는 것이다.

이처럼 특정어파일은 용어의 의미관계를 단서로 하여 색인시스템의 성능을 향상시키는 색인지원용 용어파일이라 볼 수 있다.

2. 구축방법

본 연구에서는 특정어파일을 구축하기 위해 중앙일보사에서 구축해온 50만건 가량의 신문기사DB에 등록된 색인어 400만어와 JOINS 시소러스에 수록된 용어 30만어를 분석하였다. 이때 색인어리스트와 시소러스파일에 수록된 용어들을 수작업으로 대조 분석하는 작업과 함께 주제어의 동시출현률을 분석하는 작업을 병행하였다.

그 결과 약 2천어 규모의 특정어파일을 구축할 수 있었다. 이는 실험적인 샘플파일 수준이며 본격적인 파일구축시에는 대규모의 특정어파일을 구축할 수 있을 것이다.

이처럼 특정어파일은 해당 정보기관의 데이터베이스에 등록된 색인어를 근거로 하여 경험적으로 구축할 수 있기 때문에 매우 유용한 색인도구가 될 것으로 보인다.

3. 특정어관계

특정어관계에서는 여러가지 어특징을 설정할 수 있으나 여기서는 크게 복합어 분할 범주와 개념확장 범주의 둘로 나누어 샘플파일을 만들었다.

3.1 복합어 분할

북한경제 → 북한 / 경제
부동산값 → 부동산 / 가격
한미경협 → 한국 / 미국 / 경제협력
김대통령 → 김영삼 / 대통령

3.2 개념확장(또는 특정화)

비열한전쟁 → 아르헨티나
천안문사태 → 중국 / 민주화운동
닌텐도증후군 → 전자오락
마광수 → 외설시비
현대건설 → 건설업계
국제사면위원회→ 인권단체

4. 기능

특정어파일은 III장에서 살펴본 JOINS 색인시스템에서처럼 색인 후보어로 선정된 용어들에 대해 특정어 관계로 설정된 용어들을 추가 발생시킴으로써 색인의 질을 높인다.

V. 결론 및 평가

본 연구에서 구축 실험한 특정어파일은 규모가 제한적이라는 한계가 있다. 그러나 어의적 특징을 이용함으로써 색인시스템의 성능을 향상시킬 수 있었으며 이는 시소러스 구조에서 표현하지 못한 연관개념을 보다 동적으로 표현할 수 있다는 것을 의미한다..

따라서 이같은 특정어파일은 GUI환경의 정보검색 시스템 구축시 이용자의 검색어 탐색과정을 보다 효과적으로 지원할 수 있을 것으로 보인다.

한편 앞서도 언급했듯 최근 각계의 정보관련 부서에서는 시소러스에 대해 많은 기대를 걸고 있으면서도 구축비용이 막대해 구축을 시도하지 못하는 경우가 많다. 이처럼 시소러스가 필요하면서도 구축할수 없는 기관에서 특정어파일을 동의어파일 구축과 병행하면 상당히 좋은 효과를 얻을 수 있을 것이다.