MeSH를 이용한 개념 기반 검색 엔진 시스템

* * ** *
고삼일, 박사준, 황수철, 김기태
중앙대학교 컴퓨터공학과*
인하공업전문대학**

Concept-based Search Engine System using MeSH

* * ** *
Sam-Il Ko, Sa-Joon Park, Su-Chul Hwang, Ki-Tae Kim
*Dept. of Computer Science & Engineering, Chung-Ang University
**Inha Technical Junior College

요 약

본 논문에서는 개념 기반 검색엔진 시스템(Concept-based Search Engine System)의 검색 정확도를 향상시키기 위한 방법으로 MeSH를 이용하였다. MeSH는 Medical Subject Headings의 약자로서 MEDLINE 논문의 원활한 검색을 위하여 주제어를 코드화한 것으로 이를 개념 그래프의 시소러스로 사용하여 개념 그래프의 가장 중요한 부분인 개념 추출의 정확성을 보장 하도록 하였다. 본 논문은 2003년 MeSH의 Descriptor Data의 Term 항목을 사용하여 개념과 관련이 있는 유의어를 추출 했다. 추출된 유의어로 개념 그래프를 구성한 것과, 문서 내에서의 단어 빈도수에 의하여 개념 그래프를 구성한 것의 검색 결과를 비교한 결과 MeSH를 시소러스로 사용하여 개념 그래프를 구성한 것이 훨씬 더 정확한 결과를 내는 것을 확인할 수 있었다.

1. 서론

인터넷은 정보의 바다라고 불릴 만큼 방대한 정보들을 포함하고 있다.[1] 하지만 정보를 찾는 사람의 관점에 따라 찾고자 하는 정보와는 관련이 없는 정보들로 가득 찬 쓰레기의 바다라도 불릴 정도로 의미 없는 데이터들이 쌓여있는 곳이기도 하다. 이러한 인터넷에서 원하는 정보를 정확하게 찾아 내는 것이 검색 엔진에서 가장 중요한 문제이며 이를 위하여 키워드 검색, 자연어 검색, 디렉토리 서비스, 개념 검색 등 다양한 방법들이 제시되었다. 현재 많은 검색 엔진이 사용하고 있는 키워드 검색은 사용자가 검색하고자 하는 내용(키워드)에 대해 정확하게 인지하고 있는 상태에서 검색 결과들에 대한 적합성을 판단하여 그 내용에 대한 보다 자세하고 많은 정보를 얻는 방식이다. 키워드 기반 검색에서 정확한 결과를 내기 위해서는 사용자가 찾고자 하는 내용에 대한 정확한 인지가 먼저 선행되어야 하는 것이다. 또한 키워드를 기반으로 검색을 할 경우 키워드가 속한 의미 영역에 따라 매우 상이한 의미를 가질 수 있기 때문에, 찾고자 하는 의미의 검색 결과뿐만 아니라 원하지 않는 의미 영역의 검색 결과까지 포함된 정확하지 않은 결과를 도출하게 되어 사용자가 다시 자신이 원하는 영역에 해당하는 것만 선별하여 보다 자세한 정보를 찾게 된다. 본 논문에서 제시하는 개념 기반 검색은 사용자가 내용 또는 개념은 알지만 정확한 키워드를 알지 못할 때 그 개념과 비슷한 의미의 키워드를 제시하면 그와 연관된 키워드들을 그래프로 제시하여 사용자가 원하는 정보를 보다 쉽고 정확하며 빠르게 찾을 수 있도록 한다.[2][9]

본 논문의 구성은 다음과 같다. 2장에서는 개념 기반 검색과 시소러스에 대해 살펴보고, 3장에서 이를 구현한 웹그래프 시스템을 설명한다. 4장에서는 구현된 시스템에 대한 검색 결과를 보이고, 5장에서는 결론을 맺고 추후 연구 과제를 기술한다.

2. 관련 연구

일반적인 정보 검색에서 시소러스는 용어 통제 및 탐색어의 확장이나 축소를 통해 검색 효율을 조절하는데 사용된다. 개념 기반 검색에서는 검색을 확장해 가는데 있어, 시소러스에서 연관 개념을 추출하는데 사용된다.[3] 때문에 잘 구축된 시소러스의 필요성이 요구된다.

시소러스는 수동 시소러스와 자동 시소러스, 두 가지로 크게 나누어진다. 수동 시소러스는 사람이 직접 수작업으로 구축을 한 것이다. 수동 시소러스는 다시 두 가지 형식으로 나뉘는데 하나는 Roget과 WordNet과 같은 일반 목적과 단어 기반 시소러스이다. 이것은 반대말과 동의어와 같은 의미 관계를 포함하지만 정보 검색 시스템에서는 잘 사용되지 않는다. 다른 하나는 정보 검색 지향과 구(phrase) 기반 시소러스이다. INSPEC, LCSH, MeSH와 같은 것이 여기에 해당하며 시소러스 제작자의 필요에 의해 특정 분야에 전문적이다. 이러한 시소러스는 구축하는데 많은 비용이 들고 구축 후에도 갱신이 어렵다는 문제점이 있다. 이에 반해, 자동 시소러스는 유전 알고리즘, 자기조직 네트워크등을 사용하여 문서의 내용을 기반으로 시소러스를 구축하게 된다.[4] 이것은 보통 사용되는 문서들의 집합(collection)에 의존적이다. 또한 용어의 상호 출현 빈도와 같은 것을 기반으로 구축이 되므로 사람이 의도하고자 하는 유의어나 다의어와 같은 것들을 정확하게 표현하지 못하는 문제가 있다.[5]

본 논문에서는 의학 분야의 개념을 추출하는데 있어 MeSH를 시소러스로 사용하였다. MeSH는 Medical Subject Headings의 약자로서 미국 NIH산하의 NLM(National Library of Medicine)에서 논문의 원활한 탐색을 위하여 MEDLINE에 올려 검색이 가능하도록 수년에 걸쳐 수작업으로 제작한 시소러스이다. MeSH의 구조를 이용하는데는 두 가지 주안점이 있다. 첫째는 한가지 개념에 서로 다른 용어를 사용하는 것을 정리하는 것이다. 둘째는 생명과학에 사용되는 개념을 계층구조로 만들어서 상위 개념과 하위 개념을 파악하여 적절한 단어들을 확인할 수 있도록 하는 것이다.[6] MeSH는 다양한 세부 단계로 검색을 할 수 있도록 계층적 구조를 갖는 기술어(descriptor)라 불리는 용어들의 집합으로 구성되어 있다. MeSH 기술어는 알파벳 순서와 계층적 구조 모두 구축이 되어 있다. 가장 일반적인 단계의 계층 구조에는 Anatomy, Mental Disorder와 같은 매우 광범위한 주제어가 있고 좁은 단계로 내려갈수록 Ankle, Conduct disorder와 같은 보다 자세한 주제어들을 찾을 수 있다.

3. 시스템 개요

본 논문에서 제안한 개념 기반 검색 시스템의 개요는 그림 1과 같다. 시스템은 사용자가 입력한 키워드를 질의로 받아서 키워드와 관련 있는 개념들을 추출해 낸다. 사용자는 추출된 개념 중 하나를 선택하여 그 개념을 담고 있는 웹페이지들의 URL 목록을 얻어 보다 상세한 정보를 얻는다.

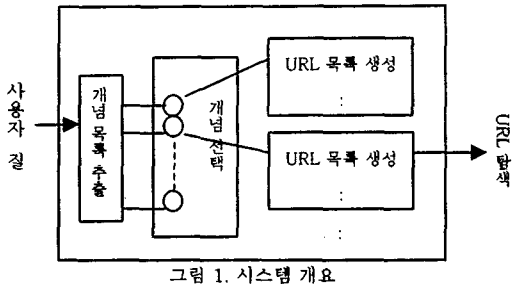

그림 1. 시스템 개요

시스템은 크게 개념 추출, 사용자 인터페이스, URL 목록 생성의 세 부분으로 나뉜다. 개념 목록은 MeSH의 Concept List 항목에서 추출하고, 추출된 개념을 사용자 인터페이스 부분에서 자바 애플릿을 통해 보여주며, 인터페이스 부분에서 선택된 개념에 대한 목록을 미리 구축된 URL인덱스를 통해 생성한다.

3.1 개념 추출

시스템은 개념 추출을 위해 MeSH2003을 시소러스로 사용했다. MeSH2003은 텍스트와 XML 두 가지 형식으로 제공되는데 본 논문에서는 XML 데이터의 Descriptor Record의 Concept List를 개념 추출에 사용했다. Descriptor Record를 담고 있는 MeSH XML 문서는 다음과 같은 구조로 되어 있다.

```
<DescriptorRecord>
  <DescriptorUI>D000005</DescriptorUI>
  <DescriptorName><string>Abdomen</string></DescriptorName>
    :
  <ConceptList>
    <Concept PreferredConceptYN="Y">
      <ConceptName><String>Abdomen</String></ConceptName>
    </Concept>
    :
  </ConceptList>
</DescriptorRecord>
```

위의 Descriptor Record 들을 개개의 파일로 분리하고, 사용자의 질의를 해시 함수를 통해 각 파일의 인덱스로 변환 후 직접 파일 내용을 읽어서 개념으로 사용한다.

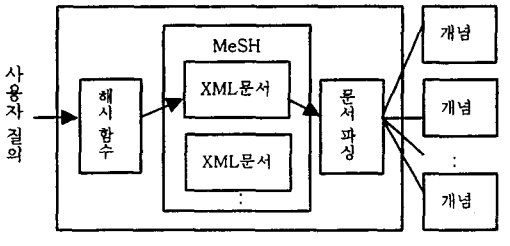

그림 2. 개념 추출

3.2 사용자 인터페이스

추출된 개념은 사용자가 알아보기 편하고 연관된 다른 개념으로 쉽게 확장 될 수 있어야 한다. 이를 위해서 자바 애플릿이 사용되었다. 각 개념을 더블 클릭하면 그와 연관된 개념이 다시 파생되고 오른쪽 버튼을 클릭하면 그 개념과 관련된 URL들이 목록으로 보여진다.[7] 이는 기존 검색엔진의 디렉토리 서비스와 유사하지만 개념간의 확장 및 이동이 훨씬 편리하다.

3.3 인덱스

선택된 개념에 대한 URL 목록을 출력하여 사용자가 선택한 개념에 대한 보다 자세한 정보를 얻을 수 있게 한다. 웹 스파이더가 웹 페이지에 대한 정보를 DB에 수집 후 인덱서를 거쳐 인덱스 파일로 저장하여 선택한 개념에 대한 인덱스 정보를 추출해 낸다.[8]

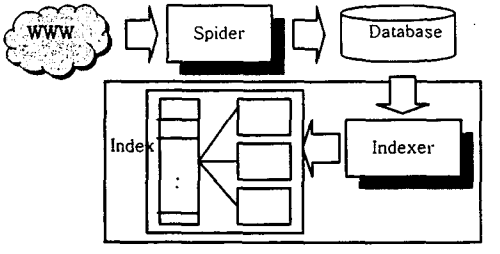

그림 3. 인덱스 생성

3.4 URL 목록

인덱스 엔진은 문서 내의 단어의 빈도수에 기반하여 다음과 같은 URL 목록을 생성한다.

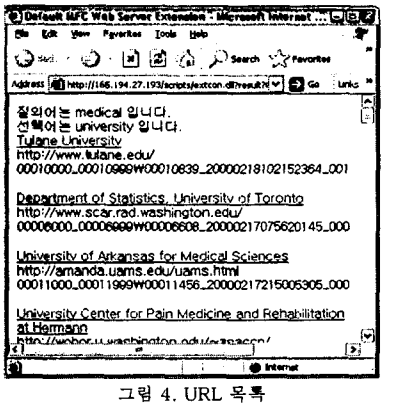

그림 4. URL 목록

4. 실험 및 평가

개념을 추출하는 데 있어 MeSH를 사용한 경우와 기존의 문서 내 단어 출현 빈도수에 기반하여 추출한 경우를 비교하였다. 실험을 하는 데 있어 의학 분야에 한해서만 검색을 하였고, 각 경우의 추출된 개념에 대해서 비교를 하였다. 이는 스파이더가 수집한 데이터가 의학 분야의 사이트를 시작점으로 하여 정보를 수집하였고, 시소러스로 사용한 MeSH가 의학 분야에 대한 것이기 때문이다.

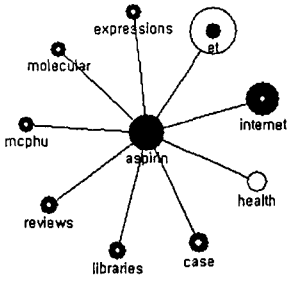

그림 5. Aspirin에 대한 단어의 빈도수에 기반한 키워드 추출

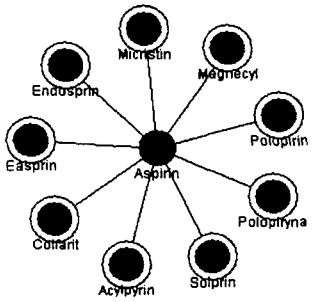

그림 6. Aspirin에 대한 MeSH 기반 키워드 추출

위의 결과를 보면 단어 빈도수에 기반한 경우 웹페이지에 자주 등장하는 internet이나 Aspirin과 관계 없어 보이는 libraries등과 같은 키워드들이 개념으로 추출이 됐다. 반면 MeSH에 기반한 경우는 Aspirin과 밀접한 관련이 있는 키워드들이 추출이 되었다. 이는 사용자가 결과 URL을 보기 위한 키워드를 제공하는 데 있어서 훨씬 더 정확한 결과를 생성할 수 있는 단서를 제공해 준다.

5. 결론 및 추후과제

시소러스에 기반한 키워드 추출이 문서 내 단어 빈도수에 기반한 키워드 추출보다 훨씬 정확한 결과를 낸다는 것을 알 수 있었다. 이는 범용성은 떨어지지만 특정 전문 분야에 있어서는 매우 정확한 결과를 도출해 낼 수 있는 가능성을 보여주는 것이라 할 수 있다. 또한 부족한 범용성은 전문 분야를 각각 하나의 개념으로 보고 전문 분야 자체를 개념으로 제시하고 거기서 다시 개념을 확장하여 해당 영역에 대한 개념을 추출해 냄으로써 충분히 극복할 수 있을 것이다. 단지 이러한 전문 분야의 시소러스를 구축하는데 많은 어려움이 따르고 모든 분야에 대한 전문 시소러스를 구축하기 어려운 현실을 감안하게 되면 자동 시소러스 구축에 대한 필요성이 대두된다.

그 외 다른 문제점이 있는데 우선 이렇게 추출한 키워드를 인덱스한 웹페이지의 URL과 어떻게 매칭시키는가에 대한 것이다. 개념이 아무리 잘 추출이 되었다 할지라도 실제 웹페이지와 연결하는데 있어서는 기존의 키워드 기반 검색 엔진이 갖는 한계를 그대로 갖고 있기 때문이다. 현재로서는 사용자가 개념은 알고 있지만 정확한 키워드를 모를 때 원하는 정보를 쉽게 찾을 수 있는 단서를 제공하는데 의의를 두고 있지만, 추출된 개념과 실제 웹 문서를 얼마나 잘 매칭시키는가가 앞으로 해결해야 할 과제이다.

참고문헌

1. Dayne Fritag, Information Extracting from HTML : Application of a General Machine Learning Approach, Information Extraction, AAAI, 1998.
2. A. Cammelli, F. Socci, A thesaurus for Improving Information Retrieval in an Integrated Legal Expert System, IEEE, 1998.
3. R. Baeza-Yates, G. Gonnet, Integrating contents and Structure in text retrieval, ACM SIGMOD Record, 25(1), 1996.
4. Hsinchun Chen, Chris Schuffels, Rich Orwig, Internet Categorization and Search : A Self-Organizing Approach, http://ai.bpa.arizona.edu/papers/som95/som95.html
5. 조민재, 웹의 개념 지식을 이용한 자동 시소러스 생성법의 설계 및 구현, 중앙대학교 92회 석사학위 논문, 1999.
6. 한국 의학 주제어 및 통제 검색어(KM-tree, Korean Medical Subject Heading) 개요. http://www.medric.or.kr/kmbase/helpMeSH.htm.
7. 양기철, 개념그래프 소개, 한국 정보과학회지 제 12권 9호, 1994.
8. 이권국, 신일수, 이상준, 김기태, 전문가 검색 엔진에서 데이터 마이닝을 이용한 개념 관계 추출, 한국 정보과학회 봄 학술발표 논문집 제 27권 1호, 2000.
9. Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Mordern Information Retrieval, Addison Wesley, 1999

참고 사이트

1. National Library of Medicine, Medical Subject Heading(MeSH), http://www.nlm.nih.gov/mesh/
2. Introduction to MeSH in XML format, http://www.nlm.nih.gov/mesh/xmlmesh.html
3. 의학연구정보센터(Medical Information Research Center, MedRIC), http://kmbase.medric.or.kr/