의미기반 정보검색

한국전자통신연구원 장명길 · 김현진 · 장문수 · 최재훈 · 오효정 · 이충희 · 허 정

1. 서 론

정보검색 기술은 90년대 후반부터 인터넷의 발전과 더불어 상업적 응용이 확대되면서 급속히 발전하고 있다. 최근에는 웹 문서의 양이 급격히 증가하면서 세계적으로는 7억 페이지 이상을 색인하는 대용량 문서 색인 기술과 함께 수만에서 수십만의 검색 결과 중에서 사용자가 원하는 의도에 맞는 정보를 정확하게 찾아주는 효과적인 검색 랭킹 기술이 요구되고 있다. 특히 웹과 같은 영역에서의 정보검색은 다양한 분야의 정보들이 서로 연결되어 있는 상황에서 빠르고 정확하게 찾아주는 점에 초점을 맞추어 기술 개발이 집중적으로 이루어지고 있다. 웹 검색 서비스에서 두각을 나타내고 있는 구글(www.google.com)의 경우, 이러한 웹의 특성을 십분 반영하여 Page Rank라는 랭킹 시스템을 도입하여 사용자 입장에서 높은 검색 효과를 얻는 서비스를 제공하고 있다[7]. 정보검색 기술은 이외에도 중복검색 결과의 제거, 메타 검색, 분산/통합 검색, 전문 검색 등 여러 이슈들에 대하여 활발한 연구가 이루어지고 있다. 또 다른 측면에서는 주제별/장르별 문서 자동분류 기술을 정보검색에 활용하여 사용자 질의 의도에 적합한 검색 결과를 제공하는 기법에 관하여 두드러진 연구 결과가 나오고 있다. 최근 새로운 검색 방식으로 선보인 애스크 지브스(www.askjeeves.com)의 경우, 자연어 문장 검색 방식이라는 새로운 형태로 사용자가 자주 묻는 질문에 대하여 7백만개 이상의 정보를 링크해 놓아, 사용자의 검색 요구에 적절히 활용하고 있다[3, 4]. 또한 아직은 일반 분야에 적용하기에 기술 수준이 미치지 못하고 있는 질문/응답형 정보검색 기술도 속속 선보이고 있으며, 제한된 분야에서는 나름대로의 실용화 가능성을 테스트하고 있다.

최근에 그 동안 연구실 프로토타입 수준에 머물렀던 자연어처리 기술이 실용적인 수준의 기술로 진전을 보이면서, 정보검색에 자연어처리 기술을 적용하여 검색 효과를 높이려는 연구가 국내외적으로 활발히 이루어지고 있다. 본 고에서는 정보검색의 효과에 큰 영향을 미치는 고유한 기법들인 가중치 부여 방법(weighting scheme), 검색 모델, 랭킹 방법 등의 기법 외에 자연어처리 기술을 이용한 의미기반 정보검색 기술에 초점을 맞추어 논의하고자 한다.

현재 대부분의 상용 정보검색 시스템들에서 사용되고 있는 자연어처리 기술은 형태소 분석이나 n-gram 방식으로 단순히 명사 단어의 키워드에 기반하여 색인하고 검색하기 때문에 문장의 내용적인 의미(semantic)를 제대로 색인에 반영하지 못하여 검색이 부정확하게 되는 결과를 낳고 있다. 본 고에서는 정보검색에서 자연어처리의 잠재성을 개척하고자하는 일련의 연구들[1, 2, 5, 9, 10, 12, 13, 14]과 맥을 같이하여, 어휘/구절/문장 의미 중의성을 처리하는 높은 수준의 자연어처리 기술을 적용하여 문서의 의미를 가장 잘 반영하는 의미색인단위의 추출에 의한 의미기반 정보검색 기술에 관하여 논하고자 한다. 또한, 사용자의 자연어 질의 의도에 가장 맞는 정답들(answer set)을 자동으로 구축하여 제공하는 지식베이스에 기초한 의미기반 정보검색 기술에 대해서도 논한다.

다음 2장에서는 현재까지 연구된 의미기반 정보검색 기술 개발의 동향에 대하여 살펴보고, 본 고에서 제안하는 의미기반 정보검색 기술을 간략히 설명한다. 이어서 3장과 4장에서는 우리가 개발한 의미기반 정보검색 기술인 자연어처리 기술을 적용한

정보검색 기술과 지식베이스의 정답문서에 기반한 정보검색 기술에 대하여 차례대로 기술한다. 5장에서는 이들 기술을 통합하여 구축한 의미기반 정보검색 시스템을 소개하고, 향후 의미기반 정보검색 기술 개발에서 보완해야할 연구 과제로 결론을 맺는다.

2. 의미기반 정보검색 기술

자연어처리 기술을 이용한 의미기반 정보검색에서는 기본적으로 자연어 텍스트로부터 적절히 색인어를 추출하여 이를 통계적인 검색 모델에 반영하여야 한다. 이를 위해서는 텍스트 문서의 분석에 자연어처리 기술을 적용하여 의미있는 색인어를 추출하는 것이 무엇보다도 중요하다. 먼저, 영어권의 경우, 자연어처리기술을 이용한 색인 기술에 관하여 지금까지 연구된 내용을 살펴본다. 90년대 초반의 자연어처리 기술을 적용한 연구 결과는 자연어처리 기술의 완성도가 떨어져 자연어처리 기술을 사용하지 않는 통계적인 정보검색 기술과 비교하여 좋지 않은 성능의 결과를 낳았다. 90년대 중반에 와서는 TREC-5부터 NLP(Natural Language Processing) SIT(Special Interest Track)가 만들어지면서 본격적으로 자연어처리기술을 이용한 정보검색의 효과 향상을 위한 다양한 연구들이 진행되었다. CLARIT[6, 13, 14]나 IRENA[1, 2] 시스템에서는 문서 텍스트의 명사구 구문분석을 통한 구 기반 색인(phrase-based indexing)이 시도되었으며, TSAs(Tree Structure Analytics)라는 구문구조 분석에 의한 색인 방법에서는 질의문과 텍스트 문서의 구문분석 결과를 트리 구문구조로 매칭하여 검색을 수행하였다[12]. 전반적으로 이 시기의 연구들은 구절 색인어(phrasal term)와 고유명사(proper name) 색인어 추출에 자연어처리 기술을 적용하는 다양한 색인 방법들에 관한 연구가 본격적으로 이루어졌다. 90년대 후반에 들어서는 FERRET 등과 같이 프레임, 스크립트 등의 지식 표현을 이용한 개념 색인(conceptual indexing)에 관한 연구도 실험적으로 시도되었다[5, 9].

국내에서의 자연어처리기술을 이용한 색인에 관한 연구는 먼저 90년대 후반에 키팩트(keyfact)라는 형태소 분석된 품사 열들로부터 문서의 내용을 [객체(object), 속성(property)]이라는 색인단위로 정규화하여 표현하고, 이를 검색에 사용하는 방법이 있었다[18, 21]. 키팩트 색인은 구문분석이라는 자연어처리기술을 적용하지 않으면서도 용언 수식 명사구나 복합 명사 색인에 효과적으로 적용되어 키워드 기반 색인 방법과 비교하여 색인 정확도에서 더 좋은 연구 결과를 보였다. 하지만, 테그 열로부터 생성될 수 있는 다양하면서 비교적 많은 키팩트를 일관성 있게 추출하는 규칙을 수작업으로 모두 기술해야 하는 문제를 해결해야만 하였다. 또한, 이 시기에는 한국어 정보검색에서 특히 중요한 복합 명사의 색인에 관하여 집중적인 연구가 진행되어 검색 성능 향상에 적잖은 기여를 하였으며, 최근에는 복합명사의 분해와 합성을 이용한 색인에 관한 연구 결과도 발표되었다[17]. 이렇듯 지금까지의 국내의 자연어처리 기술을 이용한 색인 방법에 관한 연구는 근본적으로 문서의 의미적 내용을 명사 구 수준의 색인단위로 표현하는 한계를 가지고 있었다.

의미기반 정보검색 기술의 또 다른 연구 방향으로는 웹에서의 문장형 검색 방식과 질문/응답형 검색 방식에서 나타나는 자연어처리 기술의 적용과 이를 위한 지식베이스의 구축에 관한 것이다. 하지만, 이러한 새로운 검색 방식에서의 현재의 기술 수준은 애스크 지브스에서 보는 바와 같이 사용자가 찾고자 예상되는 정보를 정답 구축 네트워크를 통하여 미리 수작업으로 구축하여 제공해야만 하며, 자연어 질의 문장의 정확한 분석과 해당하는 정답의 추출을 위해서는 구문/의미 분석의 고급 자연어처리 기술을 적용하여야 한다. 아직까지 이러한 점에 있어서 기술적인 완성이 미흡하여 일반 분야의 웹에 적용되기에는 시기 상조인 것으로 판단된다.

본 고에서는 사용자 질의 요구에 맞는 의미적으로 정확한 정보를 제시할 수 있도록 하기 위하여 의미기반 정보검색의 기술적 특성을 두 가지 측면에서 고려하였다.

첫째로 가장 중요하게 고려하는 기술적 특성은 자연어처리기술을 이용한 의미기반 색인 방법에 관한 것으로, 형태소분석 및 태깅 기술뿐 아니라 구문/의미 분석의 자연어처리 기술을 이용하여 어휘/구절/문장 의미 중의성을 해소함으로써 텍스트와 질의의 의미적인 내용을 색인에 반영하여 검색 효과를 높이려는 것이다. 의미기반 색인에서 텍스트와 질의의 의미적인 내용을 색인에 반영하기 위해서는 같은 의미를 가지는 부분들을 동일한 의미색인단위

(semantic indexing unit)로 정규화하는 과정이 필요하다. 이것은 자연어처리 기술을 적용한 의미기반 정보검색에서 매우 중요한 과정으로, 의미색인단위의 설정과 추출은 자연어처리 기술의 적용 수준에 따라 좌우되며, 이것은 색인의 정확도에 큰 영향을 미친다. 현재 대부분의 상용 시스템은 고급 자연어처리기술을 이용하지 않기 때문에 의미기반 색인이 이루어지지 않고 있다. 예를 들어, "빨간 자동차"의 의미를 반영하는 의미색인단위는 "자동차가 빨갛다", "자동차가 빨갛지 않다", "자동차가 빨갛지 않고 노랗다" 등의 여러 가지 표현에 대해서도 동일한 색인어를 만들 수 있어야 한다.

둘째로 고려하는 기술적 특성은 지식베이스에 자동 구축된 정답문서 기반 정보검색 기술에 관한 것이다. 이를 위해서는 질의에 맞는 정답을 바로 제공할 수 있는 지식베이스 구조, 정답문서를 자동으로 분류하여 구축하는 기술, 그리고 정답문서를 검색하여 질의에 해당하는 정답을 찾아 제시하는 기술이 필요하다. 본 고에서 특히 중요하게 생각하는 점은 애스크 지브스와 달리, 정보검색 대상이 되는 일반분야의 정보를 모두 포괄하는 개념망이라는 대규모 개념계층구조에 근거하여 분류되는 정답문서집합들이 문서분류기술을 이용하여 자동으로 구축된다는 점이다.

요약하면, 본 고에서 제안하는 의미기반 정보검색 기술은 텍스트 문장의 의미를 정확하게 분석하는 자연어처리기술을 적용한 의미 색인 기술과 지식베이스로 자동 구축된 정답문서 기반 검색 기술이다. 다음 3장과 4장에는 지금까지 강조한 의미기반 정보검색의 두 가지 특성을 고려하여 의미기반 정보검색의 기술적인 내용에 대하여 설명한다.

3. 자연어처리 기술을 이용한 정보검색 기술

최근, 한국어 정보검색에서 복합명사를 포함한 명사구 색인에 관한 연구들을 보면, 기존의 통계적인 방법 외에도 자연어처리 기술인 형태소 및 구문분석 등과 같은 기술을 접목하는 시도가 이루어지고 있음을 알 수 있다. 그러나, 색인이나 검색의 단위가 구단위로 확장되기는 하였으나, 그 범위가 명사를 포함하고 있는 것으로 제한하여 완전한 의미의 구단위 혹은 문장으로까지의 확장된 형태의 색인이라고 보긴 힘들다. 또한, 한국어 정보검색에서 가장 난점으로 꼽고 있는 중의성 단어 처리에 관한 연구는 아직 이루어지지 않고 있다. 중의성 단어의 경우엔, 문장 내에서의 단어의 역할을 파악하여, 의미적인 중의성을 해소해야 하는 어려움이 있기 때문에, 자연어처리 기술이 접목되지 않고서는 이를 해결할 수 없다. 따라서, 본 고에서는 한국어 자연어처리 기술 중에서 구문분석 기술을 이용한 정보검색 색인 기술과 단어 의미 중의성 해소를 통한 정보검색 기술에 대해 제안하고자 한다.

3.1 구문분석을 이용한 색인

본 절에서는 높은 정확도의 검색을 위한 고품질 색인을 위하여, 부분 구문분석과 완전 구문분석 기술을 이용한 명사구와 중심어-종속어(HM:Head-Modifier) 색인어를 추출하는 방법에 대하여 설명한다[24].

3.1.1 부분 구문분석 기술을 이용한 명사구 추출

정보검색에서 색인어를 추출하는 것은 검색 성능과 직접적인 관련이 있다. 일반적으로 문장을 대표하는 색인어로 명사들을 추출하지만 불필요한 단어들이 추출될 가능성이 높으므로 고품질의 색인을 위해서는 명사구를 색인어로 추출할 필요가 있다.

명사구를 인식하기 위해서는 구문분석을 수행하거나 별도의 구묶음(chunking) 모델을 만들어야 하는데 구문분석을 사용할 경우도 명사구만을 추출할 때는 전체 문장을 완전 분석할 필요 없이 부분 분석만으로 명사구 인식이 가능하다. 별도로 명사구만을 인식하는 방법으로는 규칙과 어휘 정보를 이용한 방법[16], 최대 엔트로피 모델을 이용한 방법[15] 등이 있다. 여기에서는 부분 구문분석을 수행한 후에 명사구로부터 색인어를 추출하는 것에 대하여 설명한다. 문장에서 명사구는 복합명사, 수식구, 병렬구, 수식절 등의 여러 가지 형태로부터 나타날 수 있다. 명사구 내의 명사들을 추출하여 색인어를 생성할 때에도 발생 순서에 의미가 있는 경우(수식구 등)와 의미가 없는 경우(병렬어구)로 나누어 생성할 수 있다.

구문분석 결과를 트리 구조로 생성하는 경우, 각 트리의 표층 구조로부터 아래와 같이 서브 트리 타입을 결정할 수 있고, 이로부터 명사구 또는 중심어

– 종속어의 색인어를 추출하게 된다.

STT_CNN : 복합명사(종이 호랑이)
STT_DET : 관형사(새 책)
STT_PAR : 명사병렬어구(철수, 영희, 민수와 만수)
STT_NDN : 속격 조사(철수의 옷)
STT_DCL : 관형어구(아름다운 영희)

구축된 부분 구문분석 결과로부터의 색인어 추출은 현재노드의 서브 트리 타입이 STT_CNN, STT_PAR, STT_NDN, STT_DCL인 경우에 이루어지며, 그 노드의 하위 노드들로부터 명사들을 추출하여 하나의 색인어로 연결시키면 된다.

3.1.2 완전 구문분석 기술을 이용한 중심어-종속어 추출

명사구 색인어를 추출하는 경우는 부분 구문분석 기술만 사용해도 되지만 중심어와 종속어를 추출하기 위해서는 전체 문장에 대한 완전 구문분석 과정이 필요하다.

완전 구문분석을 이용하는 방법 중에는 구문분석 결과의 중의성을 해소하는 경우와 해소하지 않는 경우로 나눌 수 있다. 중의성 해소를 하지 않는 경우에는 불필요한 색인어들이 너무 많이 발생하여 검색 성능을 떨어뜨릴 수 있으며, 구문분석기의 중의성 해소 정확률이 너무 낮을 경우 도리어 부정확한 색인어가 추출되어 검색 성능을 저하시킨다. 따라서, 정보검색에서 검색 효과를 얻기 위해서는 만족할 만한 수준의 구문 중의성 해소 기술이 적용되어야만 한다.

중심어-종속어 관계를 고려하면 명사구를 고려할 때보다 문장의 정확한 의미를 나타내는 색인어를 추출할 수 있으므로 높은 성능의 검색 결과를 얻을 수 있다. 반면에 구문분석기의 성능에 절대적으로 의존한다는 단점이 있다.

중심어-종속어 관계를 추출하는 과정은 구문 분석-서브 트리 타입 결정-색인어 추출의 순서로 이루어지고 서브 트리 타입을 결정하는 부분은 3.1.1절에서 설명한 것과 동일한 규칙을 사용한다. 이때 추출되는 중심어-종속어 형태는 다음과 같다.

1) 술어+명사(복합명사 포함)(V-N)
- 예 : 사람이 과자를 먹었다.
→ 먹다–사람, 먹다––자

2) 복합명사 또는 명사병렬어구(N-NIL)
- 예 : 사과, 딸기, 배를 먹다.
→ 사과-딸기+배-NIL
먹-사과, 먹-딸기, 먹-배

3) 속격 조사(N-N)
- 예: 사람의 관절을 연구한다. → 관절-사람

4) 관형절(V-N)
- 예 : 아름다운 사람을 보았다.
→ 사람-아름답다 → 아름답다-사람

4) 관형절의 경우, 관형절 정규화 과정을 거친다. 즉, "아름다운 사람"과 "사람이 아름답다."라는 두 개의 문장으로부터 추출되는 중심어-종속어 관계는 동일해야 한다. 따라서, 관형절에서 용언이 피수식 명사 단어의 서술어로 판단될 경우, 색인어 추출에서 수식어가 중심어로 추출된다.

3.2 단어 의미 중의성 해소를 통한 정보검색

문장 내에 사용된 중의성 단어의 의미를 결정하는 것은 질의에 부합하는 적당한 문서를 찾는 것과 근본적으로 밀접한 관계가 있다. 최근 단어 의미 중의성 해소(word sense disambiguation)와 관련된 많은 연구의 결과로 중의성 단어의 의미를 분별하는 기술이 크게 향상되었다. 이와 더불어 의미 중의성 해소와 정보검색의 상관관계에 대한 연구가 진행되고 있는데, 의미 중의성 해소가 정보검색 성능향상에 긍정적인 영향을 미치는 것으로 발표되고 있다[8, 11].

3.2.1 단어 의미 중의성 해소를 위한 의미정보 DB 구축

의미 중의성 해소를 위해서 근원적인 지식이 필요하다. 그 지식원(knowledge source)과 지식의 활용방법에 따라 의미 중의성 해소는 크게 세 가지로 구분된다. 교사학습(supervised learning)과 비교사학습(unsupervised learning)에 의한 중의성 해소 방법이 있고, 사전을 기반으로 한 의미 중의성 해소 방법이 있다.

본 연구에서는 사전의 뜻풀이말을 이용한 교사학습 방식으로 의미 중의성을 해소한다. 먼저 중의성 단어를 포함하는 뜻풀이말과 표제어를 추출한다. 중의성 단어와 공기관계(co-occurrence

relation)를 이루는 단어들은 중의성 단어와 의미적으로 밀접한 관계를 가진다는 가정을 기반으로 뜻풀이말 내의 중의성 단어와 공기관계를 이루는 단어들의 빈도정보를 의미정보로 이용한다.

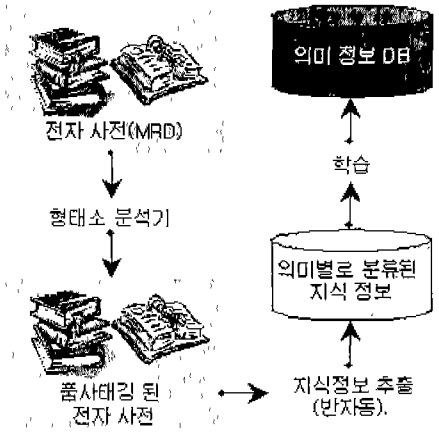

그림 1 의미정보 DB 구축 과정

의미 중의성 해소에서 원천적으로 중요한 것이 중의성 해소에 사용되는 의미정보이다. 그러므로, 의미정보의 구축은 상당히 중요한 작업이며, 의미정보의 정련성은 중의성 해소에 큰 영향을 미친다. 그림 1은 의미정보 DB를 구축하는 과정인데, 형태소분석 시 발생하는 오류를 수동으로 수정하여 품사태깅 된 전자사전을 구축한다. 지식정보 추출은 중의성을 지닌 9개의 명사[1]를 포함하는 뜻풀이말과 표제어를 대상으로 추출한다. 학습에서는 중의성 해소에 도움이 되는 명사, 형용사, 동사만을 대상으로 한다. 의미정보 DB는 3개의 품사별로 중의성 단어와 공기관계를 이루는 단어들의 빈도정보와 품사정보를 가지고 있다.

3.2.2 정보검색을 위한 단어 의미 중의성 해소 모듈

$Sim(C,S_i)$는 문장 C와 의미 S_i의 유사도를 나타내고 α와 β는 체언류와 용언류의 가중치로 $\alpha+\beta=1$이며, 실험에 의해서 결정된다. $Match(C_n,S_i)$는 문장 C에서 의미 S_i와 공기관계를 가지는 체언류(C_n)의 개수이다[23].

$$WSD(C,\ S_i)=\arg MAX_{Si}Sim(C,\ S_i) \qquad \text{수식(1)}$$

$$Sim(C,\ S_i)=\alpha\times\text{Noun}(C,\ S_i)+\beta\times\text{Pred}(C,\ S_i) \qquad \text{수식(2)}$$

$$\text{Noun}(C,\ S_i)=\text{Match}(C_n,\ S_i)\times\sum_j P(W_{vj}\mid S_i) \qquad \text{수식(3)}$$

$$\text{Pred}(C,\ S_i)=\text{Match}(C_v,\ S_i)\times\sum_j P(W_{vj}\mid S_i) \qquad \text{수식(4)}$$

$$P(W_{vj}\mid S_i)=\frac{S_i\text{내에서 }W_n\text{의 출현빈도}}{S_i\text{내에서 체언류의 총출현빈도}} \qquad \text{수식(5)}$$

정보검색 시 발생하는 단어 의미 중의성 문제를 해결하기 위해서는 크게 두 부분에서 의미 중의성 해소 모듈이 포함되어야 한다. 첫째, 색인과정에서 색인되는 단어들의 중의성을 해소하기 위한 모듈이 필요하다. 둘째, 사용자 질의 처리 시에 단어들의 중의성을 해소하기 위한 모듈이 필요하다.

정보검색에 사용되는 질의는 대체적으로 명사구나 짧은 단문임으로 이용되는 단어의 수가 적다. 이는 공기정보를 이용한 통계적인 모델에서는 질의 연산 시에 중의성 해소를 실패할 가능성이 높다는 것을 의미한다. 따라서, 질의에 포함된 단어의 중의성 해소가 어려울 경우에는 사용자 피드백에 의해 원하는 정보를 검색할 수 있도록 해야 한다. 또한, 색인 시에 의미 중의성 해소를 실패하면, 실패한 문서들만 따로 색인 할 수 있도록 새로운 키워드를 생성하여 색인한다.

표 1 질의에서 의미 중의성 해소된 경우 상위 Ranking 10위 내의 문서 중 정답문서 수

중의성 단어	의미	정답 수	중의성 단어	의미	정답 수
기관	신체부위	6	병	그릇	10
	장치	9		사람	6
	조직	10		상태	10
기구	장치	7	비	도구	0
	조직	10		기상현상	10
눈	신체부위	10		비석	1
	식물	1	신	신발	1
	기상현상	10		종교	5
다리	교각	9	차	운송수단	9
	발	9		음료	0
배	과일	10			
	운송수단	7			
	몸	5			

1) 기관, 기구, 눈, 다리, 병, 배, 비, 신, 차

표 1의 결과에서 알 수 있듯이 정보검색에서 의미 중의성 문제가 해결되면 정보검색의 정확률을 크게 향상시킬 수 있다. 하지만 일반적으로 웹 검색 질의와 같이, 질의문이 대체로 짧은 단문이거나 구인 경우에, 단어 의미 중의성 해소 문제는 정보검색에서 더욱 어려워 질 수 있다.

4. 정답문서 기반 정보검색 기술

4.1 지식베이스 구축

의미기반 정보검색에서 구축하는 지식베이스는 크게 개념망과 정답문서집합으로 나뉜다[19]. 개념망은 한국어 명사에 대해서 사전적인 의미의 상하관계를 정의한 한국어 명사 개념망을 말하며, 정답문서집합은 개념망의 개념어(concept word)가 웹상에서 어떤 주제에 대해서 사용되고 있는가에 따라 웹문서를 분류해놓은 문서 집합을 말한다. 이 정답문서집합은 사용자의 질의가 나타내는 의도에 적합한 정답을 포함하는 문서라는 뜻에서 명명된 것이다.

각 개념어에 할당되는 정답문서는 개념어의 쓰임에 따라서 분류되는데 이 분류 항목을 개념어의 의미적인 특징을 나타내는 속성(attribute)이라고 한다. 즉, 개념망과 정답문서집합은 속성에 의해서 연결된다.

다음은 개념망과 정답문서집합의 구축에 관해서 설명한다.

4.1.1 한국어 명사 개념망

한국어 명사 개념망(이후 ETRI 개념망)은 단어들의 관계를 설정한다는 점에서 기존의 시소러스와 유사하지만, 시소러스가 단어간의 관계에 대한 기준이 명확하지 않은 반면, ETRI 개념망은 사전의 뜻풀이를 중심으로 개념어들 간의 국어학적 의미관계를 연결하므로 명백한 차이가 있다[2]. 그리고, 단어의 의미관계를 표현하는 리소스로 잘 알려진 워드넷(WordNet)은 의미가 유사한 단어들의 집합(SynSet)간의 연결로써, 단어 하나하나의 개념관계를 표현하는 ETRI 개념망과는 다르다. 워드넷은 유사한 단어들이 집합을 이루고 있으므로 대용어 선택이나 다국어 번역에서의 의미 공유 등에서는 효과적이지만, 개별 단어 의미의 영역이 명확하지 않다. 한편, ETRI 개념망은 단어들의 의미 포함관계를 명확하게 나타내므로 의미 중의성 해소나 명확한 문장분석 등에서 유리하다. 우리는 상하관계에 의한 개념망을 보완하기 위하여, 동의, 유의, PART-OF, 반의 등의 관계를 추가로 정의하고 있다.

ETRI 개념망은 사전 상에 나타나는 모든 단어에 대해서, 그리고, 모든 분야에 걸쳐서 구축하는 것을 원칙으로 한다. 그러나, 1차 구축에 있어서 두 가지 제약을 가지고 구축되었다. 첫째, 1차 구축에서는 시간과 비용을 줄이기 위해 경제분야에 한해서 구축되었다. 둘째, 정보검색을 목적으로 하는 개념망 구축이므로, 국어사전에 등재되지 않더라도 경제분야에 있어서 널리 사용되는 용어에 대해서는 경제용어사전에 준해서 등록하였다. 이렇게 구축한 개념망은 약 2만개의 사전 엔트리를 가지며, 1만 5천여 개의 상하관계와 1천여 개의 동의와 유의 관계를 포함하고 있다. 내년을 목표로 진행중인 2차 구축에서는 10만 단어 수준을 예상하고 있으며, 특정분야 전문용어를 배제한 범용 목적 개념망으로 구성될 전망이다. ETRI 개념망이 구축되면, 정보검색뿐만 아니라 기계번역, 의미분석 등 한국어 처리 관련 분야에서 넓게 사용될 수 있을 것으로 예상한다.

4.1.2 정답문서집합

정답문서집합을 분류하는 속성은 각 개념어의 실생활에서의 활용도에 따라 달라진다. 따라서, 모든 개념어는 각기 다른 속성들을 가질 수 있다. 하지만, 그 쓰임에 있어서 상위어에 의해 제약을 받고, 유사하거나 동의 관계에 있는 개념어는 비슷한 쓰임을 갖는다. 그리고, 정보검색의 관점에서 웹 상에는 개념어의 주요 활용 분야에 대해서만 문서가 존재하기 때문에 한 개념어가 갖는 속성의 개수는 유한한 값이 된다.

이들 속성은 정답문서를 분류하는 기준이 되므로 각 속성들은 상호 배타적인 의미를 가져야 하며, 정답문서 기반 검색에서 검색의 키워드로서 사용되므로 사용자의 질의에 개념어와 함께 나타나기 쉬운 형태를 가져야 한다(4.2절 참조). 그림 2는 우리

2) 개념망의 상하관계는 국어학적 의미의 kind-of 관계를 주로 하며, 일부 부분-전체 및 is-a 관계도 사용한다. 그러나, 시소러스에서의 상하관계는 사전적인 의미가 아닌 필요에 따른 용례에 의한 관계가 혼재함으로써, 둔 이상의 층위를 넘어가면 의미관계가 성립하지 않는 경우가 많다.

가 정의하고 있는 지식베이스의 구조를 나타내고 있다. "불공정거래"라는 개념어는 "거래"라는 상위어를 가지며, "거래"는 "경제활동"에 속한다. 그리고, "불공정거래"는 "정의", "종류", "현황", "대책" 등의 속성을 가진다. 각 속성은 복수개의 정답문서를 가지게 되고, "불공정거래에 대한 대책은 무엇인가?"와 같은 질문에 대해 정답을 제공하게 된다.

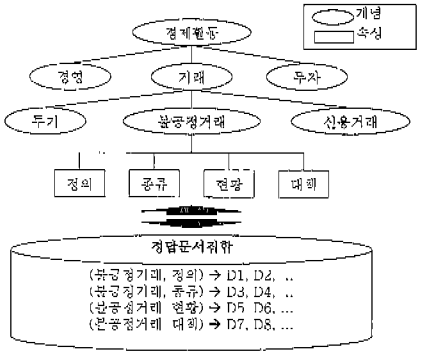

그림 2 정답문서집합을 위한 지식베이스 구조

우리는 그림 2의 하단부와 같은 정답문서집합을 구축하기 위하여, 각 개념어에 대해서 정답문서의 후보를 수집하고, 이들을 분석하여, 개념어의 속성을 정의한다. 그리고, 각 속성들에 포함된 정답문서를 결정한다. 이러한 과정은 다음과 같은 일련의 작업을 통해서 이루어진다.

① 문서 필터링 : 메타검색기를 통해 특정 개념어와 관련 있는 문서를 수집하여 관련도가 높은 문서를 추출한다.
② 속성 정의 : 수집된 문서를 분석하여 주제별로 속성을 정의한다.
③ 정답문서 수집 : 수집된 문서들에 대해서 각 속성별로 문서를 분류한다.
④ 속성 특징 추출 : 각 속성에 할당된 문서를 분석하여, 해당 속성의 특징을 나타내는 속성의 단서(clue), 즉 특징 단어나 구, 기타 요소를 추출한다. 이 특징 요소들은 구축작업의 자동화에서 분류 규칙으로 사용될 수 있다.

위와 같은 방법으로 구축되는 속성은 현재까지 구축된 개념망상의 개념어에 대하여 평균 10여 개를 가진다.

4.1.3 자동분류에 의한 정답문서집합 구축

지식베이스를 구축하는 과정은 개념어에 해당하는 문서 집합을 수집하고, 문서 집합의 특성을 나타내는 속성을 추출하는 작업이 수반된다. 그러나 이를 모두 수작업으로 구축한다는 것은 매우 어려운 일일 뿐 아니라 시간이 지남에 따라 새로운 정보로 변경되기 때문에 해당 정답문서집합을 재 구축해야 하는 경우가 발생한다. 따라서, 방대한 양의 정답문서집합을 자동으로 구축하기 위한 방법이 필요하다.

정답문서집합을 구축하기 위해선 먼저 개념어에 해당하는 문서를 수집하는 과정이 필요하다. 개념어에 해당하는 문서를 수집하기 위해 메타검색기를 활용하였다. 그러나, 메타검색기를 통해 수집된 문서는 개념어와 내용상으로 명확히 일치하지 않는 문서도 존재하므로 메타검색 결과에서 개념어와 관련있는 문서만을 추출하는 과정이 필요하다. 이것은 문서 필터링(document filtering)과 관련된 내용으로 문서 분류(document catcgorization) 기술을 활용하여 자동으로 해결할 수 있다.

수집된 문서 집합은 다시 그 성격과 내용에 따라 속성으로 나뉘게 된다. 이때 속성은 개념어의 쓰임에 따라, 즉 사용자가 해당 개념어에 대해 알고 싶어하는 분야에 따라 정의된다. 앞에서 설명한대로 개념어는 그 쓰임에 있어서 개념망 상의 상위어(parent)에 의해 제약을 받고, 유사하거나 동의 관계에 있는 개념어는 비슷한 쓰임을 갖는다. 다시 말해 같은 상위어에 해당하는 개념어들(sibling)이 가질 수 있는 속성은 유사하며 그 종류 또한 유한하다. 그러므로 각 개념어에 의해 필터링된 문서 집합을 정의된 속성 분류 체계(category)로 분류하는 방법을 통해 자동화를 할 수 있다. 또한 기 정의된 속성 분류 체계에 해당하지 않는 문서 집합에 대해 클러스터링(clustering)을 통해 새로운 속성을 제시하는데 도움을 준다.

일반적인 자동문서분류는 미리 구축된 정교한 문서 집합을 통해 분류기를 구축하고 학습된 분류 체계에 대량의 문서 집합을 할당한다. 그러므로, 학습 시 사용한 문서 집합의 성격에 영향을 많이 받게 된다. 또한 한 번 정의된 분류 체계에 대해 수정이 가해지지 않는다. 그러나 본 고에서 구축하고자 하는 정답문서 집합의 경우에는 학습을 위한 학습 문서 집합의 양이 절대적으로 작을 뿐만 아니라 정의된 분류 체계 역시 수시로 변하는 특성을 갖는다. 또한 정의된 분류 체계가 내용상으로 차이가 작은

것도 존재한다. 이러한 문제를 보완하여 정답문서 집합의 자동구축의 정확도를 높이기 위하여, 본 고에서는 두 가지 단계를 추가한다. 즉, 속성 특성을 반영하는 단서 패턴의 정의에 의한 속성 특징의 활용과 개념망에 정의된 KIND_OF 관계 활용이다. 두 가지관계에 관한 자세한 내용은 [20]을 참고하기 바란다.

이와 같이 정답문서집합의 가장 큰 특징은 정답문서의 검색 관점에서 개념어와 내용적으로 서로 다르게 구별되는 문서 집합들이 속성이라는 기준으로 분류되어 있다는 점이다. 하지만, 시간이 지남에 따라 개념어로부터 분류되는 문서 집합들이 달라질 수 있기 때문에, 수시로 새로운 속성이 생겨나고 없어지는 변화에 대응하기 위한 대책이 필요하다. 본 고에서는 자동 분류 결과 할당될 적절한 범주가 없는 문서는 미할당 문서로 취급하여 일정 개수를 모은 후 클러스터링을 실시하고, 그 클러스터링 결과를 구축자에게 제시함으로써 새로운 속성 추출을 용이하도록 하였다.

본 고에서 제안하는 정답문서집합 구축을 위한 자동분류 알고리즘은 중복할당과 미할당 기법을 통해 개념망의 정답문서 특성을 반영하고, 기계학습 결과의 오류보정을 위해 규칙 기반 분류 기법을 복합적으로 활용하는 방안(hybrid-categorization)을 제시한다. 또한 개념망에 정의된 KIND_OF 관계를 활용함으로써 정답문서 자동화 범위를 넓히고 수작업에 드는 노력을 최소화 한다. 다음은 이렇게 구축된 개념망과 정답문서집합을 통해 사용자의 질의에 해당하는 정답문서를 검색하는 방법에 대해 논의하기로 한다.

4.2 개념 속성을 이용한 정답 검색

일반적으로 사용자들이 인터넷 검색을 통해 요구하는 정보는 하나의 개념들이 가지는 포괄적이면서 방대한 내용이 아니라 그 개념들이 가지는 특정한 속성에 따라 구분되는 구체적이면서 부분적인 내용이다. 본 연구의 개념 속성을 이용한 정답 검색은 특정한 개념에 대한 포괄적인 정보를 여러 가지 속성들로 분류한 정답문서집합으로부터 사용자가 요구하는 핵심적인 정보만을 선택적으로 검색할 수 있도록 지원한다. 특히, 사용자들이 검색된 문서의 전체 내용보다는 그 문서의 특정 부분만을 요구하는 경우가 빈번하기 때문에 하나의 문서에서 사용자의 요구에 정확히 일치하는 부분만을 추출하여 제시하는 것이 이 검색의 핵심 기술이다. 여기서, 정답문서집합은 하나의 개념과 관련된 모든 문서들 중에서 특정한 속성에 따라 분류되는 문서들로 정의하며, 하나의 문서에서 사용자 질의와 가장 일치하는 부분을 정답 위치로 각각 정의한다.

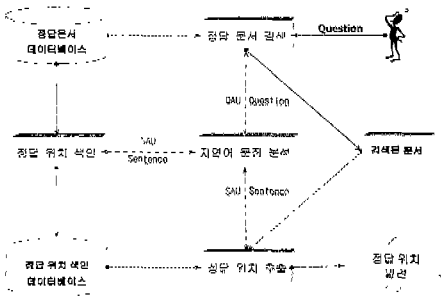

그림 3 정답 검색 시스템 구성 및 처리 과정

그림 3은 정답문서들을 검색하고 정답 위치를 추출할 수 있는 정답 검색 시스템의 구성 및 처리 과정에 대해 설명하고 있다. 이 시스템은 크게 "정답문서 검색 엔진", "정답 위치 색인 엔진" 그리고 "정답 위치 추출 엔진"으로 구성되며, 이들은 하나의 문장에서 개념 용어와 속성 용어들을 추출할 수 있도록 지원하는 "자연어 문장 분석" 모듈을 기본적으로 이용한다. "정답문서 데이터베이스"는 앞 절에서 설명된 분류 방법을 통해 구축된 모든 정답문서들을 "[개념 용어, 속성 용어] → [DocID]" 형태로 색인한 데이터베이스이며, 정답 위치 색인 데이터베이스는 "[개념 용어, 속성 용어] → [DocID] [Σ{SentNum,Weight}$_i$]"와 같이 하나의 문서에 포함된 문장들이 문서색인단위(DIU: Document Indexing Unit) [개념 용어, 속성 용어]와 어느 정도 관련이 있는지를 색인한 데이터베이스이다.

정답문서 검색 엔진은 사용자 질의를 [개념 용어, 속성 용어, 보조 용어 리스트]들의 불리언 형태로 분석한 결과(QAU: Query Analysis Unit)와 개념적으로 관련이 있는 문서색인단위로 색인된 문서들을 정답문서 데이터베이스로부터 검색한다. 예를 들어, 하나의 질의 Q = "투자신탁회사는 무엇**이며** 높은 수익의 투자신탁에는 어떤 상품이 있는가?"는 "자연어 문장 분석" 모듈에 의해 QAU= AU1 AND AU2로 분석된다. 여기서, AU1=[투자

신탁회사, 정의,{null}]이고 AU2=[투자신탁, 상품, {높다, 수익}]이다. 즉, Q에서 연결어미 "~이며"는 "AND" 연산자, '투자신탁회사'와 '투자신탁'은 개념 용어로 추출된다. 반면, "무엇이며"와 "어떤 상품이 있는가"라는 어휘패턴은 속성영역으로서 해당 개념에 대한 "정의"와 "상품"의 속성 용어로 분석되며, "높다"와 "수익"은 개념 용어를 보다 구체적으로 설명하는 주변정보이기 때문에 보조 용어로 각각 식별된다. 따라서, 정답 문서 검색 엔진은 정답문서 데이터베이스로부터 [투자신탁회사, 정의]와 [투자신탁, 상품]으로 동시에 색인된 문서들을 검색한다. 이때, 각각의 개념 용어들은 개념망을 통해 확장되며, 이 확장된 개념 용어에 대한 검색 결과 역시 사용자에게 제시된다.

정답 위치 색인 엔진은 문서에 포함된 모든 문장들에 대한 분석 결과(SAU: Sentence Analysis Unit)와 문서색인단위 DIU의 관련 정도를 평가하여 각각의 문장을 색인한다. 예를 들어, [투자신탁, 상품]으로 색인된 100번 문서의 40번 문장이 S= "그렇다면 투자신탁회사는 무엇**이며** 어떤 투자신탁 상품을 이용해야 더 많은 수익을 남길 수 있을까"라면, SAU=[투자신탁회사, 정의,{null}] AND [투자신탁, 상품, {이용, 수익}]으로 분석될 수 있다. 따라서, 이 문장은 문서색인단위 [투자신탁, 상품]와 정확히 일치하는 개념 용어와 속성 용어를 가지고 있기 때문에 높은 관련 정도로 정답위치 색인 데이터베이스에 색인된다. 여기서, QAU와 SAU에 포함된 개념 용어들 사이의 의미적 관련 정도는 ETRI 개념망에서 두 개념 용어의 거리를 통해 평가된다.

정답 위치 추출 엔진은 검색된 문서들 중에서 질의와 가장 유사한 문장을 가지고 있는 문서의 일부 내용을 제시함으로써 사용자가 검색된 문서를 전부 읽어보지 않더라도 문서에서 자신이 요구한 부분을 추출할 수 있게 한다. 즉, 사용자 질의와 문장들을 각각 QAU와 SAU 리스트로 분석한 다음, QAU와 SAU를 매칭함으로써 질의와 가장 일치하는 문장을 검색할 수 있다. 이 매칭은 SAU와 QAU의 주변 용어들까지 매칭에 참여한다는 점에서 정답 위치 색인과 다르다고 할 수 있다. 이때, 큰 문서의 모든 문장을 분석하여 질의와 매칭할 경우, 매우 많은 시간을 요구하게 된다. 따라서, 정답 위치 추출 엔진은 정답 위치 색인 정보를 이용하여 높은 관련 정도로 색인된 상위 몇 개의 문장들만을 분석하여 질의와 매칭한다는 특징을 가지고 있다.

5. 결 론

지금까지 3장과 4장에서 설명한 두 가지 방식에 의한 의미기반 정보검색 기술을 통합하여 우리는 의미기반 정보검색 시스템 *AnyQ*를 개발하였다. 본 시스템은 그림 4와 같이 정답문서 기반 정보검색 기술과 의미색인단위에 의한 정보검색 기술이 통합되어 구성되는데, 사용자 질의 문장에 대하여 정답문서 기반 검색을 1차적으로 수행하고 검색에 실패하였을 경우에 의미기반 색인 방식에 의한 검색을 적용하도록 하였다.

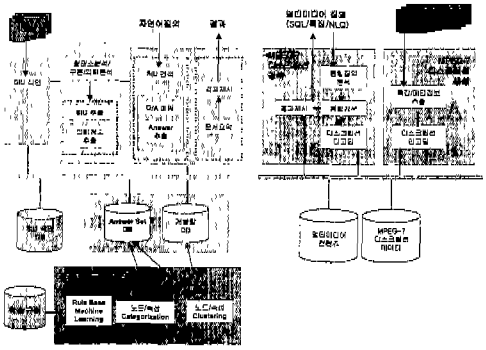

그림 4 의미기반 정보검색 시스템 AnyQ의 구성

그림 4의 의미기반 정보검색 시스템은 현재 텍스트 문서 정보검색뿐 아니라, 차세대 멀티미디어 정보검색을 위하여 MPEG-7 국제표준 문서도 다룰 수 있도록 개발하였다. 즉, 인터넷이 제공하는 방대하고 다양한 텍스트, 멀티미디어 정보들에 대하여 사용자가 의미적으로 요구하는 정보를 정확히 제공할 수 있도록 설계되었다. 멀티미디어 정보검색에 적용하는 자연어처리 기술은 멀티미디어 주석 정보에 대한 의미기반 색인, 검색 기술뿐 아니라, 색상 용어가 포함된 자연어 질의에 대한 색상 용어 키워드 추출 및 중의성 해소에 의한 주석 및 특징 통합 검색 기법이 사용되고 있다[22]. 본 고에서는 그림 4의 왼쪽 부분인 텍스트 부분의 의미기반 정보검색에 한정하여 설명하였다.

본 고에서는 지금까지 차세대 정보검색 기술로

자연어처리 기술과 지식베이스를 이용한 의미기반 정보검색 기술에 대하여 설명하였다. 하지만 현재까지 개발된 구문분석 기술을 적용한 의미기반 정보검색 기술은 조망간에 목표로 하고 있는 문장 의미분석에 의해 추출된 단순화된 QLF(quasi logic form)의 의미구조로부터 추출된 의미색인단위에 의한 의미기반 정보검색 기술로 발전시킬 계획이다. 또한, 본 의미기반 정보검색 기술이 국내 인터넷 검색, 지식관리, 포탈 서비스, 전자상거래 및 전자도서관 등의 다양한 인터넷 응용 분야에 널리 활용되기 위해서는 현재 구축 중인 지식베이스의 개념망이 일반 분야의 모든 개념어를 포함하도록 확장되어야 하고, 자동분류 기반 정답문서집합 구축 기술이 일반분야의 방대한 양의 정답문서를 구축하는데, 실제적인 테스트가 진행되어야 할 것으로 생각된다.

참고문헌

[1] A.T. Arampatzis, T. Tsoris, C.H.A. Koster and Th.P. van der Weide, "Phrase-based Information Retrieval", Journal of Information Processing & Management, Volume 34, Issue 6, Pages 693-707, November 1998.

[2] A.T. Arampatzis, T. Tsoris and C.H.A. Koster, "IRENA: Information Retrieval Engine based on Natural Language Analysis", In Proceedings of RIAO97 Computer Assisted Information Searching on Internet, pages 159-175, McGill University, Montreal, Canada, 1997. http://www.cs.kun.nl/~avgerino/IRENA.ps.Z

[3] Ask Jeeves, http://www.ask.com/docs/about/whatIsAskJeeves.html

[4] Ask Jeeves, Analyst Presentation, Nov. 1999. http://media.corporate-ir.net/media_files/NSD/askj/Piper_111799_vrob1/sld001.htm

[5] Boris V. Dobrow, N. V. Loukachevitch and T. N. Yudina, "Conceptual Indexing Using thematic Representation of Texts", TREC-6, 1997.

[6] D. Evans and C. Zhai, "Noun-Phrase Analysis in Unrestricted Text for Information Retrieval", Proceedings of the 34th Annual meeting of Association for Computer Linguistics, page 17-24, Santa Cruz, Univ. of California, June, 1996.

[7] Monika Henzinger, "Google Tutorial: Web Information Retrieval", Tutorial on Web Information Retrieval at ICDE'2000 (16th International Conference on Data Engineering), 2000. http://www.henzinger.com/monika/icde/icde-final_files/frame.htm

[8] Robert Krovetz and W. Bruce Croft. "Lexical ambiguity and information retrieval", ACM Transactions on Information Systems, 10(2):115-14, 1992.

[9] Michael L. Mauldin, "Retrieval Performance in FERRET: A Conceptual Information Retrieval System", Conceptual Information Retrieval : A Case Study in Adaptive Partial Parsing, Kluwer International Series in Engineering and Computer Science, 152.

[10] Jose Perez-Carballo and Tomek Strzalkowski, "Natural language information retrieval: progress report", Information Processing & Management, Vol. 36, Issue 1, Pages 155-178, January 2000.

[11] Mark Sanderson, "Word Sense Disambiguation and Information Retrieval", In Proceedings of ACM-SIGIR'94, pp.142-151, 1994.

[12] A.F. Smeaton, R. ODonnel and F. Kelledy, "Indexing Structures Derived from Syntax in TREC-3: System Description", In Harman, D.K. and Voorhees, E.M., editors, The Third Text Retrieval Conference (TREC-3), page 55-67, Gaithersburg, Md. 20899. Department of Commerce, NIST Special Publication, 1994.http://trec.nist.gov/pubs/trec3/papers/dublin.ps

[13] C. Zhai, X. Tong, N.M. Frayling and D.A.

Evans, "Evaluation of Syntactic Phrase Indexing CLARIT NLP Tract Report", In Harman, D.K. and Voorhees, E.M., editors, The Fifth Text Retrieval Conference (TREC-5), Gaithersburg, Md. 20899. Department of Commerce, NIST Special Publication, 1996. http://trec.nist.gov/pubs/trec5/papers/CLARIT-NLP-corrected.ps

[14] C. Zhai, "Fast Statistical Parsing of Noun Phrases for Document Indexing", In Proceedings of the Fifth Conference of Applied Natural Language Processing, Wsshington, DC, 1997.

[15] 강인호, 전수영, 김길창, "최대 엔트로피 모델을 이용한 한국어 명사구 추출", 제12회 한글 및 한국어 정보처리 학술대회 논문집, pp.127-132, 2000.

[16] 김미영, 강신재, 이종혁, "규칙과 어휘정보를 이용한 한국어 문장의 구묶음(Chunking)", 제12회 한글 및 한국어 정보처리 학술대회 논문집, pp.103-109, 2000.

[17] 원형석 외 2, 복합명사 분할과 명사구 합성을 이용한 통합 색인 기법, 정보과학회논문지:소프트웨어 및 응용 제27권 1호, 2000.1.

[18] 장명길, 김현진, 오효정, "HANTEC 3.0에서의 키팩트 기반 텍스트 검색 방법에 관한 연구", 제5회 한국 과학기술 정보인프라 워크샵 학술발표 논문집, pp. 203-221, 2000.12.

[19] 장문수, 장명길, 김현진, 오효정, 이재성, "인터넷 질의/응답을 위한 지식베이스 구축", 제12회 한글 및 한국어 정보처리 학술대회 논문집, pp 198-204, 2000.

[20] 장문수, 오효정, 장명길, "자동분류를 이용한 정답문서집합 구축", 제13회 한글 및 한국어 정보처리 학술대회 논문집, 2001.

[21] 한국전자통신연구원, 내용기반 멀티미디어 정보검색 기술개발, 정보통신부 제출 연구보고서, 1999.12.

[22] 허정, 김현진, 박성희, 최재훈, 장명길, "이미지 검색을 위한 색상질의분석", 제13회 한글 및 한국어 정보처리 학술대회 논문집, 2001.

[23] 허정, 옥철영. "사전 뜻풀이말에서 추출한 의미정보에 기반한 의미 중의성 해결", 제 12 회 한글 및 한국어 정보처리 학술대회. 2000.10. p269 - 276.

[24] 이충희, 김현진, 장명길, "구 기반 색인 시스템의 구현", 제13회 한글 및 한국어 정보처리 학술대회 논문집, 2001.

장 명 길

1988 부산대학교 계산통계학과(학사)
1990 부산대학교 계산통계학과(석사)
2000 충남대학교 컴퓨터과학과(박사수료)
1990~1998. 5 시스템공학연구소 선임연구원
1998. 6~현재 한국전자통신연구원 지식정보검색팀 팀장
관심분야:자연어처리, 정보검색, 생물정보학
E-mail:mgjang@etri.re.kr

김 현 진

1995 부산대학교 전자계산학과(학사)
1997 부산대학교 전자계산학과(석사)
1997~현재 한국전자통신연구원 지식정보검색팀 연구원
관심분야:정보검색, 한글처리, HCI 등
E-mail:jini@etri.re.kr

장 문 수

1992 고려대학교 전자전산공학과(학사)
1994 고려대학교 전자공학과(석사)
2001 동경공업대학 지능시스템과학전공(박사)
2001~현재 한국전자통신연구원 지식정보검색팀 선임연구원
관심분야:정보검색, 언어이해, 대화처리, 퍼지이론
E-mail:cosmos@etri.re.kr

최 재 훈

1994 전북대학교 전자계산학(학사)
1996 전북대학교 전산통계학과(석사)
2000 전북대학교 전산통계학과(박사)
2000~현재 한국전자통신연구원 지식정보검색팀 선임연구원
관심분야:지능형 정보검색, 멀티미디어 검색, 객체지향 데이터베이스, 소프트웨어 공학, 퍼지 이론, 자연어처리, 생물정보학
E-mail:jhchoi@etri.re.kr

오 효 정

1998 충남대학교 컴퓨터과학과(학사)
2000 충남대학교 컴퓨터과학과(석사)
2000~현재 한국전자통신연구원 지식정보검색팀 연구원
관심분야:문서자동분류, 정보검색, 자연어처리
E-mail:ohj@etri.re.kr

이 충 희

1996 한양대학교 전자계산학(학사)
1998 통신장교 중위 제대
2001 연세대학교 컴퓨터공학(석사)
2001~현재 한국전자통신연구원 지식정보검색팀 연구원
관심분야:자연어처리, 정보검색
E-mail:forever@etri.re.kr

허 정

1999 울산대학교 전자계산학(학사)
2001 울산대학교 전자계산학(석사)
2001~현재 한국전자통신연구원 지식정보검색팀 연구원
관심분야:정보검색, 자연어처리, 기계학습
E-mail:jeonghur@etri.re.kr

• 2002년도 특집주제 모집 •

2002년도 정보과학회지의 특집주제를 모집하오니 많은 제안 부탁드립니다. 간단하게 주제명만 제안하셔도 됩니다. 2001년 특집주제 내역은 본지 앞부분에 있습니다.

· 연 락 처 : 김경화 대리(한국정보과학회 사무국)
Tel. 042-588-9246/7
E-mail:khkim@kiss.or.kr

• 28회 정기총회 및 추계학술발표회 •

- 일 자 : 2001년 10월 19일(금) ~ 20일(토)
- 장 소 : 서울여자대학교
- 주 최 : 한국정보과학회
- 문 의 처 : 한국정보과학회 사무국
Tel. 02-588-9246/7, 4001/2
http://www.kiss.or.kr, E-mail:kiss@kiss.or.kr