생물학 도메인에서의 정보검색 : TREC의 Genomics Track을 중심으로

고려대학교 송영인 · 한경수 · 김상범 · 임해창*

1. 서 론

생물학 분야에서 많은 과학적인 발견을 이루어내면서 그에 관한 온라인 데이터와 정보의 량도 증가하고 있다. 게놈 서열 분석 기술이 발전하고 유전자나 단백질의 구조 인식 지원 도구 등이 개발 되면서, 생물학은 방대한 량의 정보를 다루어야 하는 데이터 집약적인 연구가 되었다. 이런 상황에서 지속적으로 증가하는 정보들에 어떻게 접근하고 또 그것을 어떻게 관리할 것인가가 생물학 연구자들에게는 큰 문제가 되었다. 이 문제를 극복하기 위해 생물학 도메인에서의 정보검색이 주목을 받고 있다.

정보검색은 TREC[1]을 중심으로 지난 수십년간 많은 연구가 진행되어 왔고 이제는 단순한 문서 검색의 수준을 넘어 질의응답, 다국어 정보검색, 웹 정보검색 등으로 연구 영역이 확장되는 추세이다. 이 추세의 한 흐름에 생물학 도메인의 정보검색이 놓여있다. 그러나 생물학 문서들은 독특한 특징을 지니고 있기 때문에, 기존의 정보검색 시스템을 생물학 도메인에 그대로 적용해서는 높은 성능을 기대하기 어렵다. 일반 문서와는 달리 생물학 문서에서는 DNA, 단백질, 조직 이름 등과 같은 생물학 용어가 매우 중요하다. 생물학 용어는 여러 단어가 모여 하나의 용어를 구성하며 그 길이가 매우 다양한 특징이 있다. 또 하나의 용어를 구성하는 개별 단어는 영문자, 숫자, 그리스/라틴어, 괄호나 하이픈과 같은 기호 등 다양한 형태를 가진다. 생물학 용어를 제대로 인식해 내지 못하면 정보검색의 성능은 형편없게 된다. 예를 들어, "protein G"라는 단백질 이름을 인식해 내지 못하여 "protein"과 "G"라는 개별 단어로 색인이 된다면 관련없는 문서가 너무 많이 검색되어 정확률이 낮아지게 될 것이다. 또 다른 문제는 동일한 용어를 지칭하기 위해 너무나 다양한 형태의 표현이 사용된다는 점이다. 예를 들어, "NF-Kappa B"라는 단백질은 "NF Kappa B", "NF kappa B", "NF kappaB", "NFkappaB" 등으로도 쓰인다. 이처럼 동일한 용어를 의미하는 상이한 표현들을 서로 같은 것으로 매칭시키지 못한다면 검색 결과의 재현율은 매우 낮을 것이다.

위와 같은 생물학 도메인의 특징을 반영한 정보검색의 연구가 필요하게 되어 2003년부터 TREC에 게노믹스 트랙(Genomics Track)이 진행되고 있다. 본 고에서는 2003년도 게노믹스 트랙을 중심으로 생물학 도메인 정보검색의 현 수준과 주요 이슈들을 살펴보고자 한다.

2. 게노믹스 트랙 소개

TREC의 게노믹스 트랙은 생물학, 특히 유전체학(Genomics) 도메인에서 정보검색 시스템의 개발과 평가를 위해 2003년도에 처음으로 시도되었다. 이 트랙은 2001년도에 제안되어 정보검색과 추출 분야 연구자들을 중심으로 논의 되었으며, 2002년 연구자들의 의견 수렴과 3번의 워크샵을 거쳐 현재와 같은 모습으로 결정되었다. 그 결과 시행 첫 해인 2003년도 TREC의 각 세부 트랙 중 두번째로 많은 29개 참여팀을 기록하는 등 성공적인 모습을 보였다.

TREC 2003의 게노믹스 트랙은 2개의 세부 태스크로 구성된다. 첫번째 태스크는 생물학 문서와 질의 환경에서의 문서 검색 태스크이며, 두 번째 태스크는 생물학 문서에서 특정 유전자의 기능과 관련된 정보를 추출하는 정보추출 태스크이다. TREC 2003에서는 정보추출 태스크보다는 문서 검색 태스크에 중점을 두고 진행되었으므로, 본 절에서는 게노믹스 트랙의 첫 번째 태스크인 문서 검색 태스크의 구성과 평가 방법에 대해 기술한다. 기본적으로 이 태스크는 문서 집합과 질의 집합, 그리고 평가를 위한 질의별 적합 문서 집합으로 구성되며 평가 방법으로는 정보검색 분야에서 전통적으로 사용되어 온 정확률(precision), 재현율(recall), 평균 정확률(mean average precision;MAP)을 그대로 사용한다.

* 종신회원

1) 미국 표준과학연구소에서 1992년부터 대량의 문서 집합을 다루는 정보검색 기술의 발전 및 연구자들간의 상호교류를 위해 개최한 학술회의(http://trec.nist.gov)

2.1 문서 집합

게노믹스 트랙의 문서 검색 태스크는 2002년 4월 1일부터 2003년 4월 1일 사이에 등록된 525,938개의 MEDLINE 문서 집합을 검색 대상 문서 집합으로 사용한다. 이 집합은 생물학, 의학, 약학 등의 분야를 다루는 저널에 실린 각 논문들에 대해 제목, 요약, 저자, 작성일자, 키워드 등 메타 정보들을 기록해 놓은 문서 집합이다. 이 문서 집합은 일반적으로 정보검색에서 사용되는 문서 집합과는 몇 가지 면에서 차이점이 있다. MEDLINE 문서는 논문의 요약과 제목을 중심으로 문서가 구성되며, 논문의 전문은 포함하지 않는다. 이로 인해 문서의 길이가 다른 문서 검색 태스크에서 사용되는 문서 길이에 비해 상대적으로 짧으며, 단일 주제에 대해 주로 서술된다.

또한 MEDLINE 문서 집합의 중요한 특징 중 하나는 생물학/의학/약학 관련 통제 어휘(controlled vocabulary)집합인 MeSH[2]의 키워드들을 전문가가 수동으로 각 문서에 부착하였다는 점이다. MeSH는 미국국립의학도서관(NLM: National Library of Medicine)에서 용어의 같은 개념에 대한 일관성 있는 검색을 제공하기 위해 사용하는 트리 구조의 통제 어휘이며, 이를 사용해 문서에 부착된 MH 필드는 해당 문서 내용과 관련된 일종의 메타 데이터로 볼 수 있다.

이 키워드들은 전문가들에 의해 수동으로 문서에 할당되었기 때문에, 실제로 게노믹스 트랙에 참여한 많은 그룹들은 검색 시스템 내부에 MeSH 키워드들을 다양한 방법으로 사용하여 검색 성능의 향상을 꾀하였다.

2.2 질의 집합

총 50개로 구성된 질의 집합은 질의별로 유전자 식별번호(ID), 종(organism) 제약, 유전자의 공식 명칭, 기호, 별칭 등 NLM의 LocusLink[3]에 등재된 유전자 정보로 구성된다. LocusLink는 유전자들에 대한 정보를 구축해놓은 데이터베이스인데, 표 1은 평가용 질의가 배포되기 전에 참가자들이 태스크에 적응할 수 있도록 미리 배포되었던 학습용 질의 집합의 3번 질의이다. 실제로 질의의 내용은 Homo sapiens 종에서의 'ets variant gene 6 (TEL oncogene)' 유전자에 대한 LocusLink 등재 정보와 일치한다.

즉, 본 태스크에서 주어지는 질의는 TREC의 다른 트랙들에서 사용되는 자연어로 구성된 질의와는 달리 항목별로 서식화되어 있다. 이렇게 구조적으로 기술된 질의에 대해, 시스템은 각 질의에서 기술된 유전자 X에 대하여 다음과 같은 자연어 질의가 입력된 것으로 간주하고 검색을 수행하여야 한다.

표 1 게노믹스 트랙 질의의 예(학습용 질의 3번)

QNUM	ID	organism	gene name type	gene name
3	2120	Homo sapiens	OFFICIAL_GENE_NAME	ets variant gene 6 (TEL oncogene)
3	2120	Homo sapiens	OFFICIAL_SYMBOL	ETV6
3	2120	Homo sapiens	ALIAS_SYMBOL	TEL
3	2120	Homo sapiens	PREFERRED_PRODUCT	ets variant gene 6
3	2120	Homo sapiens	PRODUCT	ets variant gene 6
3	2120	Homo sapiens	ALIAS_PROT	TEL1 oncogene

"유전자 X의 기초 생물학적 정보(basic biology)나 지정된 유기체에서 생성된 단백질의 기초 생물학적 정보를 주로 다루는 모든 MEDLINE 문서를 검색하라. 기초 생물학적 정보란 정상 상태나 질병 상태에서 유전자/단백질의 격리(isolation), 구조(structure), 유전적 특징(genetics), 기능(function) 등을 포함한다."

참가팀에게는 태스크의 특성을 파악할 수 있도록 질의 50개로 구성된 학습용 질의 집합과 이에 대한 적합 문서 집합이 먼저 배포되었으며, 이후에 평가를 위한 역시 50개의 평가 질의가 주어졌다. 따라서 문서 검색 태스크에 참여한 각 그룹들은 이후에 배포된 평가 질의 집합에 대해 검색을 수행한 후 그 결과를 제출하였으며, 최종 평가 종료 후 평가 질의 집합에 대한 적합 문서 집합이 배포되었다. 학습 질의와 평가 질의 집합, 그리고 각 질의 집합에 대한 적합 문서 집합은 게노믹스 트랙 홈페이지[4]에서 다운로드 받을 수 있다.

2.3 적합 문서 집합

게노믹스 트랙의 문서 검색 태스크에서는 적합 문서 집합으로 NLM의 GeneRIF 데이터를 사용하였다. NLM의 GeneRIF는 LocusLink에 등재된 유전자와 해당 유전자의 기초 생물학적 정보와 관련된 논문들에 수작업으로 정보를 부착하여 구축한 데이터베이스인데[1], 여기서 의미하는 기초 생물학적 정보란 앞 절에서 기술한 것과 동일하다. 그림 1은 GeneRIF에 정의된 MEDLINE 문서 리스트를 보여주고 있다.

2) http://www.nlm.nih.gov/mesh/meshhome.html

3) http://www.ncbi.nlm.nih.gov/LocusLink/

4) http://medir.ohsu.edu/~genomics/

실제로 검색 시스템의 성능을 평가하는데 있어서 가장 어려운 점은 질의에 적합한 정답 집합을 어떻게 구축하느냐 하는 것이다. 게노믹스 트랙은 처음 열리는 것이었으므로, 주최측은 우선 연구자들이 GeneRIF에 각 유전자들의 기초 생물학적 정보를 다루고 있는 논문들이라 해서 자발적으로 등재시켜 놓은 문서들을 기본적으로 정답 집합이라 규정하였다. 이는 몇가지 문제를 발생시킬 수 있는데, MEDLINE 문서 집합에 있는 문서들 중 질의에서 제시된 유전자의 기초 생물학적 정보를 다루고 있으나, GeneRIF에는 등재되지 않은 문서들이 많이 존재하기 때문이다. 이러한 문제는 매년 게노믹스 트랙을 준비하는 과정에서 조금씩 개선되어 갈 것이다.

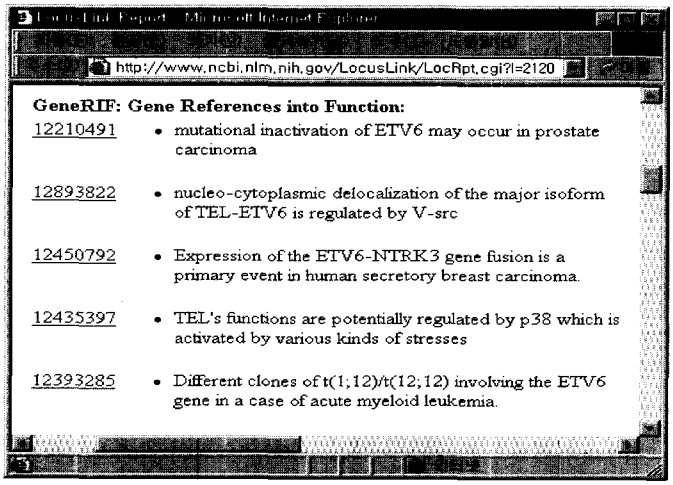


그림 1 유전자 ETV6의 기초 생물학적 정보를 다룬 것이라고 GeneRIF에 등재된 MEDLINE 논문 리스트

표 2 다양한 용어 변이의 예

다양한 약어 생성	***"epidermal growth factor receptor"*** → "Egfr", "EGF-R", "EGF-receptor" ***"adrenergic receptor, alpha 1d"*** → "alpha1D-AR", "Adra1d"
용어내 단어 도치	***"sphingolipid G-protein-coupled receptor"*** → "G protein-coupled sphingolipid receptor" ***"DNA synthesis inhibitor"*** → "inhibitors of dna synthesis"
용어내 단어 생략	***"DEAD/H (Asp-Glu-Ala-Asp/His) box polypeptide 5 (RNA helicase, 68kD)"*** → "DEAD/H box-5 (RNA helicase, 68kD)"

3. 검색성능 향상을 위한 주요 고려사항

게노믹스 트랙에서의 문서 검색 태스크는 일반적인 문서 검색 태스크와 형식면에서 상당히 유사하나, 문서의 도메인이 생물학, 의약학이라는 전문 분야로 한정되었다는 점과, 질의에 구조적인 형식으로 유전자 정보가 기술되어 있다는 점 등에서 차이를 보인다. 게노믹스 트랙에 참가한 많은 시스템들 역시 유전체학 도메인이라는 특성을 효과적으로 이용하여 검색의 성능을 향상시키고자 하였다. 본 절에서는 게노믹스 트랙에 참가했던 다수의 시스템들이 중요하게 고려하였던 사항들에 대해 기술하고자 한다.

3.1 전문 용어의 처리

다른 도메인에서 사용되는 전문 용어와는 달리, 생물학에서 사용되는 전문 용어는 여러 단어로 구성된 다어절 용어(multiword term)인 경우가 빈번하고, 공식적으로 정해진 용어(official term)의 다양한 변이 형태가 두루 사용되며, 용어의 형태도 매우 독특하다는 특징이 있다. 따라서 이러한 생물학 도메인에서의 전문 용어 특성을 고려하지 않고 일반적인 방법으로 문서나 질의를 처리하게 되면 검색성능이 심각하게 저하될 수 있다.

긴 길이의 다어절 용어는 전통적인 문서검색 시스템의 성능을 심각하게 저하시킬 수 있다. 물론 복합어의 인식 및 색인 문제는 비단 생물학 도메인에서 뿐 아니라 일반적으로 검색 엔진을 구현할 때 다루어야 할 중요한 사항 중 하나이다. 그러나 생물학 도메인에서는 질의로 사용될 가능성이 가장 높은 유전자 이름, 단백질 이름 등의 전문 용어들이 거의 대부분 다어절 용어로 구성되어

있어, 그 중요성은 기존의 정보검색 시스템에 비해 대단히 크다고 할 수 있다. 이러한 특징은 다어절 용어를 문서에서 자동으로 인식하여 처리하거나, 검색시 질의어 출현의 근접도(proximity) 등을 고려하여 문서와 질의간의 유사도를 계산하는 등의 부가적인 기술을 요구하게 된다.

더 어려운 점은 이러한 다어절 용어들이 축약, 생략, 도치 현상을 보이면서 매우 일관성 없이 사용된다는 것이다. LocusLink와 같이 상당수 유전자 명에 대해 동의어와 약어들을 기술해 놓은 데이터베이스가 공개되어 있기는 하지만, 문서 집합에 나타나는 모든 경우를 포괄하기에는 턱없이 부족하다. 이러한 문제점을 해결하기 위하여 게노믹스 트랙에 참여한 상당수의 시스템들은 용어 정규화라든가 용어 변형의 자동 생성 기법 등을 고안하였다. 또한 이 분야의 전문 용어들은 "1(2)gd-1" 와 같이 그 형태 자체도 알파벳, 숫자, 괄호, 하이픈 등이 뒤섞여 나타나기 때문에 문서 처리를 위한 토큰화 단계부터 매우 주의를 기울여야 한다. 표 2는 생물학 도메인에서 나타나는 다양한 용어 변형 현상을 보여주고 있다.

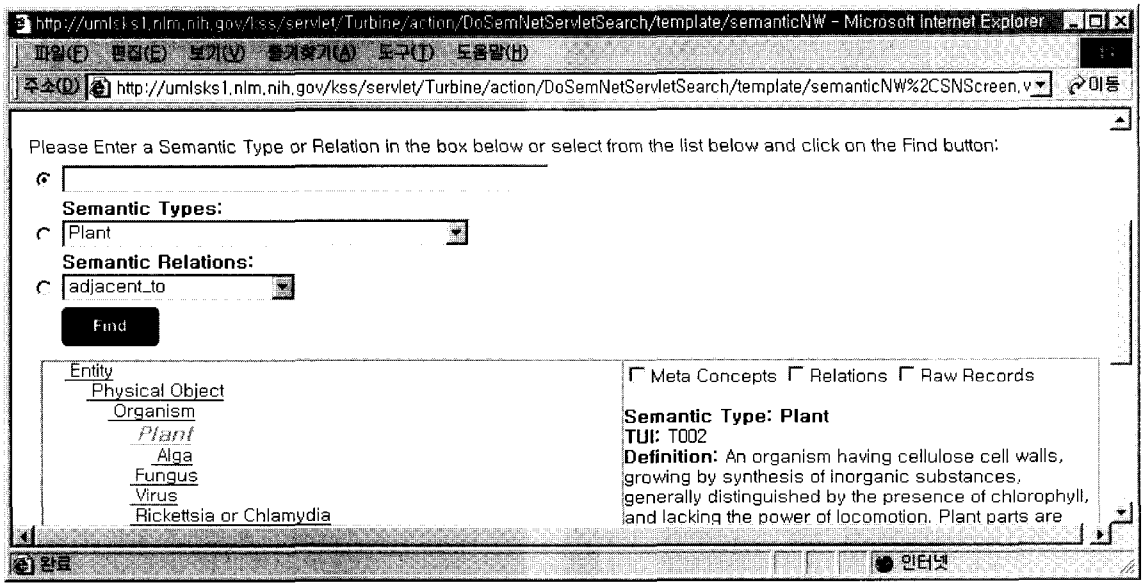


그림 2 UMLS 의미망 예 : Entity > Physical Object > Organism > Plant

3.2 생물학 지식 자원의 사용

생물, 의학 분야의 전문가들은 이미 수작업으로 많은 지식 자원들을 구축해 두었는데, 대표적인 예로 NLM에서 구축한 통합의학 언어 시스템(UMLS: Unified Medical Language System)이 그 좋은 예이다. UMLS는 의미망(Semantic Network), 메타 시소러스(UMLS Metathesaurus), 전문가 사전(SPECIALIST Lexicon)으로 구성된 지식 자원으로서 생물학/의학 분야의 다양한 어휘 및 개념 정보를 담고 있다. 그림 2는 UMLS의 의미망의 예를 보여주고 있다.

이렇게 도메인에 특화된 지식 자원은 생물학 도메인 문서 검색의 여러 가지 문제를 해결하는데 유용하다. 특히 전문 용어들에 대한 동의어(synonym)집합과 각 용어들 사이의 상하위어 관계(Is-A)에 대한 정보는 앞 절에서 언급한 생물학 용어의 다양한 변이 현상을 처리하는데 도움을 줄 수 있으며, 약어(acronym)에서 발생하는 중의성을 처리하기 위해서도 필요하다.

UMLS외에도 유전자의 기능에 관련된 정보를 담고 있는 Gene Ontology(GO)라는 온톨로지는 검색된 문서가 실제로 질의에서 주어진 유전자의 기초 생물학적 정보를 다루고 있는지 검증하는데 사용되기도 하였다.

3.3 서식화된 질의의 형식

일반적으로 문서 검색에서 사용되는 질의와는 달리, 게노믹스 트랙에서의 질의는 유전자에 대한 정보가 테이블 형태로 서식화 되어 주어진다. 이 테이블에는 질의 번호, LocusLink의 유전자 식별 번호, 유전자의 공식 명칭과 기호 및 별칭, 생산하는 단백질, 종 등의 정보가 포함되어 있으며, 이렇게 기술된 유전자의 기초 생물학적 정보를 다루고 있는 문서를 찾는 것이 시스템에게 주어진 임무가 된다. 따라서 동일한 유전자에 대해 기술된 여러 가지 정보들을 검색 과정에서 어떻게 효과적으로 활용하느냐에 따라 검색효과가 달라질 수 있다.

3.4 재순위화/필터링

유전자 정보를 질의로 하여 검색 시스템이 검색을 한 후, 마지막으로 필요한 작업은 필터링이다. 검색 후 필

터링은 다음 두 가지 목적을 위해 수행된다. 첫째로 질의에 명시된 종에 부합되는 문서만을 찾기 위함이다. 일반적으로 질의에서 주어진 유전자를 다루는 문서라 할지라도, 대개의 경우 문서마다 서로 다른 종의 영역에서 해당 유전자의 특징이나 성격을 언급하고 있다. 그러나 일반적으로 생물학자들이 특정 유전자에 대한 문서를 검색할 때는, 그 문서가 다루고 있는 종을 제약 조건으로 제시한다. 게노믹스 트랙에서의 질의도 역시 특정 종을 다루고 있는 문서만을 검색하도록 명시되어 있기 때문에, 검색된 문서의 필터링이나 가중치 조정에 의한 재순위화 작업이 필요하다.

둘째로 질의에는 유전자 정보만 명시되어 있으나, 태스크의 목적은 2.1.2에서 기술한 바와 같이 유전자에 대한 "기초생물학적 정보"를 다루고 있는 문서만을 찾는 것이다. 따라서 질의에 주어진 유전자에 대한 문서라 할지라도, 그 유전자의 기초생물학적 정보를 다루고 있지 않은 문서는 검색대상 문서에 포함되지 않는다. 따라서 검색된 문서가 기초생물학적 정보를 다루는지에 대한 여부를 기준으로 필터링할 필요가 있다. 많은 참가팀들은 유전자 서열(sequence), 활성화(activation) 등 모든 유전자의 기초 생물학을 표현하기 위하여 자주 사용되는 생물학 용어나 표현들을 수동 혹은 자동으로 구축한거나, 자동 문서 분류기를 학습시키는 방법 등을 시도하였다.

4. 참가 시스템별 주요 특징

2003년 TREC에 참가하였던 많은 시스템들은 3절에서 살펴본 주요 고려 사항들에 대해 독창적인 방법으로 해결책을 제시하고 있다. 본 절에서는 참가한 시스템들 중 대표적인 몇 가지 시스템들을 소개함으로서, 앞절에서 언급한 여러 가지 문제들을 어떻게 해결하였는지에 대해 이해를 돕고자 한다.

4.1 노이체텔 대학 시스템과 일리노이 주립대학 시스템

스위스 노이체텔 대학(University of Neuchatel)과 미국의 일리노이 대학(University of Illinois at Urbana-Champaign)에서는 도메인 종속적인 정보를 전혀 사용하지 않고, 전통적인 정보검색 방법으로 접근한 대표적인 시스템들이다[2,3]. 노이체텔 대학의 시스템은 SMART 문서 검색 시스템[5]을 사용하여 다양한 구분자와 스테밍 방법, 검색모델, 적합성 피드백, 검색 결과 퓨전 등 기존의 문서 검색에서 성공적으로 사용된 많은 기존 기법들을 게노믹스 트랙의 환경에서 다양하게 조합하여 실험하였다. 일리노이 대학의 시스템 역시 전통적인 정보검색 기법을 사용하였으나, 노이체텔 대학의 시스템과 다른 점은 노이체텔 시스템의 경우 구조적 질의의 특성을 무시하고 유전자 공식 명칭이나 별칭들을 구분 없이 하나의 질의로 사용하였으나, 일리노이 대학에서 개발한 시스템에는 게노믹스 트랙의 구조적 질의 특성을 반영해 검색 모델을 수정하는 등 질의 처리에 있어 여러 가지 아이디어가 시도되었다.

이 두 시스템은 공통적으로 도메인 종속적인 방법을 사용하지 않은 채, 전통적인 검색 모델에 기반하여 검색 성능을 향상시키는데 집중하였으며, 전체 참가팀 중 중간 정도의 평균 정확도를 보였다. 결론적으로 도메인 종속적인 정보를 사용하지 않을 경우 성능 향상에 그 한계가 상당하다는 사실을 알 수 있다.

4.2 미국국립의학도서관 시스템

처음으로 열린 게노믹스 트랙에 참여한 시스템 중 가장 우수한 성능을 보였던 시스템은 미국국립의학도서관(NLM)에서 참여한 시스템이다[4]. NLM의 시스템은 Inquery 검색 시스템을 기반으로 하고 자체적으로 개발한 생물학 도메인에서의 검색 도구(IR tools)인 SE(Search Engine)을 결합시킨 생물학 도메인에 특화된 정보검색 시스템이다.

이 시스템에서는 색인어를 추출하기 위한 토큰화 단계에서 생물학 용어의 특이성을 고려하여 하이픈(-)이나 괄호, 영어와 알파벳이 혼용된 단어처리를 위해 여러 가지 규칙을 사용하였다. 특이한 점은 Title, MeSH Heading, NameOfSubstance, Abstract 등 문서의 어떤 필드에서 질의어가 출현하였는지, 유전자 이름 형식은 어떠한지에 따라 서로 다른 가중치를 할당하여 문서를 랭킹하였다는 것이다.

질의에서 제시한 종을 다루지 않은 문서의 순위를 떨어뜨리기 위해, 제시된 종을 표현하는 단어가 문서에 나타나지 않을 경우 질의와의 유사도를 현격히 낮추었다. 기초생물학적 정보를 담고 있는 문서만을 찾기 위해서 'genetics', 'gene expression', 'sequence' 등과 같이 GeneRIF 상에서 자주 출현하는 용어를 수집, 핵심 문맥(key contenxt) 용어 목록을 작성하고, 이 목록에 포함된 단어들이 많이 출현할 수록 상위에 랭크되는 방법을 고안하였다. 그 외에도 NLM은 통제 어휘인 MeSH와 규칙 리스트에 의한 규칙 기반 검색 전략을 추가하고 생물/의학 관련 어휘 사전인 SPECIALIST Lexicon에 기반하여 처리된 어휘들의 공기 네트워크(Collocation Networks)를 베이지안 네트워크 학습 방법으로 구축하여 활용하였다.

5) 이 검색 엔진은 ftp://ftp.cs.cornell.edu/pub/smart/에서 다운로드 받을 수 있다.

NLM의 시스템이 사용한 여러 가지 방법들 중, 질의에 제시된 종을 표현하는 단어가 문서에 나타나지 않을 경우 질의와의 유사도를 떨어뜨리는 전략이 대단히 효과적이었다.

4.3 캐나다 국립연구소 시스템

캐나다 국립연구소(NRC : National Research Council of Canada)팀 역시 다양한 도메인 종속적인 기법들을 사용하여 좋은 성능을 보여주었다[5]. 이 시스템은 문서 집합을 데이터베이스에 저장하고 질의어가 나타난 모든 문서들을 1차로 검색한다. 다음 단계로 전문가가 작성한 규칙들을 사용하여 질의에서 주어진 유전자 이름들이 취할 수 있는 변이 형태들을 모두 질의에 추가하여 확장하는 한편, 약어 중의성 해소를 통해 질의에 주어진 약어가 출현한 문서라 할지라도 그 약어가 질의에서 주어진 약어와 다른 전문 용어의 약어일 경우 그 문서들은 리스트에서 제외시킨다. 예를 들어 "mitral valve prolapse", "microvascular pressure", "Midwifery Ventouse Practitioners"는 서로 다른 전문 용어이지만 약어로는 모두 MVP로 표기된다. 약어 중의성 해소 기술을 사용할 경우, 만일 "mitral valve prolapse"에 대한 약어 MVP가 질의로 주어졌다면, 1차 검색에서 검색된 문서들 중 "microvascular pressure"나 "Midwifery Ventouse Practitioners" 등의 약어로 사용된 MVP가 포함된 문서들은 리스트에서 제거할 수 있게 된다.

위의 두 작업이 완료되면 의사 적합성 피드백을 2회 수행한다. 이때 각 회별로 전체 문서 리스트의 20% 이상에서 출현한 단어들 중 적어도 5개를 추가하여 문서 리스트를 다시 구축한다. 이렇게 구축된 최종 문서 리스트에서 질의에 명시된 종을 표현하는 용어가 없는 문서들은 모두 제거하고, 전통적인 tf • idf 방법에 의한 문서 랭킹을 수행하여 문서들을 정렬하게 된다. 마지막으로, 기초생물학 정보를 표현하는 구(phrase)를 GeneRIF 데이터를 사용하여 학습한 뒤, 이러한 구가 제목에 출현하였을 경우 순위를 높여주는 방법으로 검색의 정확률 향상을 꾀하였다.

다른 시스템과 비교했을 때 NRC 시스템에서 특이한 점은 약어 중의성 해소와 학습을 사용한 기초생물학 관련 구를 사용하여 순위를 조정하는 단계로 볼 수 있지만, 이들 모두 성능 개선에 큰 도움을 주지는 못했다. 반면 NLM의 시스템과 마찬가지로 질의에 명시된 종을 표현하는 용어가 없는 문서들을 제거하였던 것이 평균 정확률을 크게 향상시키는 것으로 나타났으며, 용어 어형 확장 방법도 성능 개선에 크게 기여하였다.

4.4 버클리대학 시스템

버클리대학(University of California, Berkeley)의 BioText팀에서 개발한 시스템은 n-그램 오버랩이라는 방법을 사용하여 말뭉치로부터 유전자 명칭들이 변형되는 패턴들을 추출한 뒤 전문가가 수동으로 변형규칙을 만들어냈다는 특징이 있다[6]. n-그램 오버랩은 말 그대로 유전자의 공식 명칭과 문자 오버랩이 많이 일어나는 문자열을 말뭉치로부터 뽑아내는 것을 말한다. 이렇게 뽑아낸 문자열들은 유전자 명칭의 변형일 가능성이 크기 때문에, 전문가는 이를 참조하여 적용률이 높은 유전자 명칭 변형규칙을 효율적으로 구축할 수 있게 된다.

또 다른 주된 특징은 기초생물학적 정보에 관련된 문서를 찾기 위해 자동 문서 분류 기법을 사용했다는 점이다. 자동 문서 분류를 위한 학습 집합은 질의로 사용되지 않은 50개의 유전자와 관련된 GeneRIF 등재 문서를 사용했으며, 문서 내 출현한 단어의 어휘 정보와 MeSH 필드에 할당된 키워드를 자질로 사용한 단순 베이지안 학습방법을 사용하였다. 이 방법은 성능 향상에 도움을 주기는 하였으나 그 향상 정도는 크지 않았는데, 이는 지도학습에 의한 자동 문서 분류기를 학습하기 위해 GeneRIF 만을 사용하여 구축한 학습 집합 자체에 매우 오류가 많기 때문이라고 결론을 내었다.

4.5 액손톨로직 시스템

좋은 성능을 보여준 시스템 중 하나인 액손톨로직사(Axontologic, Inc.)에서 개발한 시스템은 유전자 이름의 동의어들과 그 변형들을 사용하는 질의 확장 기법에 초점을 두었다[7]. 이 시스템은 수동으로 구축된 생성문법을 사용하여 유전자 이름 변이를 자동으로 생성한 뒤 초기 문서 검색을 수행한다. 그리고 이 결과는 문서가 유전자와 유전자의 기능에 얼마나 적합한가에 따라 재순위화 되는 과정을 거치는데, 이를 위해 기능 표현어 목록(function term list)을 구축하였다. 기능 표현 용어 목록이란 유전자의 기능과 관련된 단어 리스트를 말하는데, 학습 질의/문서 집합과 Gene Ontology(GO)와 같은 생물학 도메인 자원을 분석해 수동으로 구축한 "inhibit", "cleave", "activation", "regulation" 등의 단어들이 목록에 수록되어 있다.

이를 사용하여 시스템은 초기 검색 결과를 재순위화하는데, 각 문서에서 유전자 이름과 기능 표현어가 한 문장 내에서 얼마나 함께 자주 출현했는가, 또 그 두 종류의 단어가 얼마만큼 떨어져서 출현했느냐에 따라 가중치를 부여하는 방식으로 진행된다. 즉, 유전자 이름과 기능 표현어가 함께 출현한 문서는 게노믹스 트랙에서 요구하는 기초생물학적 정보를 포함할 가능성이 크다는 가정인 것이다. 그 외에도 유전자 이름이 각 기능 표현

어의 주어나 목적어로 사용될 경우도 높은 가중치를 부여하며, 최종적으로 질의에서 주어진 종과 관련된 MeSH 키워드가 문서에서 출현하였는지 살펴 문서 필터링을 수행한다.

4.6 캘리포니아 주립대학 시스템

캘리포니아 주립대학에서 개발한 시스템은 정확한 질의-문서간 매칭을 위해 다양한 규칙을 사용했다는 점이 특징이다[8]. 이들은 유전자 이름을 정규화하거나 유전자 이름 변종을 생성하기 위해 10개의 규칙을 정의했으며, 이와는 별도로 단일 어절의 길이가 긴 유전자 이름에 대해 두개의 문자열로 분할하여 탐색하는 방법과 다양한 방식으로 약어를 생성하는 규칙등을 추가 사용하였다. 또한 "METALLOTHIONEINIII"와 "METALLOTHIONEIN3"의 경우처럼 단어의 일부분만 일치하지 않는 경우에도 질의-문서 매칭을 수행하기 위해 단어의 문자열이 80% 일치할 경우, '유사하다'라고 판정하는 80% 유사도 규칙 등 다양한 경우를 고려하여 검색을 수행하였다.

이 검색 시스템은 유전자의 이름이 부분적으로 매칭되는지 아니면 전체가 매칭되는지에 따라 다른 가중치를 할당하며, 특히 문서 제목에 유전자 이름이 출현한 경우 별도의 부가적인 가중치를 주어 문서를 랭킹한다.

5. KUBIOIR 생물학 문서검색 시스템

본 장에서는 생물학 도메인에서의 정보검색 시스템에 대해 좀 더 자세히 살펴보기 위하여 고려대학교 자연어처리 연구실에서 개발한 KUBIOIR 시스템에 대해 설명하고자 한다. 이 시스템은 3장에서 설명한 대부분의 사항들을 고려하여 개발되었으며, 게노믹스 트랙에 참여한 시스템들 중 상위권에 해당하는 좋은 성능을 보였다.

KUBIOIR은 생물학 용어 인식기, 색인어 추출 규칙 등 도메인 어휘 특성을 고려한 색인 방법, 다양한 유전자 이름 변이와 유전자 용어의 어휘 특성을 반영한 질의 처리 및 질의-문서 가중치 부여 방법 등을 그 특징으로 한다. 또한 종과 기초생물학적 정보에 대한 조건을 만족시키기 위해, 초기 검색 후 문서 재순위화와 필터링 과정을 수행한다[9].

이 시스템은 크게 문서를 처리하여 색인을 생성하는 색인 모듈과 사용자의 질의를 처리하여 내부 질의로 변환하는 질의 분석 모듈, 문서와 질의를 매칭하여 검색을 수행하는 순위화 모듈, 그리고 재순위화와 문서 필터링을 포함한 후처리 모듈 등으로 구성된다. 그림 3은 KUBIOIR 생물학 문서 정보검색 시스템의 구조를 보여준다.

다음 각 절에서는 KUBIOIR 시스템에서 사용한 색인어 추출 방법, 질의-문서 랭킹 모델, 초기 검색 후 후처리 방법에 대해 기술한다.

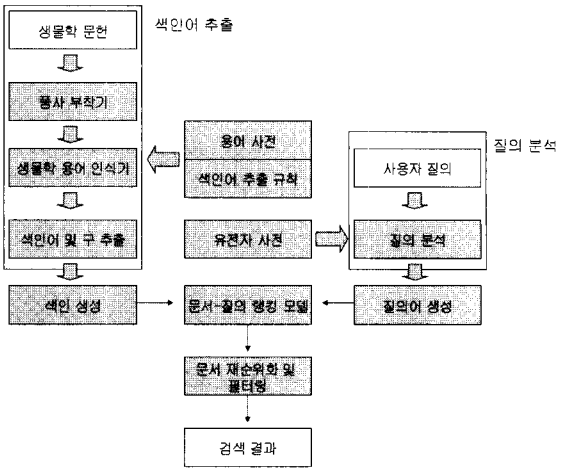


그림 3 KUBIOIR 시스템 구성도

5.1 색인어 추출

KUBIOIR 시스템은 색인어 추출을 위해 토큰화(Tokenization), 불용어 제거, 어근화(stemming) 등을 수행한다. 생물학 용어의 특성을 반영하기 위해 간단한 색인어 추출 규칙과 생물학 용어 경계 인식기를 사용한 구(phrase) 추출 방법을 사용한다. 하이픈이나 괄호 등 거의 모든 기호들은 색인어 추출을 위한 토큰화 과정에서 모두 구분자로 사용되며, 구분자를 통해 인식된 토큰열에 대해 다음과 같은 규칙이 적용된다.

"단어 w1이 단일 알파벳으로 구성되거나 2음절 이내의 숫자로 구성되고, 좌우에 인접한 단어 w2가 그렇지 않은 단어일 경우, w2가 색인어로 추출되고 w1과 w2이 결합하여 정규화된 w1:w2 가 역시 색인어로 추출된다."

예를 들어, 문서에서 "G protein", "G-protein", "protein G"와 같이 동일한 전문 용어가 서로 다른 형태로 출현했을 경우, 위 규칙에서 w2에 해당하는 "protein"이 우선 색인어로 추출되고, w1에 해당하는 "G"와 결합된 "G:protein"이 역시 추출된다. 이 규칙은 단일 알파벳이나 숫자를 불용어로 처리하지 않으면서도, 질의와 무관한 문서가 지나치게 많이 검색되어 나오는 현상을 막을 수 있다. 실제로 "G"라는 1음절 알파벳은 "protein G"에서 뿐 아니라 "PROTEINAS G", "Pyronine G" 등 전문용어에서 빈번하게 출현함을 알 수 있다.

5.2 생물학 용어 인식기에 기반한 색인어 추출

전문 용어를 좀 더 정확하게 처리하기 위하여 본 시스템에서는 문서에서 용어 경계를 자동으로 인식한 후 간단한 구 색인어 추출을 수행한다. 구 색인어 추출은 품

사 자동 부착, 용어 경계 인식, 구 생성의 세 단계로 구성되어 있다.

구 색인을 위해 사용하는 품사 자동 부착기는 HMM 기반의 품사 부착기를 사용하되, 생물학 도메인에서의 정확성 향상을 위하여 생물학 용어에 관한 학습 자질과 생물학 도메인 학습문서 집합[6]을 추가로 사용하였다. 이렇게 해서 품사가 부착된 문서는 생물학 개체명 인식기에 입력된다. SVM을 사용하여 개발된 이 인식기는 내부적으로 용어 경계 인식과 개체명 분류의 두 단계로 나뉘는데[10], 본 검색 엔진에서는 대량의 문서를 처리해야 하기 때문에 용어 경계 인식까지만 수행하도록 수정되어 결합되었다. 마지막으로 인식된 각 다어절 용어에 대해서 다어절 용어 내부의 모든 단어 바이그램이 추출된다. 물론 인식된 구를 하나로 보아 색인어로 사용할 수도 있으나, 본 시스템에서는 검색 정확도를 향상시키는 구 색인의 장점을 유지하면서도, 용어 내 일부 단어가 생략되거나, 도치되는 경우에도 부분 매칭이 가능하므로, 인식기의 오류에 민감하지 않고 견고하다는 장점이 있다.

물론 인식된 용어 내의 모든 개별 단어들과 용어로 인식되지 않은 모든 단어들은 독립적으로 색인된다.

표 3 게노믹스 트랙에 참여한 25그룹의 시스템 중 상위 10개 성능

NLM	0.4165
NRC	0.3941
UC Berkeley	0.3753
University of Waterloo	0.3534
Axontology Inc.	0.3173
California State University San Marcos	0.3079
University of Edinburgh & Stanford University	0.3015
Korea University	0.2980
Tarragon Consulting Corporation	0.2837
IBM Corp.	0.2823

5.3 질의어 가중치 할당

KUBIOIR은 Okapi 시스템[11]의 BM25 가중치 함수 중 질의 내 질의어 가중치를 독창적인 방법으로 할당한다. 게노믹스 트랙에서의 질의는 유전자의 공식 이름과 약어, 그리고 별칭 등이 나열되어 있는데, 직관적으로 질의에서 제시한 유전자의 약어 하나가 문서에 1회 출현하는 것과, 여러 단어로 구성된 공식 유전자 명칭이 1회 출현하는 것은 동일한 가치를 갖는다. 그러나, 기존의 정보검색 모델을 그대로 사용할 경우 다어절의 공식 유전자 명칭이 출현한 문서가 약어가 출현한 문서보다 훨씬 높은 질의와의 유사도를 갖게 된다. 이는, 공식 유전자 명칭에 들어있는 모든 단어의 가중치가 유사도에 모두 더해지기 때문이다. 이러한 현상을 방지하기 위하여, 본 시스템에서는 유전자의 공식 명칭, 약어, 별칭 등 각 항목에 고정된 질의 내 가중치를 주고, 각 항목 내의 단어들은 주어진 가중치를 어절수로 나눈 가중치를 갖는다. 즉, 특정 질의에서 약어 단어가 1의 가중치를 갖는다면, 유전자의 공식 명칭을 이루는 4개의 단어에는 각각 0.25의 가중치가 할당된다.

또한, 정보검색에서 역문서 빈도(inverse document frequency)의 개념과 유사한 역질의 빈도(inverse query frequency)라는 개념을 사용하여 유전자의 명칭들 중에서도 다른 유전자들과 차별성을 갖는 단어들에 더 높은 가중치를 줄 수 있도록 하였다. 실제로 기존의 역문서 빈도도 이와 유사한 기능을 해 주지만, MEDLINE이라는 말뭉치가 유전자학에 관련된 문서만을 다루지 않고 생물학, 의학, 약학과 관련된 전 분야를 포괄하기 때문에 이중 극히 일부인 유전체학 도메인에 종속적인 검색 엔진을 구축하는데는 이 말뭉치에서 얻어낸 역문서 빈도가 부적절한 상황이 생긴다.

따라서 본 시스템에서는 웹에서 수집한 약 2만여 개의 유전자 명칭 목록으로부터 역질의 빈도를 계산한다. 게노믹스 트랙에서는 오직 유전자의 명칭만이 질의로 입력되기 때문에, 가능한 모든 질의집합을 수집한 유전자 명칭 목록으로 가정할 수 있다. 이렇게 수집된 목록으로부터 계산된 역질의 빈도는 어떤 단어가 유전자 명칭을 구별할 수 있는 변별력이 어느 정도인지에 대해 MEDLINE 문서 집합에서 계산해낸 역문서 빈도보다 더 적절한 값을 제공한다.

5.4 문서 필터링과 문서 재순위화

본 시스템에서는 마지막으로 기초생물학 정보를 다루

6) 학습문서 집합으로는 일본 동경대에서 진행중인 GENIA 프로젝트의 산출물중 하나인 말뭉치를 사용하였다. 이 말뭉치는 http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/에서 다운로드 받을 수 있다.

면서 질의에서 제시한 종과 일치하는 문서만 걸러내기 위해 별도의 필터링 및 재순위화를 수행한다. 우선 검색된 문서 중에 질의에서 제시된 종에 해당하는 MeSH 키워드는 없으나, 다른 종에 해당하는 MeSH 키워드가 수록되어 있는 문서들은 모두 제거하였다. 바꾸어 말하면 질의에서 제시된 종이 MeSH 키워드에 수록되어 있던가, 아니면 어떤 종에 관련된 MeSH 키워드도 수록되어 있지 않은 경우만 검색 결과로 제시된다.

또한 질의 유전자의 기초생물학적 정보에 적합한 문서를 검색하기 위하여 전문가에 의해 수집된 생물학에서 발생하는 이벤트와 관련된 160여 개의 동사 목록을 사용하여, 이러한 이벤트 동사들이 빈번하게 출현한 문서일수록 순위가 높아질 수 있게 하였다.

표 3은 고려대학교 시스템을 포함하여 2003년 게노믹스 트랙에 참가한 25개 그룹 중 좋은 성능을 보인 상위 10개 그룹 시스템의 성능(평균 정확도)을 보여주고 있다.

6. 결 론

생물학과 전산학이 결합된 생명정보학(Bioinformatics)은 최근 몇 년간 대단히 각광을 받아왔으며, 특히 2001년 말에 인간 유전자 지도가 완성된 이후 인간의 삶을 풍요롭게 할 수 있는 생명공학 기술들은 전산학 기술과 결합되어 그 발전 속도가 날이 갈수록 빨라지고 있다. 이러한 기술 발전의 원동력은 생물학이나 의학에서 쏟아져 나오는 수많은 연구 결과들을 체계적으로 정리해 두어 연구자들이 빠르고 편리하게 정보를 얻어낼 수 있는 환경에서 나올 수 있는데, 이를 위해 미국에서는 국가적인 차원에서 오래전부터 투자를 계속해 오고 있다. 본 고에서 중점적으로 다룬 TREC의 게노믹스 트랙은 이렇게 생물학자들을 지원할 수 있는 정보검색 시스템을 개발하기 위한 여러 가지 아이디어와 알고리즘들을 공유하고 토론하기 위한 학술회의로서 북미, 유럽, 아시아 등 세계 각국의 연구그룹들로부터 많은 관심을 모으고 있다.

향후 생명과학이나 공학에 종사하는 연구원들을 지원할 정보검색 시스템은 좀 더 지능적인 문헌 분석과 추론이 가능하여 연구자들이 실험을 기획하거나 새로운 사실을 발견하는데 걸리는 시간을 획기적으로 단축시켜줄 수 있도록 하는 방향으로 발전될 것이다. 국내에서 정보검색을 연구하는 학교와 연구소, 그리고 기업들이 올해 두 번째로 열리게 될 게노믹스 트랙에 많이 참여하여 외국 유수의 연구 그룹들과 유익한 토론의 장을 가질 수 있기를 기대한다.

참고문헌

[1] J. Mitchell, A. Aronson, et al., "Gene Indexing: Characterization and Analysis of NLM's GeneRIFs," Proc. AMIA Symposium. 2003

[2] S. Jacques, R. Yves, P. Laura, "Report on the TREC-2003 Experiment:Genomic and Web Searches," The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. NIST. 2003

[3] Z. ChengXiang, T. Tao, F. Hui, S. Zhidi, "Report on the TREC-2003 Experiment : Genomic and Web Searches," The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. NIST. 2003

[4] M. Kayaalp,A. Aronson,et al., "Methods for accurate retrieval of MEDLINE citations in functional genomics," The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. NIST. 2003

[5] B. deBruin, J. Martin, "Finding gene function using LitMiner," The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. NIST. 2003

[6] G. Bhalotia, P. Nakov, et al., "BioText team report for TREC 2003 Genomics Track," The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. NIST. 2003

[7] H. Richard, W. Larry, "Recognizing Gene and Protein Function in MEDLINE Abstracts," The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. NIST. 2003

[8] G. Rocio, F. Tasnim, "REGEN: Retrieval and Extraction of Genomics Data," The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. NIST. 2003

[9] Y.I. Song, K.S, Han, H.C. Seo, S.B. Kim, H.C. Rim, "Biomedical Text Retrieval System at Korea University," The Twelfth Text REtrieval Conference: TREC 2003, Gaithersburg, MD. NIST. 2003

[10] K.J. Lee, Y.S Hwang, H.C. Rim, "Two-Phase Bio-medical NE Recognition based on SVMs," In ACL'03 nlbio workshop, 2003.

[11] S.E. Robertson, S Walker, "Okapi/Keenbow at TREC-8," In The Eighth Text REtrieval Conference, 2000

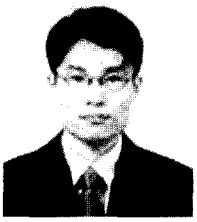
송 영 인

2001 고려대학교 컴퓨터학과 학사
2003 고려대학교 컴퓨터학과 석사
2003~현재 고려대학교 컴퓨터학과 박사 과정
관심분야 : 정보검색, 자연어처리, 생물정보학
E-mail : sprabbit@nlp.korea.ac.kr

김 상 범

1998 고려대학교 컴퓨터학과 학사
2000 고려대학교 컴퓨터학과 석사
2000~현재 고려대학교 컴퓨터학과 박사 과정
관심분야 : 정보검색, 자연어처리, 기계학습
E-mail : sbkim@nlp.korea.ac.kr

한 경 수



1998 고려대학교 컴퓨터학과 학사
2000 고려대학교 컴퓨터학과 석사
2000~현재 고려대학교 컴퓨터학과 박사 과정
관심분야 : 정보검색, 문서요약, 자연어처리
E-mail : kshan@nlp.korea.ac.kr

임 해 창

1990 Texas 주립대학 컴퓨터학과 박사
1991~현재 고려대학교 컴퓨터학과 교수
1998. 5~2000. 5 정보과학회 한국어정보처리연구회 운영위원장
2001~현재 ACM Transaction on Asian Language Information Processing Associate Editor
관심분야 : 자연어처리, 정보검색, 생물정보학
E-mail : rim@nlp.korea.ac.kr