# 한국어 정보검색연구를 위한 시험용 데이터 모음(KTSET) 개발

김 재군*, 김 영환*, 김 성혁**
*한국통신 S/W연구소 인공지능팀
**숙명여대 문헌정보학과

# Development of the Data Collection(KTSET) for Korean Information Retrieval Studies

Jaigun Kim*, Youngwhan Kim*, Sunghyuk Kim**
*KT S/W Lab, AI team, **SookMyung Univ

## 요약

정보검색분야의 여러 기술들을 연구하고 이 결과들을 실험 평가하기 위해서는 모든 연구자들이 공동으로 사용할 수 있는 시험용 데이터 모음(Test Data Collection)이 필요하다. 외국에서는 이미 오래전부터 각 분야별 시험용 데이터 모음들을 준비하여 검색시스템의 개발 및 객관적인 성능평가에 이용하여 왔는데 국내에서는 아직까지 이러한 시험용 데이터 모음이 개발되지 못한 실정이다.

본 연구는 한국어 정보검색 기술연구 활성화에 기여하기 위하여 한국어 정보검색 기술 연구결과의 성능평가에 공동으로 활용할 수 있는 국내 최초의 시험용 데이터 모음인 KTSET을 개발하였다. KTSET은 정보과학회와 정보관리학회지의 논문지 및 학술대회 논문집으로부터 추출된 1,053개의 논문과 이를 검색대상으로 한 50개의 자연어질의어로 구성되었으며 대상문서들과 질의어 각각에 대한 색인결과와 질의어와 대상문서들간의 적합도 정보를 제공한다.

## 1. 서론

사회의 발전에 따라 정보가 다양해짐으로써 필요한 정보를 신속, 정확하게 검색해 주는 정보시스템이 필요하게 되었다. 외국에서는 이미 1950년대부터 정보검색시스템을 만들어 사용해 왔으며 우리나라에서도 1980년대에 들어서부터 정보검색시스템들이 개발되어 여러 분야에 활용되기 시작했다. 최근에는 S/W, H/W기술이 놀라울 정도로 발전하였으며 이들이 통신기술과 결합되어 아주 다양한 정보서비스 시스템들이 생활 곳곳에 활용되고 있다. 따라서 정보검색 기술이 독자적인 정보검색시스템 구축뿐만 아니라 다양한 형태의 정보서비스 시스템 구축에 중요 핵심기술로 필요하게 되었다.

1970년대 미국을 중심으로 각 분야별 시험용 데이터 모음(Test Data Collection)들이 개발되어 정보검색 연구의 여러분야에 활용되어 새로운 방

법론 개발이나 이들의 객관적인 성능평가에 크게 기여하였다. 결국 이들 데이터 모음들을 이용한 실험 결과로부터 현재의 기술 수준과 필요한 기술의 개발에 대한 판단의 근거를 제공받았다. 하지만 이들 데이터 모음은 한국어 정보검색 연구에는 직접적인 도움이 되지 못하며 이를 위한 별도의 데이터 모음이 필요하지만 국내에는 아직까지 개발된 데이터 모음이 없는 실정이다.

본 연구는 시험용 데이터 모음(KTSET)을 개발하여 한국어 정보검색분야를 연구하는 모든 관련자들이 서로 공유하며 각자의 목적에 맞게 사용하여 정보검색 기술연구의 활성화를 도모하자는 것을 목적으로하고 있다. 개발된 KTSET은 크게 문서모음, 질의어 모음 그리고 질의어와 문서간의 적합도 정보 부분으로 구성되어 있다. 문서모음은 총 1,053건의 논문을 대상으로 하고 있으며 각각은 국·영문 저자, 논문제목, 서지사항, 초록, 분류기호, 색인어등 14개 항목으로 구성되어 있다. 질의어 모음은 1,053개의 문서를 검색대상으로 하는 50개의 자연어 질의어 샘플과 이들 각각에 해당하는 불리안 질의어로 구성되어 있다. 그리고 적합도 정보는 50개의 질의어 각각에 대해서 모든 문서를 대상으로 한 적합도가 주어져 있다.

## 2. 배경

### 2.1 정보검색시스팀의 평가

시험용 데이터 모음은 정보검색연구의 여러가지 평가에 주로 활용된다. 이러한 평가들 중에서 첫번째로 들 수 있는 것은 자동색인기에 대한 평가다. 데이터 모음에 있는 문서들을 대상으로 자동색인을 실시한 후에 데이터 모음내에 미리 준비된 각 문서에 대한 색인어들과 그 결과를 비교하면 된다.

두번째로는 자연어질의 처리기의 성능 평가이다. 데이터 모음에 있는 각 자연어 질의어에 대해 얼마나 정확하게 불리안 질의어로 변환하는가에 대한 평가이다. 이 경우에도 질의 처리기의 변환 결과와 데이터 모음내에 미리 준비된 각 자연어 질의에 대응된 불리안 질의어와 비교하면 된다.

세째로는 검색기에 대한 성능 평가다. 검색에 대한 성능 평가는 검색모델이 무엇이냐에 따라 성능 평가 방법과 기준이 달라지지만 모델 종류와 상관없이 필요한 정보는 적합성이다. 적합성은 문헌이 나타내려는 속성과 그것을 이해하려는 사용자간에 얼마나 효과적인 정보전달이 발생했는지를 나타내는 척도로서 내용상의 일치정도를 나타낸다. 이 적합성 정보는 전문가들에 의해 판단된 데이터 모음내의 각 질의어와 각 문서들간의 적합도로 표현되며 주로 0(완전 부적합)에서 1(완전 적합)사이의 값으로 표현된다. 만약 Boolean모델을 사용했을 때는 대상 문서가 적합한 부류와 부적합한 부류 2가지로 구분되며 이 때는 적합도의 적당한 경계치를 기준으로 양분하면 된다. 이때 사용되는 성능 평가기준은 정확률(precision)과 재현율(recall)이다. [1] 만약 Vector Model, FuzzySet Model 또는 Extended Boolean Model

과 같이 질의어와 문서간의 적합도를 계산하여 검색결과를 순서화 알고리즘(Ranking Algorithm)에 의해 제공하는 경우에는 적합도 정보를 그대로 사용하여 데이터 모음내의 순서와 검색결과의 순서와의 상관관계(Spearman Correlation Coefficients)를 사용하면 된다. [2]

이러한 세가지 부류의 평가외에도 하부저장구조의 효율성, 문서부류(Document Classification), Relevance Feedback, 시소러스 자동구축등의 여러가지 연구에 대한 실험 평가에도 활용할 수 있다.

그 동안 국내에서는 이와같이 공동으로 사용하여 상호 연구결과를 객관적으로 비교하고 분석할 수 있기 위한 한국어 시험용 데이터 모음이 개발 보급되지 않아서 각기 독자적으로 조그만 규모의 실험 데이터를 만들어 사용하여 왔기 때문에 정보검색 분야 연구발전에 상당한 저해요인으로 작용하였다. 이제부터는 개발된 KTSET이 널리 보급되어 정보검색 시스템의 개관적인 성능평가에 공동으로 활용되었으면 한다.

## 2.2 외국의 사례

70년대초 기계역학과 항공공학 분야의 1398개의 문서를 포함하는 최초의 시험용 데이터 모음(CRANFIELD TEST COLLECTION)이 미국에서 만들어져 정보검색 관련자들 사이에 널리 사용된 후 여러 분야의 시험용 데이터 모음들이 나오기 시작했다. 그 중 대표적인 예로서 정보과학 분야에서 프로그램 디버깅과 관련된 ADI 와 문헌정보학(Library Science)분야의 CISI와 LISA, 컴퓨터 분야의 CACM, 의학 분야의 NLM 과 MED, 우주항공 분야의 CRAN등을 볼 수 있다. 이들 중 대부분은 5개의 화일로 구성되어 있다. 제목, 저자, 본문, 색인어등의 원시 텍스트정보를 갖고있는 화일(*.all)과 자연어질의를 담은 화일(*.qry), 문서 벡터와 질의문 벡터로 구성된 화일(*.npl), 자연어 질의에 해당하는 변환된 불리안 질의어 셋과 이를 확장한 질의어 셋을 포함한 화일(*.bln) 그리고 마지막으로 질의와 문서들간의 적합성 정도를 보여주는 화일(*.rel)로 구성되어 있다. <표 1>은 대표적인 시험용 데이터 모음들에 포함되어 있는 문서의 수와 질의문의 수를 비교한 것이다.

| 데이터 모음 | 주 제 | 문 헌 수 | 질 의 수 |
|---|---|---|---|
| ADI | 정보과학 | 82 | 35 |
| CACM | 전산학 | 3200 | 64 |
| CISI | 문헌정보학 | 1460 | 76 |
| CRAN | 항공학 | 1400 | 225 |
| LISA | 문헌정보학 | 6004 | 35 |
| MED | 의학 | 1033 | 30 |
| NLM | 의학 | 3078 | 155 |
| NPL | 전자공학 | 11429 | 100 |
| TIME | 일반 사설 | 423 | 83 |

<표 1> 대표적 시험용 데이터 모음들의 문서 및 질의어 수 비교

## 3. KTSET 개발

위 <표 1>에서 보듯이 절반 이상의 시험용 데이터 모음이 1500개 이하의 문서로 구성되었으며 질의어의 갯수도 100개 이하인 것들이 대부분이다. KTSET 개발에 있어서도 이를 참작하여 문서와 질의어의 수를 각각 1,053개와 50개로 결정했다. 원시 문서화일과 자연어 및 불리안 질의어 모음 그리고 적합성 판정 결과 외에 문서의 분류를 위해 CRCS(Computing Review Classfication Strucure)분류표를 전기통신용어사전을[4] 기준으로 번역하여 이를 기준으로 입력 문서들을 분류하였다. 참고로 CRCS는 Computing Reviews 잡지내에서 사용된 분류표로 전산분야 문서에 관해서는 상세히 분류하고 있다. 다음은 각각의 구성 화일에 대한 설명이다.

### 3.1 한국어 문서 모음 구성

<id> 1038
<title> 요약 결집에 근거한 2단계 Signature 화일 방법
<author> 이준호, 장재우, 이윤준
<affiliation> 한국과학기술원 전산학과
<language> 한국어
<journal> 정보과학회논문지
<issn> 0258-9125
<year> 1989
<volume> 16
<number> 5
<pages> 445-455
<abstract> 본 논문에서는 많은 양의 데이타에 대한 다중-키 접근을 효율적으로 지원하는 새로운 signature 화일 방법을 제안한다. 제안하는 방법은 요약 결집과 단어 분별력 개념을 이용하여 Sacks-Davis가 제안한 2단계 접근 방법을 개선한다. 고분별력을 갖는 단어들에 대하여 역 화일로써 별도의 효율적인 접근 방법을 구성하고, 이를 요약결집에 이용함으로써 보다 좋은 검색 성능을 얻는다.
<etitle> Two Level Signature File Method Based on Signature Clustering
<eauthor> Joon Ho Lee, Jae Woo Chang, Yoon Joon Lee
<eabstract> In this paper, we propose a new signature file method which provides multikey accesses to a large amount of data. The proposed method is based on signature clustering and term discrimination so that we may improve the two-level signature file method designed by Sacks-Davis et al. We can achieve better retrieval performance by two techniques; constructing a separate and efficient access method (inverted file) for terms with high discriminatory power, and clustering similar signatures on the basis of these terms.
<classification> H.2.2.1 D.4.3.1
<keywords> Signature 화일 방법, 다중-키 접근, 요약 결집, 단어 분별력 Sacks-Davis, 2단계 접근 방법
<notes>

<표 2> 문서모음 화일 구성

대상문서의 수집을 위해 우선 범위를 전산분야의 한국정보과학회(KISS, Korea Information Science Society)와 정보학분야의 한국정보관리학회(KSIM, Korean Society of Information Management)의 논문지를 선택하였으나 문서의 갯수가 519개로 예상치 이하였기때문에 한국정보과학회 93년 춘

계·추계 학술발표대회 논문집에서 534개를 추가하여 총 1,053개를 선정하였다. 논문의 전체 내용대신에 초록만으로 그 내용을 대표하였고 고유번호, 국·영문 제목, 국·영문 저자, 영문 초록, 소속, 게재지, 년도, 페이지등 게재지에 기록된 기타자료와 함께 색인어휘들과 분류표에 의한 분류번호도 저장되었다. 수작업에 의한 색인작업은 참여한 주제전문가 4명이 전체를 분담한 후에 자연어 형태로 먼저 대상 색인어를 뽑아 토론과 협의를 거쳐 가장 적절한 10여개의 색인어를 선정하였다. 문서분류도 복수분류를 허용하여 평균 2,3개의 분류번호가 주어졌다. <표 2>는 하나의 문서에 대한 문서 모음 화일의 구성 예를 보인 것이다.

### 3.2 질의어 모음 구성

준비된 문서 데이터에 대한 적합성 검증을 위한 50개의 자연어 질의어(KTSET.nlq)가 작성되었다. 이들은 실제 사용자의 정보검색 요구에 의해 작성된 것이 아니라 문서 데이터의 분류 분포에 의해 어느 특정 부류의 문서만 반복해서 조회되지 않게 분포에 맞추어 질의어를 조절하여 작성하였다. 각 질의어도 내용이 너무 일반적이거나 특정적이지 않게 2~4개의 주제 탐색어가 추출될 수 있게 하였다. 또는 각각 질의어에 대해 주제 탐색어를 논리연산자와 자연어 질의에서 표현하고자 하는 의미에 맞게 결합하여 불리안 질의어 모음(KTSET.bq1)을 만들었다.

```
nlq 31 = 데이타베이스의 질의처리기에 관한 논문
nlq 32 = 하이퍼텍스트 또는 하이퍼미디어 시스템에서의 항해 기법에 관한 연구
nlq 33 = 인공지능 기법을 이용한 전문가 시스템의 응용에 관한 연구
nlq 34 = 자연언어처리중 한국어 형태소 해석에 관한 연구
nlq 35 = 결함허용시스템에 관한 연구

        <자연어 질의어 샘플>

'bq31 = '데이타베이스' and '질의처리기
'bq32 = ('하이퍼텍스트 or '하이퍼미디어') and '항해기법'
bq33 = '인공지능' and '전문가시스템' and '응용'
bq34 = '자연언어처리' and '한국어' and '형태소해석'
bq35 = '결함허용시스템'

        <불리안 질의어 셋 1 샘플>

bq31 = ('데이타베이스' or 'Database' or 'DB')
       and ('질의처리기' or 'Query Processor')
bq32 = ('하이퍼텍스트' or 'Hypertext' or '하이퍼미디어' or 'Hypermedia)
       and ('항해기법' or 'Navigation')
bq33 = ('인공지능' or 'Artificial Intelligence' or 'AI') and
       ('전문가시스템' or 'Expert System') and ('응용' or 'Application')
bq34 = ('자연언어처리' or 'Natural Language Processing' or 'NLP') and
       ('한국어' or 'Korean')and ('형태소해석' or 'Morphological Analysis')
bq35 = '결함허용시스템' or 'Fault Tolerent System'

        <불리안 질의어 셋 2 샘플>
```

<표 3> 질의어 모음 예

또한 불리안 질의어의 각 주제 탐색어를 동의어 및 관련어 또는 영어표기로 확장한 확장 불리안 질의어 모음(KTSET.bq2)도 준비하였다. <표 2>는 각 질의어 모음에서 50개 질의어 중 31번째부터 각각 다섯개의 질의어 샘플을 보여준다.

### 3.3 적합도정보 구성

앞서 언급한 바와 같이 적합성이란 문서가 나타내려는 속성과 그것을 이해하려는 사용자간의 정보전달 관계를 표현하는 척도로서 주관적인 판단에 의해 이루어지는 것이나 여기에서는 네 명의 전문가가 각 질의어에 대한 적합문서를 선정한 후, 네 명 모두의 의견이 적합하다고 판정한 경우 적합성 척도를 1로하고, 3명이 일치한 경우 0.75, 2명는 0.5, 1명은 0.25인 값을 부여하여 객관적인 수치의 적합성을 표현하였다. 전체적으로 3명 이상의 일치율이 93%로 대부분을 차지하였다. 질의어에 대한 적합문서 및 그 적합정도의 표시는 <표 4>와 같이 각 질의어 번호에 대하여 해당하는 적합문서 번호와 적합도 값으로 구성되어 있다.

| 질의어 | 문 서 | 적합도 |
|---|---|---|
| 27 | 0987 | 1.00 |
| 27 | 0988 | 1.00 |
| 27 | 0994 | 1.00 |
| 28 | 0958 | 1.00 |
| 28 | 0962 | 1.00 |
| 28 | 0974 | 1.00 |
| 28 | 0985 | 1.00 |
| 29 | 0948 | 1.00 |
| 29 | 0954 | 1.00 |
| 29 | 0962 | 0.75 |
| 29 | 0971 | 1.00 |
| 29 | 0985 | 0.75 |
| 30 | 0996 | 1.00 |
| 30 | 0997 | 1.00 |

<표 4> 적합도정보 구성 예

## 4. KTSET 분석

KTSET은 1985년 이후의 한국정보과학회 논문지내의 401개 논문과 창간호 이후의 한국정보관리학회 논문지 116개 논문을 비롯해서 한국정보과학회 93년도 춘,추계 학술발표대회 논문집에서 발췌한 534개 논문으로 구성되어 있다. 이들 1,053개의 논문에서 표제를 제외한 초록으로부터 초록내 어절수와 초록이 포함한 색인어에 대한 분석 결과를 <표 5>에 정리하였다. 또한 색인어에 대해 복합어 구성분포와 한.영 혼합분포도 표시하였다. <표 5>에 나타난 바와 같이 초록당 평균 약 11개의 색인어를 포함하고 이들 색인어는 단일어에 비해 2어절로 구성된 복합어가 더 많이 사용되었음을 알수있다.

(단위 : 어절수)

| | 어절 | 색인어 |
|---|---|---|
| 총 수 | 76,756 | 11,305 |
| 초록당 평균 | 73 | 10.74 |
| 초록내 최소 | 9 | 4 |
| 초록내 최대 | 176 | 29 |

복합어 구성분포

| | 평균수 |
|---|---|
| 단일어 | 3.20 |
| 2어절 복합어 | 4.55 |
| 3어절 복합어 | 2.17 |
| 4어절 복합어 | 0.82 |

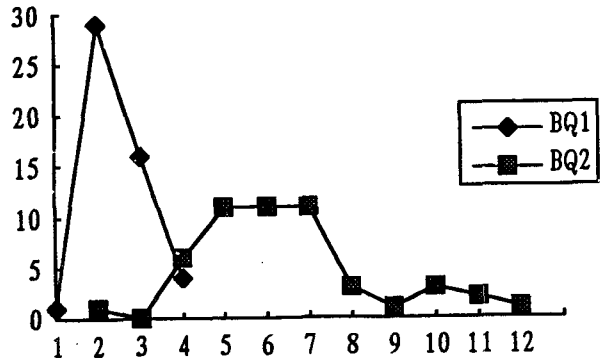
한.영 혼합 분포

| | 평균수 |
|---|---|
| 한글 | 7.58 |
| 영어 | 2.35 |
| 한영 혼합 | 0.81 |

<표 5> KTSET내 1,053개 초록의 어절 및 색인어 분석

영문 색인어도 상당히 높은 비율로 사용되었고 '바다 DBMS'나 'MIN 연산'과 같은 영·한 혼용 색인어도 전체적으로 800개 이상 사용되었음을 보여주고 있다.

50개의 자연언어 질의에 대해 두가지 형태의 불리안 질의가 만들어졌는데 기본개념(KTSET.BQ1)과 확장개념(KTSET.BQ2)을 각각 포함하고 있다. 확장 불리안 질의어는 기본개념의 유사어나 관련어 영문표기, 약어등을 포함한다. 50개의 기본 개념 불리안 질의어가 총123개의 어절로 구성되어 질의어당 평균 2.46개인 반면 확장개념의 질의어는 총 319개의 어절로 구성되었다. 이는 평균 한 어절이 2-3개의 개념으로 확장되었음을 보여준다. <표 6>은 이 두가지 형태의 질의어들의 어절수 분포를 나타낸 것이다.




<표 6> 불리안 질의어 셋의 어절분포

이러한 질의어에 대한 문서의 적합성 판별은 객관성을 유지하기 위해서 네명의 주제 전문가가 각기 적합성 평가를 하여 서로의 일치도에 따라 그 통계치를 추출하였다. 50개의 질의에 대하여 총 711개의 적합 문서가 추출되어 질의어당 평균 14개 문서가 적합 판정되었다. 총 추출문서 중 84%가

네명 모두 일치함(1.00)을 보였고 세명이 일치한 경우(0.75)가 12%, 두명의 일치(0.50)가 3%로 이 세가지 경우가 전체중의 99% 이상을 보였다.

## 5. 결론

방대한 정보를 처리하여 필요한 정보를 신속, 정확하게 제공하여 주는 정보검색시스팀의 성능 평가와 정보검색 연구결과의 객관적인 평가를 위하여 외국에서는 이미 80년도 부터 시험용 데이터 모음을 준비하여 시스팀의 개발 및 객관적인 성능 평가에 이용하여 왔으나 국내에서는 아직 아무런 데이터 모음이 없는 상태에서 KTSET의 개발은 한국어 정보처리연구 활성화에 큰 도움을 주리라 생각한다.

한국 정보과학회 논문지 및 학술발표회논문집(Proceeding) 그리고 한국정보관리학회 논문지로부터 1,053개의 논문을 선정해 문서 데이터를 준비하였다. 자동색인기의 성능시험을 위해 각 문서은 네명의 주제 전문가에 의해 각각 수동색인이 되었다. 번역된 CRCS분류표에 의해 모든 문서들을 분류하여 분류된 각 그룹들이 공평하게 검색될 수 있도록 50개의 자연어 질의어들을 만들었다. 불리안 질의 변환기를 테스트할 수 있도록 두가지 형태의 불리안질의어 셋도 생성하였는데 첫번째는 자연언어로부터 기본개념만 추출하여 그 사이에 논리연산자를 결합시켜 만들었고 두번째는 이 기본개념들을 유사어나 관련어등으로 확장시키었다. 마지막으로 이러한 50개의 질의어에 대해 각 문서의 적합성 테스트를 하여 각 문서와의 적합도를 구하였다. 결과를 표현하기위한 양식은 전산분야 테스트 데이터 모음인 CACM의 것을 대부분 수용하였다.

이러한 KTSET은 현재 한국통신에서 무상 보급중으로 인터넷 상에서 '147.6.2.181' 주소로 'Anonymous FTP'를 하여 디렉토리 '/pub/KTSET' 에서 필요한 화일들을 찾을 수 있다. 앞으로 관련분야 연구원의 많은 협조로 데이터 건수의 확장 및 주제분야의 확장이 이루어져 한국어 정보검색분야의 시험용 데이터 모음의 표준으로 발전하여 널리 활용되기를 기대한다.

## 참고문헌

[1] G Salton, J McGill, Introduction to Modern Information Retrieval, McGraw-Hill 1983

[2] Y.W Kim and J.H Kim, " A Model of Knowledge Based Information Retrieval with Hierarchical-Concept Graph," Journal of Documentation, Vol 46, No 2, June 1990, pp 113-136

[3] William B Frakes and R Baeza-Yates, 'Information Retrieval:Data structure & Algorithms' Prentice Hall, 1992

[4] 전기통신용어사전 , 한국전자통신연구소, 1985