

정보 검색 시스템 평가를 위한 균형 테스트 컬렉션 구축

Construction of a Balanced Test Collection for Evaluation of Information Retrieval Systems

맹성현(Sung-Hyon Myaeng)*, 이석훈(Suk-Hoon Lee)**, 이준호(Joon-Ho Lee)***,
이용봉(Eung-Bong Lee)****, 송사광(Sa-Kwang Song)*****

목 차

- | | |
|------------|--------------|
| 1 서 론 | 5 적합성 판정 |
| 2 문헌 집합 | 6 테스트 컬렉션 분석 |
| 3 질의 집합 | 7 결 론 |
| 4 후보 문헌 생성 | |

초 록

검색 시스템들의 평가를 위해 국내에서도 테스트 컬렉션에 관한 여러 연구가 진행되어왔다. 그러나 그 규모나 대상 분야가 편중되어 있고 질의 및 문헌 특성의 균형 등에 대한 고려가 반영되어 있지 않아 평가 결과를 객관화하기는 사실상 어려운 실정이다. 본 논문에서는 분야별, 사용자별 균형을 고려한 대규모 테스트 컬렉션인 HANTEC에 대해 기술한다. HANTEC 테스트 컬렉션은 총 12만 건의 문헌집합으로 구성되었는데 일반, 사회과학, 과학기술 각 분야별 4만 건 씩으로 특정 분야에 편중되지 않도록 하였고 질의집합도 각 분야별 10개씩 30개로 구성하였다.

ABSTRACT

There has been some research in Korea on test collections for evaluation of information retrieval (IR) systems. The test collections constructed as an outcome from the research have provided a starting point and opportunities to test Korean IR systems in an objective manner. However, they are well short of the standard practice in the broader IR community in that they are small in their size and usually unbalanced in terms of the characteristics of the documents and the queries (such as the subject domains). In this article, we describe our research effort to alleviate this problem and the resulting test collection, called HANTEC (Hangul TEst Collection). HANTEC is balanced in terms of the subject domains, document lengths, and user types, and currently consists of 120,000 documents divided into three groups: general area, social science area and science/technology area. The 30 queries in the collection are grouped into the same three areas in one dimension and into three distinct user groups in the other dimension.

키워드: 테스트 컬렉션, 정보 검색 시스템 평가, HANTEC

* 충남대학교 컴퓨터과학과 교수
** 충남대학교 통계학과 교수
*** 숭실대학교 컴퓨터학부 조교수
**** 충남대학교 문헌정보학과 조교수
***** 전자통신연구원
■ 논문 접수일 : 1999년 6월 1일

1 서 론

인터넷의 보편화로 인한 방대한 정보의 유통으로 인해 정보검색 시스템의 중요성은 날로 중요해지고 있고 이와 연관된 응용 시스템도 속속 개발되고 있다. 정보검색 시스템은 대용량의 텍스트 정보로부터 사용자가 필요로 하는 정보 즉 적합한 문헌을 검색해서 제공해 주는 시스템인데 그 성능의 평가에 있어 타 정보시스템(예:DBMS)과 다른 기준을 사용해 오고 있다. 검색 속도, 사용자 편의성 등의 일반적인 평가 기준이외에 검색 신뢰도¹⁾(effectiveness)를 사용하는데, 이는 검색시스템이 사용자 의도가 충분히 반영되지 않은 검색질의를 사용하여 문헌의 적합성(relevance)을 내부적으로 계산하여 검색결과를 결정하기 때문이다.

이러한 검색 신뢰도는 미국 및 유럽을 중심으로 발전해온 지난 30여년간의 정보검색 시스템 관련 연구에서 가장 중요한 지표로 사용되어왔고 현재도 실제 시스템의 성능을 평가하는 학술대회(예:ACM SIGIR 혹은 TREC)에서 지속적으로 사용되고 있다. 이러한 신뢰도에 기반을 둔 평가를 하기 위해서는 체계적으로 구축된 도구 즉 테스트 컬렉션(test collection)이 필요한데, 이는 주로 질의 집합, 대용량 문헌집합, 적합성 판단집합 등으로 이루어져 있다. 이러한 테스트 컬렉션은 정보검색 분야의 연구뿐만 아니라 사용자 집단이 상용화 시스템의 성능을 평가하여 적절한 시스템을 선택하는 데에도 매우 중요한 역할을 하므로 상용 시스템의 발전과 궁극적으로는 정보 유통에 있어서의 경쟁력 강화에도 필수적인 역할을 한다.

외국의 경우 검색엔진의 평가를 위해 소규모 테스트 컬렉션이 1960년 말부터 산발적으로 구

축되어 사용되어 오다가 정보검색 대상 데이터의 규모가 기하급수적으로 증가하면서 1990년 초부터 대용량 테스트 컬렉션의 구축이 시작되어 현재는 기가바이트 수준 이상의 실험데이터를 사용해서 검색 시스템을 평가하는 것이 이미 일반화되어 있다. 미국의 경우는 TREC(Text REtrieval Conference)에서 매년 컬렉션을 사용하여 실험실에서 개발된 시스템 뿐만 아니라 상용 시스템도 평가하여 그 결과를 발표하고 있다. 1997년에 발표한 TREC-6는 약 2백 50만 건의 문헌과 50개의 질의어를 포함하고 있다(Ellen M. Voorhees 외 1997). 이 컬렉션은 NIST(National Institute of Standards and Technology)가 주축이 되어 학계 전문가를 중심으로 구축되었고 매년 그 규모 및 종류를 증가 시켜가고 있다. 일본의 경우도 테스트 컬렉션의 중요성을 인식하여 정부 기관인 NACSIS(National Center for Information Systems)가 주관이 되어 대규모 컬렉션 구축 사업을 추진중이다. 또한 NTT Data Corporation에서 BMIR-J1과 BMIR-J2라는 컬렉션을 개발하였는데 BMIR-J1은 600건의 문헌과 60개의 질의로 구성되었고 BMIR-J2는 경제학 및 공학분야에서 5080건의 신문기사와 60개의 질의어를 포함하고 있다(Tsuyoshi Kitani 외 1998).

국내에서도 이러한 테스트 컬렉션에 대한 관심이 높아지고 있으나 아직 검색시스템의 성능평가에 있어 신뢰도에 근거한 객관적인 평가가 일반화되어 있는 상태는 아니다. 그 이유가 여러 가지가 있을 수 있으나 결정적인 이유로는 평가결과에 결정적인 영향을 주는 테스트 컬렉션의 구축이 아직 본격화 되어 있지 않다는 점을 들 수 있

1) 본 논문에서는 efficiency와의 혼동을 피하기 위하여 effectiveness를 신뢰도로 표기함.

다. 다행히 1994년에 정보과학회 논문을 대상으로한 KT-SET 테스트 컬렉션(김성혁 1994)이 구축되었는데 이는 30개의 매우 단순한 질문과 단지 1,053개의 학회 논문만을 포함하고 있다. 1995년에는 13,315건의 과기처 연구보고서를 대상으로 한 KRIST 컬렉션(이준호 외 1995)이 구축되었는데 생명과학, 의용전자공학, 기계공학 등을 주요 대상 분야로 하고 있다. 1996년에는 KT-SET을 확장하여 4,414건의 문헌과 50개의 자연어 및 불리언 질의를 포함한 KT-SET 2.0(K.S. Choi 외 1996)을 구축하여 컬렉션에 논문 뿐만 아니라 신문기사와 저널을 포함하여 확장하려는 시도가 있었다.

이러한 테스트 컬렉션을 이용하여 검색 시스템들의 기본적인 평가는 가능하나 그 규모가 작고 대상분야가 편중되어 있을 뿐만 아니라 질의 및 문헌 특성의 균형 등에 대한 고려가 반영되어 있지 않아 평가의 결과를 일반화하기는 사실상 어려운 실정이다. 영어권 문헌의 경우 대규모 컬렉션으로 시스템 평가가 이루어지면서 소규모 컬렉션으로 평가한 과거의 결과를 재심사하여야 하는 상황이 발생한 것을 볼 때, 한국어 문헌 정보검색의 경우에도 어느 정도 수준이상의 규모로 구축된 테스트 컬렉션을 사용하는 것이 시스템 혹은 관련기술의 정확한 평가를 위해 필수적이다. 또한 TREC이 시작된 후 정보검색 기술의 발전 속도가 눈에 띄게 발전되어 왔다는 점을 고려할 때 국내에서도 TREC에 버금가는 테스트 컬렉션을 구축하는 것은 기술개발 자체에 대한 투자만큼 중요하다.

본 논문에서는 한국어 문헌 검색 시스템 평가를 위한 테스트 컬렉션 구축에 대한 연구결과에 주안점을 두고 있는데, 특히 특정분야에 편중되지 않은 균형 잡힌 테스트 컬렉션의 개발을 위한

기반 구축 과정을 기술한다. 2장에서는 범용적이고 균형적인 문헌집합의 선정과정과 특성에 대해 기술하고 3장에서는 질의 집합의 선정 및 특성에 대해 설명한다. 선정된 문헌집합에서 후보 문헌을 추출하는 과정을 4장에서 언급하고 적합성 판정 방법 및 절차를 5장에서 설명한다. 7장에서는 본 연구에서 구축된 HANTEC(HANgul TEst Collection) 테스트 컬렉션의 통계적인 특성을 분석하고 8장에서 결론을 맺는다.

2 문헌 집합

정보 검색용 테스트 컬렉션은 일반적으로 문헌집합, 질의 집합 그리고 각 질의에 대한 적합 문헌 리스트로 구성된다. 이들 중 검색의 대상이 되는 문헌 집합은 테스트 컬렉션 구축에 있어서 가장 기본적인 요소이다. 문헌 집합의 구성에 있어서 고려해야 할 사항은 다양한 분야 및 크기의 문헌들로 문헌 집합을 구성해야 한다는 것이다. 정보검색 기술은 대개 통계적인 방법이나 언어처리 기술을 사용하는데, 이들은 모두 문헌 종류에 많은 영향을 받는다. 또한, 문헌과 질의의 유사도 계산에 핵심적 역할을 하는 가중치 기법들 (weighting schemes) 중 일부는 특정 크기의 문헌들에 높은 유사도를 부여하는 특성을 지니고 있기 때문에 가중치 기법의 성능 평가를 위해서라도 다양한 크기의 문헌들로 문헌 집합을 구성하는 것이 바람직하다.

따라서 본 연구에서는 개발하는 테스트 컬렉션의 문헌 집합을 일반, 사회과학, 과학 기술 분야에 속하는 120,000건(약 244MB)의 다양한 크기의 문헌들로 구성했고, 이를 각 분야별로 40,000건씩 균등하게 선정하여 특정 분야에 편

〈표 1〉 문헌 통계량

문헌 집합	문헌 개수	평균	표준 편차	최대 바이트 수	최소 바이트 수
한국 일보	22,000	1,742.0	952.6	11,764	52
웹 문헌(com, gov)	18,000	3,616.8	9,411.1	361,882	82
한국 경제신문	39,480	1,230.9	1,001.1	25,361	73
한국 여성개발원 계재 논문	110	53,021.9	35,876.1	288,714	9,024
경북 도의회 회의록	410	60,608.7	51,468.8	246,112	1,918
과기처 지원 연구보고서	10,000	1,689.2	739.6	5,532	206
해외 과학기술 동향	18,000	2,082.0	1,538.4	130,582	193
학술논문 서지사항	12,000	806.9	294.3	2,822	272

중되지 않고 고른 분포를 갖도록 하고 있다. 특히 문헌 집합은 짧게는 수십 바이트에서 길게는 수십만 바이트까지 매우 다양한 문헌들로 이루어져 있어서 검색 알고리즘의 강건성을 테스트할 수 있도록 하였다. 이러한 시도를 통해 특정 문헌 형태(예: 초록)에 국한시켜 개발된 검색 알고리즘이 특별히 좋은 평가를 받는 불공정성을 제거함으로써 평가결과가 현재의 정보검색 환경을 반영 할 수 있다.

각 분야별 문헌 집합의 구성 정보는 다음과 같다.

◆ 일반 종합 분야

-1994년에 발행된 한국일보 기사: 22,000건

-gov 확장자를 갖는 웹 페이지: 9,000 건

-com 확장자를 갖는 웹 페이지: 9,000 건

◆ 사회 과학 분야

-1994년에 발행된 한국 경제 신문 기사:

39,480 건

-한국 여성개발원이 발행한 정기간행물

여성연구에 게재된 논문: 110 건

-경북 도의회 회의록: 410 건

◆ 과학 기술 분야

-과기처지원 연구보고서: 10,000 건

-연구개발 정보센터에서 발간한 해외과학기술 동향: 18,000 건

-논문 서지 사항: 12,000 건

문헌 집합 중 웹 문헌이나 신문기사의 경우는 중복된 문헌들이 다수 존재하므로 이를 문헌을 각 문헌간의 유사도를 기준으로 제거하였다. 이 중복문헌 제거에서는 SMART시스템을 한글화한 검색 시스템이 사용되었고 이때 사용된 문헌과 질의의 가중치 부여 기법은 atc-atc²(이준호 외 1995)를 사용하였다. 본 연구에서 사용된 120,000 건의 문헌집합은 이러한 중복 문헌 제거 과정을 거친 후 구축된 것이다.

표 1은 각 문헌집합별 통계적 특징을 정리한 것으로 각 문헌 집합별로 문헌 개수와 문헌크기

```

<num> 01
<title> 월드컵 축구 유치
<desc> 한국의 2002년 월드컵 축구 유치 활동 내용
<narr> 한국의 2002년 월드컵 축구 유치를 위한 국내외적인 활동이나 한국개최에 대한 회원국의 반응을 포함한 정보는?
<query> 2002년 월드컵 축구 피파 FIFA 회원국 한국 개최 주최 유치전략 홍보 활동
<num> 10
<title> 여성의 정치 참여
<desc> 여성의 정치참여를 위한 정책 결정 과정
<narr> 여성의 정치참여를 위한 정책을 다루는 기관과 정책 내용을 담은 문헌은?
<query> 여성 여당 야당 정치참여 복지 지위 정책 법

```

〈그림 1〉 질의의 구성

를 보여 주는데, 문헌크기는 바이트 단위로 평균, 분산, 최대값, 최소값 등의 통계정보를 기술하였다. 이 통계 수치는 문헌별 바이트 수에 따라 계산되어진 것으로 각 집단의 분포를 대표하는 값들이다. 이들은 각 문헌집합의 문헌별 텍스트 길이의 분포를 예측할 수 있게 하므로 검색 및 평가 결과를 분석하는데 유용한 정보를 제공한다.

3 질의 집합

3.1 질의 형태

질의의 형태는 TREC-6의 Topic Statement 형태를 따르되 <query> 항목이 추가되어 5개의 부분으로 이루어져 있다. TREC의 경우 질의 형태는 TREC-1부터 TREC-4까지 단순해져 가는 경향이 있었으나 TREC-4 질의의 모호성 문제가 제기되어 TREC-5와 TREC-6에서는 <num>, <title>, <desc>, <narr>의 4개 태그를 사용해오고 있다(Ellen M. Voorhees 외 1997).

본 연구에서 사용된 질의는 <num>, <title>, <desc>, <narr>, <query>의 5개로 구성되며 각각 질의 번호, 질의 제목, 질의 설명, 질의 해설, 질

의 단어 리스트를 나타낸다. 질의 중에서 <title>과 <desc>는 실제 검색 시스템이 사용하여 내부 질의를 생성할 수 있도록 한 부분이고 <narr>은 적합문헌을 판별하는 기준을 기술한 것이다. 이 부분은 적합성 판단의 판정자가 검색된 문헌 집합 중에서 적절한 문헌을 판별하는 기준을 제공하는 것을 주 목적으로 하고 있으나 검색 시스템도 내부질의 생성에 사용할 수 있다. 이는 <title>과 <desc>만으로는 질의의 모호성 해소가 안되는 경우가 많아 적합성 평가 시에 평가자 간의 일관성이 없을 수 있기 때문이다. 마지막으로 <query>는 검색 시스템의 내부질의 생성을 도와주기 위한 부분으로 관련된 어절의 집합으로 구성되어 있으며 앞의 4개의 태그에 포함되지 않은 어절이라도 검색을 보충할 단어라면 <query>에 포함된다.

2) 질의어의 기중치 부여식이 주로 용어 빈도, 장서 빈도, 정규화로 나뉘는데 atc는 용어 빈도로 확장된 정규용어 빈도(Augmented normalized term frequency: $0.5 + 0.5 \frac{tf}{maxtf}$)를 사용하고, 장서 빈도로는 역문서 빈도가 곱해진 용어 빈도($\ln \frac{N}{n}$)를 사용하며, 정규화는 코사인 정규화(Cosine normalization: $\frac{1}{\sqrt{\sum_{vector} w_i^2}}$)를 사용한다.

그림 1은 본 연구에 사용된 질의의 예이다.

3.2 질의 생성 기준

질의 생성은 분야별, 사용자별로 분류하여 최종적으로 30개를 생성하였다. 표 2와 같이 질의는 분야별로 일반, 과학기술, 사회과학 3분야로 나뉘어지는데 각 분야별 질의어의 비율은 1:1:1로 배열되었고 사용자별로는 일반인, 전문가, 청소년 3그룹으로 나뉘며 이들 각 분야별 분포는 4:3:3이 되도록 구성되었다. 이와 같이 질의 집합 생성 과정에서 분야별, 사용자별 균형을 고려함으로써 질의가 특정 분야 또는 사용자 그룹에 편중됨으로써 발생할 수 있는 테스트 컬렉션의 불균형 현상을 없애고자 하였다.

질의 주제설정 및 구성 시에 고려한 내용은 다음과 같다. 먼저 일반 종합 분야는 일반적 주제를 위하여 일간신문의 국내외 10대뉴스, 분야별(중소기업, 연예, 문화, 스포츠, 경제) 10대뉴스 분야의 문헌을 분석하여 질의를 추출하였다. 청소년들의 주제탐색은 주로 문화(연예, 스포츠)와 과학분야의 문헌에서 수행되었고 합당한 문헌이 몇 개 없을 수 있는 특수한 질의를 의도적으로 생성, 포함시켰다. 다음으로 과학기술분야는 KRIST 테스트 컬렉션(이준호 외 1995)에 사용되었던 질의 중 전문적인 것과 신문기사 등에서도 나올 수 있

는 류(類)의 잘 알려진 전문적인 질의를 추출하여 각 질의의 예상되는 결과문헌을 고려하여 질의를 선택하였다. 학술논문 서지사항과 해외기술동향에서는 임의추출방식으로 추출한 문헌을 바탕으로 질의를 추출하였다. 마지막으로 웹 문헌, 여성 개발원 게재 논문 그리고 경북도의회 회의록에서도 임의적으로 자료를 탐색하여 각 자료의 특성을 잘 나타내는 문헌을 기준으로 질의 설명과 질의 해설을 구성하였다. 이와 같은 방법으로 각 문헌집합에서 추출된 질의는 총 62개이었다.

3.3 질의 집합 선정

질의 집합은 1단계로 62개를 위에서 기술한 방법으로 생성한 후 각 질의의 “품질”을 평가하여 최종 30개의 질의를 선정하였다. 이러한 여과과정을 거치는 이유는 질의의 성격이나 분야에 따라 적합한 문헌의 수가 다르므로 그 균형을 맞추는 데 있다. 예를 들어 컬렉션에 나타나는 적합한 문헌이 극소수일 것으로 예측되는 경우나 적합한 문헌이 너무 많아 검색기의 비교평가에 도움이 안 되는 질의는 제외하였다. 최종적으로 선정된 질의 집합은 위에서 설명한 사용자의 관심 분야별 균형 기준도 맞추었다. 비록 각 분야별 질의 개수가 10개로 제한되었지만 이렇게 함으로써 분야별 문헌특성에 따른 검색기의 편중정도를 측정

(표 2) 사용자 그룹별 질의 분포

사용자 \ 분야	일반 종합	과학 기술	사회 과학
사용자			
일반인	4	4	4
영역 전문가	3	3	3
청소년	3	3	3

할 수 있는 기반을 마련하였다.

최초에 작성된 62개의 질의는 각각 충남대 검색기를 통한 1차적인 검색 테스트에 사용되었는데, 이때 검색 결과 문헌수가 극소수인 경우나 너무 많아 검색기의 성능평가 비교에 도움이 되지 않는 극단적인 질의는 최종 질의의 집합에서 제외되었다. 이와 같은 과정을 거쳐 조정된 30개의 질의 집합은 표 2에 나타난 비율에 따라 일반, 사회과학, 과학기술, 각 분야별로 10개씩 분포되어 있고 이는 다시 사용자별로 일반인, 전문가, 청소년 각각 4,3,3개의 질의를 포함한다.

4 후보 문헌 생성

각각의 질의에 대한 적합 문헌 리스트의 생성을 위한 가장 확실한 방법은 각각의 질의에 대하여 테스트 컬렉션에 포함된 모든 문헌들을 사람이 읽고 적합성 여부를 판단하는 것이다. 그러나 이 방법은 문헌의 수가 많은 경우에 대단히 많은 시간을 요구하므로 현실적으로 거의 불가능하다.

보다 현실적인 방법으로 각각 다른 특성을 가진 다수의 검색 시스템을 사용하여 검색을 수행하고, 각각의 시스템에 의해 높은 순위를 부여 받은 문헌들에 대해서만 적합성 여부를 판단하는 방법이 제안되었다(Harman D. 1993). 이 방법의 주안점은 특성이 다른 다수 시스템들에 의해 검색된 문헌의 합집합은 컬렉션 내에 존재하는 거의 모든 적합문헌을 포함할 것이라는 가정이다. 사용자의 적합성 판단작업이 전체 컬렉션이 아닌 이 집합에 국한되므로 컬렉션이 큰 경우 현실적으로 가능한 방법으로 알려져 있다. 이 방법은 풀링 방법(pooling method)이라고 불리며, 테스트 컬렉션 구축 시 적합 문헌 리스트 생성에 효

과적인 방법으로 알려져 있다.

풀링 방법을 적용하기 위해서는 다수의 검색 시스템이 요구된다. 본 연구에서는 다양한 검색 결과를 생성하기 위해 검색기는 충남대와 숭실대에서 제공한 검색기를 사용하되 각 검색기의 색인 방식, 가중치 부여 방식, 적합성 피드백(relevance feedback)의 포함여부 등의 환경을 변화시켜 다양한 후보문헌 집합을 얻었다. 이렇게 여러가지 방식을 조합함으로써 총 38개의 검색결과를 생성하였는데, 비록 2개의 기본 시스템으로부터 모든 결과가 생성되었지만 각각은 매우 상이한 검색 시스템으로 작용하였다.

충남대 검색기를 사용한 경우 적합성 피드백을 사용하지 않은 1차 결과와 이를 사용한 2차 결과 각각 10의 후보문헌 집합을 생성하여 총 20개의 후보 문헌 집합을 얻었다. 먼저 1차 결과로서 색인 방식과 질의 구성 방법의 가능한 조합들을 이용하여 검색을 실시하였는데 색인 방법으로는 코퍼스 단어 패턴 기반의 TRIE 구조 자체 색인기(장동현 외 1996)를 이용한 방법과 한성대의 HAM형태소 분석기(강승식 1995)를 이용한 방법 2가지를 사용하였다. 검색모델은 벡터 공간모델을 이용하였다.

질의어 구성 방법은 질의어가 <title>, <desc>, <narr>, <query> 4개 태그로 나뉘어지므로 각 태그의 가능한 모든 조합(15개)을 사용하였다. 앞 장에서 기술한 바와 같이 각 질의의 타당성을 위하여 질의 초안을 사용하여 일차 검색을 실시한 후 검색된 문헌을 연구자가 직접 확인하는 과정을 거쳤다. 이 과정에서 너무 애매하거나 적절한 문헌이 검색될 확률이 매우 낮은 질의는 각각 질의 단어를 구체화시키거나 첨가하여 수정하였다. 앞의 두 색인방법과 조합함으로써 총30개의 검색 결과집합을 얻었다. 이들 결과집합은 문헌간 검

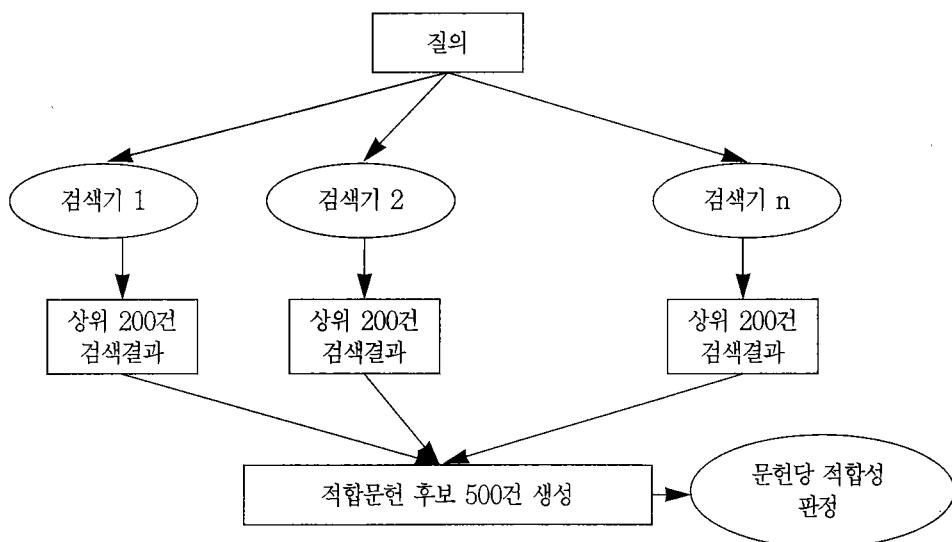
색결과의 유사 정도와 결과문헌 수에 따라 HAM 색인방법을 사용한 5개의 집합과 TRIE를 사용한 5개의 집합으로 압축하여 총 10개 집합을 얻었다. 이 때 사용된 질의어 조합은 {<title><desc><narr><query>} , {<title><desc><query>} , {<title><narr><query>} , {<desc><narr><query>} , {<narr><query>}의 5가지 방법이다.

2차 결과는 적합성 피드백을 이들 각 결과 집합에 적용하여 생성하였다. 적합성 피드백을 위해 1차 결과로써 얻어진 결과 리스트로부터 적합한 문헌을 4개 수동으로 선택하여 이 문헌을 검색 시 질의에 추가 시켰고 결과적으로 또 다른 10개의 검색 결과집합을 얻었다. 결국 1, 2차 모두 20개의 검색 집합을 얻을 수 있었는데 이들은 각각 검색결과의 상위 200개 문헌들로 이루어진 후보 문헌 집합이다.

승실대의 검색기는 코넬에서 개발된 SMART를 한글화한 검색기인데 마찬가지로 1차 검색은

적합성 피드백을 적용하지 않은 상태에서 수행하였고 적합성 피드백을 사용하여 2차검색을 실시하였다. 1차 결과는 bigram 색인법(이준호 1996)과 형태소 단위 색인법을 이용하여 색인을 하였고 각 색인방법에 가중치 부여 방법에 변화를 주는 방법(이준호 1995)을 사용했다. 결과 집합은 색인방법 2개와 가중치 부여 방법 3개의 조합으로 6개를 얻을 수 있었다. 2차 결과는 각 색인 방법에 추가로 적합성 피드백을 적용하였는데 확률 피드백 방법과 Rocchio 피드백 방법(Joon Ho Lee 1997)을 사용하여 12개 집합을 추가로 얻을 수 있어 총 18개의 집합을 생성하였다. 충남대 검색 결과와 마찬가지로 각 검색 결과 집합은 질의당 200개의 문헌으로 구성되었다.

충남대와 승실대 검색기를 통해 위와 같이 검색된 38개의 후보 문헌 집합은 풀링 과정을 통해 각 질의 당 500개의 최종 결과 문헌을 얻었다. 그럼 2는 후보문헌 생성과정을 나타내고 있다. 풀링하는 과정은 먼저 각 결과집합은 동일한 가



〈그림 2〉 후보문헌 생성을 통한 최종 적합성 판정

중치 값을 갖는다는 가정하에 1부터 38까지의 각 집합을 임의의 순서로 배열한 후 각 집합의 문헌을 각 집합에서의 랭크 순으로 추출하여 나간다. 이때 동일한 문헌이 이미 추출된 경우는 최종 결과집합에 추가하지 않으며 총 500개 문헌이 될 때까지 반복하여 실시한다.

5 적합성 판정

적합성 판정 과정은 각 질의 당 500건의 검색 문헌들에 대하여 질의 단위로 다음과 같은 과정을 거쳐 실시하였다.

임시 결과 집합을 평가하는 가평가 과정을 통해 평가자의 시각이 일관되도록 훈련시킴으로써 평가자 간의 시각의 차이를 최대한 줄이도록 하였다.

총 10명의 평가자를 2명씩 조를 구성하여 각 조가 6개 질의에 대한 문헌 총 3000개를 평가하되 평가자 2인은 상호 독립적으로 평가하도록 하였다. 즉 특정질의에 대해 하나의 문헌을 2인의 평가자에 의해 적합성이 판정되도록 하였다. 이때 각 질의에 대한 관점 차이의 극복을 위해 두 평가자 간에 질의에 대한 충분한 토의를 거친 후 상호 독립적인 평가작업을 실시하였다.

평가 방식은 '적합'과 '부적합'으로 나누는 종래의 방식에서 적합정도를 1(부적합), 2(약간적합), 3(다소적합), 4(적합), 5(매우적합)의 다섯 가지로 나누도록 하였다. 이러한 방식을 판정하는데 시간이 많이 걸리는 어려움을 갖고 있지만, 적합정도를 보다 섬세하게 나타낼 수 있다는 장점과 더불어 평가자간에 발생할 수 있는 관점이나 생각의 차이가 양극화 되는 위험을 막을 수 있는 보호 장치 역할을 하게 된다.

각 평가자의 평가 결과는 원시 결과로써 확보해 놓은 상태에서 두 평가자 간의 이견이 있는 문헌(최도 5점 중 3점 이상 차이가 나는 문헌)에 대해서는 상호 협의를 통해 조정하도록 하였다. 결국 조정하기 이전의 원시 평가 결과와 평가자 이견 조정 후의 수정 평가 결과를 모두 다 확보하였다. 원시 평가 결과와 수정 평가 결과는 테스트 컬렉션 평가 시에 데이터로 사용된다.

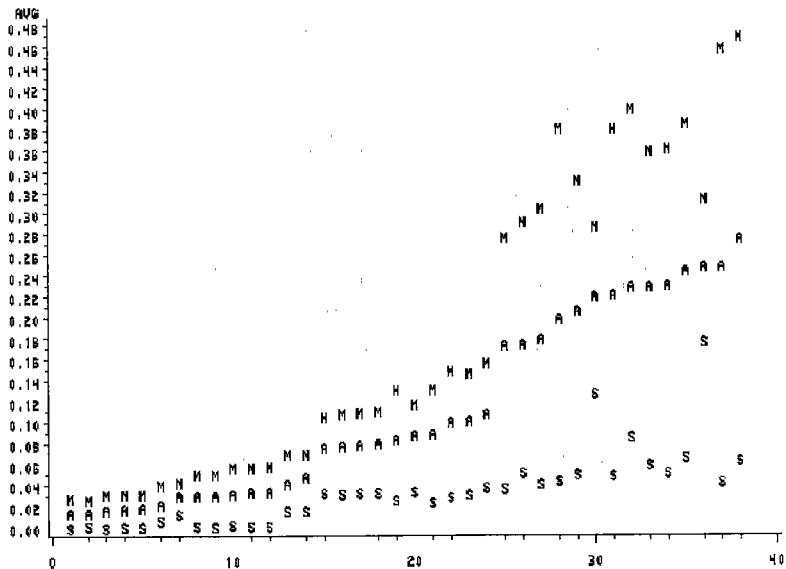
각 질의에 해당하는 문헌에 대한 적합성 평가 시, 평가자가 <narr>에서 사용된 템파이어들 간의 불리언 논리(AND, OR, NOT)에 특히 신중을 기하도록 지시하였다. 그럼 1의 질의의 구성을 보면 1번 질의의 <narr>문장에서 '~활동이나 한 국가최애~'라는 부분을 볼 수 있는데 이때 '~이나'는 OR 논리로 간주하여 두 조건 중 하나만을 만족하는 문헌을 적합한 문헌으로 평가하였다. 10번 질의의 <narr>문장에서는 '~기관과 정책~'이라는 부분에서 '~과'를 AND 논리로 간주하여 두 조건을 모두 만족시키는 문헌을 적합한 문헌으로 간주하였다.

6 테스트 컬렉션 분석

적합성 판정 과정을 거쳐 확보된 30개의 각 질의별 최종 적합문헌 정보를 이용해 테스트 컬렉션 평가 과정을 거쳤다. 표 3은 원시 적합판정 자료를 바탕으로 적합 기준의 차이에 따른 적합 문헌 개수를 나타낸 것이다. 평가자가 적합 정도를 5점 척도, 즉 1(부적합)부터 5(매우 적합)까지의 점수를 부여하는 방식을 사용했기 때문에 기술의 정확률과 재현율에 기반한 평가 방법을 사용하기 위해서는 각 문헌의 적합여부를 최종적으로 결정하기 위해 임계기준을 설정해야만 한다. 따라서

〈표 3〉 적합 기준의 차이에 따른 적합 문현 개수

질의	H2	H3	H4	H5	L2	L3	L4	L5
1	78	50	32	22	58	34	24	11
2	90	51	34	6	63	30	10	.
3	31	24	20	13	24	21	14	7
4	40	27	23	6	26	21	11	3
5	14
6	50	7	3	2	42	4	2	1
7	69	16	5	.	29	8	2	.
8	163	98	77	48	110	82	65	32
9	127	44	27	15	46	28	20	9
10	78	36	31	19	36	30	20	9
11	122	55	34	27	60	43	32	15
12	7
13	131	8	1	1	22	2	1	1
14	24	11	5	3	14	6	3	1
15	83	32	21	19	40	26	21	18
16	31	18	11	3	23	16	7	1
17	66	28	14	5	42	25	10	3
18	28	12	8	4	17	8	6	2
19	44	23	17	10	25	17	12	7
20	39	24	19	8	23	19	14	4
21	68	19	.	.	25	11	.	.
22	120	20	4	.	21	6	2	.
23	60	15	4	4	24	9	3	2
24	64	28	15	4	31	17	8	1
25	66	23	6	1	32	12	2	.
26	169	95	40	13	123	52	21	3
27	21	4	1	.	5	1	.	.
28	37	12	5	.	18	6	2	.
29	35	7	4	2	12	4	2	2
30	91	35	13	3	45	18	5	3
평균 (%)	13.64	5.87	3.51	2.07	7.40	3.97	2.45	1.29



〈그림 3〉 각 시스템별 정확률(Precision)

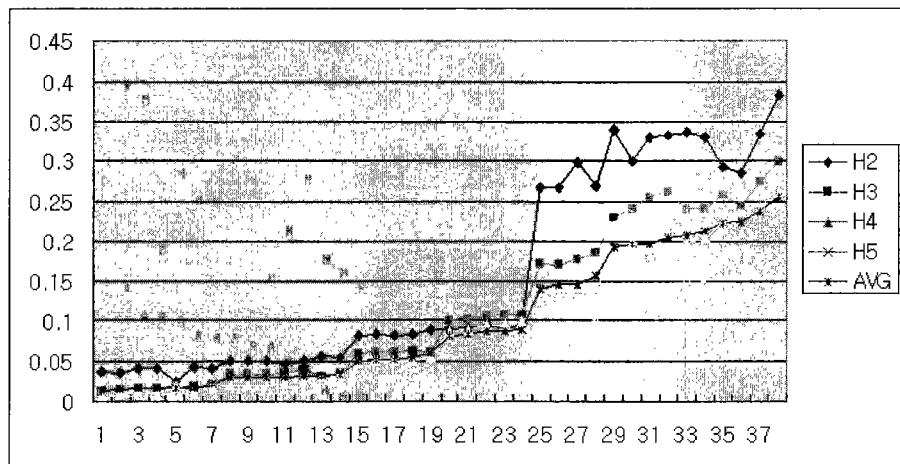
표 3에서 L2, L3, H2, H3 등의 숫자부분은 임계값을 의미한다. 즉 L2는 2이상의 평점을 얻은 문헌을 적합한 문헌으로 보았을 때의 적합문헌의 개수를 표현한 것이고, H2의 경우도 평가점수가 2이상인 문헌들을 적합문헌으로 결정한 것이다. 여기서 문자 L은 두 평가자가 부여한 점수 중 낮은 것을 그 문헌의 평점으로 간주한 것이고 문자 H는 반대로 높은 평점을 사용한 것이다. H3 열의 숫자는 두 평가자의 평가 점수 중 3이상인 문헌 개수를 의미하는 것이다. 숫자가 나타나 있지 않고 마침표로 표시된 셀은 해당하는 문헌의 개수가 0임을 의미한다. 표 3의 하단에 있는 평균은 30개 질의에 대해 각 기준별 적합 문헌수의 평균을 백분율로 나타낸 것이다.

테스트 컬렉션의 생성에 있어서 가장 중요하고 또한 논란의 여지가 많은 부분이 적합성 판정의 주관성 개입에 관한 논의이기 때문에 이 현상을 조사하기 위한 방안으로 표 3에 제시한 8개의 적

합성 판정 결과에 대하여 각 시스템이 30개의 질의를 통하여 반응한 8개의 평균 정확률(average precision)의 평균값(A), 최대값(M), 최소값(S)을 각각 구하여 그림 3으로 나타내었다. 평균 정확률이라 함은 각 기준별로 질의 30개 각각에 대한 평균 정확률을 합산하여 이를 산술 평균한 값을 의미한다. 이때 가로축은 38개의 시스템 중 평균이 가장 낮은 것부터 올림차순으로 정렬한 것이다.

그림 3의 결과에서 확인할 수 있는 것은 38개 시스템³⁾중에서 약 24개 시스템(주로 좌측에 위치한 시스템들)은 적합성 판정의 기준이 바뀌어도 정확률의 변화 폭에 큰 변화를 나타내지 않는 반면에 나머지 14개의 시스템(주로 우측에 위치한 시스템들)의 경우는 적합성 판정기준에 대단히

3) 엄밀히 말하면 다양한 기능 조합을 통해 생성된 38개의 후보 문헌 집합

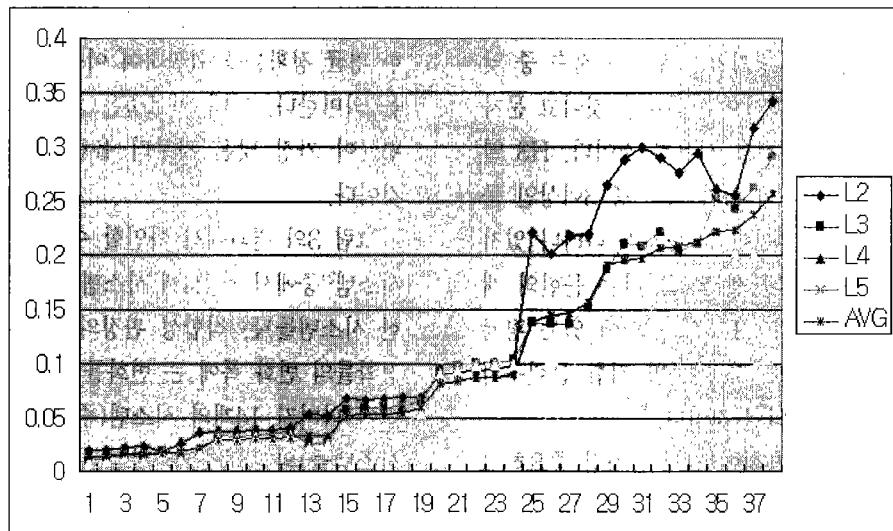


〈그림 4〉 평균 정확률과 H2, H3, H4, H5 기준 정확률과의 비교

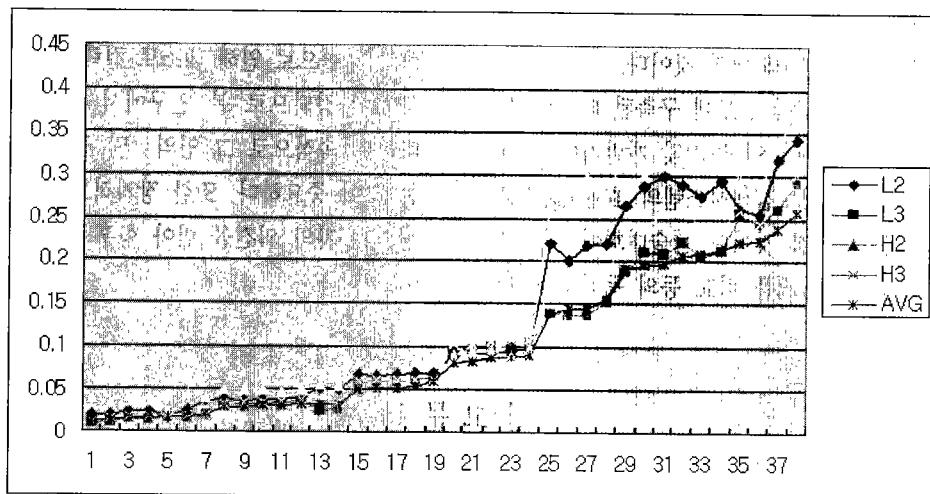
민감한 반응을 보이고 있는 것을 알 수 있다. 따라서 적합성 판정 기준에 대한 민감성은 테스트 컬렉션 자체의 특성이라기 보다는 각 시스템의 특성으로 이해될 수 있다고 판단된다.

한편 '적합'과 '부적합'의 경계가 시스템 평가에 미치는 영향을 판단하기 위하여 다음과 같은

실험을 하였다. 8개의 적합성 판정결과 각각을 바탕으로 하여 얻은 시스템 성능 순위의 상관관계를 Kendall의 상관계수(John Neter 외 1990)로 산출하였는데 평균 0.948의 높은 결과를 얻었다. 따라서 본 연구에서 구성한 테스트 컬렉션에서는 '적합'과 '부적합'의 경계를 어디에 두는가



〈그림 5〉 평균 정확률과 L2, L3, L4, L5 기준 정확률과의 비교



〈그림 6〉 평균 정확률과 L2, L3, H2, H3 기준의 정확률과 비교

에 따른 결정이 시스템의 성능을 순위로써 평가하는데 민감한 영향을 미치지 않는다는 결론을 내릴 수 있다.

그림 4는 평균 정확률 순으로 정렬된 38개 시스템 각각의 평균 정확률과 H2, H3, H4, H5 4개의 정확률을 겹은선 그래프로 표현한 것이다. 검색 시스템들의 랭크가 어떤 기준을 선택하느냐에 따라 다소 차이가 생기는 경우가 있으나 거의 대부분의 경우 시스템 간의 순위가 그대로 유지되는 것을 확인할 수 있다. 그림 5의 경우는 H대신 L에 적용한 결과인데 이 그래프도 그림 4와 유사한 모양을 나타내고 있다.

그림 6은 H2, H3와 L2, L3를 평균 적합도와 비교하기 위해 도시한 그래프이다. 이 그래프에 도시된 4개 기준 중에 다른 기준에 비해 상대적으로 평균 적합도 그래프와 유사한 시스템 순위를 갖는 기준을 구별하기란 쉽지 않다. 따라서 적합 여부를 5점 척도로 하였을 때 어떤 점을 경계로 적합, 부적합을 결정할 것인가는 확정적으로 결정하기 어렵다. 그러나 L2, L3 또는 H2, H3 기

준이 L4, L5, H4, H5 기준에 비해 시스템간의 성능비교를 더 뚜렷이 한다는 것을 확인할 수 있다.

7 결 론

본 연구에서는 문헌 집합의 선정에서부터 질의 집합의 선정에까지 특정 분야에 치우치지 않은 균형 잡힌 테스트 컬렉션을 구축하였다. 정보검색 분야가 지속적으로 발전하면서 보다 정확하고 객관적인 방법으로 시스템을 평가할 수 있는 환경에 대한 중요성은 더욱 높아질 것으로 판단된다. 정보검색 시스템의 응용 분야가 다양해지면서, 새로운 언어적, 구조적, 영역적 특성을 지닌 문헌 집합을 테스트 컬렉션에 추가하는 것이 필요할 것으로 전망된다. 추가적으로 단순한 문헌 검색 기능을 초월하여 문단 검색, 구조화 문헌 검색, 정보 추출, 정보 요약 등의 새로운 기능을 시험할 수 있는 테스트 컬렉션의 구축에 대한 사항도 향후 중요한 이슈로 등장할 것이다. 따라서 새

로운 문헌집합 추가 및 다양한 문헌 형태의 테스트 컬렉션 구축이 연구되어질 것이다.

본 연구를 통해 당해 연도에 구축된 테스트 컬렉션은 정보검색 분야 연구자 및 개발자에게 자유롭게 배포되어 정보검색시스템의 효율성 측정 목적으로 사용되어질 것이며 학술대회에서의 연구 결과 발표 혹은 제품 비교 등의 목적으로 활용되어 질 수 있다. 본 연구에서의 시간 및 예산 상의 제한으로 인해 테스트 컬렉션의 영역 및 규모가 한정되므로 본 연구에서 제안하는 계획에 따라 지속적으로 그 영역 및 규모를 확장 시키는 것이 중요하다. 또한 장기적으로 테스트 컬렉션 사용자의 의견을 받아 향후 구축에 반영시켜야 한다.

참 고 문 헌

- 강승식. 1995. 한국어 자동 색인을 위한 형태소 분석 기능, 『한국정보과학회 봄 학술대회 발표논문집 제22권, 제1호』, 929-932
- 김성혁 외. 1994. 자동색인기 성능시험을 위한 Test Set 개발, 『정보관리학회지』, 11(1), 82-101
- 이준호. 1995. 다중 가중치 기법을 이용한 검색 효과의 개선, 『정보관리학회지』, 12(2)
- 이준호. 1996. 한글 문서의 효과적인 검색을 위한 n-Gram 기반의 색인 방법, 『정보관리학회지』, 13(1)
- 이준호, 최광남, 한현숙, 김종원, 남성원. 1995. 정보검색을 위한 KRIST테스트 컬렉션의 개발, 『정보관리학회지』, 12(2), 225-232
- 장동현, 맹성현. 1996. 효율적인 색인어 추출을 위한 복합명사 분석방법, 『제8회 한글 및 한국어 정보처리 학술대회』.
- Choi K. S., Park Y. C., Kim J. K., Kim Y. W. 1996. Development of the Data Collection Ver. 2.0 for Korean Information Retrieval Studies (KTSET 2.0), Presented at The Workshop on Information Retrieval with Oriental Languages, June 28-29
- Harman, D. 1993. Overview of the 1st text retrieval conference, *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 36-48
- Kitani Tsuyoshi, Yasushi Ogawa, etc. 1998. "Lessons from BMIR-J2: A Test Collection for Japanese IR Systems", *SIGIR '98, Melbourne, Australia*.
- Lee Joon Ho. 1997. "Combining Multiple Evidence from Different Relevance Feedback Methods", *Proceedings of the Fifth International Conference on Database Systems for Advanced Applications*, Melbourne, Australia
- Neter John, William Wassermen, and Michael Kanter. 1990. *Applied Linear Statistical Models*, 3rd ed. Irwin Inc.
- Voorhees, Ellen M. Donna Harman. 1997. "Overview of the Sixth Text REtrieval Conference(TREC-6)", *The Sixth Text REtrieval Conference*.