정보검색 테스트 컬렉션 구축 및 유효성 평가+

맹성현*, 장동현*, 송사광*, 김지영*, 이석훈**, 이준호***, 이응봉****, 서정현*****

*충남대학교 컴퓨터과학과, **충남대학교 통계학과, ***숭실대학교 컴퓨터학부
****충남대학교 문헌정보학과, *****연구개발정보센터

Construction of an Information Retrieval Test Collection and its Validation

Sung Hyon Myaeng*, Dong-Hyun Jang*, Sa-Kwang Song*, Ji-Young Kim*
Seok-Hoon Lee**, Joon-Ho Lee***, Eung-Bong Lee****, Jeong-Hyun Seo*****
*Dept. of Computer Science, Chungnam National University
**Dept. of Statistics, Chungnam National University
***School of Computing, Soongsil University
****Dept. of Library & Information Science, Chungnam National University
*****KORDIC

요 약

본 논문은 정보검색 시스템 평가에 필요한 한국어 문서집합 구축과 적합 문서 리스트(relevance file) 생성에 관한 기법을 문서 수집과정부터 평가작업까지 상세히 기술한다. 문서집합은 일반, 사회과학, 과학기술 분야에서 각각 4만 건으로 영역별로 균등히 구축하였으며, 질의 집합도 각 분야에 대해 10개씩 할당하여 총 30개의 질의 집합을 생성하였다. 또한 질의집합은 사용자의 수준을 고려하여 일반인, 영역 전문가, 중고등학생에 해당하는 질의를 생성함으로써 특정 영역, 특정 사용자에 독립적인 문서집합 및 질의집합을 구축하고자 하였다. 생성된 질의를 사용하여 여러 검색기에서 총 38가지의 방법으로 검색을 실시하였으며, 검색결과를 바탕으로 각 질의당 500개의 문서로 이루어진 후보 결과집합을 만든 후 이들을 대상으로 각 질의에 대한 문서의 적합성 평가를 실시하였다. 이 과정을 통해 생성된 적합문서 집합의 유효성을 보이기 위해 후보 문서 리스트 이외의 문서집합에서 적합문서가 존재할 가능성을 확인하였는데 그 방법으로 후보 리스트의 개수 증가에 따른 적합문서 개수의 변동 추세를 알아보았다. 현재 질의 개수를 50개로 확장하는 방향으로 테스트 컬렉션 구축에 대한 연구를 진행 중에 있으며, 일본 NACSIS와의 질의 교환을 통해 질의 개수를 확장할 뿐만 아니라 일본어 질의 또는 한국어 질의에 대해서 한국어 문서, 일본어 문서를 각각 검색할 수 있는 한일 교차언어 문서검색 환경을 구축하고 있다.

+본 연구는 연구개발 정보센터의 지원으로 수행하였음.
*송사광은 ETRI에 근무중임

1. 서 론

컴퓨터를 이용하여 대용량의 문서로부터 사용자가 필요로 하는 정보, 즉 적합한 문서를 효율적으로 검색할 수 있는 정보검색 시스템에 관한 많은 연구가 이루어져 왔으며, 관리할 정보의 양이 기하급수적으로 증가하고 있는 오늘날에는 효율적인 정보검색 시스템에 대한 요구는 더욱 절실하다. 1960 년대부터 정보검색이라는 연구분야가 확립된 외국과 달리 국내에서는 1990 년대 중반 이후부터 다양한 한국어 정보검색에 관한 연구가 활발히 진행되고 있다. 그러나, 범용적인 한국어 정보 시스템의 평가를 위한 체제가 제대로 갖추어져 있지 않아 검색시스템의 신뢰성을 평가하기가 어려운 실정이다[1].

정보검색 시스템에 대한 검색 신뢰도는 이 분야의 연구에서 가장 중요한 지표로 사용되어 왔다. 검색 신뢰도에 기반을 둔 평가를 하기 위해서는 질의 집합, 대용량의 문서집합, 적합문서 판단으로 구성된 테스트 컬렉션(test collection)이 필요하다. 이러한 테스트 컬렉션은 정보검색 분야의 연구뿐만 아니라 사용자 집단이 상용화 시스템의 성능을 평가하여 적절한 시스템을 선택하는데도 중요한 역할을 한다.

본 논문에서는 한국어 문서 정보검색 시스템의 성능 평가의 기준이 될 수 있는 새로운 테스트 컬렉션 개발에 관한 내용을 기술하고 있다. 과거에 구축된 몇몇 한국어 테스트 컬렉션과 비교하여 본 연구에서 구축된 컬렉션이 갖는 특성은 다음과 같다.

첫째, 문서 영역을 일반, 사회과학, 과학기술 3 분야로 나누어 각 분야별로 4 만 건의 문서를 수집함으로써 특정 분야에 편중되지 않도록 하였고 컬렉션의 크기도 대폭 증가시켰다.

둘째, 질의를 일반, 사회과학, 과학기술 3 분야로 나누어 10 개씩 할당함으로써 질의 종류의 균형을 맞추었다.

셋째, 문서를 신문, 연구보고서, 논문, 회의록, 웹 문서로 구축함으로써 문서의 크기와 종류를 다양화하였다.

2. 관련 연구 현황

외국의 경우 1960 년대부터 소규모의 테스트 컬렉션이 구축되기 시작했으며, 정보검색 대상 데이터의 규모가 기하급수적으로 증가하면서 1990 년대에는 대용량의 테스트 컬렉션이 구축되어 오고 있는데 현재는 기가바이트 수준의 실험데이터를 사용해서 정보검색 시스템을 평가하고 있다.

미국은 NIST(National Institute of Standards and Technology)가 주축이 되어 학계 전문가를 중심으로 1991 년부터 TREC 테스트 컬렉션을 구축해 오고 있다. 1998 년에 발표한 TREC-7 은 1,634,243 건의 문서와 350 개의 질의로 구성되어 있다[2]. TREC(Text Retrieval Conference)에서는 매년 컬렉션을 사용하여 비상용 시스템 뿐만 아니라 상용 시스템을 다양한 방법으로 평가하여 그 결과를 발표하고 있다.

일본의 경우도 테스트 컬렉션의 중요성을 인식하여 정부 기관인 NACSIS(National Center for Science Information Systems)가 주관이 되어 대규모 컬렉션 구축 사업을 추진중이다. 1998 년 현재 약 33 만 건의 문서집합으로 대부분 학회 논문 요약으로 구성되어 있고 100 개의 질의가 있다. 일본 문서집합의 경우 대부분의 문서가 일어와 영어 병행 코퍼스로 이루어져 있어 교차언어 검색분야에 적용 가능하다.

또한 NTT Data Corporation 에서는 BMIR-J1 과 BMIR-J2 라는 컬렉션을 개발하였는데 BMIR-J1 은 600 건의 문서와 60 개의 질의로 구성되었고, BMIR-J2 는 5080 건의 신문기사와 60 개의 질의를 포함하고 있다[3]. 분야는 경제와 공학부분으로 한정되어 있고 질의를 다양한 형태로 분류하였다.

국내에서도 이러한 테스트 컬렉션에 대한 관심이 높아지고 있으나 평가결과에 대한 신뢰에 결정적인 영향을 주는 테스트 컬렉션의 사용이

보편화되어 있지 않은 실정이기 때문에 검색시스템의 성능평가에 있어 신뢰도에 근거한 객관적인 평가가 일반화되어 있는 상태는 아니다.

1994년에 국내에서 구축된 KT-SET[4] 테스트 컬렉션은 정보과학회 논문을 대상으로 하고 있으며 30개의 단순 질의와 1,053개의 학회 논문 초록으로 구성되어 있다.

1995년에 13,315건의 과기처 연구보고서를 대상으로 한 KRIST[5] 컬렉션이 구축되었는데 주로 생명과학, 의용전자공학, 기계공학 등을 주요 대상 분야로 하고 있고 TREC 질의와 유사한 형태의 30개의 질의가 구축되었다.

1996년에는 KT-SET을 확장하여 KT-SET 2.0[6]이 구축되었는데 4,414건의 문서와 50개의 자연어 및 불리언 질의로 구성되었으며 컬렉션에 논문, 신문기사, 저널을 포함하였다.

이와 같이 국내에서 개발된 테스트 컬렉션을 이용하여 검색 시스템들의 기본적인 평가가 가능하지만 그 규모가 작고 대상 분야가 편중되어 있으며 질의 및 문서 특성의 균형 등에 대한 고려가 반영되어 있지 않아 평가 결과에 대한 통계적 유의성 부여나 결과에 대한 일반화가 사실상 어려운 실정이다.

영어권 문서의 경우 대규모 컬렉션으로 시스템 평가가 이루어지면서 소규모 컬렉션으로 평가한 과거의 결과를 재평가하여야 하는 상황이 발생한 것을 볼 때 한국어 문서 정보검색의 경우에도 일정 수준이상의 규모로 구축된 테스트 컬렉션을 사용하는 것이 시스템 혹은 관련기술의 정확한 평가를 위해 필수적이다. 또한 TREC이 시작된 후 정보검색 기술이 급속도로 발전되어 왔다는 점을 고려할 때 국내에서도 TREC과 같은 테스트 컬렉션을 구축할 필요가 있다.

3. 문서 집합 선정 및 수집

정보 검색용 테스트 컬렉션을 일반적으로 문서집합, 질의집합, 각 질의에 대한 적합문헌 리스트로 구성된다. 이들 중 검색의 대상이 되는 문서집합은 테스트 컬렉션 구축에 있어서 가장 기본적인 요소이다. 다양한 분야 및 상황에서의 검색 시스템 평가를 위해서는 질의의 종류와 검색 대상 문서의 영역 및 언어적 특성을 고려하는 것이 필수적이다. 이를 통해 다양한 분야에 적용되는 정보검색 시스템의 성능을 여러 각도에서 테스트하여 평가할 수 있게 되는데, 본 연구에서는 다음의 두 가지 측면을 중요하게 고려하였다.

첫째, 다양한 분야의 문서들로 문서 집합을 구성하였다. 이는 분야마다 어휘나 문장의 특성들이 모두 다르고 통계적인 특성이 다르므로 문서의 다양성을 통해 검색기의 성능을 전체적으로 공정하게 혹은 분야별로 시험할 수 있기 때문이다.

둘째, 다양한 크기의 문서들로 문서집합을 구축하였다. 정보검색 기술의 기본이 되는 가중치 기법 중 일부는 특정 크기의 문서들에 높은 유사도를 부여하는 특성을 지니고 있기 때문에 문서 크기가 편중되어 있는 경우 특정 검색기의 성능을 과대 혹은 과소 평가할 수 있기 때문이다.

본 연구에서 개발한 테스트 컬렉션의 문서집합은 일반, 사회과학, 과학기술 분야에 속하는 12만 건의 다양한 크기의 문서들로 구성되어 있다. 이들은 각 분야별로 4만 건씩 균등하게 선정되어 특정 분야에 편중되지 않고 고른 분포를 가지고 있다. 각 분야별로 4만건의 문서가 구축되어 있으므로 통계적 유의성을 잃지 않으면서 검색기가 가지는 문서 분야별 특성도 비교 평가할 수 있는 기반을 마련하였다. 또한 문서 집합들은 수십 바이트에서 수십만 바이트까지 매우 다양한 크기의 문서들로 이루어져 있어서 검색 알고리즘을 다양한 각도에서 테스트 할 수 있도록 하였다. 즉, 초록과 같은 특정 문서 형태에 국한시켜 개발된 검색 알고리즘이 특별히 좋은 평가를 받는 불공정성을 제거하도록 노력하였다. 구축된 테스트 컬렉션의 분야별 문서정보는 [표 1]과 같다.

문서 집합 중 웹 문서나 신문기사의 경우 중복된 문서들이 다수 존재하므로 이들 문서를 각

문서간의 유사도를 기준으로 제거하였다. 이 과정은 검색시스템을 사용하여 자동적으로 수행되었으며 문서간 80% 이상의 유사도를 갖는 문서들의 경우 한 건만 제외하고 모두 문서집합에서 제외하였다.

[표 1] 분야별 문서 정보

분야	문 서 내 용	건수
일반 종합	1994 년 한국일보기사	22,000
	gov 확장자를 갖는 웹문서	9,000
	com 확장자를 갖는 웹문서	9,000
사회 과학	1994 년 한국경제신문 기사	39,480
	한국여성개발원 정기간행물 논문	110
	경북도의회 회의록	410
과학 기술	과기처지원 연구보고서	10,000
	해외과학기술 동향	18,000
	논문 서지 사항	12,000

4. 질의집합 생성

질의의 형태는 TREC-6 의 Topic statement 형태를 따르되 <query> 항목이 추가되어 5 개 부분으로 이루어져 있다. TREC 의 경우 질의 형태는 TREC-1 부터 TREC-4 까지 단순해져 가는 경향이 있었으나 TREC-4 질의의 모호성 문제가 제기되어 TREC-5, TREC-6 에서는 <num>, <title>, <desc>, <narr>의 4 개 태그를 사용하고 있다[2].

본 연구에서 사용된 질의는 <num>, <title>, <desc>, <narr>, <query> 5 개로 구성되며 각각 질의번호, 질의제목, 질의설명, 질의해설, 질의단어 리스트를 나타낸다. 질의 중에서 <title>과 <desc>는 실제 검색 시스템이 사용하여 내부 질의를 생성할 수 있도록 한 부분이고 <narr>은 적합문서를 판별하는 기준을 기술한 것이다. 이 부분은 적합성 판단의 판정자가 검색된 문서 집합 중에서 적절한 문서를 판별하는 기준을 제공하는 것을 주목적으로 하고 있으나 검색시스템이 내부질의 생성에도 사용할 수 있다. 적합성을 판단하는 평가자가 <title>과 <desc> 정보만으로는 질의의 모호성 해소가 안 되는 경우가 많아 평가의 일관성이 없을 수 있기 때문에 <narr>이 필요하다. 마지막으로 <query>는 검색 시스템의 내부질의 생성을 도와주기 위한 부분으로 관련된 어절의 집합으로 구성되어 있으며 앞의 4 개 태그에 포함되지 않은 어절이라도 검색을 보충할 단어라면 <query>에 포함된다. [그림 1]은 본 연구에서 구축한 질의의 예이다.

<num> 01
<title> 월드컵 축구 유치
<desc> 한국의 2002 년 월드컵 축구 유치 활동 내용
<narr> 한국의 2002 년 월드컵 축구 유치를 위한 국내외적인 활동이나 한국개최에 대한 회원국의 반응을 포함한 정보는?
<query> 2002 년 월드컵 축구 피파 FIFA 회원국 한국 개최 주최

[그림 1] 질의 예

질의생성은 분야별, 사용자별로 분류하여 최종적으로 30 개를 생성하였다. 분야별로는 일반, 과학기술, 사회과학 3 분야로 나뉘어지는데 각 분야별 질의어의 비율은 1:1:1 이며, 사용자별로는 일반인, 전문가, 청소년 3 그룹으로 나뉘며 이들 각 분야별 분포는 4:3:3 이 되도록 구성하였다.

주제설정 및 질의 구성 시에 고려한 내용은 다음과 같다. 먼저 일반분야는 일반적 주제를 위하여 일간신문의 국내외 10 대 뉴스, 분야별(중소기업, 연예, 문화, 스포츠, 경제) 10 대뉴스들을 기본으로 앞에서 정한 사전기준을 가급적 따르도록 하였다. 청소년들의 주제탐색은 주로 문화(연예, 스포츠)와 과학분야에서 수행하였고 적합한 문서의 개수가 적은 특수한 질의를 의도적으로 생성, 포함시켰다. 다음으로 과학기술분야는 KRIST 테스트 컬렉션에 사용되었던 질의 중 아주 전문적인 것과 신문기사 등에서도 비록 낮은 확률을 갖고 나올 수 있는 종류의 잘 알려진 전문적인 것을 추출하였고, 학술논문 서지사항과 해외기술동향에서는 임의추출방식으로 추출한 문서로부터 질의 설명을 작성하였다. 마지막으로 웹 문서, 여성개발원 게재 논문 그리고 경북도의회 회의록은 임의적으로 자료를 탐색하여 각 자료의 특성을 잘 나타내는 문서를 기준으로 질의설명과 질의해설을 구성하였다.

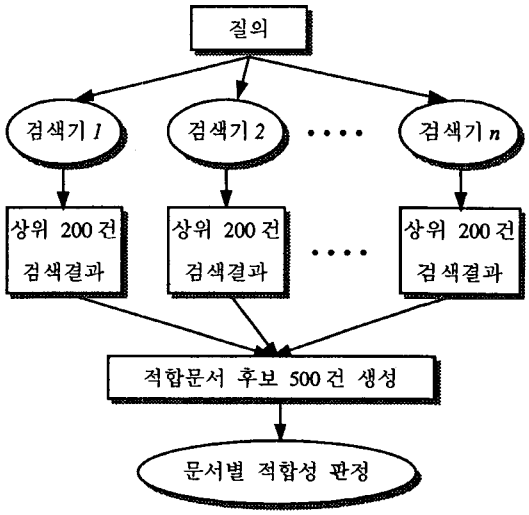
질의집합은 1단계로 62개를 생성한 후에 각 질의의 품질을 평가한 후 최종 30개의 질의를 선정했다. 이러한 여과과정을 거치는 이유는 질의의 성격이나 분야에 따라 적합한 문서의 수가 다르므로 그 균형을 맞추는데 있다. 예를 들어 적합한 문서가 극소수인 경우나 적합한 문서가 너무 많아 검색기의 성능평가에 도움이 안 된다고 판단되는 질의는 제외하였다. 최종적으로 선정된 질의 집합은 위에서 설명한 사용자의 관심분야로 균형 기준도 맞추었다.

5. 후보문서 생성

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

보다 현실적인 방법으로 다수의 검색 시스템을 사용하여 검색을 수행하고 각각의 시스템에 의해 높은 순위를 부여 받은 문서들에 대하여 적합성 여부를 판단하는 방법이 제안되었다[7]. 이 방법의 주안점은 특성이 다른 다수 시스템들에 의해 검색된 문서의 합집합에 컬렉션 내에 존재하는 거의 모든 적합문서가 포함될 것이라는 가정이다. 사용자의 적합성 판단 작업이 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 효과적인 방법으로 알려져 있다.

풀링 방법을 적용하기 위해서는 다수의 검색 시스템이 요구된다. 본 연구에서는 다양한 검색결과를 생성하기 위해 검색기는 충남대와 숭실대에서 개발한 검색기를 사용하되 각 검색기의 색인, 가중치 부여방식, 적합성 피드백(relevance feedback) 등의 환경을 변화시켜 38개의 후보문서 집합을 생성하였다.



[그림 2] 적합문서 판별 과정

[그림 2]는 후보문서 생성과정을 나타내고 있다. 풀링하는 과정은 먼저 각 결과집합은 동일하다는 가정하에 1부터 38까지의 각 집합을 임의의 순서로 배열한 후 각 집합의 문서를 랭킹 순으로 추출하여 나간다. 이 때 동일한 문서가 이미 추출된 경우는 최종 결과집합에 추가하지 않으며 총 500개 문서가 될 때까지 반복하여 실시한다.

6. 적합성 평가

적합성 판정은 각 질의당 500건의 검색된 문서들에 대하여 질의 단위로 다음과 같은 과정을 거쳐 실시하였다.

첫째, 임시 결과 집합을 평가하는 방법의 가평가 과정을 통해 평가자의 시각이 일관되도록 훈련시켰고 평가자 간의 시각의 차이를 최대한 줄이도록 하였다.

둘째, 총 10명의 평가자를 2인 1조로 구성하여 각 조가 6개 질의에 대한 문서 3000개를 평가하되 평가자 2인은 상호 독립적으로 평가하도록 하였다. 즉, 특정질의에 대한 하나의 문서를 2인의 평가자에 의해 적합성이 판정되도록 하였다.

이 때 각 질의에 대한 관점의 차이 극복을 위해 두 평가자 간에 질의에 대한 충분한 토의를 거친 후 상호 독립적인 평가작업을 실시하였다.

셋째, 평가방식은 적합, 부적합으로 나누는 종래의 방식에서 적합정도를 1(부적합), 2(약간 적합), 3(다소 적합), 4(적합), 5(매우 적합)의 다섯 가지로 나누도록 하였다. 이러한 방식은 판정하는데 시간이 많이 걸리는 어려움이 있지만, 적합의 상태를 보다 섬세하게 나타낼 수 있다는 장점과 더불어 평가자간에 발생할 수 있는 관점이나 생각의 차이가 양극화되는 위험을 막을 수 있는 보호장치 역할을 하게 된다.

넷째, 각 평가자간의 평가결과는 원시결과로써 확보해 놓는다. 또한 두 평가자 간의 이견이 있는 문서(3 점 이상 차이가 나는 문서)에 대해서는 상호 협의를 통해 조정하도록 하였다. 결국 조정하기 이전의 원시 평가 결과와 평가자 이견 조정 후의 수정 평가 결과를 모두 다 확보하였다. 원시 평가결과와 수정평가 결과는 테스트 컬렉션 평가시에 데이터로 사용된다.

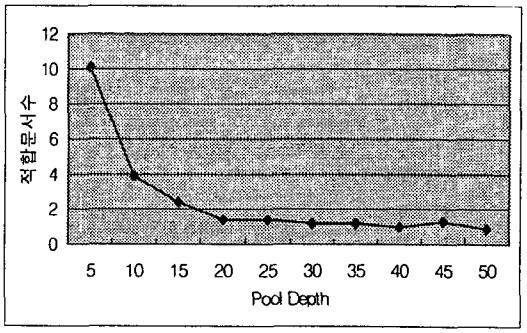
다섯째, 각 질의에 해당하는 문서에 대한 적합성 평가시 평가자가 <narr>에서 사용된 탐색어들 간의 불리언 논리(AND, OR, NOT)에 특히 신중을 기하도록 하였다.

7. 유효성 평가

테스트 컬렉션 구축에 있어서 가장 난해한 부분은 적합문서 리스트 구축인데, 그 이유는 적합성 평가시 평가자의 의견이 다른 경우가 빈번하게 발생하고, 컬렉션에 속한 모든 문서에 대해서 적합성 판정을 할 수 없기 때문이다. TREC 를 대상으로 실험한 연구 결과를 살펴보면 적합문서 리스트외에도 적합한 문서가 있을 가능성이 있기 때문에 재현률 평가시 과대 평가 될 수 있다는 보고가 있다[8].

신뢰성 있는 적합문서 리스트를 구축하기 위한 방법으로는 풀링기법을 적용할 때 적합성 평가 대상이 되는 후보문서 리스트의 수를 새로운 적합문서가 나오지 않을 때까지 증가시키는 방법이 있다. 이 방법은 질의별 특성에 따라서 풀의 깊이(pool depth)를 달리할 수도 있다. 다른 방법으로는 후보문서 리스트 구축에 참여하지 않은 새로운 시스템을 대상으로 후보문서 리스트를 구축한 후, 그 중에 새로운 적합문서가 발견되는지 관찰하는 방법이다. 후자의 방법은 다수의 검색기가 있어야 하는 환경 구축의 어려움으로 인해 본 연구에서는 전자의 방법을 이용하여 적합문서의 유효성을 알아보았다.

5 장에서 기술한 바와 같이 38 개의 시스템이 생성한 적합문서 후보 리스트를 대상으로 풀의 깊이를 점차적으로 증가시키면서 적합문서를 관찰한 결과를 [그림 3]에서 보여주고 있다. 본 연구에서는 풀 깊이를 최대 50 개까지만 관찰하였는데, 그림에서 Pool Depth 는 후보리스트 생성에 참여한 각 시스템에서 추출한 상위 문서의 개수를 의미하며, 적합 문서의 수는 모든 질의에 대한 평균 적합문서 개수를 의미하는데 동일한 문서는 제외하였다. 예를 들어, 그림에서 풀의 깊이가 10 인 경우, 평균적으로 각 질의에 적합한 문서는 약 4 개 존재하게 된다. 그림에서 보듯이 풀의 깊이가 50 인 경우 약 1 개의 적합문서가 존재하기 때문에 풀의 깊이를 더 증가시킬 경우 새로운 적합문서가 존재할 가능성이 있다. 이 부분에 대한 보완 연구가 현재 진행 중에 있으며, 적합문서 추출에 있어서도 새로운 시스템을 추가시킬 계획이다.



[그림 3] 후보문서 개수에 따른 적합 문서 수

8. 결론 및 향후연구

본 연구에서는 문서집합의 선정에서부터 질의 집합의 선정에까지 특정분야에 치우치지 않는 균형 잡힌 테스트 컬렉션을 구축하였다. 정보검색 분야가 지속적으로 발전하면서 보다 정확하고 객관적인 방법으로 시스템을 평가할 수 있는 환경에 대한 중요성은 더욱 높아질 것으로 판단되며 정보검색 시스템의 응용 분야가 다양해지면서 새로운 언어적, 구조적, 영역적 특성을 지닌 문서집합을 추가하는 것이 필요할 것으로 전망된다. 또한 단순한 문서검색 기능을 초월하여 문단검색, 구조화 문서검색, 정보추출, 정보요약 등의 새로운 기능을 시험할 수 있는 테스트 컬렉션의 구축에 대한 사항도 향후 중요한 이슈로 등장할 것이다.

향후 연구사항과 진행중인 사항은 다음과 같다.

첫째, 새로운 문서집합 추가 및 다양한 문서형태의 테스트 컬렉션 구축이 지속적으로 이루어져야 한다. 현재까지 구축된 컬렉션의 크기는 약 240M bytes로 수십 기가 바이트에 이르는 외국의 경우와 비교해볼 때 작은 규모로 계속해서 문서집합과 질의집합의 수를 증가시킬 필요가 있다.

둘째, 테스트 컬렉션의 신뢰도를 더욱 향상시킬 필요가 있다. 적합문서 리스트 생성과정에서 각 질의에 대해 모든 문서의 적합성을 판정한 것이 아니기 때문에 적합성 판정 대상이 되는 후보문서 리스트에서 제외된 문서 중에 적합한 문서가 존재할 가능성이 있다. 현재 적합 문서에 대한 신뢰도를 확률적인 접근 방법으로 제시할 수 있는 방법과 이를 통해 최적의 후보 리스트 개수를 산출할 수 있는 방법을 연구 중에 있다. 또한 새롭게 추가되는 질의에 대한 적합문서 생성시 기존에 사용하지 않은 시스템을 참여시킬 계획이다.

셋째, 과학기술 분야의 질의를 20개 추가시켜 현재 30개인 질의 수를 50개로 증가시키는 연구가 진행중이다. 결국 과학기술 분야 30개가 됨으로써 이 분야에 대한 평가의 통계적 유의성이 높아지게 되는데, 향후 타 분야도 동일한 수준으로 증가할 계획이다.

넷째, 일본과의 공동 연구를 통한 교차언어 검색 컬렉션 구축이 진행 중에 있다. 일본 NACSIS와의 질의 교환을 통해 질의 개수를 확장할 뿐만 아니라 한일 교차언어 문서검색 환경을 구축하고 있다.

본 연구를 통해 구축된 테스트 컬렉션은 정보검색 분야 연구자 및 개발자에게 자유롭게 배포되어 정보검색 시스템의 신뢰도 측정 목적으로 사용되어질 것이며 학술대회에서의 연구 결과 발표 혹은 제품 비교 등의 목적으로 활용되어질 것이다. 현재까지는 테스트 컬렉션의 영역 및 규모가 한정되어 있어 지속적으로 확장시킬 필요가 있으며, 테스트 컬렉션 사용자의 피드백을 받아 향후 구축에 반영시켜야 한다.

참고문헌

[1] 맹성현, 이석훈, 이준호, 이응봉, 송사광, "정보검색 시스템 평가를 위한 균형 테스트 컬렉션 구축," 한국정보관리학회지, 제 16 권, 제 2 호, 1999.

[2] Ellen M. Voorhees, Donna Harman, "Overview of the Seventh Text Retrieval Conference(TREC-7)", The Seventh Text Retrieval Conference(TREC-7).

[3] Ysuyoshi Kitani, Yasushi Ogawa, etc., "Lessons from BMIR-J2: A Test Collection for Japanese IR System," SIGIR'98 Melbourne, Australia.

[4] 김성혁, "자동색인기 성능시험을 위한 Test Set 개발". 정보관리학회, 1994.

[5] 이준호, 최광남, 한현숙, 김종원, 남성원, "정보검색을 위한 KRIST 테스트 컬렉션의 개발," 한국정보과학회, 1995.

[6] K.S.Choi, Y.C.Park, J.K.Kim, Y.W.Kim, "Development of the Data Collection Ver. 2.0 for Korean Information Retrieval Studies(KTSET2.0)," Presented at The Workshop on Information Retrieval with Oriental Languages, June 28-29, 1996.

[7] Harman D., "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, 1993.

[8] Justin zobel, "How Reliable are the Results of Large-Scale Information Retrieval Experiments?," Proceedings of the 21st Annual International ACM SIGIR Conference, 1998.