한국어 테스트 컬렉션 HANTEC의 확장 및 보완+

김지영*, 장동현*, 맹성현*, 이석훈**, 서정현***, 김 현***
*충남대학교 컴퓨터과학과
**충남대학교 통계학과
***연구개발정보센터
{jykim, dhjang, shmyaeng}@cs.cnu.ac.kr

Extension and Validation of Hangul Text Collection(HANTEC)

Ji-Young Kim*, Dong-Hyun Jang*, Sung Hyon Myaeng*, Suk-Hoon Lee**,
Jeong-Hyun Seo***, Hyun Kim***
*Dept. of Computer Science, Chungnam National University
**Dept. of Statistics, Chungnam National University
***Korea Research & Development Information Center

요 약

HANTEC1.0은 12만 건의 문서집합과 30개의 질의집합, 그리고 각 질의에 대한 적합문서로 구성된 정보검색용 한글 테스트 컬렉션이다. 본 연구에서는 HANTEC1.0의 확장 및 보완하기 위해 과학기술분야 20개의 질의를 추가하였는데, 질의 추가를 위해서 일본 NACSIS 테스트 컬렉션의 질의를 번역하여 사용함으로써 한일 교차언어 검색환경을 조성하고자 하였다. 추가된 각 질의에 대해서는 여러 검색기에서 총 41가지 검색방법으로 검색한 후, 각 검색조합의 상위 50개 문서로 구성된 중간 결과집합을 만들었으며, 이를 대상으로 적합성판정에 대한 평가기준 및 절차 교육이 이루어진 평가자가 각 질의에 대한 적합성평가를 실시하였다. 이렇게 구축된 HANTEC 테스트 컬렉션의 적합문서 집합의 객관적 품질 평가와 시스템 성능평가를 위하여 통계적인 방법을 적용하므로써 공신력있고 일반화된 테스트 컬렉션을 구축하고자 하였다. 현재 HANTEC2.0은 검색분야 연구자 및 개발자에게 자유롭게 배포 중이며 정보검색 시스템의 신뢰도 측정을 목적으로 하는 학술대회의 연구결과 발표 및 제품 비교 등에 활용되어질 것이다.

1. 연구배경

정보의 디지털화로 정보가 기하급수적으로 증가함에 따라 대용량의 문서로부터 사용자가 필요로하는 정보를 효율적으로 검색할 수 있는 정보검색 시스템에 대한 중요성은 날로 증가되고 있다. 이와 관련하여 다양한 한글 정보검색 시스템과 응용 시스템이 개발되고 상용화되고 있으나 한글 검색시스템 평가 기준의 부재로 인해 한글 검색시스템의 신뢰성을 평가하기 어려운 실정이다[1].

미국의 경우, TREC(Text Retrieval Conference)에서는 1991년부터 매년 컬렉션을 사용하여 비상용 시스템 뿐만 아니라 상용 시스템도 평가하여 그 결과를 발표하고 있다. 이 컬렉션은 NIST(National Institute of Standards and Technology)가 주축이 되어 학계 전문가를 중심으로 구축되었고 매년 그 규모와 종류를 증가시키고 있다.

일본의 경우, 정부기관인

+ 본 연구는 연구개발 정보센터의 지원으로 수행하였음.

NACSIS(National Center for Science Information Systems)가 주관이 되어, 1999년 학회논문 요약을 대상으로 한, 약 33만건의 문서집합과 100개의 질의로 구성된 테스트 컬렉션을 구축하여 지속적으로 개발 중에 있으며, 일영 병행 코퍼스를 구축하여 교차언어 검색이 가능하도록 하였다[3]. 또한, NTT Data Corporation에서는 BMIR-J1과 BMIR-J2라는 컬렉션을 개발하였는데, BMIR-J1은 600건의 문서와 60개의 질의로 구성되었고, BMIR-J2는 5080건의 신문기사와 60개의 질의를 포함하고 있다[4].

유럽의 경우, 유럽의 언어들에 있어서의 정보검색 시스템 평가를 위해 CLEF(Cross-Language Evaluation Forum) 테스트 컬렉션을 구축되었는데, 유럽 언어들에 대한 다국어 검색, 교차언어 검색, 문서검색을 평가하기 위한 것으로, 367,763개의 다국어 문서집합과 25개의 다국어 질의로 구성되어 있다.[5]

국내의 경우는 1994년에 KT-SET 테스트 컬렉션이 구축되었는데, 30개의 단순질의와 1053개의 학회논문 초록을 포함하고 있다[6].

1995년에 구축된 KRIST 컬렉션은 13,315건의 과기처 연구보고서와 30개의 질의로 구성되었고, 주로 생명과학, 의용전자공학, 기계공학 등을 대상으로 하고 있다[7]. 1996년에는 KT-SET이 확장되어 4,414건의 문서와 50개의 자연어 및 불리언 질의로 구성된 KT-SET2.0이 구축되었는데, 논문, 신문기사, 저널 등을 포함하였다[8]. 1997년에 구축된 계몽사 컬렉션은 23,113건의 문서와 46개의 질의로 구성되었고 문서는 분야별로 계층적으로 분류되어 있다.

이와 같이 국내의 경우, 컬렉션의 규모가 작고 대상 분야가 편중되어 있으며 질의 및 문서의 특성을 고려하지 않아 정보검색 시스템을 객관적으로 비교 평가하기에는 어려움이 있었다.

이런 상황을 극복하기 위해 1998년도에 구축된 HANTEC1.0은 질의 및 문서간의 분야별 균형을 고려한 12만 건의 문서, 30개의 질의, 그리고 각 질의에 대한 적합문서집합으로 구성된 국내에서 가장 큰 규모의 테스트 컬렉션[1][9]이다. 그러나 테스트 컬렉션의 중요성 및 신뢰도를 고려하여 그 품질에 대한 검증 및 확장을 통해 테스트 컬렉션의 완성도를 높일 필요가 있다. 이는 영어권 문서의 경우 대규모 컬렉션으로 시스템 평가가 이루어지면서 소규모 컬렉션으로 평가한 과거의 결과를 재심사하여야 하는 상황이 발생한 것을 볼 때, 한국어 문서 정보검색의 경우에도 어느 정도 수준이상의 규모와 신뢰도를 갖춘 테스트 컬렉션을 사용하는 것이 시스템 혹은 관련기술의 정확한 평가를 위해 필수적이기 때문이다.

2. 문서집합

정보 검색용 테스트 컬렉션에서 검색의 대상이 되는 문서집합은 테스트 컬렉션 구축에 있어서 가장 기본적인 요소이다. 본 연구에서는 HANTEC1.0의 12만 건의 문서집합을 그대로 사용하였는데, 이 컬렉션은 다음과 같은 두 가지 측면이 고려되었다.

첫째, 다양한 분야의 문서들로 문서집합으로 구성되었다. 이는 분야마다 어휘나 문장의 특성들이 모두 다르고 통계적인 특성이 다르므로 문서의 다양성을 통해 검색기의 강건성을 시험할 수 있기 때문이다.

둘째, 다양한 크기의 문서들로 문서집합이 구축되었다. 이는 정보검색 기술의 기본이 되는 가중치 기법들 중 일부는 특정 크기의 문서들에 높은 유사도를 부여하는 특성을 지니고 있기 때문에 문서의 크기가 편중되어 있는 경우 특정 검색기의 성능을 과대 혹은 과소 평가할 수 있기 때문이다.

HANTEC1.0의 문서집합은 일반, 사회과학, 과학기술 분야에 속하는 12만 건의 다양한 크기의 문서들로 구성되어 있다. 이들은 각 분야별로 4만 건씩 균등하게 선정하여 특정 분야에 편중되지 않고 고른 분포를 가지고 있으며, 문서의 길이도 짧게는 수십 바이트에서 길게는 수십만 바이트까지 매우 다양하게 구성되어 검색 알고리즘의 강건성을 테스트 할 수 있도록 구축되었다.

3. 질의집합

HANTEC에서 사용한 질의는 <num>, <title>, <desc>, <narr>, <quer>의 5개의 태그로 구성되며 각각 질의 번호, 질의 제목, 질의 설명, 질의 해설, 질의 단어 리스트를

나타낸다.

[그림 1]은 HANTEC2.0에서 사용된 질의의 예로 HANTEC1.0과 같은 형식으로 이루어졌다.

<num> 06
<title> 단어 열
<desc> 텍스트에서의 단어 열(collocation) 자동추출에 관해서
<narr> 텍스트에서 관용표현과 같이 고 빈도로 같이 출현하는 단어 열(collocation)을 자동추출 하는 방법에 관해서 보고하고 있는 문헌이 검색요구를 충족한다.
<quer> 단어 단어열 자동추출 텍스트 관용표현 자연어처리 관용어

[그림 1] HANTEC2.0에서 사용된 질의 예

HANTEC1.0에서 생성된 질의는 일반, 사회과학, 과학기술 3분야로 나뉘어 있으며, 사용자별로 일반인, 전문가, 청소년 3그룹으로 나뉘어 구성되었다.

HANTEC2.0에서는 HANTEC1.0이 현재 지니고 있는 분야별 균형과 질의 난이도별 균형 등을 유지하면서 과학기술 분야의 질의 20건을 추가하여 과학기술 분야 질의 수를 총 30개로 확장함으로써 과학기술 분야만을 독립적으로 평가할 수 있도록 했다. 또한, 추가되는 질의는 일본 NACSIS에서 구축한 테스트 컬렉션의 질의를 번역하여 사용함으로써 향후 한일공동 교차 언어 검색 테스트 컬렉션을 구축할 수 있는 환경을 조성하였다.

이를 위해 NACSIS 질의 83개를 한글로 번역하고, 그 중에서 HANTEC 컬렉션과 관련있는 질의 79개를 선정하였다. 이렇게 선정된 79개의 질의는 각각 충남대 검색기를 통해 일차 검색 결과를 생성하는데 사용되었으며, 그 결과를 토대로 각 질의의 품질을 평가하였다. 각 질의마다 적합문서수가 극소수이거나 너무 많아 검색기의 평가에 도움이 안되는 질의는 제외하였는데, 이러한 여과 과정을 거치는 이유는 질의의 성격이나 분야에 따라 적합한 문서의 수가 다르므로 그 균형을 맞추는 데 있다. 본 연구에서는 최종 20개의 질의를 선정하여, 총 50개의 질의로 구성된 질의 집합을 생성하였다.

4. 적합문서집합

테스트 컬렉션 구축에 있어 가장 중요한 요소는 각각의 질의에 대한 적합문서집합의 생성이다. 적합문서집합을 생성하기 위해 테스트 컬렉션에 포함된 모든 문서의 적합성 여부를 판단하는 것이 현실적으로 불가능하므로 보다 현실적인 방법으로 다수의 검색 시스템을 사용하여 검색을 수행하고, 각각의 시스템에 의해 높은 순위를 부여 받은 문서들에 대하여 적합성 여부를 판단하는 방법이 있다[10].

폴링 방법(Pooling method)이라고 불리는 이 방법은 서로 다른 다수의 시스템에 의해 검색된 문서들 중에서 상위 K개의 검색결과의 합집합이 컬렉션 내에 존재하는 거의 모든 적합문서를 포함하고 있다고 가정하기 때문에 사용자의 적합성 판단 작업이 전체 컬렉션이 아닌 이 집합에 국한된다. 따라서 컬렉션이 클 경우 많은 시간과 노력을 줄일 수 있다.

본 연구에서 추가된 20개의 질의는 HANTEC1.0의 구축에 사용된 충남대와 숭실대 검색기 외에 연구개발정보센터의 크리스탈과 상용 시스템인 다센21 검색기를 추가하여 총 41가지의 검색방법을 통해 후보문서를 생성하였다. 최종 결과문서집합을 생성하기 위해 41개의 검색방법으로 얻어진 각 후보문서집합은 풀링을 하게 되는데, 풀링하는 과정은 먼저 각 결과집합은 그 품질이 동일하다는 가정 하에 각 집합을 임의의 순서로 배열한 후 각 집합의 문서를 랭킹 순으로 상위 50개의 문서를 추출한다. 이 때 동일한 문서가 이미 추출된 경우는 최종 결과 집합에 추가하지 않는다. 이렇게 생성된 최종 문서집합은 교육된 평가자들을 통해 5점 척도(1-부적합, 2-약간적합, 3-다소적합, 4-적합, 5-매우적합)로 적합성 평가가 실시 되도록 하였다.

5. 테스트 컬렉션 분석

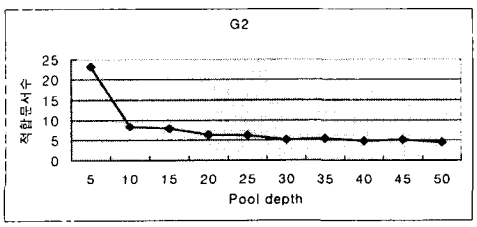
테스트 컬렉션 구축에 있어서 가장 난해한 부분이 적합문서집합을 생성하는 것인데, 그 이유는 적합성 평가 시 평가자의 의견이 다른 경우가 빈번하게 발생하고, 컬렉션에 속한 모든 문서에 대해서 적합성 판정을 하는 것이 불가능하다. 또한, 풀링 방법을 적용할 경우, 상위 K개의 문서들에 대해서만 적합성 판단이 이루어지므로 이때 제외된 문서 중에 적합한 문서가 존재할 가능성이 있기

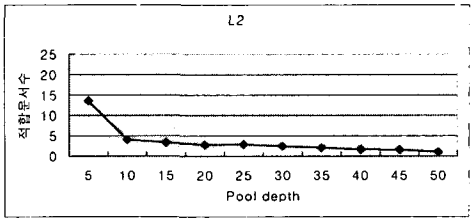
때문이다. TREC 을 대상으로 실험한 연구 결과를 살펴보면 적합문서집합 이외에도 적합한 문서가 있을 가능성이 있기 때문에 재현률 평가 시 과대 평가될 수 있다는 보고가 있다[11].

신뢰성 있는 적합문서집합을 구축하기 위한 방법으로 풀링 기법을 적용할 때 적합성 평가 대상이 되는 후보문서집합에 포함되는 문서의 수를 새로운 적합문서가 나오지 않을 때까지 증가시키는 방법이 있다. 풀 깊이(Pool depth)를 조정하는 이 방법은 질의별 특성에 그 깊이를 달리할 수도 있다. 다른 방법으로 후보문서집합 생성에 참여하지 않은 새로운 시스템을 대상으로 후보문서집합을 생성한 후 그 중에 새로운 적합문서가 발견되는지 관찰하는 방법이 있다. 후자의 방법은 다수의 검색기가 있어야 하는 환경 구축의 어려움이 있어 본 연구에서는 전자의 방법을 이용하여 적합문서의 유효성을 알아보았다.

5.1. 풀 깊이(Pool Depth)별 분석

각각의 질의에 대하여 다양한 검색기로부터 얻은 후보문서집합을 대상으로 풀의 깊이를 점차적으로 증가시키면서 풀 깊이가 50일 때까지 적합문서 수의 변화를 관찰하였다. 이때 적합문서는, [그림 2]는 추가된 20개의 질의에 대해 풀 깊이가 증가함에 따라 새롭게 발견되는 적합문서 수를 평균한 값을 G2 와 L2 에 대해서 보여주고 있다. 이때 숫자 2 는 5 점 척도에 따른 적합정도를 기준으로 2 점 이상인 문서를 나타내며, 문자 G 와 L 은 두 평가자 간의 판정의 차이가 있는 경우 이중 높은 점수를 선택했을 때(G)와 낮은 점수를 선택했을 때(L)를 의미한다. G2 를 보면 풀 깊이가 30 인 경우 20 개 질의에 대해서 평균적으로 약 5 개의 새로운 적합한 문서를 발견할 수 있음을 알 수 있다.





[그림 2] Pool depth 에 따른 적합 문서의 변화 추이도

풀 깊이와 추가되는 적합문서 수와의 관계를 구체적으로 살펴보기 위해, 풀 깊이를 p 라 하고 추가되는 적합문서 수를 n 이라 할 때 식 $ln(n+1)=A+B*ln(p)$를 풀 깊이와 그에 대응하는 추가된 적합문서 수에 적합시켜 계수 A,B 를 추정할 수 있었고, 그 때의 적합정도를 R^2라는 척도로 다음과 같이 구할 수 있었다.

적합기준	A	B	R^2
G2	4.7469	-0.4903	78%
L2	4.3817	-0.6542	85%

위의 식을 사용하여 풀 깊이가 51 에서 100 까지 늘어날 때 각각의 예상되는 적합문서 수를 살펴보면 [표 1]와 같다.

[표 1] 풀 깊이를 증가했을 경우 예상되는 적합문서 수

풀 깊이	적합기준에 따른 적합문서 수	
	L2	G2
1-50 (실제치)	709	1534
51-55 (예상치)	25	77
56-60 (예상치)	23	74
61-65 (예상치)	21	71
66-70 (예상치)	20	68
51-100 (예상치)	191	651

비록 가설이긴 하지만 이러한 분석결과는 현재 테스트 컬렉션의 적합성 판정이 불완전하다는 것을 의미한다. 그러나 최근의 연구에 의하면 비록 모든 적합문서가 식별되지 않은 상황에서도 이러한 테스트 컬렉션이 시

스템의 상대적인 우열을 가리는 데는 문제가 없다고 알려져 있으므로[2], 이러한 목적으로 사용하는데는 무리가 없을 것으로 보인다.

5.2. 시스템 성능평가에 미치는 영향

본 연구에서는 [2]의 결과를 확인하여 HANTEC 테스트 컬렉션의 신뢰도를 검증하기 위해 시스템 성능평가 시 제기될 수 있는 두 가지 문제에 대해서 조사해 보았다.

첫째, 적합성 판정결과에 대하여 각 시스템이 20 개의 질의를 통하여 반응한 41 개의 시스템에 대하여 L2 와 G2 를 사용하여 평균 정확도를 구한 후, 그 결과를 바탕으로 풀 깊이를 달리 했을 때에 시스템의 순위변화를 살펴보았다. [표 2]는 풀 깊이를 30, 40, 50 으로 하였을 때 각각 얻은 41 개 시스템 출력 순위를 보여주고 있다. [표 2]에서 볼 수 있듯이 풀 깊이가 30 에서 40 으로, 40 에서 50 으로 증가하여도 시스템의 순위에는 거의 변화가 없음을 알 수 있다

풀깊이 / 검색시스템ID	G2			L2		
	30	40	50	30	40	50
1	36	36	36	24	26	26
2	32	34	34	26	27	27
3	33	32	32	31	31	31
4	34	33	33	25	25	24
5	22	22	22	22	23	25
6	37	37	37	35	36	36
7	16	17	18	18	18	18
8	28	28	28	29	29	30
9	26	29	30	32	32	32
10	28	30	31	36	35	35
11	35	35	35	28	24	23
12	31	31	29	34	33	33
13	21	20	20	30	30	29
14	25	25	26	33	34	34
15	29	28	28	20	20	20
16	20	21	21	27	28	28
17	18	19	19	21	21	22
18	30	27	24	23	22	21
19	27	26	27	19	19	19
20	24	24	25	37	37	37
21	10	10	10	11	11	11
22	6	6	6	6	6	6
23	13	13	13	14	13	13
24	14	14	14	13	14	15
25	39	39	39	39	39	39
26	38	38	38	38	38	38
27	4	4	5	3	3	3
28	2	2	2	4	4	4
29	7	8	8	9	10	10
30	9	9	9	10	9	9
31	5	5	4	5	5	5
32	12	12	12	12	12	12
33	11	11	11	8	8	8
34	41	41	41	40	40	40
35	40	40	40	41	41	41
36	3	3	3	2	2	2
37	1	1	1	1	1	1
38	8	7	7	7	7	7
39	19	18	17	17	17	17
40	17	16	16	16	16	16
41	15	15	15	15	15	14

[표 2] 풀 깊이의 차이에 따른 시스템 순위

앞에서 살펴본 바로 깊이를 더 늘릴 경우 적합한 문서가 나올 가능성이 존재하지만, [표 2]의 결과를 볼 때 시스템 평가에는 거의 영향을 미치지 않는다는 것을 알 수 있으며, 시스템의 성능평가에 있어 시스템 순위는 풀 깊이와는 관계가 거의 없다는 것으로서 기존에 연구된 내용과 일치되는 것을 알 수 있다. 그러나 이 결과는 적합성 판정에 참여한 시스템들의 경우에 나타난 것이기 때문에, 적합성 판정에 참여하지 않은 시스템의 경우에도 이러한 결과가 나올지에 대해서는 추후 조사 연구가 필요하다.

둘째, HANTEC 구축에 사용한 검색기들 중 특정 검색기를 제외시켰을 경우 적합문서의 수에 어떤 영향을 주는지를 살펴보았다. 다음 [표 3]는 각각의 검색기를 제외했을 경우 적합문서 수를 보여주고 있다.

[표 3] 특정 검색기를 제외했을 때의 적합문서수

전체	적 합 문 서 수	
	G2	L2
	1534	709
검색기1 제외	1093(71%)	529(74%)
검색기2 제외	1125(73%)	564(79%)
검색기3 제외	1508(98%)	697(98%)
검색기4 제외	1519(99%)	703(99%)

위의 결과는 검색기 1 과 검색기 2 의 경우 전체 적합문서를 찾아내는데 각각 약 29%, 27% 정도로 HANTEC 테스트 컬렉션의 적합문서집합을 구성하는데 많은 영향을 주고 있으며 검색기 3 과 4 는 새로운 적합문서를 찾는데는 별 영향을 주지못하고 있다는 것을 알 수 있다. 다시 말해서 HANTEC 테스트 컬렉션에서 사용한 검색기 1 과 검색기 2 만으로도 적합문서집합을 구성하는 데는 문제가 없다고 할 수 있다.

이 결과는 이러한 분석을 좀 더 많은 검색기를 동원하여 수행할 경우, 어느 한계점 이상으로 새로운 검색기를 사용하여도 테스트 컬렉션의 품질에 영향을 주지 않는다는 결론을 내릴 수 있는 실마리가 될 수 있다.

5. 결론 및 향후 연구

본 논문에서는 체계적인 정보검색 시스템

평가 체제를 마련하기 위한 작업으로 기존의 HANTEC1.0을 확장 및 보완하여 12만 건의 문서집합과 50개의 질의, 각 질의에 대한 적합문서집합으로 구성된 테스트 컬렉션 HANTEC2.0으로 확장하였다. 먼저, 질의 확장을 위해서 질의와 문서의 영역 및 문서 형태에 대한 균형을 고려하여 과학기술분야 질의를 20개 추가하였는데, 추가된 질의로는 일본 NACSIS 테스트 컬렉션 질의를 번역하여 사용함으로써 한일 교차언어 검색환경을 조성하였고, 테스트 컬렉션의 신뢰도를 객관적으로 확립하기 위해 통계적인 방법을 제시하고 그 결과를 제시하였다. 이렇게 확장 및 보완된 HANTEC 정보검색 테스트 컬렉션은 그 규모와 품질면에서 볼 때, 외국의 테스트 컬렉션 수준에 크게 뒤지지않을 뿐만 아니라, 국내의 테스트 컬렉션 수준을 크게 향상시켰다고 할 수 있다.

현재 HANTEC2.0은 검색분야 연구자 및 개발자에게 자유롭게 배포 중이며 정보검색 시스템의 신뢰도 측정을 목적으로 하는 학술대회의 연구결과 발표 및 제품 비교 등에 활용되어질 것이다.

앞으로 한국어 정보검색 시스템의 신뢰도 평가에 있어 공신력 있는 표준으로 삼을 수 있도록 하기 위해 8만 건의 문서를 추가하여 총 20만 건의 문서집합으로 확장하고 있으며, 보다 많은 검색기의 참여시키고 다양한 통계적인 방법을 사용하여 확장 및 보완할 계획이다.

6. 참고 문헌

[1] 맹성현, 이석훈, 이준호, 이응봉, 송사광, "정보검색 시스템 평가를 위한 균형 테스트 컬렉션 구축," 한국정보관리학회지, 제6권, 제2호 1999.

[2] http://trec.nist.gov/

[3] Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, Soichiro Hidaka, Jun Adachi, "The NTCIR Workshop:the First Evaluation Workshop on Japanese Text Retrieval and Cross-Lingual Information Retrieval," Proc. of IRAL'99, Taipei, Taiwan.

[4] Tsuyoshi Kitani, Yasushi Ogawa, etc. "Lessons from BMIR-J2: A Test Collection for Japanese Text Retrieval and Cross-Lingual Information Retrieval," Proc.of IRAL'99, Taipei, Taiwan.

[5] http://galileo.iei.pi.cnr.it/DELOS/CLEF/clef.html

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

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

[8] 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.

[9] 맹성현, 이석훈, 송사광, 박혁로, "정보검색 시스템 평가를 위한 균형 테스트 컬렉션 구축," Proc.of KOSTI'98.

[10] Harman D., "Overview of the 1st Text Retrieval Conference", Proc. of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp36-48, 1993.

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