정보검색 기술 평가

맹성현
충남대학교 컴퓨터과학과

Evaluation of an Information Retrieval Techniques

Sung Hyon Myaeng
*Dept. of Computer Science, Chungnam National University

1. 서 론

1960 년대부터 확립된 후 꾸준히 발전해 온 정보검색 분야는 90 년대에 들어 인터넷이 보편화되고 텍스트 정보의 양이 기하급수적으로 증가하면서 관련 연구 및 시스템 개발이 급증하는 양상을 보이고 있다. 국내에서는 1990 년대 중반부터 다양한 한국어 정보검색에 관한 연구가 활발히 진행되고 있고 검색 시스템의 수요자 층이 넓어지고 있다. 그러나, 한국어 정보검색 시스템 혹은 관련 기술의 평가를 위한 체제가 제대로 갖추어져 있지 않아 연구자와 일반 사용자 모두 혼란한 상태에 놓여있는 실정이다.

정보검색 시스템에 대한 평가는 효율성(efficiency), 신뢰성(effectiveness), 가용성(ease of use) 등 다양한 차원에서 이루어져야 한다. 이러한 지표에 의한 평가를 어떻게 할 것인가 하는 문제는 광의의 정보검색 시스템이 가질 수 있는 구체적인 기능에 따라 또 달라진다. 예를 들면, 변화하는 사용자 질의에 대한 문서검색(retrospective search) 기능, 고정된 질의에 대한 문서 여과(filtering) 기능, 문서분류(categorization) 기능, 문단검색(passage retrieval) 기능, 사실검색 (fact retrieval) 기능, 심지어는 요약 기능 등은 모두 별도로 평가되어야 한다.

정보검색과 관련된 다양한 평가 대상 기능 및 지표 중 현재까지 국내외적으로 가장 잘 알려지고 연구되어 왔으며 또 기술 개발의 원동력이 되어 온 것은 일반 문서검색을 대상으로 한 신뢰도[1] 평가이다. 검색 신뢰도에 기반을 둔 평가를 하기 위해서는 질의 집합, 대용량의 문서집합, 적합문서 판단으로 구성된 테스트 컬렉션(test collection)이 필요하다.

2. 관련 연구 현황

외국의 경우 영국과 미국을 중심으로 1960 년대 후반부터 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[1]. TREC(Text Retrieval Conference)에서는 매년 컬렉션을 사용하여 실험실 시스템 뿐만 아니라 상용 시스템을 다양한 방법으로 평가하여 그 결과를 발표하고 있다.

---

[1] Effectiveness 로 알려져 있는 이 지표를 "효율성"으로 번역하는 경우도 있으나 이는 efficiency 와 혼돈이 되므로 의미적으로 보다 가까운 "신뢰도"를 사용한다.

일본의 경우도 테스트 컬렉션의 중요성을 인식하여 정부 기관인 NACSIS(National Center for Science Information Systems)가 주관이 되어 대규모 컬렉션 구축 사업을 추진중이다. 1999 년 현재 학회 논문 요약으로 구성되어 있는 약 33 만 건의 문서집합과 83 개의 질의로 이루어진 컬렉션을 가지고 있고 TREC 과 유사한 평가 워크샵을 1999 년도 8 월에 수행하였다[2]. 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.

국내에서는 1994 년에 최초로 KT-SET[4] 테스트 컬렉션이 구축되었는데 정보과학회 논문을 대상으로 하고 있으며 30 개의 단순 질의와 1,053 개의 학회 논문 초록으로 구성되어 있다. 외국의 테스트컬렉션에 비해 그 규모나 품질면에서 크게 떨어지지만 국내에서 정보검색 기술의 객관적인 평가에 대한 인식을 심어주었고 최초의 공개된 컬렉션이라는 면에서 큰 의미를 갖는다.

1995 년에 13,315 건의 과기처 연구보고서를 대상으로 한 KRIST[5] 컬렉션이 구축되었는데 주로 생명과학, 의용전자공학, 기계공학 등을 대상 분야로 하고 있고 TREC 질의와 유사한 형태의 30 개의 질의가 구축되었다. 이는 TREC 의 방법론에 따른 구축과정을 거쳐 국제적 수준의 과학적인 평가체제를 갖추려는 시도로 평가된다.

1996 년에는 KT-SET 을 확장하여 KT-SET 2.0[6]이 구축되었는데 4,414 건의 문서와 50 개의 자연어 및 불리언 질의로 구성되었으며 컬렉션에 논문, 신문기사, 저널을 포함하여 KT-SET 이 가지고 있던 문제점을 부분적으로 보완하였다

1998 년부터 시작된 HANTEC(HANgul TEst Collection)[2] 과제는 컬렉션 규모를 크게 향상시키고 문서 및 질의의 종류를 다양화하면서 적합성 판정의 품질을 제고하는데 목표를 두고 진행되고 있다[7, 8]. 총 12 만 건의 문서와 50 개의 질의를 갖춘 이 컬렉션은 1999 년도 말에 일차 완성되어 국내 연구자들에게 공개될 예정이며 그 특징은 다음과 같다.

- 문서 영역이 일반, 사회과학, 과학기술 3 분야로 나누어져 있고 각 분야별로 4 만 건의 문서가 수집되어 균형이 잡혀있다.
- 질의는 일반, 사회과학, 과학기술 3 분야에 각 10 개씩 할당되어 있고, 1999 년도에 과학기술분야 자체의 완결성을 위해 20 개의 질의가 추가되었다.
- 문서 컬렉션이 신문, 연구보고서, 논문, 회의록, 웹 문서로 이루어져 있어 문서 크기 및 종류의 다양성을 갖추고 있다.

3. 평가체제 구축에 대한 제안

영어권에서 TREC 의 대규모 컬렉션을 사용하여 시스템 평가를 수행하면서 얻은 결론 중의 하나는 소규모 컬렉션으로 평가한 과거의 결과를 재평가해야 한다는 것이다. 그 이유로는 소규모 컬렉션의 경우 문서나 질의의 종류 및 영역 등이 매우 제한되어 있어 다양한 실제 상황에 대한 대표성이 약하고, 평가 결과에 대한 통계적 유의성이 결여되어 있다는 점이 가장 크다.

이러한 경험을 토대로 볼 때 한국어 문서 정보검색의 경우에도 일정 수준이상의 규모와 다양성을 갖춘 테스트 컬렉션을 사용하는 것이 시스템 혹은 관련기술의 정확한 평가를 위해 필수적이다. 또한 TREC 을 통한 공식적인 평가가 시작

[2] 본 과제는 연구개발정보센터(KORDIC)의 지원으로 수행되었으며 컬렉션의 공식이름은 추후에 변경될 수 있다.

된 후 정보검색 기술이 급속도로 발전되어 왔다는 점을 고려할 때 국내에서도 이와 유사한 평가체제를 갖추는 것이 필요하다. 이러한 평가체제를 통해 적합성 판정에 필요한 후보 문서 생성을 하는 경우, 다양한 시스템이 참여하게 되므로 적합성 판정의 품질이 좋아 질 가능성도 높다.

평가체제에 필수적인 것은 역시 신뢰성과 대표성이 큰 테스트 컬렉션을 구축하는 것이다. 다양한 상황에 적합한 질의를 지속적으로 생성하고 문서의 종류도 지속적으로 늘려가서 다양한 검색 상황에 적합한 평가를 가능하게 하여야 한다. 예를 들어 웹 문서만을 취급하는 평가 컬렉션을 구축한다면 상용 시스템의 품질을 평가가 가능할 뿐만 아니라 새로운 기술의 상용화도 촉진시킬 수 있을 것이다.

평가체제 구축의 성패를 좌우하는 요소는 크게 두 가지로 요약될 수 있다. 첫째, 평가용 테스트 컬렉션 및 평가 결과에 대한 신뢰성이다. 공정한 평가 및 결과에 대한 적절한 해석을 통해서만 많은 연구개발자 및 상용 시스템의 참여를 기대할 수 있고 평가 환경을 정착시킬 수 있을 것이다. 이를 위해서는 전문가 그룹을 형성하여 컬렉션 및 평가 과정에 대한 모니터링을 수행하고, TREC 이나 일본의 NTCIR 과 같은 국제적인 평가회의를 수행하는 것이 바람직 할 것이다. 둘째, 평가용 테스트컬렉션의 지속적인 구축과 평가체제의 유지이다. 테스트컬렉션 구축은 문서 수집, 적합성 판단, 질의 생성 등 많은 자원을 요구하는 작업이므로 장기간에 걸친 예산이 필요할 뿐만 아니라, 정보검색의 기능이 점점 다양해 지고 고급화 되면서 새로운 기능에 대한 평가체제 및 컬렉션이 지속적으로 추가되어야 한다.

참고문헌

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

[2] Noriko Kando , "Overview of IR Tasks at the First NTCIR Workshop," in Proc. of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition, Kando & Nozue (eds), 1999.

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

[8]. 맹성현 외 6 명. "정보검색 테스트 컬렉션 구축 및 유효성 평가", 한글 및 한국어정보처리 학술대회, 1999.