한텍(HANTEC) 테스트 컬렉션 적합성 정보 재평가 및 보완[1)]

Review and Make Up of HANTEC Test Collection Relevant Information

강현규, 장형일, 박경일, 김현태, 염성욱, 나동열*, 최호섭**, 윤화묵**
건국대학교, 연세대학교*, 한국과학기술정보연구원**

Kang Hyun-Kyu, Jang Hyeong-Il, Park Kyung-Il, Kim Hyun-Tae, Yeom Sung-Wook, Ra Dong-Yeol*, Choe Ho-Sup**, Yoon Hwa-Mook**
Konkuk Univ., Yeonsei Univ.*, KISTI**

요약

정보검색 시스템 평가를 위한 한글 테스트 컬렉션인 한텍(HANTEC) 테스트 컬렉션 2.0이 배포되었다. 한텍 2.0은 12만건의 문서 집합과 50개의 질의 집합, 그리고 각 질의에 대한 적합성 정보로 구성되어 있다. 테스트 컬렉션에서 중요한 부분 중의 하나인 적합성 정보가 한텍에서는 풀링(pooling) 방법으로 구축되었다. 보다 더 정확한 정보검색 시스템의 평가를 위해서는 무엇보다도 정확한 적합성 정보가 중요하다. 따라서 현재 구축된 12만건 중 2만건을 대상으로 풀링방법이 아닌 수동방법으로 적합성 정보를 재평가함으로서 풀링방법의 유용성과 현재 배포된 한텍 테스트 컬렉션의 정보검색 평가용으로서의 유용성 여부를 확인 하고자 한다. 수동 적합성 정보 판정를 위한 도구를 만들었으며 적합성 판정 기준을 정하여 적합성을 판정하였다. 한텍과의 적합성 정보 비교 평가를 함으로서 풀링방법 및 현재 배포된 한텍 적합성 정보의 유용성을 비교 확인 하였다. 앞으로 2만 데이터에 대한 수동 적합성 판정 결과를 이용한 정보검색 시스템 신뢰도 측정에도 사용될 수 있을 것이다.

Abstract

HANTEC 2.0 (A Korean Test Collection) is distributed for evaluation of information retrieval systems. HANTEC 2.0 is consists of 120,000 documents, 50 topics(queries) and relevant information. The relevant information is constructed by pooling methods. The relevant information is very important for evaluation of information retrieval systems. So we would like to review of the relevant information by manual method. It will be show validation of pooling method and HANTEC relevant information. We make tool for manual review of relevant information and review of that. We review of relevant information between manual relevant information and HANTEC's. We review of pooling method and HANTEC relevant information. The manual relevant information will be use evaluation of information retrieval systems.

I. 서 론

최근에 정보검색 시스템이나 웹 등의 검색엔진들이 많이 존재하고 있다. 정보검색 시스템의 질적 성능을 객관적이고 공정하게 평가하기 위해 외국에서는 오래전부터 테스트 컬렉션을 구축하고 정보검색 시스템의 질적 성능(효율성)을 평가하기 위해 질의 집합 및 적합성 정보를 같이 제공하고 평가를 하고 있다[3-5]. 우리나라에서도 1994년도 이후에 여러 종류의 테스트 컬렉션들이 존재하여 왔다[1,2].

그러나 국내의 경우 그 테스트 컬렉션의 규모가 그리 크지 않고 정확하게 정보검색 시스템의 질적 평가를 하기 위해서는 많이 부족하였다. 그러던 와중에 1998년도에 비교적 중규모의 HANTEC(한텍) 2.0이라는 테스트 컬렉션이 개발 되었다[2]. 이는 보통 중규모 이상에서 적합성 정보를 수동으로 구축 할 수 없기 때문에 일반적으로 사용되는 풀링(pooling)방법을 사용하여 적합성 정보를 구축 하였다. 그러나 그 이후에 테스트 컬렉션 구축이 중단되었으며 현존하는 많은 정보검색 시스템의 질적 평가를 하기에는 활성화 되지 않은 측면이 있다.

따라서 지난 수년간 중단된 한글 정보검색 테스트 컬렉션의 개선 및 정제를 통하여 한국어 정보 검색 시스템의 질적 비교 평가를 위한 기반을 구축하기 위하여 HANTEC 2,0 테스트 컬렉션 일부의 적합성 정보를 재평가 하고 보완함으로서 테스트 컬렉션의 품질 향상 및 방향을 도모 하고자 한다. 기 구축된 HANTEC 2,0 테스트 컬렉션[6] 12만 건 중 2만건을 대상으로 적합성 정보를 수동 검토하고 기 구축된 적합성 정보와의 상관성을 분석하여 현재 배포된 한텍 테스트 컬렉션의 정보검색 평가용으로서의 유용성 및 구축 방법의 정당성를 확인하여 추후 테스트컬렉션 구축에 참고 하고자 한다.

본 논문의 2장에서는 기존 국내외 테스트 컬렉션 구축 현황

1) 본 연구는 KISTI의 지원으로 수행하였음.

을 설명하고 3장에서는 HANTEC 테스트 컬렉션에 대하여 설명한다. 4장에서는 HANTEC의 적합성 정보를 수동으로 개선 및 정제 하기 위해 사용된 HANTEC 평가 도우미 프로그램을 설명한다. 5장에서는 HANTEC 적합성 정보 재평가 및 보완을 위한 기준 및 보완에 대해 설명한다. 아울러 6장에서는 적합성 정보 분석 평가 및 결과들을 통계를 이용하여 설명하고 마지막으로 7장에서 결론을 맺는다.

II. 기존 테스트 컬렉션

기존 테스트 컬렉션을 나누자면 크게 외국의 테스트 컬렉션과 국내의 테스트 컬렉션으로 나눌 수 있다. 외국의 경우 1990년 초부터 데이터의 규모가 기하급수적으로 늘어나기 시작하여, 이 데이터의 규모를 따라가기 위해 대규모 테스트 컬렉션이 구축 사용 되었다[3,4].

미국의 경우, 1992년도부터 TREC(Text REtrieval Conference)[4]에서는 실험실에서 개발 된 시스템 외에도 상용 적으로 사용되는 시스템까지 평가를 하여 그 결과를 발표하고 있다. NIST(National Insitute of Standards and Technology)와 여러 학계 전문가들이 구축, 매년 규모를 증가 시켜가고 있다.

일본의 경우 NTT Data Corporation에서 600건의 문서와 60개 질의어를 사용한 BMIR-J1과 5,080건의 문서와 60개의 질의어를 사용한 BMIR-J2라는 컬렉션을 개발하였다. 또한 일본에서도 테스트 컬렉션의 중요성을 인식, NACSIS(일본의 정부기관)에서 주관이 되어 테스트 컬렉션을 구축 사업을 추진하고 있다[5].

국내에서도 이러한 테스트 컬렉션이 몇몇 개발 되었다[1,2]. 먼저 1994년에 구축된 KT-SET테스트 컬렉션은 30개의 질의어와 1,053개의 학회 논문 초록을 사용하여 구축 하였다. 이 테스트 컬렉션의 질의어는 매우 단순한 질문으로 이루어 져 있었으며, 문서의 양도 매우 적은 편이었다. 1995년에는 KRIST컬렉션이 구축이 되었다[7]. 이 테스트 컬렉션은 데이터의 분야를 생명공학, 의용전자공학, 기계공학 등의 분야를 주요 대상으로 하였다. 이 테스트 컬렉션은 13,315건의 문서와 질의 및 적합성 정보로 구성되어 있다.

1996년에는 단순했던 KT-SET테스트 컬렉션을 확장 하여 KT-SET 2.0을 개발하였다. 이 컬렉션은 50개의 자연어 및 불리언 질의와 4,414건의 문서를 사용하여 구축 되었다. 이 컬렉션의 특징은 기존에는 논문 초록에서만 데이터를 사용하였는데, KT-SET 2.0에서는 신문기사와 저널을 포함하여 확장하였다는 점이다. 1997년에 구축된 계몽사 테스트 컬렉션은 23,113건의 문서와 46개의 질의 및 적합성 정보로 구성되었으며, 문서는 백과사전 성격의 전분야별로 구성되어 있다.

이와 같이, 국내의 테스트 컬렉션들은 미국의 TREC에 비하여, 컬렉션의 규모가 작고 대상 분야가 넓지 못하다는 점이 있어 정보검색 시스템을 평가하기에는 어려움이 있어왔다.

III. 한텍(HANTEC) 테스트 컬렉션

1. HANTEC 1.0

HANTEC 1.0[1]은 테스트 컬렉션 구성에 가장 기본인 문서 집학과, 질의어 집합, 각 질의어에 적합한 문서 리스트로 구성되어 있다. HANTEC 1.0은 가중치 기법(weighting schemes)를 사용, 좀 더 좋은 평가를 얻기 위해 다양한 크기의 문서들로 구성하였다.

HANTEC 1.0의 문서 구성을 보면 일반, 사회과학, 과학기술분야에 속하는 120,000건의, 짧게는 수십 바이트에서 길게는 수십만 바이트까지 다양한 크기의 문서들로 이루어져 있다. 각 분야별로는 40,000건씩 균등하게 구성이 되어 있다. 일반분야 40,000건에는 1994년 발행된 한국일보 기사 22,000건과 gov 확장자를 가지고 있는 정부의 웹페이지 9,000건, com 확장자를 가지고 있는 웹페이지 9,000건으로 이루어져 있다. 사회과학분야는 1994년에 발행된 한국경제신문 기사 39,480건과 한국여성개발원 발행 정기간행물 논문 110건, 경북도의회 회의록 410건으로 이루어져 있다. 과학기술 분야는 과기처지원 연구보고서 10,000건과 해외과학기술동향 18,000건, 학술논문 서지사항 12,000건으로 이루어져 있다.

질의어는 TREC-6의 형태에 <query>를 추가한 형태이며, 질의어 분포는 일반분야, 과학분야, 사회과학분야 비율을 1:1:1로 하였으며, 사용자별 일반인 4명, 전문가 3명, 청소년 3명으로 질의어를 구성하였다. 즉, 총 30개의 질의어로 구성되어져 있다. 후보 문서 생성은, 각 질의어당 충남대, 숭실대의 검색기에서 나온 결과들에 적합성 피드백을 적용하여 문서 500개를 추출하였다. 이를 2명씩 5조가, 각 조 6개 질의어(3000개 문서)를 적합정도 1(부적합)~5(매우 적합)까지의 점수로 평가 하였다.

2. HANTEC 2.0

HANTEC 2.0[2, 6]에서는 HANTEC 1.0에서 사용하던 문서들을 그대로 사용하였으며, 질의어 형식도 1.0의 형식을 유지 하였다. 다만, 질의어의 개수를 조정, 과학기술 분야 질의어 20개를 추가하여 총 과학기술분야 30개 질의어로 구성, 전체 질의어 50개로 구성하였다.

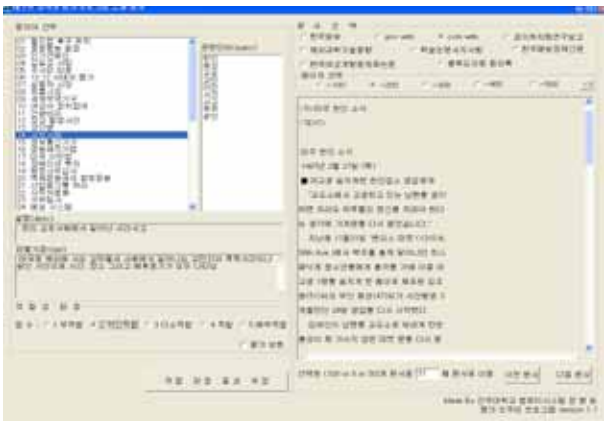
적합문서 집합을 생성을 위해서 기존 1.0의 충남대, 숭실대 검색기에 크리스탈[7]과 다센21 검색기를 추가하여 41가지 검

색방법을 사용하여 후보 문서를 생성하였고 이 후보 문서들을 다시 풀링(pooling)과정을 거쳐 임의의 순서로 배열 후 상위 50을 추출, 풀 깊이 조정 방법을 사용하였다.

Ⅳ. HANTEC의 적합성 정보 개선 및 정제를 위한 도우미 프로그램

HANTEC의 적합성 정보를 수동으로 개선 및 정제 하기 위해 HANTEC 평가도우미 프로그램을 사용하였다. 평가는 HANTEC의 문서 120,000건 중 20,000을 추출 평가하였다. 한국일보(1-2,500), gov web(1-2,500), com web(1-2,500), 한국경제신문(1-4,480), 한국여성개발원게재논문(1-110), 충북도의회 회의록(1-410), 과기처지원연구보고(1-2,500), 해외과학기술동향(1-2,500), 학술논문서지사항(1-2,500) 등 총 9개 분야로 나누었다. 분야별 균형 데이터를 유지하고 각 문서의 1-2,500번까지 2,,500건을 기준으로 하고 2,500건 미만은 모든 문서를 포함 하였다.

먼저 이 프로그램은 질의어 부분, 문서 부분, 평가 부분로 나누어진다. 평가 인원들이 모든 정보들을 수작업으로 평가를 한다는 점을 고려, 위의 3부분을 그림 1처럼 한 인터페이스로 구성하였다[11].



▶▶ 그림 1. 평가도우미 프로그램 인터페이스

그림 1에서 보는 것처럼, 질의어 부분은 프로그램 좌측 상단에 위치하고 있다. 이 질의어 부분에는 HANTEC의 질의어 50개를 표현하기 위해 리스트 박스를 사용하여 질의어의 번호와 질의어 title을 사용하여 질의어들을 표현하였다. 그 옆 박스에서는 HANTEC의 TREC-6과의 질의어에서의 차이점인 <query>의 내용들을 표현하였다. 그리고 그 하단부분에는 적합문서를 판별하는 기준을 기술한 <narr>부분을 표현하였으며, 그 하단에는 실제 검색에서 시스템이 사용하여 내부 질의를 생성할 수 있게 해주는 <desc>부분을 표현 하였다. <query>, <narr>, <desc> 부분은 리스트 박스에서 질의어 번호를 선택했을 시, 그에 해당하는 <query>, <narr>, <desc>이 표현 되도록 되어 있다.

문서 부분은 20,000건의 문서들을 표현하기 위해 각 문서들을 검토 분야별로 나누었다. 각 분야는 그림 1의 우측 상단에 라디오 버튼을 사용하여 표현하였다. 평가자는 자신이 맡은 분야를 선택 한 후, 평가할 문서의 그룹을 정하게 된다. 문서의 그룹이란, 한 분야 당 문서의 크기가 크기 때문에, 빠른 평가를 위하여 문서를 그룹으로 나누었는데, 충북도의회 회의록은 문서 9개씩, 한국여성개발원게재논문은 16개씩, 나머지는 100개씩으로 하였다. 이 부분은 그림 1의 분야별 라디오 버튼 하단에 위치하고 있다. 문서의 그룹을 선택하게 되면 그 그룹의 첫 번째 문서가 그림 1의 우측 중간에 위치한 텍스트 박스에 출력되어진다. 그림 1의 우측 하단에는 문서의 앞뒤 이동과 그룹 내 번호 이동에 관한 인터페이스이다.

평가 부분은 그림 1의 질의어 파트의 하단에 위치하고 있다. 평가 점수는 그림 2에서와 같이 HANTEC에서 사용한 1(부적합)~5(매우적합)을 사용하였다. 그리고 각 평가자가 평가를 하다가 평가를 내리기 예매한 문서를 보류로 선택, 따로 저장하게 되는데 그림 2의 5매우적합 버튼 바로 하단에 표현하였다.

▶▶ 그림 2. 평가도우미 프로그램 적합평가 파트 인터페이스

평가자는 프로그램을 실행하여 먼저 문서부분에서 분야를 선택하고, 문서의 그룹을 선택하여 문서를 연 후 이 문서에 관련되는 질의어를, 질의어 부분의 리스트 박스에서 선택 후, 이 질의어의 title, narr, desc, query를 참조하여 평가를 하게 된다. 평가 점수를 정하게 되면 평가자는 그림 2의 평가 부분의 자신이 판정한 점수 버튼을 선택 한 후 저장 버튼을 클릭하여 저장 하게 된다. 만일 평가를 내리기 예매한 문서가 나오게 되면, 그림 2의 평가보류 버튼을 클릭 후 저장을 하게 되면 보류 문서로 따로 저장이 되어, 평가자 전원이 모인 회의 자리에서 보류 문서를 보고 후 모두의 의견을 수렴하여 평가를 내리게 된다.

Ⅴ. 한텍 적합성 정보 재평가 및 보완

기존의 한글 정보검색시스템 평가 기준으로서 한텍

(HANTEC 2.0) 컬렉션[5]의 신뢰성 여부를 판단하기 위해 각 분야별 2,500 건의 문서, 전체 12만건의 문서중 2만건의 문서를 대상으로 적합성 정보의 개선 및 정제를 하였다.

한텍에서 2인 1조로서 문서의 적합성을 판단하여 적합 점수를 낮게 평가한 것(L 레벨)과 높게 평가한 것(G 레벨) 두 개로 양분하여 처리한 것과는 달리 각 1인이 2개 분야별 각 2,500 건의 문서에 대해 평가를 하였다. 기존의 한텍 컬렉션이 2인 1조로 평가가 진행 된 것에 비해 한명이 분석을 함으로서 신뢰도와 정확성, 객관성이 부족해지는 것을 예방하기 위해 매주 모임에서 본인의 정확한 판단 여부가 어려운 문서(보류 문서)들에 대해 논의하여 종합 평가하도록 하였다.

1. 평가 기준

문서의 적합성을 판단하는 근거로서 질의어, 관련 단어, 질의 설명(Description) 그리고 나레이션을 참조하게 되지만, 이는 같은 문서에 대해 다양한 평가를 할 수 있기 때문에 명확한 평가 기준을 세웠다.

첫째, 질의어는 질의 설명(Description) 에 포함되므로 문서를 평가할 때, 질의 설명을 기준으로 적합성를 평가 할 것인지에 대한 결정을 했다. 질의어 자체가 해석에 따라 다른 의미를 가지는 경우가 발생 하므로 이때 질의 설명에 쓰인 '~와' 와 같은 단어가 쓰인 경우 AND 개념으로 '~이거나' 와 같은 단어가 쓰인 경우는 OR 개념으로 생각하여 평가하였다. OR 개념일 경우 앞쪽만 맞아도 4점 이상이지만, AND 개념일 경우 반드시 앞문장과 다음 문장이 질의 설명과 일치해야 4점 이상을 획득할 수 있다.

둘째, 관련 단어의 경우 직접적인 관련이 없다는 생각아래 문서 평가 시 중요시 여기지 않았다. 검색 시스템을 이용하는 사람들이 모두 검색하려는 내용에 대해서 전문적인 지식을 가지고 있다고 보기 어렵기 때문에 질의어가 아닌 관련 단어만 포함된 문서가 검색되었을 경우, 그 문서는 유효성이 없다고 생각하였다. 실제로 관련 단어가 질의 설명(Description) 에 포함되는 경우도 많지 않았다.

셋째, 나레이션의 경우 질의 설명 내용에 부합되는 문서 즉 적합하다고 여겨지는 문서에 한하여 그 적합 정도를 평가하는데 사용하였다. 나레이션 역시, 질의 설명과 같이 '~와', '-이거나' 같은 연결 단어를 AND, OR 로 치환하여 일정한 평가 기준을 가질 수 있도록 하였다.

넷째, 위의 3가지 처리과정을 토대로 문서를 읽는 이의 주관적인 판단 근거아래 본래 이상으로 확장된 내용으로 받아들여지는 것을 막고 전체적인 질의의 의미 및 내용에 준하여 판정하게 하였다. 실제로 '국내의 재난, 해난 등 예보 시스템에 대한 운용의 문제점과 해결방안' 과 같은 질의 설명의 경우 문서의 내용이 태풍을 예보하는 것이라도 질의 설명의 '운용의 문제점과 해결방안' 이 나오지 않는 경우 태풍예보만으로 무리하게 문서의 내용을 확장하여 적합문서로 취급하지 않도록 했다. 또한 기존의 문서 일부 내용 대해 평가자가 잘 알고 있다고 하더라도 그 정보를 가지고 적합을 평가하지 못하게 하였다. 가령 '컴퓨터기법이 사용된 영화의 제작과정이나 흥행에 관한 정보' 의 경우 영화 제목만 가지고 본인의 경험을 토대로 컴퓨터 기법이 사용되었다고 생각하기에는 무리가 있다.

위와 같은 명확한 평가 기준을 토대로 객관적으로 정확한 정합성 평가가 이루어지도록 하였다.

2. 평가 보완

한텍 컬렉션의 신뢰성 여부를 판단하면서 자주 논의된 내용은 적합 정도를 평가하는데 나레이션이 큰 몫을 차지함에도 불과하고 나레이션의 내용이 적합 정도를 평가하기에 부족하고 애매한 부분이 많다는 것이다.

그중 가장 대표적인 것은 '국내 유통시장에 대한 국내외 대기업 진출 현황' 이라는 질의 설명에 대해 '국내 유통시장에 진출한 국내 대기업의 사업내용이나 국내 유통시장 진출을 위하여 기업합병 및 인수를 시도하고 있는 외국 기업들의 사업내용에 관한 문서는?' 나레이션이다. 이 경우 두 가지 해석의 문제를 가지게 되는데, 첫째는 유통이라는 단어의 범위를 어디까지 확장하고 제한할 것인가이다. 유통에 포함되는 수많은 사건들(시설확장, 점포확대)을 명확히 규정하지 못해 명확한 문서의 평가 기준을 세우기 어려웠다. 이러한 유통이라는 단어의 범주에 대해 나레이션에서 정확히 명시하여 정확한 평가를 할 수 있도록 도와야 한다. 둘째는 나레이션 자체에서 같은 국내 유통 시장에 대해, 국내 기업의 경우는 사업내용, 외국 기업의 경우 기업합병과 인수를 시도하는 기업의 사업내용으로 나뉘어 평가가 애매하다. 외국 기업의 경우 꼭 인수합병을 시도하는 기업에 한해 문서의 적합여부를 판단해야 하는 것인지, 국내 기업의 경우와 마찬가지로 인수합병 내용이 없어도 사업내용만 가지고 평가해야 하는지 기준을 마련하는데 상당히 어려움이 많았다.

나레이션의 역할은 평가를 돕기 위해 질의 설명에서 부족한 용어의 명확한 정의와 올바른 적합정도를 평가할 수 있는 도움 설명이다. 다음과 같은 경우 나레이션의 역할이 명확하게 나온 예이다. '데이터 구동 화상처리시스템의 제안' 라는 질의 설명에 대해 '데이터 구동 화상처리 데이터 플로우 프로세서에 의한 화상처리가 중심적인 내용인 문헌이 검색요구를 충족한다. 데이터 플로우 프로세서 그 자체에 관한 논문이더라도, 화상처리를 예로서 열거하는 등, 화상처리에 조금이라도 관련하는 문헌도 부분적으로 검색요구를 충족한다. 데이터 구동이 아

닌 것과 화상처리를 하지 않는 것은 검색요구를 충족하지 않는다.' 라는 나레이션을 제공하여, 정확한 평가를 할 수 있도록 여러 가지 조건과 상황에 대해서 명시하고 있다.

검색시스템 평가 기준으로서 HANTEC 컬렉션을 사용하기 위해서는 HANTEC 컬렉션 자체가 신뢰할만한 평가 기준을 가져야 한다. 이를 위해서 제공되는 질의어와 질의 설명, 관련 단어와 나레이션의 역할이 명확히 규정되어야 하고, 용어의 범위를 정확히 규정하여 불필요하게 확장된 평가로 인해 오해가 생기지 않도록 해야 한다. 이러한 부분들의 평가는 보류 판정을 한 후 전체가 모이는 매주 회의를 통하여 논의하고 종합 토의를 거쳐서 판정 하였다. 종합 토의된 내용을 기반으로 판정의 견고성을 유지하고 판정의 피드백을 거치도록 하였다.

VI. 적합성 정보 분석 평가 및 결과

1. 적합성 정보의 수집

적합성 정보에 대한 수집은 다음에 소개할 프로그램의 결과를 토대로 진행 하였다.

1.1 프로그램 소개

적합성 정보에 대한 수집에 쓰인 프로그램은 각 파트별 G2와 L5 사이의 평가에 맞는 적합 문서의 수를 나타내고, 그 문서의 수에서 HANTEC 적합성 결과와 LAB(수동 평가 한 실험실의 적합성 정보)의 결과 사이에 얼마만큼의 차이를 보이는 지를 숫자와 백분율로 나타내고 있다.

프로그램의 작성은 Visual C# .Net을 사용하였고, 프로그램의 실행을 위해서는 .Net Framework 2.0 이상의 버전을 필요로 한다.

1.2 실행 화면 및 사용법

▶▶ 그림 3. 프로그램 실행화면

위의 그림 3은 프로그램의 실행 화면 이며, 사용법은 다음과 같다.

프로그램의 실행 시 필요한 파일은 다음의 3가지 이다.

1) HANTEC의 적합성 정보가 저장된 파일(Hantec.txt)
2) LAB의 적합성 정보가 저장된 파일(LAB.txt)
3) 각 파트(두 분야 5,000개의 문서)별 적합문서(HANTEC 제공)가 담긴 파일(Docs.txt)

위의 3가지 파일이 존재 하는 디렉터리를 지정한 후, 통계 버튼을 누르게 되면 해당 문서들에 대한 적합 정보의 비교(문서 수의 차이, 백분율)를 수행하고, 실행화면에서 보는 것처럼 사용자에게 결과를 보여준다. 결과는 LAB의 작업을 기준으로 하였으며, 화면의 아래에 표시된 표에서 Lab 과 Hantec 항목은 각각 LAB의 작업결과 및 HANTEC 제공 적합 결과를 나타낸다. 그리고, Differ는 적합 문서 수의 차이를 나타내었고, 문서의 수가 LAB의 결과를 초과할 경우 붉은색으로 나타내었고, 미만일 경우에는 푸른색으로 나타내었다. 마지막, Ratio 항목은 LAB의 결과를 100으로 보고 그것에 대한 HANTEC의 비율을 백분율로 나타낸 것이다. HANTEC의 결과가 100%를 초과할 경우 붉은색으로 나타내었고, 100%이하일 경우에는 푸른색으로 나타내었다.

프로그램의 실행 후에는 각 평가 수준(G2 와 L5)에 해당하는 하위 디렉터리가 생성되며, 그 디렉터리에는 각 평가 수준에서 HANTEC과 LAB의 결과에서 중복 평가된 문서(두 결과 모두에 포함되어 있는 문서)들을 제공하며, 중복 평가를 제외한 나머지 문서들의 집합 파일을 제공한다.

2. 적합성 정보의 분석 및 평가

앞서 소개한 프로그램에서 얻어진 해당 문서에 대한 적합성 정보에 대해 LAB의 결과와 HANTEC의 결과를 비교 분석하여 HANTEC이 제공하는 적합성 정보에 대한 연관성 평가를 하고자 한다.

2.1 각 파트별 적합성 정보

[표 1] 각 파트별 LAB 결과 기준 백분율(단위 : %)

	G2	G3	G4	G5	L2	L3	L4	L5
Part1	175.7	71.91	71.87	48.97	78.94	50.56	35.93	30.61
Part2	73.01	48.27	46.47	46.15	32.14	24.82	22.53	15.38
Part3	263.9	137.7	64.28	32.60	123.4	62.28	32.85	2.173
Part4	141.3	100.0	117.3	84.84	78.44	68.18	80.43	60.60

위의 표를 보면 낮은 기준(적합도 2)에서 높은 기준 (적합도 5)으로 올라 갈수록 대체적으로 비율이 적어지는 것을 볼 수

있다. 높은 기준(G/L4, G/L5)의 경우는 비율이 현저히 낮다. 지금부터는 평가 기준들 중에서 가장 포괄적이면서 정확한 평가의 기준이 될 수 있는 HANTEC L2와 LAB의 적합도 2점 이상을 중점적으로 비교 분석할 것이다.

2.2 각 파트별(HANTEC 제공) 적합성 정보의 비교 분석 표

[표 2] LAB 적합도 2이상과 HANTEC L2에 대한 적합도 개수 차이 및 비율(비율 : 단위 %)

	LAB	HANTEC	차 이	비 율
Part1	95	75	-20	78.94
Part2	252	81	-171	32.14
Part3	158	195	+37	123.4
Part4	116	91	-25	78.44

표 2를 보면, LAB의 적합도 2점 이상 기준 대비 HANTEC L2의 차이 및 비율을 볼 수 있다. 차이에서 음수는 LAB의 적합도 2점 기준 대비 HANTEC L2의 부족 개수를 의미 하며 양수는 초과 개수를 의미한다. 비율은 LAB의 적합도 2점 기준 대비 HANTEC L2의 개수 비율이다.

LAB(적합도 2점 이상)과 HANTEC(L2)의 차이 부분을 분석 해 보자.

먼저, HANTEC과 LAB의 결과에서 공통된 부분(적합 문서)을 제외하고 나머지 부분에 대한 검토가 필요하다. 이유는 같은 적합도 문서들을 제외하여야 차이점을 쉽게 찾아 낼 수 있기 때문이다.

[표 3] 각 파트별 공통문서(HANTEC과 LAB 결과에 공통으로 적합으로 판별된 문서들)를 제외한 적합 문서의 수

	LAB	Hantec	[공통문서 제외]			
			LAB		Hantec	
Part1	95	75	40	42.11%	24	32.00%
Part2	252	81	186	73.81%	18	22.22%
Part3	158	195	36	22.78%	65	33.33%
Part4	115	91	38	33.04%	14	15.38%
공통문서제외평균비율			42.94%		25.73%	

표 3을 보면 공통문서(HANTEC과 LAB 결과에 공통으로 적합으로 판별된 문서들)를 제외한 적합 문서의 수를 볼 수가 있다. Part1의 경우 공통문서를 제외한 적합 개수가 LAB에서는 40개 Hantec에서는 24개가 존재한다는 의미이다. 이제 이 결과를 재검토 하여 차이점을 찾아보도록 하겠다.

2.3 각 파트별 적합성 정보 재검토 결과

[표 4] 각 파트별 HANTEC 재검토 결과 관계없음 비율

	전체	관계 없음	비 율(단위 : %)
Part1	75	24	32.00
Part2	81	18	22.22
Part3	195	65	33.33
Part4	91	14	15.38
평균			25.73

표 4는 재평가 판별 기준을 적용한 파일(표 2에서 나타난 결과)에 대해 HANTEC의 결과를 재검토(적합도 재평가) 한 것이다. 관계없음의 평균 비율이 25.73% 이다. 약 1/4이 재평가 판별기준에 적합하지 않았다.

[표 5] 각 파트별 LAB 재검토 결과 중 4/5점의 비율

	전체	4/5점	비 율(단위:%)
Part1	40	22	50.00
Part2	186	40	21.50
Part3	36	12	33.33
Part4	38	16	42.10
평균			36.73

표 5는 표 3에서 나타난 결과를 토대로 LAB의 적합 정보를 재검토 한 결과이다. 결과 중에서 2점과 3점은 HANTEC의 시스템과 판별 기준의 차이(견해 차이)라고 생각 할 수 있다. 하지만, LAB의 입장에서 재평가 판별기준의 적합성 정보에 대한 신뢰도가 높은(적합도 점수가 높은) 항목인 4점(적합), 5점(매우적합)의 평균 비율은 36.73% 이다. 즉 재평가 기준에 상당히 부합되는 적합성 정보가 HANTEC에는 들어가 있지 않다라는 의미이다.

2.4 HANTEC의 적합성 정보 비교 분석

앞의 표 2에서 LAB과 HANTEC의 해당 문서에 대한 적합 개수가 상당한 차이를 보이고 있다. 아울러 표3에서 공통문서를 제외한 차이의 개수도 많은 차이를 보이고 있다. 표4에서 보는 바와 같이 재평가 판별 기준을 적용한 HANTEC의 결과를 재검토(적합도 재평가) 결과 관계없음의 평균 비율이 25.73% 이다. 표5는 재평가 판별기준의 적합성 정보에 대한 신뢰도가 높은(적합도 점수가 높은) 항목인 4점(적합), 5점(매우적합)의 평균 비율이 36.73%으로 재평가 기준에 상당히 부합되는 적합성 정보가 HANTEC에는 들어가 있지 않았다.

위의 결과를 종합해보면 수동 검토한 LAB의 결과와 HANTEC의 적합 결과가 상당한 차이를 보이고 있다. 이는 우선 서로의 평가 기준이 달라서 나타날 가능성도 있다. 그러나 표4의 결과에서 보는바와 같이 공통문서를 제외한

HANTEC 적합 개수의 25,73%는 관계없는 적합 정보가 HANTEC에 들어가 있다. 이것은 적합판정에 있어 평가 기준 차이로 인해 문제가 있거나 적합 데이터 오류일 수 있다. 또한 표5의 공통문서를 제외한 LAB 4점 또는 5점에 해당하는 적합 개수의 36.73%가 HANTEC에는 포함되어 있지 않다. 이는 나머지 부분(63.27%)은 견해차이라 하더라도 36.73%는 적합 가능성이 매우 크다 할 수 있다. 또 하나는 풀링(pooling)방법에 있어서 상위 문서들로 올라오지 않는 문서 일 수 있다. 그래서 적합성 판정 시 검토되지 않은 문서 일 수 있다[9]. 즉 풀링 방법에 사용된 검색엔진의 성능에 문제가 있거나 풀링 방법의 근본적이면서도 일반적인 문제일 수 있다.

VII. 결 론

본 논문에서는 기 배포된 HANTEC 2.0 테스트 컬렉션의 12만건 문서 중 2만건을 대상으로 해당 질의에 대한 적합성 정보의 수동 판정을 행하였다. 2만건에 대한 2점 이상의 수동 판정과 해당 문서에 대한 HANTEC의 적합성 정보 L2를 비교 검토한 결과 HANTEC의 경우 LAB과의 일치한 부분을 제외한 25.73%가 관계없는 것으로 나타났다. 이는 평가 기준의 차이로 적합성 판정에 있어 문제가 있거나 적합정보 축적 오류 일 수 있다.

또한 일치한 부분을 제외한 LAB 부분 중 4-5점에 해당하는 36.73%가 HANTEC의 적합성 정보에는 들어가 있지 않았다. 이는 적합성 정보를 다루는 방법인 풀링(pooling) 방법의 문제이거나 사용된 검색시스템의 문제 일 수 있다. 또는 HANTEC의 적합성 정보 평가 기준이 매우 달라서 생긴 문제일 수 있다.

HANTEC과 LAB의 일치하지 않는 적합성 정보 평균이 각각 25.73%와 42.94%이다. 이의 차이는 생각보다 크다 하겠다. 따라서 테스트 컬렉션을 구축 할 시 대량을 컬렉션을 대상으로 현재는 풀링 방법 밖에는 대안이 없지만 검색엔진을 신중하게 선택을 해야 하며 무엇보다도 적합성 판정에 대한 정확하고 명확한 기준들을 만들어서 객관적이고 철저하게 적합성 판정을 해야 할 것이다.

▌참 고 문 헌▐

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

[2] 김지영, 장동현, 맹성현, 이석훈, 서정현, 김현, “한국어 테스트 컬렉션 HANTEC의 확장 및 보안”, 제12회 한글 및 한국어 정보처리 학술대회, pp.210-215, 2000.

[3] Harman D., "Overview of the 1st text retrieval conference", Proc. of 16th ACM SIGIR, pp.36-48, 1993.

[4] TREC's Homepage, "http://trec.nist.gov/"

[5] 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, 1999.

[6] KORDIC, 한글 테스트 컬렉션(HANTEC) version 2.0, CD, 연구개발정보센터.

[7] 이준호, 최광남, 한현숙, 김종원, 남성원, “정보검색을 위한 KRIST 테스트 컬렉션 개발”, 한국정보과학회, 1995.

[8] Kristal Homepage, "http://www.kristalinfo.com/"

[9] Justin Zobel, "How Reliable are the Results of Large-Scale Information Retrieval Experiments?", Proc. of the 21th ACM SIGIR, 1998.

[10] KORTERM, 정보검색시스템의 성능 평가를 위한 Test 시범 DB 구축 및 평가방법론 - 한국어 질의/응답시스템 테스트컬렉션 구축 - "http://www.korterm.or.kr/ ksurimal/report/정보검색평가방법론.htm"

[11] 정동원 역, Beginning Visual C++, Ivor Horton, 정보문화사. 2006.