INEX Book Search 트랙의 실험 고찰

Task Review of INEX Book Search Track

박 미 성(Mi-Sung Park)*

〈 목 차 〉

초 록

본 연구의 목적은 XML 검색 추진체인 INEX의 실험트랙 중에서 Book Search 트랙의 실험을 소개하고 실험방법을 분석함으로 XML기반의 디지털화된 도서 검색에 대한 관심과 더 많은 연구를 이끌어내고자 함이다. 이를 위해 첫째, INEX 실험트랙, 실험트랙 참여방법, 절차 및 참여기관을 상세히 소개한다. 둘째, INEX의 다양한 실험트랙 중, 디지털화된 도서검색을 위한 Book Search 트랙을 소개하기 위해, 이 트랙의 테스트컬렉션과 실험태스크, 실험결과 제출방법 및 평가방법을 제공한다. 셋째, 2007년에 시작된 INEX Book Search 트랙의 연구 논문들을 분석해봄으로 향후 연구과제들을 밝혀본다. 본 연구가 INEX 실험트랙에 대한 관심과 디지털화된 도서검색에 대한 관심을 국내에서 불러일으키는 계기가 되기를 기대해 본다.

키워드: INEX, Ad hoc 트랙, 도서검색트랙, 테스트컬렉션, 도서검색태스크

ABSTRACT

The purpose of this paper is to grow more interest and to forster research in full-texts retrieval of digitized books area through the review of Book Search Track and the analysis of research methods. First, this paper introduces the INEX tracks, the registration of INEX, the task process and the participating organizations. Second, to introduce the Book Search Track of all INEX tracks, this paper provides an overview of the test collection, the tasks, the task and submission guidelines and evaluation results of the Book Search Track's. Third, through paper review of the Book Search track that was lunched in 2007 as part of the INEX initiative, this paper presents the future research subject. This study expects that the readers are attracted by INEX tracks and full-texts retrieval of digitized books in korea.

Keywords: INEX, Ad Hoc Track, Book Search Track, Test Collection, Book Retrieval Task

* 경북대학교 중앙도서관 전산관리팀 팀장(mspark@knu.ac.kr)
• 접수일: 2009년 11월 26일 • 최초심사일: 2009년 11월 30일 • 최종심사일: 2009년 12월 26일

Ⅰ. 서 론

인터넷과 웹의 발달 초기에는 HTML이 정보표현과 문서교환의 수단으로 큰 각광을 받았다. 그 이유는 누구나 쉽게 정보를 표현하고 사용할 수 있다는 장점 때문이었다. 하지만 HTML은 태그가 고정되어 있고 새로운 태그를 정의할 수 없어서 다양한 콘텐츠의 내용을 기술하는데 많은 한계를 드러냈다. 이러한 한계를 극복하기 위해, W3C(World Wide Web Consortium)는 차세대 웹 문서 기술 표준 언어인 XML(eXtensible Markup Language)을 제안하였다.

현재 XML은 정보표현능력이 풍부하고 확장성이 뛰어나 이미 웹상에 분산되어 있는 정보표현 및 교환수단으로 크게 각광받고 있다. XML문서는 기존 문서와는 달리 하나의 문서에 내용정보와 다양한 구조정보를 함께 가지고 있어서 중간 매개체 없이도 문서의 데이터를 사람이 쉽게 이해할 수 있고 별도의 프로그램 없이 웹 브라우저만으로도 인식이 가능하다는 장점을 가진다.

이러한 특징과 장점으로 XML은 B2B 전자상거래뿐만 아니라 콘텐츠 제공자간의 콘텐츠 관리, 기관 내부에 존재하는 다양한 형태의 이종 애플리케이션 간의 연동, 데이터웨어하우징 등에서 특히 주목받고 있으며, 최근 디지털도서관에서 목록 및 원문 생산을 위해서 널리 사용되고 있고, 과학데이터 레포지터리 및 웹상에서도 XML문서들이 넘쳐나고 있다.

이처럼 XML문서들이 널리 활용되면서 XML문서를 검색할 수 있는 검색시스템에 대한 관심도 당연히 커질 수밖에 없다. 기존의 XML을 지원하지 않는 검색시스템들은 문서에서 제공하는 키워드를 기반으로 내용정보에 대한 검색을 실시함으로 너무 많은 자료를 결과로 검색해 준다. 이는 사용자가 더 적합한 문서를 찾기 위해 내용을 일일이 검토하던지 아니면 시스템적으로 적합성 피드백을 수행하던지 등의 더 적합한 문서를 선별해 내기 위한 별도의 많은 시간과 노력을 요구하게 된다.

반면에 XML의 다양한 논리적 구조정보를 활용할 수 있는 검색시스템은 기존 검색시스템처럼 내용정보 속의 키워드를 검색하는 대신에 '필드'와 같이 페이지의 특정 부분을 의미 있는 태그로 정의하여 사용할 수 있기 때문에 해당 태그만을 대상으로 검색케 하여 관련 없는 페이지의 검색은 피하고 더 정확하고 적합한 정보를 결과로 제공해 줄 수 있다.

이러한 특징을 가지는 XML문서의 대량 증가와 함께 XML문서의 효과적인 검색을 위한 연구들이 요즘 활발히 수행되고 있다. 특히 XML문서의 다양한 색인 전략이나 브라우징 및 검색 전략에 대한 연구가 촉진되고 있는데, 이러한 연구 분야에 큰 역할을 담당하고 있는 기관이 INEX (Initiative for the Evaluation of XML retrieval)[1)2)]이다. INEX는 2002년에 발족된 XML 검색평가를 위한 추진체로, XML 검색 및 성능평가에 관한 연구를 활발히 수행하고 있고, XML검색

1) 2002-2007년 INEX 홈페이지, 〈http://inex.is.informatik.uni-duisburg.de/〉 [cited 2009. 10. 11].
2) 2008-2009년 INEX 홈페이지, 〈http://www.inex.otago.ac.nz/〉 [cited 2009. 10. 11].

에 관심을 가지고 있는 전 세계의 많은 대학 및 연구소들도 뜻을 모아 연구에 동참하고 있다.

INEX는 XML문서의 저장 및 색인, 검색 순위화 전략개발 등 XML검색 분야의 발전을 촉진하고 객관적인 실험환경을 만들기 위해 대용량의 테스트 컬렉션(Test Collection)을 구축하고 있다. 테스트 컬렉션 구축 시, 객관성을 보장받기 위해 전 세계 실험 참여자들이 토픽개발이나 적합성평가에 직접 참여한다. 참여자는 함께 구축한 테스트컬렉션이라는 통제된 실험환경에서 실험을 거쳐 각자의 시스템에 대한 객관적인 성능을 평가받게 된다.

해마다 INEX실험에 참여할 수 있는 연구트랙들이 조금씩 변하는데 2008년과 2009년에는 Ad Hoc, Book,[3] Efficiency, Entity Ranking, Interactive(iTrack), Question Answering(QA@INEX), Link-the-Wiki, XML Mining트랙이 오픈되었다.

이 트랙들 중에서 본 논문에서 중점적으로 다루고자 하는 Book Search 트랙[4]은 2007년에 처음 시작되었다. 대용량의 도서컬렉션에서 정보를 찾는 것은 곧 구조화된 문서검색인 XML검색의 한 응용 분야이고, 이용자 정보요구에 적합한 도서의 부분에 직접적인 접근(Direct access)을 가능하게 하는 것은 이용자에게 많은 이익을 가져다준다는 생각으로 시작된 트랙이다.

Book Search 트랙은 XML구조의 대용량 도서저장, 색인, 검색을 위한 인프라 개발 및 도서의 특징을 활용한 정보검색 기술개발에 초점을 두고 있는데, 궁극적인 목표는 이용자 요구에 적합한 도서에 랭킹(Ranking)을 부여할 수 있는 도서만의 특징적인 랭킹전략을 조사하는 것이다. 저자에 의해 제공되는 권말색인과 도서관 목록정보인 메타데이터(MARC) 연계 등을 통해 도서만이 가지는 다양한 특징들을 개발하고 활용하여 도서의 검색성능을 높이는 것이다. 아울러 도서검색을 위한 다양한 이용자인터페이스(UI) 이슈 및 이용행태 연구도 이 트랙의 연구 분야이다.

본 논문에서는 전 세계적으로 XML 정보검색에 관심을 가진 그룹, 기관 및 개인 연구자들이 참여하고 있는 INEX의 연구트랙 중에서 특히 디지털도서관과 밀접한 관련이 있는 Book Search 트랙을 상세하게 소개하고 지금까지 나온 연구들의 실험방법을 분석해 본다.

목적은 도서관의 핵심 콘텐츠인 도서 중 디지털화된 도서검색에 관심을 갖게 하고, 아울러 실험절차를 상세히 소개함으로 실험에 참여할 수 있는 방법을 전하며, 현재까지 INEX Book Search 트랙에서 진행된 실험방법을 분석해 봄으로써, 향후 도서검색을 위한 다양한 접근법을 모색해 보고자 한다. 이를 통해 디지털도서관의 전자책(E-Book) 검색과 기타 기반 기술에 관심을 가진 국내의 학교, 기관 및 연구자들의 관심과 참여를 이끌어 내고, 국내 디지털도서관의 전자책(E-Book) 검색기술 발전에 도움이 되고자 한다.

---

3) 2007-2008년까지 Book Search Track으로, 올해 2009년에는 Book Track으로 오픈됨.
4) INEX Book Track, 〈http://www.inex.otago.ac.nz/tracks/books/books.asp〉 [cited 2009. 10. 11].

Ⅱ. 관련 연구

1. INEX 실험트랙

INEX 실험트랙이 매년 조금씩 바뀌는데, 2009년에 오픈된 트랙은 Ad Hoc, Book, Efficiency, Entity Ranking, Interactive(iTrack), Question Answering(QA@INEX), Link-the-Wiki, XML Mining 트랙이다.

각 트랙별로 실험방법 및 절차는 유사하나 가장 큰 차이점은 트랙의 특성 및 목적에 따라 사용하는 실험태스크와 테스트컬렉션이 다르다는 것이다. 테스트컬렉션이란 전통적으로 IR(Information Retrieval) 평가를 위해 사용되는 것으로, 세 개의 부분 즉, 문서집합, 정보요구의 집합인 토픽(Topic), 적합성평가(relevance assessments)로 구성된다.

많은 트랙 중에서 현재 가장 활발히 진행되고 있는 INEX의 메인트랙은 Ad Hoc 트랙[5]이다. Ad Hoc 트랙은 각 참여자들이 대용량의 일반 XML 문서를 대상으로 검색실험을 하는 트랙으로, 각 트랙의 실험태스크도 해마다 조금씩 다르다. Ad Hoc 트랙의 경우, 〈표 1〉과 같이 2007/2008년에는 3개의 태스크가 수행되었고, 2009년에는 4개의 태스크가 수행되었다.

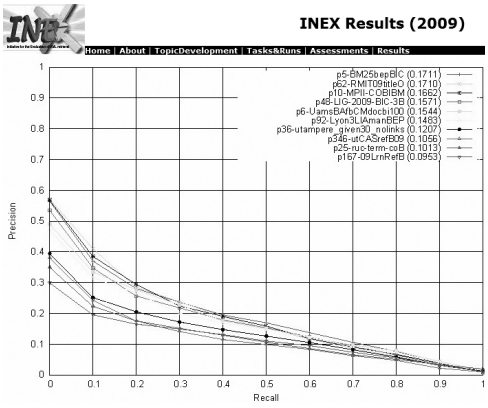
〈표 1〉 최근 3년간 Ad Hoc트랙의 실험태스크

2007년 Task[6]	2008년 Task[7]	2009년 Task
Focused Task Relevant in Context Task Best in Context Task	Focused Task Relevant in Context Task Best in Context Task	Focused Task Relevant in Context Task Best in Context Task Thorough Task

위 〈표 1〉에서 Focused Task는 실험문서 전체를 대상으로 이용자 질의 즉, 토픽에 가장 적합한 엘리먼트(elements) 또는 단락(passages)을 포함하는 기사의 순위리스트(ranked list)를 제출하는 태스크이다. Relevant in Context Task는 각 기사(article)당 가장 적합한 엘리먼트(elements) 또는 단락(passages)의 순위를 결과로 제출하는 태스크이다. 그리고 Best in Context Task는 토픽에 가장 적합한 베스트 엔트리를 가진 기사들을 결과로 제출하는 태스크이고, Thorough Task는 2009년 새롭게 추가된 태스크로 컬렉션 내에서 엘리먼트(elements)의 적합성(relevance)을 평가하는 태스크이다.

5) Ad Hoc Track, 〈http://www.inex.otago.ac.nz/tracks/adhoc/adhoc.asp〉 [cited 2009. 10. 11].
6) Norbert Fuhr et al., "Overview of the INEX 2007 Ad Hoc Track," *Pre-poceedings of INEX 2007*(2007), pp.1-22.
7) Jaap Kamps et al., "Overview of the INEX 2008 Ad Hoc Track," *INEX 2008 Workshop Pre-proceedings* (2008), pp.1-28.

각 트랙의 특정 태스크 참여자는 실험태스크의 시나리오에 맞게 각자의 시스템으로 동일한 테스트컬렉션을 가지고 실험을 수행하여 실험결과를 태스크별 제출지침에 따라 제출한다. INEX는 각 참여자의 제출결과를 비교 분석해서 성능에 대한 순위를 제공한다. 2009년 트랙 중 Ad Hoc트랙 태스크의 결과만 현재 오픈되었는데, 그 중 Best in Context Task의 결과를 살펴보면 다음 〈그림 1〉과 같다. 그림 오른쪽 부분에 기관별 순위가 나와 있다.



#	MAgP	Institute	Institute	Run
1	0.1711	5	Queensland University of Technology	BM25bepBIC
2	0.1710	62	RMIT University	RMIT09titleO
3	0.1706	5	Queensland University of Technology	BM25AncestorBIC
4	0.1671	5	Queensland University of Technology	ANTbigramsbepBIC
5	0.1664	5	Queensland University of Technology	ANTbigramsBOTbepBIC
6	0.1662	10	Max-Planck-Institut Informatik	MPII-COBIBM
7	0.1661	5	Queensland University of Technology	BM25BOTbepBIC
8	0.1657	5	Queensland University of Technology	BM25ArticleBIC
9	0.1652	5	Queensland University of Technology	BM25BOTancestorBIC
10	0.1651	5	Queensland University of Technology	ANTbigramsBOTancestorBIC
11	0.1645	5	Queensland University of Technology	ANTbigramsAncestorBIC
12	0.1608	62	RMIT University	RMIT09title
13	0.1604	5	Queensland University of Technology	BM25BOTarticleBIC
14	0.1591	5	Queensland University of Technology	ANTbigramsBOTarticleBIC
15	0.1584	5	Queensland University of Technology	ANTbigramsArticleBIC
16	0.1571	48	LIG	LIG-2009-BIC-3B
17	0.1568	48	LIG	LIG-2009-BIC-1B
18	0.1544	6	University of Amsterdam	UamsBAfbCMdocbi100
19	0.1483	92	University of Lyon3	Lyon3LIAmanBEP
20	0.1297	92	University of Lyon3	Lyon3LIAautoBEP
21	0.1207	36	University of Tampere	utampere_given30_nolinks
22	0.1177	10	Max-Planck-Institut Informatik	MPII-COBIRF
23	0.1175	6	University of Amsterdam	UamsBSfbCMsec2docbi100art1
24	0.1056	346	University of Twente	utCASrefB09
25	0.1013	25	Renmin University of China	ruc-term-coB
26	0.0953	167	School of Electronic Engineering and Computer Science	09LrnRefB
27	0.0923	25	Renmin University of China	ruc-sm-coB
28	0.0917	60	Saint Etienne University	UJM_15490
29	0.0913	36	University of Tampere	utampere_auth_40_top30_low
30	0.0904	60	Saint Etienne University	UJM_15506
31	0.0884	48	LIG	LIG-2009-BIC-2B
32	0.0852	346	University of Twente	utCASeleB09
33	0.0801	36	University of Tampere	utampere_auth_40_top10_low
34	0.0795	60	Saint Etienne University	UJM_15508
35	0.0671	167	School of Electronic Engineering and Computer Science	09HcaBst
36	0.0653	167	School of Electronic Engineering and Computer Science	09LrnHcaB
37	0.0000	72	University of Minnesota Duluth	umd_bep_1

〈그림 1〉 2009년 Ad Hoc트랙의 Best in Context Task 실험결과[8)]

Book Search 트랙의 경우도 2007년, 2008년, 2009년 태스크가 조금 차이가 나는데, 2009년 태스크는 Book Retrieval Task, Focused Book Search Task, Active Reading Task, Structure Extraction Task로 4개의 태스크가 오픈되었는데, Ⅲ장에서 상세히 소개할 것이며, 그 외의 트랙은 어떤 실험을 하는 트랙인지 여기서 간략히 소개만 하겠다.

Efficiency 트랙[9)]은 실(real) 데이터와 실 질의에 대해 XML검색 접근법의 효과성(effectiveness)과 효율성(efficiency)을 평가하는 태스크이다. 이 태스크에서는 질의를 처리하는데 드는 하드웨어 비용(즉, CPU, 메모리, 디스크 등)과 소요된 시간(즉, Runtime, CPU 시간 등)을 측정하여 효율성(efficiency)을 평가하고, R-Precision(재현율-정확률) 등의 효과성(effectiveness)을 평가한다.

Entity-Ranking[10)] 트랙은 토픽에 대해 적합한 문서나 엘리먼트(elements)를 결과로 제출하는

8) Ad Hoc Track의 한 태스크인 Best in Context Task의 2009년 검색 성능 실험 결과 페이지로 2009년 11월 11일 배포되었음.
〈http://www.inex.otago.ac.nz/tracks/adhoc/results.asp?task=bic&year=2009〉 [cited 2009. 11. 11].

9) Efficiency Track, 〈http://www.inex.otago.ac.nz/tracks/efficiency/efficiency.asp〉 [cited 2009. 10. 18].

10) Entity-Ranking Track,

대신에 엔티티(entity)를 추출하는 검색시스템의 기술을 평가하는 트랙이다. Interactive 트랙(iTrack)[11]은 검색 시, 이용자 행태를 분석하는 트랙으로 XML문서 내의 엘리먼트(elements)와 이용자가 상호작용할 때 이용자 행태를 연구함으로 이용자 환경에서 효과적인 엘리먼트(elements) 검색에 대한 접근법을 개발하는 태스크이다.

QA 트랙[12](QA@INEX Track)은 위키피디어를 사용하여 학술적 질문에 답변하는 태스크로, 질문 유형도 분석하여 적절한 단락(passages)이나 XML 엘리먼트(elements)를 검색하여 답변을 구성해야 한다. 질의의 유형도 간결한 것과 복잡한 것 두 가지 유형으로 나뉘는데, 간결한 질의의 경우에는 짧은(Short) 답변 즉 질의에 대한 답변을 포함하고 있는 엘리먼트(elements)나 단락(passages)을 순위하하여 리스트만 제공하면 된다. 복잡한 질의의 경우에는 긴(Long) 답변 즉 복수개의 문서 속에 있는 몇 개의 단락(passages)을 합쳐서 최대 500words 범위내로 답변을 재구성하여 결과를 제공해야 한다. 이 태스크는 단락(passages)추출에 의한 자동요약시스템에 도움을 줄 수 있다.

Link-the-Wiki 트랙[13]은 문서간의 링크를 발견하는 태스크로 텍스트를 분석하여 콜렉션 내 문서의 시작문서(anchor-text)에서 BEP(Best Entry Point)까지의 인(incoming)링크와 아웃(outgoing)링크의 집합을 추천해 주는 태스크이다. 여기서 시작문서(anchor-text)는 목적(target)문서의 특정 위치(position)로 링크되고, BEP(Best Entry Point)은 참고해야할 자료를 읽기 시작해야 할 위치를 의미한다.

XML Mining 트랙[14]은 기계학습(Machine Learning)기술로 구조화된 데이터를 분류하고 클러스터링 하는 기술을 개발하는 태스크이다. Classification 태스크와 Clustering 태스크가 있는데, 문서들을 클러스터들로 그룹화하기 위해 비통제-분류 기술을 활용하는 태스크이다. 이 태스크에서 서로 다른 클러스터의 수는 100, 500, 1000, 2500, 5000, 10000이다.

해마다 테스트컬렉션의 문서집합이 추가되거나 바뀌는데, Ad Hoc 트랙의 문서집합은 2006년 이전까지 IEEE Computer Society의 1995-2004년간 출판물에 포함된 복잡한 XML 구조를 지닌 논문기사 약 17,000건을 사용하였고, 2006년 이후부터는 위키피디어 문서가 실험문서로 지정되었으며, 2007년과 2008년에는 659,338건의 위키피디어 기사를 사용했고, 2009년에는 문서가 많이 추가되어 이미지가 없는 위키피디어 컬렉션의 2,666,190개 문서를 대상으로 실험이 진행되었다. 현재 Book Search 트랙을 제외한 나머지 모든 트랙은 위키피디어 컬렉션을 테스트컬렉션 문서집합으로 사용하고 있고, Book Search 트랙은 XML로 구성된 약 5만권의 도서를 테스트컬렉션 문서집합으로 사용하고 있다.

〈http://www.inex.otago.ac.nz/tracks/entity-ranking/entity-ranking.asp〉 [cited 2009. 10. 18].

11) Interactive Track, 〈http://www.inex.otago.ac.nz/tracks/interactive/interactive.asp〉 [cited 2009. 10. 18].

12) QA Track, 〈http://www.inex.otago.ac.nz/tracks/qa/qa.asp〉 [cited 2009. 10. 18].

13) Link-the-Wiki Track, 〈http://www.inex.otago.ac.nz/tracks/wiki-link/wiki-link.asp〉 [cited 2009. 11. 1].

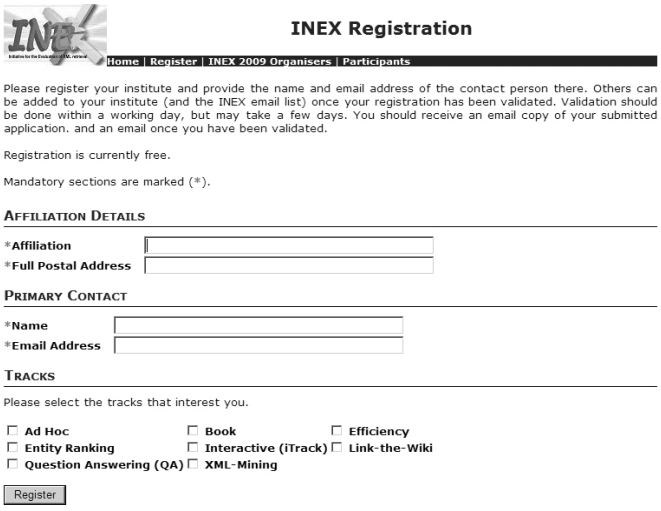
14) XML Mining Track, 〈http://www.inex.otago.ac.nz/tracks/wiki-mine/wiki-mine.asp〉 [cited 2009. 11. 1].

2. 실험트랙 참여방법 및 실험절차

가. 실험트랙 참여방법

INEX 실험트랙 참여[15] 자격에 특별한 제한은 없고 누구나 관심을 가진 기관 및 그룹, 개인이 등록(Register)[16]하여 참여할 수 있다. 등록을 하게 되면 INEX 등록 관리자가 등록 검증과정을 거쳐 승인을 하고, 승인 후 아이디와 패스워드를 이메일(E-mail)로 보내온다.

등록 화면은 다음 〈그림 2〉와 같고 등록 시, 관심 있는 트랙에 체크를 하면, 관심트랙의 일정에 따라 이루어지는 실험에 관한 모든 내용을 이메일(E-mail)로 제공받을 수 있고, 또한 관심트랙의 참여자들이 실험을 진행하면서 발생된 문제점 및 의견을 상호 교환할 수 있다.



〈그림 2〉 INEX 참가 등록 화면[17]

아이디와 패스워드를 부여받으면 그 때부터 정식회원이 되어 INEX에 로그인을 할 수 있게 되고, 로그인이 되면 INEX 전체 실험트랙 목표, 오픈태스크 및 일정, 각 트랙별 테스트 컬렉션 등 모든 관련정보를 확인할 수 있다.

정식 회원이 되면 각 참여자는 크게 세 가지 활동을 수행해야 한다. 첫째, 참여 트랙의 일정에 맞추어 배포되는 토픽 개발지침을 확인하고 포맷에 따라 토픽을 개발하여 제출해야 한다. 둘째, 각자의 검색시스템을 통해 태스크를 수행하고 태스크 제출포맷에 따라 검색결과를 제출하여야 한다.

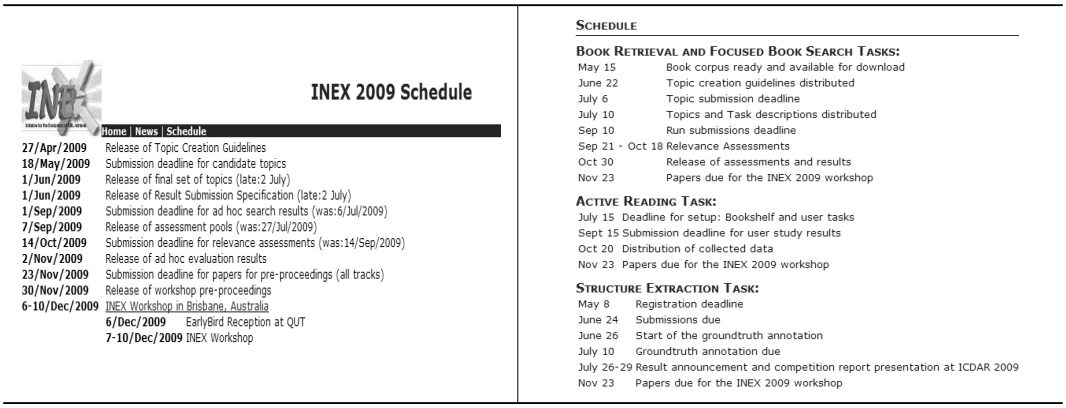
15) INEX 소개 및 참여 안내 페이지, 〈http://www.inex.otago.ac.nz/about.html〉 [cited 2009. 11. 1].
16) INEX 로그인 및 등록 연결 페이지, 〈http://www.inex.otago.ac.nz/login.asp〉 [cited 2009. 11. 1].
17) INEX 회원 등록 페이지, 〈http://www.inex.otago.ac.nz/people/register.asp〉 [cited 2009. 11. 1].

셋째, 여러 참여자가 제출한 토픽으로 적합성평가에 참여해야 한다.

회원 가입 후 요구되는 모든 활동에 전혀 참여하지 않으면 실험에 필요한 서비스를 제공받을 수 없다고 명시하고 있다. 하지만 INEX가 요구하고 있는 모든 과정에 참여할 여건이 되지 않는다면 참여 가능한 토픽개발 및 적합성 평가 등에만 참여해도 무방하다.

나. 실험절차

실험절차는 보통 INEX 실험일정에 따라 진행되며, 실험일정은 트랙마다 약간의 차이가 있지만, 2009년에 공개된 일정은 〈그림 3〉과 같다. 왼쪽 부분은 전체 일정이고, 오른쪽 부분은 Book Search 트랙의 일정이다.



INEX 2009 Schedule

Home | News | Schedule

27/Apr/2009	Release of Topic Creation Guidelines
18/May/2009	Submission deadline for candidate topics
1/Jun/2009	Release of final set of topics (late:2 July)
1/Jun/2009	Release of Result Submission Specification (late:2 July)
1/Sep/2009	Submission deadline for ad hoc search results (was:6/Jul/2009)
7/Sep/2009	Release of assessment pools (was:27/Jul/2009)
14/Oct/2009	Submission deadline for relevance assessments (was:14/Sep/2009)
2/Nov/2009	Release of ad hoc evaluation results
23/Nov/2009	Submission deadline for papers for pre-proceedings (all tracks)
30/Nov/2009	Release of workshop pre-proceedings
6-10/Dec/2009	INEX Workshop in Brisbane, Australia
6/Dec/2009	EarlyBird Reception at QUT
7-10/Dec/2009	INEX Workshop

SCHEDULE

BOOK RETRIEVAL AND FOCUSED BOOK SEARCH TASKS:

May 15	Book corpus ready and available for download
June 22	Topic creation guidelines distributed
July 6	Topic submission deadline
July 10	Topics and Task descriptions distributed
Sep 10	Run submissions deadline
Sep 21 - Oct 18	Relevance Assessments
Oct 30	Release of assessments and results
Nov 23	Papers due for the INEX 2009 workshop

ACTIVE READING TASK:

July 15	Deadline for setup: Bookshelf and user tasks
Sept 15	Submission deadline for user study results
Oct 20	Distribution of collected data
Nov 23	Papers due for the INEX 2009 workshop

STRUCTURE EXTRACTION TASK:

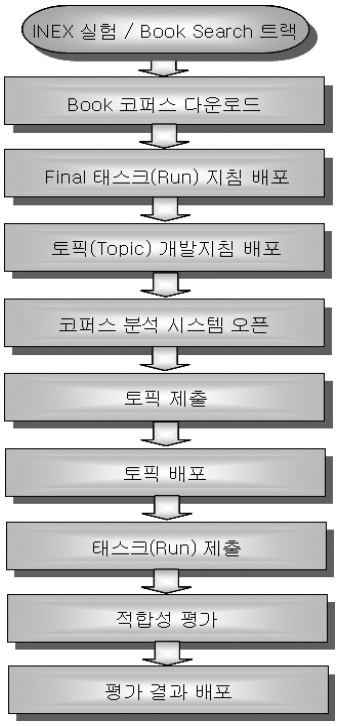
May 8	Registration deadline
June 24	Submissions due
June 26	Start of the groundtruth annotation
July 10	Groundtruth annotation due
July 26-29	Result announcement and competition report presentation at ICDAR 2009
Nov 23	Papers due for the INEX 2009 workshop

〈그림 3〉 INEX 2009 Schedule[18] 및 Book Search Schedule[19]

보통 트랙의 시작이 4월 말경인데, Book Search 트랙의 검색 태스크인 Book Retrieval 태스크와 Focused Book Search 태스크의 일정은 전체 일정에 비해 조금 늦게 시작되어 2009년에는 5월 중순에 시작되었다. 일정에 따르면 5월에 도서코퍼스를 다운받고, 6월 말에는 토픽개발지침이 배포되어, 7월 초에 지침에 따라 토픽을 개발하여 제출하고, 제출된 토픽들은 수집되어 참여자에게 재-배포 된다. 그리고 곧 태스크제출지침이 배포되어 9월 중순까지 실행(Run)[20]을 제출하도록 되어 있다. 그리고 9월 중순에서 10월 중순까지 적합성 평가를 진행하고, 10월 말경에 평가결과를 제공받도록 되어 있다. 하지만 2009년에도 연초에 공개된 전체 일정보다 조금씩 더 늦어지고 있다.

18) INEX 2009 Schedule, 〈http://www.inex.otago.ac.nz/schedule/schedule.asp〉 [cited 2009. 10. 11].
19) INEX 2009 Book Search Schedule, 〈http://www.inex.otago.ac.nz/tracks/books/books.asp〉 [cited 2009. 10. 11].
20) run이란, 참여자가 동일한 테스트컬렉션으로 자신만의 검색기법을 통해 색인방법, 질의정의, 문서정의 등을 다르게 다양한 방법으로 정의하여 실험을 수행하되, 그 중 한 가지 방법으로 실험을 수행하여 결과를 제출하는 것을 한 개의 run이라 한다. 참여자는 보통 여러 개의 run을 제출할 수 있다.

일정에 따른 INEX 실험절차를 Book Search 트랙을 중심으로 정리하면 〈그림 4〉와 같다. Book Search 트랙의 각 실험절차에 나타나는 태스크의 실행(Run), 토픽(Topic), 코퍼스(Corpus)에 대한 상세한 내용은 다음 Ⅲ장에서 상세히 다룬다.



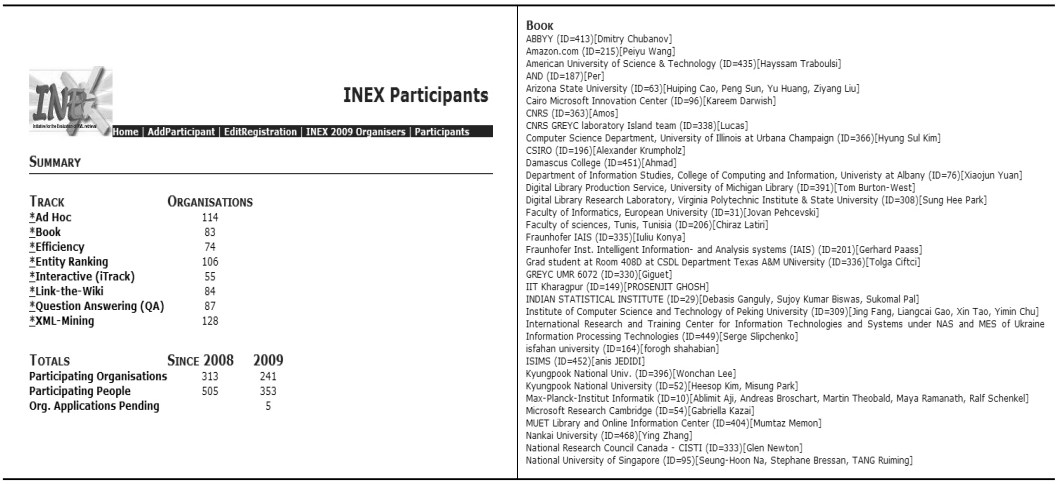
〈그림 4〉 INEX Book Search 트랙 실험절차

3. 참여기관

참여기관은 각 트랙별로 다른데, 자세한 것은 INEX 홈페이지 People메뉴의 Participants[21]에서 확인이 가능하다.

〈그림 5〉의 왼쪽 부분을 보면 각 트랙별로 참여기관을 표시하고 있는데, 2008년 참여기관수는 총 313개 기관이며 2009년에는 241개 기관으로 감소하였다. 그리고 참여인원도 2008년 505명에서 2009년 353명으로 감소하였다. Book Search 트랙의 총 참여기관수는 83개 기관인데, 그림의 오른쪽 부분이 Book Search 트랙의 참여기관 리스트의 일부이다. 각 기관마다 고유의 ID를 부여받게 되며, 실험결과 제출 시에도 ID로 각 태스크를 구분하게 된다.

21) 2009년 INEX 트랙 참여 기관 수, 〈http://www.inex.otago.ac.nz/people/participants.asp〉 [cited 2009. 11. 4].



INEX Participants

Home | AddParticipant | EditRegistration | INEX 2009 Organisers | Participants

SUMMARY

TRACK	ORGANISATIONS
*Ad Hoc	114
*Book	83
*Efficiency	74
*Entity Ranking	106
*Interactive (iTrack)	55
*Link-the-Wiki	84
*Question Answering (QA)	87
*XML-Mining	128

TOTALS	SINCE 2008	2009
Participating Organisations	313	241
Participating People	505	353
Org. Applications Pending		5

BOOK
ABBYY (ID=413)[Dmitry Chubanov]
Amazon.com (ID=215)[Peiyu Wang]
American University of Science & Technology (ID=435)[Hayssam Traboulsi]
AND (ID=187)[Per]
Arizona State University (ID=63)[Huiping Cao, Peng Sun, Yu Huang, Ziyang Liu]
Cairo Microsoft Innovation Center (ID=96)[Kareem Darwish]
CNRS (ID=363)[Amos]
CNRS GREYC laboratory Island team (ID=338)[Lucas]
Computer Science Department, University of Illinois at Urbana Champaign (ID=366)[Hyung Sul Kim]
CSIRO (ID=196)[Alexander Krumpholz]
Damascus College (ID=451)[Ahmad]
Department of Information Studies, College of Computing and Information, Univeristy at Albany (ID=76)[Xiaojun Yuan]
Digital Library Production Service, University of Michigan Library (ID=391)[Tom Burton-West]
Digital Library Research Laboratory, Virginia Polytechnic Institute & State University (ID=308)[Sung Hee Park]
Faculty of Informatics, European University (ID=31)[Jovan Pehcevski]
Faculty of sciences, Tunis, Tunisia (ID=206)[Chiraz Latiri]
Fraunhofer IAIS (ID=335)[Iuliu Konya]
Fraunhofer Inst. Intelligent Information- and Analysis systems (IAIS) (ID=201)[Gerhard Paass]
Grad student at Room 408D at CSDL Department Texas A&M UNiversity (ID=336)[Tolga Ciftci]
GREYC UMR 6072 (ID=330)[Giguet]
IIT Kharagpur (ID=149)[PROSENJIT GHOSH]
INDIAN STATISTICAL INSTITUTE (ID=29)[Debasis Ganguly, Sujoy Kumar Biswas, Sukomal Pal]
Institute of Computer Science and Technology of Peking University (ID=309)[Jing Fang, Liangcai Gao, Xin Tao, Yimin Chu]
International Research and Training Center for Information Technologies and Systems under NAS and MES of Ukraine
Information Processing Technologies (ID=449)[Serge Slipchenko]
isfahan university (ID=164)[forogh shahabian]
ISIMS (ID=452)[anis JEDIDI]
Kyungpook National Univ. (ID=396)[Wonchan Lee]
Kyungpook National University (ID=52)[Heesop Kim, Misung Park]
Max-Planck-Institut Informatik (ID=10)[Ablimit Aji, Andreas Broschart, Martin Theobald, Maya Ramanath, Ralf Schenkel]
Microsoft Research Cambridge (ID=54)[Gabriella Kazai]
MUET Library and Online Information Center (ID=404)[Mumtaz Memon]
Nankai University (ID=468)[Ying Zhang]
National Research Council Canada - CISTI (ID=333)[Glen Newton]
National University of Singapore (ID=95)[Seung-Hoon Na, Stephane Bressan, TANG Ruiming]

〈그림 5〉 2009년 INEX 참여기관

Ⅲ. Book Search 트랙

1. 테스트컬렉션(Test Collection)

테스트컬렉션은 앞에서도 잠시 언급했듯이 전통적으로 IR(Information Retrieval) 평가를 위해 사용되는 통제된 실험환경이라 할 수 있다. 대용량의 테스트컬렉션은 객관적인 실험환경을 보장함과 동시에 검색결과에 대한 신뢰성을 보장해 준다.

테스트컬렉션은 세 개의 부분 즉, 문서집합, 정보요구의 집합인 토픽(Topic), 적합성평가(relevance assessments)로 구성되는데, Book Search 트랙의 각 구성요소를 살펴보면 다음과 같다.

가. 문서집합(Book Corpus)

Book search 트랙의 문서집합인 도서코퍼스는 Microsoft Live Book Search와 Internet Archive[22]에 의해 비영리목적으로 제공되었다. 대부분 저작권의 유효기간(out-of-copyright)이 지난 1930년 이전 도서로, 2007년 코퍼스는 42,090책(210G)의 도서가 Djvu.xml 포맷으로 마크데이터와 함께 제공되었고, 2008년에 8,200여 책이 추가되어 50,239책이 제공되었다. 그리고 2009년 코퍼스는 2008년과 동일하다. 연도별 XML도서 원문구조는 〈그림 6〉과 같다.

22) Internet Archives 홈페이지, 〈http://www.archive.org/〉 [cited 2009. 11. 14].

2007년 Book Corpus[23] : 42,090책

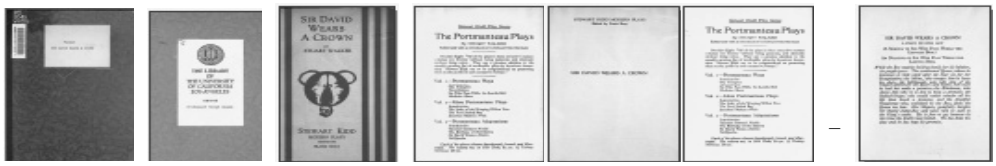
```
<DjVuXML>
 <BODY>
  <OBJECT data="file.." ...>
   <PARAM name="PAGE" value="..">
   [...]
   <REGION>
    <PARAGRAPH>
     <LINE>
      <WORD coords="..."/>
      <WORD coords="..."/>
     </LINE>
    </PARAGRAPH>
   </REGION>
   [...]
  </OBJECT>
  [...]
 </BODY>
</DjVuXML>
```

2008년/2009년 Book Corpus[24] : 50,239책

```
<document>
 <page pageNumber=''I-N'' label=''PT_CHAPTER'' [...]>
  <region regionType=''text'' [...]>
   <section label=SEC_BODY'' [...]>
    <line [...]>
     <word val=''Moby'' [...]/>
     <word val=''Dick'' [...]/>
    </line>
    <line [...]>
     <word val=''Herman'' [...]/>
     <word val=''Melville'' [...]/>
    </line>     [...]
   </section>   [...]
  </region>     [...]
 </page>        [...]
</document>
```

〈그림 6〉 Book Corpus XML 구조

2007년 도서구조와 2008/2009년 도서구조가 다른데, 2008/2009년 도서구조는 2007년 도서구조를 BookML(ocrml.xml) 포맷으로 변환한 것이다. BookML은 Document Layout Team of Microsoft Development Center Serbia에 의해 개발된 포맷으로 목차(ToC) 엔트리를 포함하고 있는 등 2007년보다 더 풍부한 구조정보를 제공해 주고 있다. 2008/2009년 도서구조를 자세히 살펴보면 도서를 하나의 〈document〉 태그로 간주하고, 각 페이지는 〈page〉태그로, 〈page〉태그는 또 몇 개의 〈region〉태그로 구성되며, 각 〈region〉태그는 여러 개의 〈line〉 태그들로 구성되며, 〈line〉 태그는 〈word〉태그[25]들로 구성되어 있다. 한권의 샘플 도서 이미지인 〈그림 7〉과, 그 중 페이지의 일부를 확대하여 나타낸 도서구조의 내용인 〈그림 8〉을 통해 살펴보면 이미지와 원문의 구조를 확인할 있다.

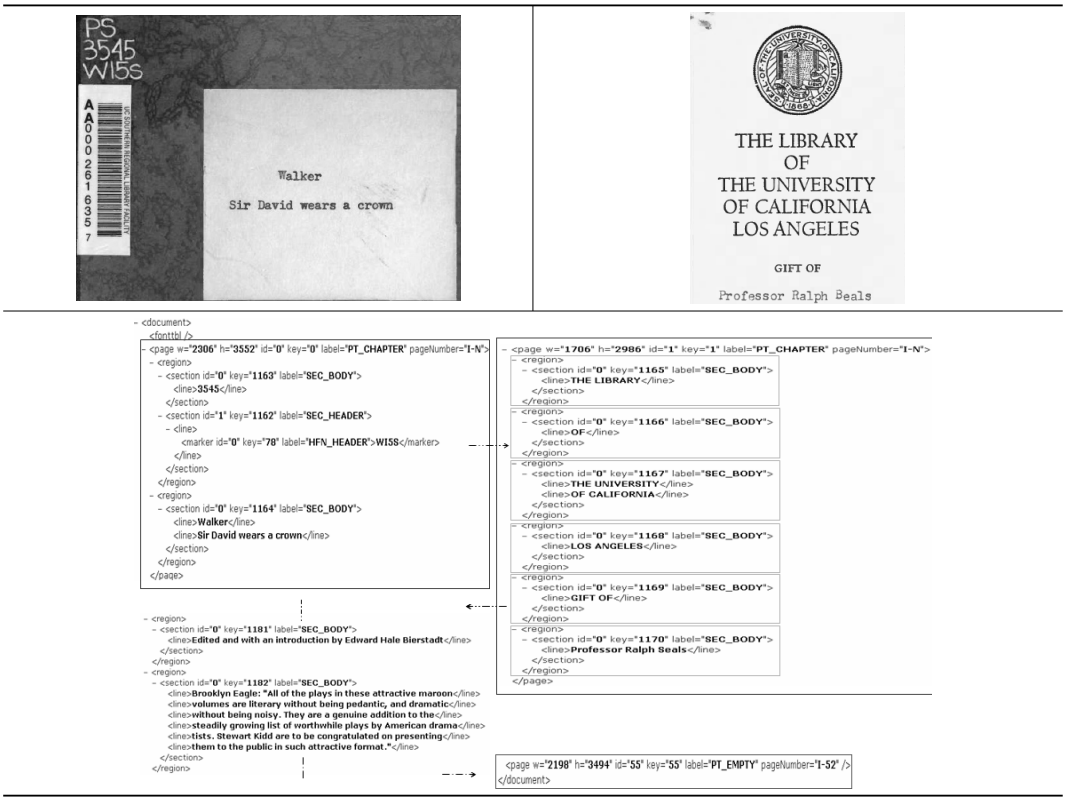


〈그림 7〉 "Sir david wears a crown"[26] 도서의 책 표지 등 일부 이미지

23) Gabriella Kazai and Antoine Ducet, "BookSearch'07 : INEX 2007 Book Search Track Overview," *Pre-poceedings of INEX 2007*(2007), p.178.

24) Gabriella Kazai, Antoine Ducet, and Monica Landoni, "Overview of the INEX 2008 Book Track," *INEX 2008 Workshop Pre-proceedings*(2008), p.169.

25) 〈word〉 태그와 그의 attribute(속성)들은 코퍼스 배포 시, 코퍼스의 크기를 줄일 목적으로 제거되어 배포되었다. 〈word〉 태그를 제거함으로써 코퍼스 크기가 400Giga에서 50Giga로 줄어들었음.



〈그림 8〉 "Sir david wears a crown"의 도서 상세 구조

Book Search 트랙의 문서집합인 도서 코퍼스(Book Corpus)는 위의 샘플도서와 같은 형식으로 구성된 도서 50,239권이 제공되었고, 부가적 정보로 마크데이터도 함께 제공되었다. 위 샘플 도서의 마크데이터는 다음 〈그림 9〉와 같다.

00624nam 2200217I
450000100080000000500170000800800410002501000170006603500200008303500160010304000180011905000250013710000300016224500510019226000490024330000340029249000460032691000110037293500140038391000090039734270861987091912433O.0770401s1922 ohuag 000 0 eng a 22006721 a(OCoLC)02853676 907-AMS-8412 aDLCclULdCLU0 aPS3545.A54bS47 19221 aWalker, Stuart,dd. 1941.10aSir David wears a crown,cby Stuart Walker ... aCincinnati :bStewart Kidd company,c[c1922] a47 p.billus. (music)c19 cm.0 aStewart Kidd modern plays, ed. by F. Shay arcp 46 aMC4172233 aMARS

〈그림 9〉 "Sir david wears a crown"의 마크(MARC)

26) "Sir david wears a crown" 도서의 원문 URL, 〈http://www.archive.org/details/sirdavidwearscro00walkiala〉 [cited 2009. 11. 14].

나. 토픽(Topics)

토픽이란 이용자들의 정보요구를 표현한 것인데, 정보요구의 집합은 검색시스템의 질의로 이용된다. 토픽들은 INEX가 제공하는 토픽개발지침에 따라 각 참여기관들에 의해 만들어진다. 2007년에 30개의 토픽이 배포되었는데, 그 중 일부 토픽은 Live search Books의 질의 로그로부터 추출된 것이고, 나머지는 참여기관이 토픽개발지침에 따라 직접 만든 것이다. 그리고 2008년에는 40개의 토픽이 새롭게 추가되었다. 2007년, 2008년의 토픽포맷은 동일하나, 2009년 토픽포맷은 좀 복잡하게 바뀌었다. 2009년에 제출된 총 토픽 수는 16개이다.

다음 〈그림 10〉은 2007년, 2008년 토픽의 DTD(Document Type Definition), 토픽샘플 그리고 각 토픽의 구성요소에 대한 부가 설명이다.

```
<!ELEMENT inex_topic (title,description,narrative)>
<!ATTLIST inex_topic
track CDATA #REQUIRED
task CDATA #REQUIRED
topic_id CDATA #REQUIRED
ct_no CDATA #REQUIRED
>
<!ELEMENT title (#PCDATA)>
<!ELEMENT description (#PCDATA)>
<!ELEMENT narrative (task,infneed)>
<!ELEMENT task (#PCDATA)>
<!ELEMENT infneed (#PCDATA)>
```

※ 그림 설명
1) 좌측 상단 : 토픽 DTD
2) 좌측 하단 : 토픽 샘플[27]
3) 우측 하단 : 토픽 구성요소 설명

```
<title>The constellation</title>
<description>I am looking for books and its pages talking about star
groupings in astronomical aspect.</description>
<narrative>
<task>I want to find detailed information on constellations including
their names, figure, pattern, direction, place in the sky, and stories
behind them to present in my astronomy class at middle school as
a result of homework.</task>
<infneed>I am looking for constellation name and their figure, pattern,
direction, place in the sky and stories behind them. Any texts
containing information mentioned above (i.e., figure, pattern,
direction, place in the sky) are relevant. And stories about each
constellation from different countries are also relevant. However,
for other uses of constellations and any information about its
history, magnitude and distance are irrelevant.</infneed>
</narrative>
```

<title>	- 검색엔진에서 사용할 검색 질의 표현 - 이용자 정보 요구 내용의 요약
<description>	- 정보요구에 대한 자연어 정의
<narrative>	- 정보요구에 대한 상세 설명 : 각 요소(element)가 적합/부적합인지 기술 - narrative : 도서 속의 특정 text fragment가 요구를 충족/충족하지 않은지를 결정하기 위해 정보요구에 관한 명백하고 정확한 기술을 포함해야 함 - 적합성 평가 : narrative로만으로 이루어짐 - narrative는 찾고자하는 정보가 무엇인지, 정보요구에 대한 정황과 동기를 설명 : 정보를 왜 찾아야 하는지 : 어떤 work-task가 해결에 도움을 줄 수 있는가 등 - narrative는 <task>와 <infneed>를 두 부분을 포함
<task>	- 정보 요구를 위한 정황, 배경, 동기 열거 - 어떤 정보를 찾고 있는지 태스크 기술
<infneed>	- 찾고 있는 정보에 대한 상세 설명 - 무엇이 적합/부적합한지에 대한 상세 설명

〈그림 10〉 토픽 DTD, 샘플 및 구성 요소 설명

토픽을 개발할 때, 몇 가지 고려사항이 있는데 첫째, 토픽은 반드시 코퍼스의 주제를 커버해야 한다는 것이다. 둘째, 토픽은 실제 정보요구를 정확하게 반영해야 한다. 셋째, 토픽이 다양한 주제를 다루되 너무 광범하거나, 너무 좁은 개념이어서는 안 된다.

위의 고려 사항을 만족하는 토픽을 개발하려면 코퍼스의 내용을 어느 정도 파악해야 정확한 토픽을 만들 수 있다. 그러므로 이를 돕기 위해, 앞의 〈그림 4〉 INEX Book Search 트랙 실험절차에

27) 본 논문의 저자가 개발하여 제출한 2008년 토픽 중 하나이다.

서 토픽개발지침을 배포한 후, 코퍼스 분석을 위한 검색시스템[28]을 오픈해 준다. 아울러 토픽개발지침의 부록 부분에 Microsoft Live Book Search의 질의 로그로부터 추출된 질의 집합을 제공해 준다. 코퍼스 규모가 너무 커서, 토픽을 만들 때 이 부분을 참조하여 토픽의 주제를 찾아나가면 토픽개발에 도움이 된다.

2009년 Book search 트랙의 토픽개발지침[29]에는 2007년, 2008년과 다른 큰 변화가 생겼는데, 토픽개발 시, 위키피디어 기사를 사용하라는 것이다. 이유는 위키피디어 기사들이 종종 기사의 토픽과 관련 있는 도서의 리스트들을 포함하고 있어서 기사의 특정 부분에 관련된 도서들을 인용할 수도 있고, 이는 실용적인 토픽개발에 좋은 정보를 줄 수 있다는 것이다. 그리고 위키피디어 기사를 브라우징 함으로써 검색을 위해 사용될 더 적합한 주제, 소-주제(aspect) 및 용어를 선정하는데 통찰력을 제공해 줄 수 있기 때문이다.

다음 〈그림 11〉은 2009년에 변경된 토픽포맷이다. 기존 토픽과 거의 유사하고 박스 속의 부분이 새롭게 추가된 내용이다. 〈wikipedia-title〉, 〈wikipedia-url〉, 〈wikipedia-text〉태그는 차례로, 토픽에 관련된 위키피디어 기사의 타이틀, url, 내용부분을 기술한 것이다. 〈aspect〉태그는 토픽을 소 관점으로 나누어 기술한 것이다. 〈aspect〉태그는 소 관점을 좀 더 상세히 기술하기 위해 다시 〈title〉, 〈narrative〉, 〈wikipedia-text〉태그로 나누어 기술된다. 토픽개발 시에 토픽은 적합한 도서가 적어도 2개 이상이 되어야 하고, 가능한 20개 이상 갖도록 만들되 각 토픽 당 적어도 2개의 소-주제(aspects)를 갖도록 권하고 있다.

<topic>
<task> My task is to add citations to the Wikipedia article on Donations of Alexandria referring to books and parts of books relevant to the topic and specific aspects of the topic. </task>
<title> Donations of Alexandria </title>
<description> I am looking for information on the political statement by Mark Antony that is referred to as the Donations of Alexandria in which he distributed lands held by Rome and Parthia amongst Cleopatra and their children. </description>
<narrative> Any information relating to the donations, their motivation and political background, and their consequences are relevant. Data on what donations and titles were given to whom are relevant. Details of the ceremony when the donations were announced are also relevant. Octavian's response and the political situation that arose as a result of the donations, leading to the civil war, are all relevant.
</narrative>
<wikipedia-title> Donations of Alexandria </wikipedia-title>
<wikipedia-url> http://en.wikipedia.org/wiki/Donations_of_Alexandria </wikipedia-url>
<wikipedia-text> [text of the Wikipedia page] </wikipedia-page>

<aspect>
<title> The Donations </title>
<narrative> Of relevance are the details of the donations, i.e., the ceremony and the particulars of who got what. </narrative>
<wikipedia-text> [text of the "The Donations" section from the Wikipedia page] </wikipedia-text>
</aspect>

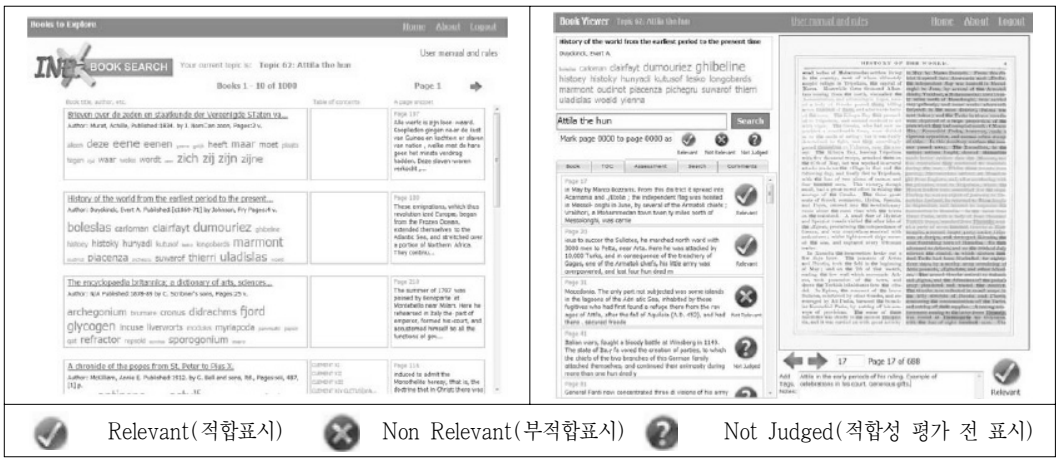
<aspect>
<title> Consequences </title>
<narrative> Any information that details how the donations were received and what reactions they generated, in particular how Octavian reacted. Descriptions of the political atmosphere are relevant as well as events leading to the civil war. The civil war itself is however not relevant. </narrative>
<wikipedia-text> [text of the "The Donations" section from the Wikipedia page] </wikipedia-text>
</aspect>
</topic>

〈그림 11〉 2009년 변경된 토픽 포맷 예[30]

28) INEX Book search system, 〈http://www.booksearch.org.uk/Default.aspx〉 [cited 2009. 11. 8].
29) 2009년 토픽개발지침 문서, 〈http://www.inex.otago.ac.nz/tracks/books/TopicDevGuideImages/INEX2009BookTrack-TopicDevGuide.pdf〉 [cited 2009. 11. 8].

다. 적합성 평가

Microsoft Research Cambridge가 개발한 Book Search System은 앞에서 언급했듯이 토픽개발 전에 오픈되어 토픽개발에도 사용되고, 적합성평가에도 사용된다. 이 시스템은 〈그림 12〉와 같이 직접 도서의 내용을 검색 및 브라우징 하는 기능을 제공하고, 매 페이지를 읽어보고 해당 페이지가 토픽의 내용에 적합한지, 적합하지 않은지 표시할 수 있는 기능도 함께 제공해 준다. 2008년 적합성평가의 경우에 많은 참여자의 참여를 유도하기 위해 Book Explorers' Competition[31]을 개최하였다. 방법은 참가신청을 한 후, 로그인을 하면 ID별로 토픽 리스트가 나타난다. 참가자는 토픽 리스트를 살펴본 후, 검색해 보고 싶은 토픽을 최소 한 가지 이상 선택하면 해당 토픽 당 1000개의 도서 리스트가 나타난다. 도서 리스트들은 참가자가 적합성 판단을 해야 할 도서목록들이다. 참가자는 표시된 해당 도서와 도서별로 나타난 해당 페이지를 읽어보고 그 내용이 토픽에 적합한 부분인지 아닌지 적합성 유무를 체크함으로 적합성평가가 이루어진다. 적합성평가는 토픽 당 1000개의 도서를 평가해야 하므로 매우 힘든 작업 중 하나이다.

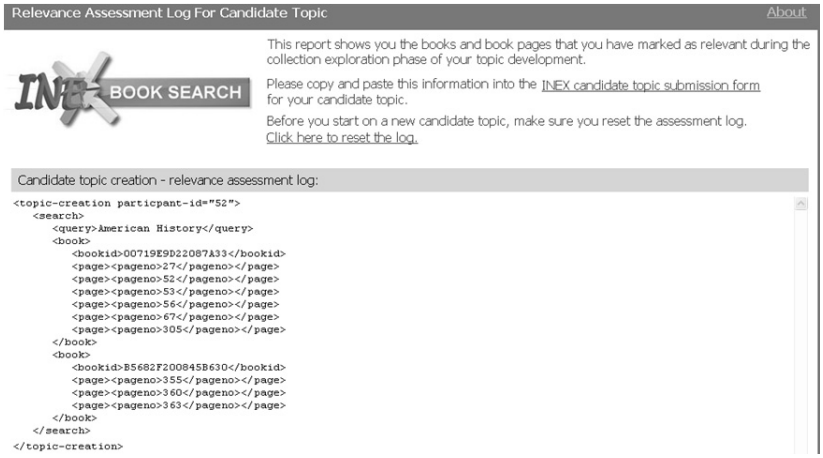


〈그림 12〉 적합성 평가

이상과 같이 적합성평가가 이루어지면 시스템은 각 참가자가 적합하다고 판단한 도서와 도서의 각 부분에 대해 〈그림 13〉과 같이 시스템 로그를 남기게 된다. 이는 향후에 IR 평가(Evaluation)를 위해 사용될 유용한 정보이다.

30) *Ibid.*, p.3. ※ 참고 : INEX 토픽개발지침 문서에서 그대로 인용된 예제이나 원 문서에 오류가 있음. 토픽 포맷 박스 내 셋째 줄 〈/wikipedia-page〉 태그는 〈/wikipedia-text〉로 바뀌어야 함.

31) Book Explorers' Competition은 2008년 12월 1일에서 15일까지 진행되었고, 많은 참여를 유도하기 위해 Microsoft Research가 HDD, USB, Mouse 등 다양한 상품을 후원하였음.



〈그림 13〉 적합성 평가 시스템 로그

2. 실험태스크 및 실행(Run) 제출지침

Book Search 트랙은 Ⅲ장에서 설명한 테스트컬렉션(Test Collection)이란 통제된 실험환경 하에 검색실험을 수행한다. Book Search 트랙의 목표는 디지털화된 도서에 대한 서비스 즉, 도서 브라우징, 검색, 주석달기 등의 기능을 제공하는 기술을 개발하고 평가하며, 제공된 서비스에 대한 이용자 관련 이슈 및 행태를 연구하는 트랙이다. 검색실험을 위한 태스크도 해마다 조금씩 다른데, 2007년, 2008년, 2009년 태스크를 살펴보면 〈표 2〉와 같다.

〈표 2〉 2007/2008/2009년 Book Search 트랙의 실험태스크

2007년	2008년	2009년
Book Retrieval Task Page in Context Retrieval Task	Book Retrieval Task Page in Context Retrieval Task Structured Extraction Task Active Reading Task	Book Retrieval Task Focused Book Search Task Structure Extraction Task Active Reading Task

〈표 2〉의 태스크들은 Book Search 트랙의 연도별 목표수행을 위한 태스크인데, Book Retrieval Task와 Focused Book Search Task(or Focused Book Search Task)는 검색 태스크이고, Structured Extraction Task는 목차 및 도서구조를 추출하는 태스크이고, Active Reading Task는 이용자행태를 연구하는 태스크이다. Structured Extraction Task와 Active Reading Task는 간단히 정의만 살펴보고, 두 개의 검색 태스크에 대해서는 태스크의 목표와 실험 시나리오 및 태스크, 태스크 수행 후의 결과제출포맷 등에 대해 상세히 살펴본다.

가. Book Retrieval Task

이 태스크는 검색시스템에서 토픽에 가장 적합한 도서의 리스트를 순위화하여 결과를 제출하는 도서 단위 검색 태스크이다.

(1) 목표

특정 도메인 도서에 대해 Book-Specific IR 기술과 표준 IR기술을 비교하고 성능을 확인하는 태스크이다. 즉 Book-Specific IR 기술 적용이 표준 IR기법에 비해, 특정 도메인에 대한 색인 및 검색에 개선된 비용(cost-증가된 처리 시간 및 저장 공간 등)과 이익(Benefit-개선된 검색 효율성)을 가져오는지 측정하는 것이다.

(2) 사용자 시나리오(User Scenario)

주어진 토픽에 대해 연구 목적, 강의 목적, 또는 재미거리로 읽을 만한 도서 그리고 구매하기 위한 도서, 도서관에서 빌릴만한 도서의 Reading List나 Reference List를 구축하기 위해 도서를 찾는 것이다.

(3) 태스크 기술(Task Description)

사용자 정보요구 즉, 토픽에 적합하다고 판단한 도서의 순위화 된 리스트를 결과로 제출한다. 도서 도메인에 한정되지 않은 IR기술 및 Book-Specific IR기술을 사용할 수 있다. 그리고 참여자는 태스크를 수행하되 single run[32] 또는 pairs of runs[33]을 제출 가능하되 최대 10개까지 제출할 수 있고, 최소 1개의 automatic run[34]은 의무적으로 제출하여야 한다.

각 토픽에 대해, 토픽의 타이틀 부분만 사용할 수 있는 의무적 automatic run을 제외한 다른 실행(run)들은 토픽의 어떤 부분이라도 사용해서 결과를 제출할 수 있다. 한 가지 주의할 것은 relevance(적합성) 평가를 위해 주어진 토픽 당 적합한 도서 중에서, 상위에 랭크된 최대 1,000책까지 포함해서 제출할 수 있다.

(4) 제출 포맷(Submission Format)

참여자는 각자의 검색시스템으로 태스크를 수행하고, 그 결과를 제출해야 하는데 제출 포맷은 태스크제출지침[35]에 따라야 한다. 제출 DTD와 예시를 보이면 〈그림 14〉와 같다.

---

32) single run이란, Standard IR 방법 또는 Book-Specific IR 방법 중 한 가지 유형만을 적용한 결과를 제출하는 실험.
33) pairs of runs이란, Standard IR 방법의 결과와 Book-Specific IR 방법 즉, 도서의 특징 정보를 활용하여 특별히 튜닝된 방법을 사용한 결과를 함께 제출하는 것.
34) automatic run이란, 검색을 위해 인간의 간섭 없이 토픽의 title 부분만 사용하여 실험한 run.
35) 2009년 태스크제출지침 문서.

제출 DTD

```
<!ELEMENT bs-submission (topic-fields, description, topic+)>
<!ATTLIST bs-submission
participant-id      CDATA      #REQUIRED
run-id              CDATA      #REQUIRED
paired-run-id       CDATA      #REQUIRED
task                (book-retrieval) #REQUIRED
query               (automatic | manual) #REQUIRED
result-type         (book)      #REQUIRED
retrieval-type      (non-specific | book-specific) #REQUIRED
>
<!ELEMENT topic-fields EMPTY>
<!ATTLIST topic-fields
title               (yes|no) #REQUIRED
description         (yes|no) #REQUIRED
narrative           (yes|no) #REQUIRED
aspects             (yes|no) #REQUIRED
>
<!ELEMENT description (#PCDATA)>
<!ELEMENT topic (book+)>
<!ATTLIST topic topic-id CDATA #REQUIRED >
<!ELEMENT book (bookid, rank?, rsv?)>
<!ELEMENT bookid   (#PCDATA)>
<!ELEMENT rank     (#PCDATA)>
<!ELEMENT rsv      (#PCDATA)>
```

제출 예시

```
<bs-submission participant-id="25" run-id="BM25F-With-ToC-BackOfBookIndex-Streams"
               paired-run-id="BM25" task="book-retrieval" query="automatic"
               result-type="book" retrieval-type="book-specific">
 <topic-fields title="yes" description="no" narrative="no" aspects="yes"/>
 <description>BM25F using 2 streams extracted from the table of contents and the back-of-book
               index sections of books. The rest of the book content is ignored. Parameters of
               BM25F were trained using RankNet.</description>
<topic topic-id="01">
 <book>
  <bookid>300A5334B2869F47</bookid>
  <rank>1</rank>
 </book>
 <book>
  <bookid>BAD598FB0A7D02E2</bookid>
  <rank>2</rank>
 </book>
 <book>...</book>
...
</topic>
<topic> ... </topic>
</bs-submission>
```

〈그림 14〉 Book Retrieval Task의 제출 DTD[36] 및 예시

제출 DTD에 대해 부가설명을 하면 @participant-id는 참여기관이 부여받은 기관식별 아이디이고, @run-id는 한 기관에서 보내진 모든 제출결과를 식별해 주는 실행(run) ID로, 의미를 부여해서 사용하되 가능한 짧은 이름을 사용하길 권한다. @paired-run-id는 현재 제출한 실행(run)이 pairs of runs일 경우의 식별 실행(run) ID이다. 만약 single run을 제출했다면 paired-run-id='NA'로 처리한다. @task는 태스크를 식별하는 것으로 이 태스크에서는 'book-retrieval'만 사용 가능하다. @query는 검색 query에 대한 스펙인데 검색 query를 토픽으로부터 자동으로 구축 시 'automatic'이고, 수동으로 구축 시 'manual'로 세팅한다. @result-type은 결과 유형에 대한 스펙으로 이 태스크에서는 'book'으로 세팅한다. @retrieval-type은 검색 유형에 대한 스펙인데 Standard IR방법인 경우는 'non-specific'으로 세팅하고 Book-specific 특징 및 알고리즘을 사용한 방법의 경우에는 'book-specific'으로 세팅한다. topic-fields는 검색 질의어를 구축하기 위해 사용된 토픽필드에 대한 스펙으로, 'title', 'description', 'narrative', 'aspects'가 될 수 있다. description은 실행(run)을 생성하기 위해 적용된 검색 접근법에 관한 상세 기술인데, 이것은 추후 결과 비교 및 분석에 도움을 준다. topic은 주어진 토픽에 적합하다고 평가된 순위화 된 도서들의 리스트를 포함한 것으로 적합성 내림 값의 순으로 정렬되며, 각 토픽 당 최대 1000개를 리턴 한다. @topic-id는 토픽식별 ID이며, book은 순위화가 있는 각 도서에 대한 정보이고, bookid는 도서를 식별하기 위한 ID로 도서의 XML소스(마크 메타파일과 함께)가 있는 디렉터리의 이름이며, rank는 도서의 적합성 순위 정보이다.

〈http://www.inex.otago.ac.nz/tracks/books/2009BookTrack-SearchTasksGuide.pdf〉 [cited 2009. 11. 15].

36) 제출 DTD에서 topic-fields의 애트리뷰터 리스트에 aspects 부분을 제외하면 2008년의 제출 DTD와 동일함.

나. 2009 Focused Book Search Task(2007/2008 Page in Context Retrieval Task)[37]

Book Retrieval Task가 도서단위의 태스크라면 이 태스크는 토픽에 가장 적합한 도서의 부분(page, passage, element 등)을 검색해내는 태스크이다.

(1) 목표

디지털화된 도서를 대상으로 Focused 검색 접근법에 대한 응용을 조사하는 것으로 Ⅱ장에서 언급한 Ad Hoc 실험트랙과 문서집합(코퍼스)이 다를 뿐, Relevant in Context 태스크와 유사하다. 이는 도서내의 도서 부분에 적합성 순위를 매기는 것이다.

(2) 사용자 시나리오(User Scenario)

이 태스크의 시나리오는 주어진 토픽에 대해 적합한 정보를 포함하고 있는 도서의 부분을 직접 찾아주는 것이다. 찾고자하는 정보가 책 속에 숨겨져(hidden) 있을 수도 있고 메인 주제로 드러나 있을 수도 있는데, 이용자는 적합한 도서의 부분을 직접 지시해 주기를 기대하고 있다.

(3) 태스크 기술(Task Description)

주어진 토픽에 적합한 정보를 포함하고 있는 도서의 부분을 확인하고 순위화 하여 도서별로 그룹화하여 결과를 제출한다. 즉, 사용자에게 순위화된 도서리스트와 각 도서별 적합한 부분의 순위 정보를 제공하는 것이다. 2009년 토픽에 aspect 부분이 추가되었기 때문에 2009년의 이 태스크에서는 질의로 토픽의 aspect만 사용하도록 하였다.

도서의 순위는 도서의 부분(page, passage, element)의 최고점수(best score) 또는 평균점수(average score) 또는 그 어떤 다른 점수 계산 방법으로도 산출할 수 있으며 적합성 내림차순으로 순위화 하여 제출한다. Run을 최대 10개까지 제출가능 하되 1개의 automatic run과 1개의 manual run[38]을 의무적으로 제출해야 한다. 제출 시, 유의할 점은 상위 1,000책에 대해, 각 도서별로 도서 부분의 순위화 정보를 제공하되 중첩되지 않는 도서 페이지(book page), 단락(passages), 엘리먼트(elements)의 순위리스트(ranked list)를 적합성 내림차순으로 정렬해야 한다. 그리고 결과제출 유형도 혼합될 수 없다. 즉 페이지(page), 단락(passages), 엘리먼트(elements) 중 하나를 선택해서 제출해야 한다.

(4) 제출 포맷(Submission Format)

이 태스크의 제출포맷은 〈그림 15〉와 같고, Book Retrieval Task와 유사한데, 가장 큰 차이점

37) 2007/2008년의 Page in Context Retrieval Task와 2009년의 Focused Book Search Task는 명칭이 다르나, 동일한 태스크임.

38) 사람의 간섭을 허락하는 run.

은 Book Retrieval Task가 도서ID, rank(순위)를 제출하는데 비해, 이 태스크는 도서ID와 rank(순위), 도서부분(/document[1]/page[1]/title[1], /document[1]/page[1]/section[1])과 도서부분의 순위를 함께 제출해야 한다는 것이다. 도서 부분을 기술할 때는 주어진 Xpath문법 규칙[39]에 맞게 기술해야 한다.

제출 DTD

```
<!ELEMENT bs-submission (topic-fields, description, topic+)>
<!ATTLIST bs-submission
participant-id          CDATA       #REQUIRED
run-id                  CDATA       #REQUIRED
task                    (book-ad-hoc) #REQUIRED
query                   (automatic | manual) #REQUIRED
result-type             (element | passage | page)      #REQUIRED
>
<!ELEMENT topic-fields EMPTY>
<!ATTLIST topic-fields
aspect-title            (yes|no) #REQUIRED
aspect-narrative        (yes|no) #REQUIRED
topic                   (yes|no) #REQUIRED
>
<!ELEMENT description (#PCDATA)>
<!ELEMENT topic-aspect (book+)>
<!ATTLIST topic-aspect
        aspect-id CDATA #REQUIRED >
<!ELEMENT book (bookid, rank?, rsv?, result+)>
<!ELEMENT result ((path|passage), rank?, rsv?)>
<!ELEMENT bookid    (#PCDATA)>
<!ELEMENT path      (#PCDATA)>
<!ELEMENT passage EMPTY>
<!ATTLIST passage
        start           (#PCDATA) #REQUIRED
        end             (#PCDATA) #REQUIRED
>
<!ELEMENT rank      (#PCDATA)>
<!ELEMENT rsv       (#PCDATA)>
```

제출 예시

```
<bs-submission participant-id="25" run-id="BM25F-Focused-PageLevelRetrieval-With-ToC-
                BackOfBookIndex-Streams" task="book-ad-hoc" query="automatic"
                result-type="page">
 <topic-fields aspect-title="yes" aspect-narrative="yes" topic="no"/>
 <description>BM25F using 2 streams extracted from the table of contents and the back-of-book
                index sections, indexing and retrieval only at page level, no relevance
                propagation</description>
<topic-aspect aspect-id="01">
  <book>
    <bookid>384D10DAEA4E34A8</bookid><rank>1</rank>
    <result><path>/document[1]/page[27]</path><rank>1</rank></result>
    <result><path>/ document[1]/page [122]</path><rank>2</rank></result>
    <result><path>/ document[1]/page [5]</path><rank>3</rank></result>
    ...
  </book>
  <book>
    <bookid>5AFEE130174076E3</bookid><rank>2</rank>
    <result><path>/ document[1]/page [531]</path><rank>1</rank></result>
    <result><path>/ document[1]/page [14]</path><rank>2</rank></result>
    ...
  </book>
  <book>...</book>
...
</topic-aspect>
<topic-aspect> ... </topic-aspect>
</bs-submission>
```

〈그림 15〉 2009년 Focused Book Search Task의 제출 DTD 및 예시

다. Structured Extraction Task

이 태스크는 검색 태스크가 아니라, 도서의 구조정보를 추출해내는 태스크이다.

(1) 목표

이 태스크의 목표는 하이퍼링크된 목차를 생성하기 위해 디지털화된 도서들로부터 구조정보를 유도해 내기 위한 자동기술을 테스트하고 비교하는 것이다. 현재 디지털 및 OCR 기술은 이미지 원문으로부터 최소한의 구조정보 즉, 페이지(page)나 패러그래프(paragraph) 정도는 OCR로 식별 및 마크업 되지만, 더 정교한 구조정보인 장(chapter), 절(section) 등은 인식하지 못한다. 이 태스크는 도서원문으로부터 좀 더 깊은 구조정보를 추출해내는 기술을 개발하기 위함이다.

(2) 태스크 기술(Task Description)

이 태스크는 OCR(DjVu XML 포맷), PDF, JPEG 이미지 파일로부터 정보를 사용하여 디지털화된 도서의 목차를 구축하는 것이다. 목차 구축을 위해 다른 장르 및 스타일의 도서 100책을 샘플컬렉션 즉 테스트 셋으로 사용한다. 그리고 테스트 셋 도서의 스캔된 페이지의 OCR output,

39) 2009년 태스크제출지침 문서, pp.7-8.

PDF 파일, JPEG이미지를 활용할 수 있다. 참여자는 10개의 실행(runs)까지 제출할 수 있고, 각 실행(run)은 테스트 셋의 모든 도서 목차를 포함해야 한다.

라. Active Reading Task

이 태스크는 이용자가 독서를 할 때 도서와 이용자 간의 이용행태를 연구하는 트랙이다. 즉 이용자가 전자책(e-book)을 어떻게 브라우징하고 또 어떤 편리함 때문에 어떤 시나리오를 가지고 접근하는지, 그리고 전자책(e-book)을 읽기 위한 하드웨어 또는 소프트웨어 툴이 이용자에게 어떻게 독서와 관련된 다양을 활동을 지원할 수 있는지를 조사하고 연구하는 태스크이다.

3. 평가결과(Evaluation Result)

각 참여자는 자신의 실험결과에 대한 객관적인 성능을 INEX에서 제공하는 공식 평가결과를 통해 확인할 수 있다. INEX는 각 트랙마다 평가척도(Evaluation Measure)에 대한 기술문서를 각 참여자에게 제공한다. 2009년 Ad Hoc트랙의 경우는 이미 INEX 홈페이지에 공식 평가결과를 발표했다.

하지만 Book Search 트랙의 경우에는 2008년 Structure Extraction task의 평가척도 기술문서[40]와 결과는 배포되었지만 그 외에, Book Retrieval task, Page in Context task, Active Reading에 대한 평가척도는 아직 배포되지 않았는데, 확인하면 〈그림 16〉과 같다.

Book Search 트랙의 Structure Extraction task도 보통 다른 트랙처럼 3가지 척도 즉, Precision(정확률), Recall(재현율), F-Measure가 사용되었다. 정보검색시스템의 일반적인 정의로 Precision(정확률)이란, 검색된 문헌들 중 검색된 적합문헌의 비율로 시스템이 검색한 문서가 얼마나 적합한가? 즉, 부적합한 문헌을 검색해 내지 않을 능력으로 검색의 정확성을 평가하는 척도이다. Recall(재현율)은 적합한 문헌들 중 검색된 적합문헌의 비율로 시스템이 적합문헌을 검색해 낼 능력 즉, 검색의 완전성을 평가하는 척도이다.

그리고 F-Measure는 Precision(정확률)과 Recall(재현율)을 복합적으로 반영하는 단일가 척도이다. 이 척도는 보통 재현율과 정확률이 한 쌍으로 성능을 나타내기 때문에 두 개 이상의 시스템 성능을 비교할 경우, 어느 시스템의 성능이 나은지 판단하기 어렵다. 이런 경우에 주로 사용되는 척도이다. Book search 트랙의 Structure Extraction task도 이 세 가지 척도를 적용하여 다음 〈표 3〉과 같이 재-정의하여 사용하였다.

40) 2008년 Structure Extraction 태스크의 평가척도(Evaluation Mesure) 기술문서, 〈http://www.inex.otago.ac.nz/tracks/books/INEXBookTrackSEMeasures.pdf〉 [cited 2009. 11. 20].

41)

BOOK RETRIEVAL TASK

Coming Soon...

PAGE IN CONTEXT TASK

Coming Soon...

STRUCTURE EXTRACTION TASK

The aim of this task is to test and compare automatic techniques for deriving structural informatio from digitized books and building a hyperlinked table of contents.

EVALUATION DOCUMENTS

Description of the Evaluation Measures of the Structure Extraction track

EVALUATION DATA

Groundtruth file ("ideal" tables of contents)
Evaluation Tool

RESULTS

Each RunID is linked to further details on the evaluation of this run.

RunID	Participant	F-measure (complete entries)
MDCS	Microsoft Development Center Serbia	53,47%
MDCS_NAMES_AND_TITLES	Microsoft Development Center Serbia	52,59%
MDCS_TITLES_ONLY	Microsoft Development Center Serbia	23,24%
HF_ToC_prg_Jaccard	Xerox Research Centre Europe	10,27%
HF_ToC_prg_OCR	Xerox Research Centre Europe	10,18%
HF_TPF_ToC_prg_Jaccard	Xerox Research Centre Europe	10,10%
HF_ToC_lin_Jaccard	Xerox Research Centre Europe	5,05%

42)

General Info

Participant Id		Some ID
Run Id		Some ID
Task		book-toc
Toc Creation		automatic
Toc Source		full
Source files	XML	Yes
	JPG	No
	PDF	No
Description		Description here...

Results

	Precision	Recall	F-Measure
Titles	x%	x%	x%
Levels	x%	x%	x%
Links	x%	x%	x%
Complete entries	x%	x%	x%
Entries	x%	x%	x%

43)

General Info

Participant Id		125
Run Id		MDCS
Task		book-toc
Toc Creation		automatic
Toc Source		full
Source files	XML	Yes
	JPG	No
	PDF	No
Description		In the first step each page in a book is assigned with a logical page number. Then, all TOC pages are detected and TOC structure extraction is performed on each. After all words in TOC pages are labeled the words are grouped into entries. Each entry is given a depth level according to the entry clustering. Also, each entry is assigned with a link by searching for the entry title on the target page (even if the entry does not have page number; if it does a logical page numbers are used).

Results

	Precision	Recall	F-Measure
Titles	75,51%	78,45%	76,12%
Levels	57,01%	59,54%	57,52%
Links	70,28%	72,78%	70,79%
Complete entries	53,09%	55,14%	53,47%
Entries disregarding depth	70,28%	72,78%	70,79%

〈그림 16〉 Book Search 트랙의 평가결과

〈표 3〉 Structure Extraction task의 평가척도 및 공식

평가척도	공식
Precision	$\frac{\sum Correctly\ recognized\ entities\ X}{\sum Entities\ recognized\ as\ X}$
Recall	$\frac{\sum Correctly\ recognized\ entities\ X}{\sum Entities\ X\ ini\ the\ ground\ truth\ file}$
F-Measure	$\frac{2\times \mathrm{Precision}\times Recall}{\mathrm{Precision}+Recall}$

41) 2008년 Book Search 트랙의 평가결과 페이지, 〈http://www.inex.otago.ac.nz/tracks/books/Results.asp〉 [cited 11. 20].
42) Book Search 트랙의 Structure Extraction 태스크 평가 결과를 위한 score sheet.
43) Book Search 트랙의 Structure Extraction 태스크 참여기관 ID 125의 run id MDCS의 평가결과 예시, 〈http://www.inex.otago.ac.nz/tracks/books/Results.asp?action=score&run=MDCSNamesTitlesAdditions.EVAL.html〉 [cited 11. 20].

현재 가장 아쉬운 점은 검색태스크의 평가척도가 아직도 배포되지 않은 것인데, 이는 아마 테스트컬렉션의 적합성평가가 많이 확보되지 않아 현재의 적합성평가만으로는 아직까지 객관적 평가가 힘든 상황이기 때문이라 생각된다. 향후 Book search 트랙의 검색태스크에 많은 관심과 참여로 객관성을 확보할 만한 크기의 적합성평가가 이루어진다면 검색태스크의 평가결과가 조속히 배포될 수 있으리라 생각한다.

Ⅳ. 2007년, 2008년 Book Search 트랙의 실험방법 분석

지금까지 INEX 실험을 통해 구조화된 도서 전문검색을 위해 전통적인 검색모델 적용 및 Book-Specific 특징을 이용한 검색모델개발 연구가 거의 없었다. 2007년 INEX Book Search 트랙에 참여하여 검색 실험을 한 논문이 2편, 그리고 2008년 12월에 발표한 pre-proceeding에 2편의 논문이 나와 있다. 현재까지 발표된 4편 논문의 실험방법을 분석해 보면 다음과 같다.

첫 번째 논문은 CMIC(Cairo Microsoft Innovation Center)의 실험논문[44]으로, 검색모델은 언어모델과 추론망 모델을 결합한 확률모델을 사용했다. 이 연구에서는 Book Retrieval Task(이하 BR이라 함)과 Page in Context Task(이하 PiC라 함)에 참여하여 일반적인 IR검색 기법을 적용한 실험과 일반 IR기법에 도서의 고유특징을 활용한 실험을 수행했다.

BR 태스크의 경우에 세 가지 pairs of runs를 수행했는데 일반적인 IR기법의 실행(run)들로 세 가지 즉, ① 각 도서의 전체 내용을 한 문서로 정의한 실험, ② 전체내용에 blind relevance feedback(상위 25개 도서에서 30개 용어 추출)을 추가 적용한 실험, ③ 하나의 페이지를 문서로 간주하여 상위 5000개 페이지를 추출하여 순위화 된 개별 페이지의 스코어 합을 도서의 순위로 정하는 실험을 수행하였다. 그리고 도서 고유 특징을 적용한 세 가지 실행(run)들로는 ① 책 속의 모든 헤드를 문서로 간주한 실험, ② 목차(앞에서 3000문자)와 색인페이지(마지막 10페이지)를 문서로 간주한 실험, ③ 목차와 색인 페이지와 blind relevance feedback(상위 25개 도서에서 20개 용어 추출)을 문서로 간주한 실험을 비교 수행하였다.

PiC 태스크의 경우에는 7개의 실행(run)들을 수행했으며, 모든 경우 문서를 하나의 페이지로 간주하여 질의 구성 방법을 7가지 즉, ① title, ② title+blind relevance feedback(상위 25개 문서의 20개 용어 추출), ③ title+description, ④ title+description+blind relevance feedback, ⑤ title+description+narrative, ⑥ title+description+narrative+blind relevance feedback, ⑦ 수동으로 재구성한 질의를 사용하여 7개의 실행(run)들을 수행하였다.

---

44) W. Magdy and K. Darwish, "CMIC at INEX 2007 : Book Search Track," *Pre-Proceedings of INEX 2007* (2007), pp.197-199.

다른 두 번째, 세 번째 논문은 UC Berkeley의 2007년[45]과 2008[46]년에 수행한 INEX Book Search 트랙의 실행(run)들에 관한 논문이다.

Larson은 기존 TREC2에서 사용하였던 알고리즘인 Logistic 회귀(확률)모델을 일부 수정하여 도서검색을 위해 사용하였다. 즉, 문서(D)와 질의(Q)간의 적합할 확률을 나타내는 $P(R|Q, D)$에서 문서(D)를 대신하여 컴포넌트(C)[47]개념을 채택하여 $\log O(R|C, Q)$[48]을 계산하도록 기존 알고리즘을 변경하여 사용하였다.

2007년과 2008년 실험의 차이점은 2007년 실험에서는 Book Collection의 마크 데이터용 EVIs[49](Entry Vocabulary Indexes)와 blind 피드백을 적용한 실험을 하였는데, 도서의 Full-text의 데이터베이스 셋업의 문제점으로 인해 페이지레벨의 색인을 수행하였고, 페이지 내의 패러그래프 단위의 실험, 마크 데이터를 통한 직접 검색, Fusion Operator를 통해 앞의 실험결과를 결합해 보는 실험을 수행하였다.

반면에 2008년 실험에서는 같은 알고리즘을 사용하되, EVIs를 사용하지 않았고 전체 컬렉션에 대해 단일색인을 수행하여 실험을 하였다. 2008년 BR 태스크에서 세 개의 실행(run)들을 제출하였는데, ① MARC 데이터만 사용한 실행(run), ② 도서 전체 내용을 대상으로 한 실행(run), 세 번째는 MARC와 도서전체 내용을 함께 통합하여 사용한 실행(run)을 수행하여 제출하였다.

네 번째 논문은 RMIT University[50]의 실험논문으로, 이 실험에서는 4가지 방법의 질의 구성으로 도서전체를 대상으로 한 4가지 실행(run)들과 페이지레벨을 대상으로 검색을 수행한 4가지 실행(run)들을 수행하여 총 8개의 실행(run)들을 제출하였다.

4가지 질의는 topic의 구성 요소인 title, description, task, infneed 중 title과 infneed를 개별 또는 결합을 통해 질의를 구성하였다. 즉 ① title의 words ② title과 infneed의 words ③ title의 words를 AND로 구성 ④ title의 words+일부 수동으로 용어를 선택하여 질의를 구성하였다.

색인을 위해 Zettair search engine[51]을, 검색을 위해 BM25[52]을 사용하였다. 페이지 레벨의

45) R. R. Larson, "Logistic Regression and EVIs for XML Books and the Heterogeneous track," *Pre-Proceedings of INEX 2007*(2007), pp.185-196.

46) R. R. Larson, "Adhoc and Book XML Retrieval with Cheshire," *INEX 2008 Workshop Pre-proceedings* (2008), pp.194-204.

47) 고정된 문서와 달리 도서의 페이지, 페이지 내의 패러그래프, 섹션 등이 될 수 있는 가변적인 특정 단위를 말함.

48) $\log O$를 취하면 유효하지 않은 확률 값을 피할 수 있다고 함.

49) EVIs 개발 목적 : 도서관 목록 MARC레코드의 자동 분류를 용이하게 하여 검색에 사용하기 위해 개발되었고, MARC내에 사용되는 태그를 그룹화하여 names, pauthor, title, subject, topic, lcclass, doctype, localnum, ISBN, publisher, place, date, lang 필드로의 매핑을 정의하고 있음.

50) Mingfang Wu, Falk, Scholer, and James A. Thom, "RMIT University at the INEX Book Search Track," *INEX 2008 Workshop Pre-proceedings*(2008), pp.205-207.

51) RMIT대학에서 TREC에서 사용한 검색 엔진, 〈http://www.seg.rmit.edu.au/zettair/〉.

52) TREC-3에서 사용한 알고리즘.

실행(run)에서는 도서의 순위화를 위해 두 가지 전략을 사용하였는데, ⓐ 질의 당 상위 3000페이지를 검색하여 책 당 검색된 페이지의 백분율로 계산하여 도서를 순위화, ⓑ 질의 당 상위 3000페이지를 검색하여 책 당 검색된 연속 페이지의 최대 수를 근거로 순위화 하는 전략을 사용하였다. 페이지 레벨로 수행한 4개의 실행(run)들에서 첫째 실행(run)은 질의 ①과 도서 순위화 전략 ⓐ, 둘째 실행(run)은 질의 ①과 도서 순위화 전략 ⓑ, 셋째 실행(run)은 질의 ④와 도서 순위화 전략 ⓐ, 넷째 실행(run)은 질의 ④와 도서 순위화 전략 ⓑ의 방법을 적용하여 실험하였다.

2007년과 2008년에 참여한 실험방법들을 요약하여 정리하면 〈표 4〉와 같다. 4개의 논문 모두

〈표 4〉 2007/2008년 Book Search 트랙의 실험방법 정리

	CMIC	Larson' 2007	Larson' 2008	RMIT
검색모델	-언어모델+추론망 결합	-회귀(확률)모델 TREC2 LR알고리즘	-회귀(확률)모델 TREC2 LR알고리즘	-확률모델 BM25 알고리즘
실험시스템	-Indri search toolkit	-Cheshire II system[53]	-Cheshire II system	-Zettair search engine (색인) -Okapi BM25(검색)
색인단위	-도서전체 -페이지 -도서의 모든 헤드 -목차+색인페이지	-페이지 -패러그래프(복수개의 엘리먼트 또는 컴포넌트) -MARC	-도서전체 -MARC	-도서전체 -페이지
검색(질의) 확장	-Blind Feedback(상위 25개 문서에서 용어 20개 추출)	-MARC의 EVIs -Blind Feedback	MARC의 EVIs 사용하지 않음	-
수행태스크 (BR의 검색 단위는 도서, PiC의 검색 단위는 페이지임)	-BR : 3개 pairs of runs (1) 일반 IR기법 ① 도서전체를 문서로 간주한 run ② 도서전체와 Blind Feedback을 문서로 간주한 run ③ 페이지를 문서로 간주 & 상위 5000페이지의 점수 합산하여 도서 랭킹 산정 (2) 도서특징적용 IR기법 ① 책속의 모든 헤드를 문서로 간주한 run ② 목차(앞의 3000문자) & 색인페이지(마지막 10페이지)를 문서로 간주한 run ③ 목차+색인+Blind Feedback을 문서로 간주한 run -PiC : 7개 runs • 색인 : 페이지 • 질의구성 : 7가지 ①/②/③/④/⑤/⑥/⑦	-BR/PiC : 10개 runs[54] : Fusion Operator로 MARC와 EVIs의 다양한 결합을 시도하여 최대 제출 runs 10개까지 수행하였다고 함.	BR : 3개 runs ① MARC 데이터만 사용한 run ② 도서 전체를 사용한 run ③ MARC+도서전체를 사용한 run	-BR : 4개 runs (도서 전체를 대상으로 질의 구성 4가지 방법으로 4개 runs 수행) • 색인 : 도서전체 • 질의구성 : ①/②/③/④ -PiC : 4개 runs • 도서순위화 전략 2가지 방법 : ⓐ /ⓑ(본문내용참조) • 질의 구성 : ①/②/③/④ • 4개 runs : 질의 ① & 전략 ⓐ : 질의 ① & 전략 ⓑ : 질의 ④ & 전략 ⓐ : 질의 ④ & 전략 ⓑ

53) 2005년 Ad hoc 트랙 평가에서 good retrieval performance로 평가 받은 시스템이라고 함.
54) Fusion Operator와 10가지 run 각각에 대한 정의가 논문에 기술되지 않았음.

	CMIC	Larson' 2007	Larson' 2008	RMIT
질의구성	① title ② title+Blind Feedback ③ title+description ④ title+description+ Blind Feedback ⑤ title+description+ narrative ⑥ title+description+ narrative+Blind Feedback ⑦ 수동으로 질의 구성	-	-	① title의 words ② title과 infoneed의 words ③ title의 words의 AND 연산 ④ title의 words+수동으로 용어 선택

각자의 방법으로 실행(run)들을 제출하였지만 2007년의 경우는 평가시스템의 미완성으로, 2008년 논문의 경우에도 적합성 평가 및 최종 결과가 나오지 않아 결과를 비교할 수는 없음을 밝히고 있다.

Book Search 트랙의 코퍼스가 대용량이고, Ad Hoc트랙에 비해 연구가 진행된 시간도 얼마 되지 않았으며, 여러 가지 실험환경에 제한이 많아 올해 2009년 실험도 지연되고 있다. 검색에 관심 있는 분들이 이 논문을 통해서 INEX Book Search 트랙에 대해 좀 더 잘 알게 되고, 더 많은 관심과 참여로 다양한 검색실험을 할 수 있다면, 도서를 대상으로 다양한 검색 모델 적용을 통한 성능 비교와 평가시스템 구축이 조속히 이루어질 수 있으리라 생각한다.

V. 결론 및 향후 연구과제

본 논문에서는 XML 검색 추진체인 INEX의 실험트랙, 실험에 참여하는 방법, 절차 및 참여기관 등에 대한 전반적인 내용을 소개하였다. INEX 실험트랙 중에서 특히 2007년에 처음 시작된 Book Search 트랙을 중점적으로 살펴보았는데, 그 내용으로 Book Search 트랙의 테스트컬렉션과 실험태스크, 실험결과 제출방법 및 평가방법에 대해 살펴보았다.

Book Search 트랙의 테스트컬렉션은 IR평가를 위해 전통적으로 사용되는 통제된 실험환경으로 문서집합(Book Corpus)과 이용자 정보요구의 집합인 Topics, 적합성 평가로 구성됨을 보았다. Book Search 트랙의 실험태스크에는 4가지 태스크가 있는데, Book Retrieval 태스크와 Focused Book Search 태스크(또는 Page in Context 태스크)는 도서검색 태스크였고, Structure Extraction 태스크는 도서의 구조를 추출하는 태스크이고, Active Reading 태스크는 독서할 때 이용자와 도서 간의 이용행태를 연구하는 태스크였다. 이 연구로 INEX Book Search 트랙의 실험에 대한 전반적인 내용 이해와 디지털화된 도서를 대상으로 하는 연구 분야를 알 수 있었다. 마지막에는 2007년부터 현재까지 진행된 검색태스크의 연구논문들의 실험방법들을 분석해 보았는데 이는 디지털화된 도서를 대상으로 진행되고 있는 연구방법들을 알게 하고, 향후 도서검색을 위한 다양한 접근법을

모색하는데 도움이 되었다.

Book Search 트랙의 연구가 시작된 지 이제 겨우 3년이 되었고, 향후 전자도서관, 디지털도서관, 멀티미디어도서관을 넘어서 가상현실 전자책(E-Book) 도서관이 구현되기 위해서 풀어야 할 무수한 연구과제들이 있으리라 생각된다. 이 트랙이 시작되면서 관심을 가지고 부분적으로 실험에 참여해 본 결과 우선 해결되어야 할 과제들이 있어 몇 가지 짚어본다.

첫째, Book search 트랙의 기본 인프라 구축의 애로점을 해결하는 것이다. 이 트랙의 실험문서는 대용량의 코퍼스 즉 디지털화된 도서 5만권으로 기본 원문만 해도 매우 큰 용량이며, 검색실험을 위해서는 원문 이상의 색인도 생성되어야 한다. 국내에서도 Microsoft 및 도서관 패키지 개발사 등 유관기관과의 연계를 통해 좀 더 나은 실험환경을 구축하여 참여자들이 실험환경에 대한 제약을 극복할 수 있는 방법들이 모색되어야 한다. 둘째, 도서의 검색 성능을 보장하려면 코퍼스의 구조 추출 및 특성 분석이 우선되어야 한다. 도서 코퍼스만의 특징적인 특성과 상세구조를 분석하여 잘 활용할 수 있다면 높은 성능의 도서 검색시스템을 구현할 수 있으리라 생각한다. 셋째, 지금까지 진행된 연구 방법은 Ad Hoc트랙 즉 일반문서에 적용된 모델들을 극히 제한적으로 적용한 실험이기에 앞으로 도서 코퍼스에 다양한 정보검색모델을 적용해 보는 실험이 요구된다. 이를 통한 객관적인 다양한 방법의 성능 비교로 도서에 적합한 모델 개발이 필요하다. 넷째, 현재 적합성평가 정보가 부족하여 Book Search 트랙의 검색태스크에 대한 IR평가가 제대로 이루어지지 못하고 있다. INEX 실험에 좀 더 많은 관심과 참여로 어느 정도 크기의 신뢰성 있는 적합성평가 정보구축으로 조속히 평가시스템이 구축되어야 한다. 다섯째, 도서(Book) 토픽을 검색시스템에 적합한 질의형태로 변환하는 다양한 기법 연구가 필요하다.

끝으로, Book Search 트랙이 올해로 3년째 실험연구가 진행되고 있으나 실험코퍼스가 대용량이고, 여러 가지 실험환경 인프라 구축에 어려움이 많고, 참여자나 관심 등이 부족한 상황이다. 이 논문을 계기로 우리 국내의 연구자 및 디지털도서관 구축을 위한 프로그램 개발업체, 전자책(E-Book) 개발업체들이 Book Search 트랙에 대한 관심을 갖고, 많은 참여자가 생겨 장기적 비전을 갖고 함께 연구할 수 있었으면 좋겠다는 바램을 가져본다.

〈참고문헌은 각주로 대신함〉