개인화 검색시스템 평가에 관한 연구

김광영○, 최호섭, 진두석, 김진숙
한국과학기술정보연구원
{kykim,hschoe,dsjin,jinuk}@kisti.re.kr

A Study of Personalized Retrieval System Evaluation

KwangYoung Kim○, Ho-Seop choe, Dusuk Jin, Jinsuk Kim
Korea Institute of Science and Technology Information

요 약

본 논문에서는 주제별 분류기반의 개인화 검색시스템의 평가를 위해서 기존의 한글 정보 검색시스템 평가를 위해서 사용하는 한글 테스트 컬렉션(HANTEC v2.0)을 사용하였다. 주제별 분류기반의 개인화 검색시스템의 평가를 위해서 첫째, 한글 테스트 컬렉션을 한국일보-40075 문서분류 테스트 컬렉션을 이용하여 주제별 분류를 수행 하였다. 둘째, 한국일보-40075 문서분류 테스트 컬렉션의 분류 체계에 다라 한글 테스트 컬렉션의 문서들을 kNN 분류기를 이용하여 분류를 수행하였다. 마지막으로 구축된 컬렉션을 이용하여 주제별 분류기반의 개인화 검색시스템의 성능 평가를 수행하였다.

1. 서 론

대부분의 검색시스템은 사용자의 검색 의도를 반영하지 않고 문서 검색을 하고 있다. 사용자에게 적합한 다양한 서비스들을 제공하기 위해서는 우선 사용자의 성향을 분석해야 한다. 그렇지만 사용자의 주요 관심 정보를 파악하는 것은 어렵다.

사용자의 성향 정보를 분석하기 위해서 일반적으로 사용자 프로파일을 만들어 사용한다. 사용자 프로파일을 생성하는 방법에는 사용자가 직접적으로 개인 정보나 관심 정보를 이용하는 명시적(explicit)인 방법과 사용자의 행동을 통해 암시적(implicit)으로 개인 성향 정보를 추론하는 방법으로 나눌 수가 있다. 사용자의 명시적인 표현을 통해서 프로파일을 이용할 경우에는 빠르고 정확하게 사용자의 성향을 분석할 수 있지만 대부분의 사용자들은 이것을 번거롭게 생각하며 동적인 취향 변화를 반영하기 어렵다는 단점이 있다.

현재 일반적인 검색 엔진에서는 서로 다른 사람이 똑 같은 질의어를 보낼 경우에 똑 같은 결과를 제시한다. 이것은 정보 요구가 서로 다른 사람들에게는 적합하지 않다.

대형 웹 포털이나 전문 검색시스템에서 이러한 개인화 검색서비스를 제공하기에 현실적으로 많은 어려운 점들을 가지고 있다. 현재 대형 웹 포털 사이트들도 이러한 문제점들을 극복하기 위해서 개인화된 다양한 서비스들을 제공하고 있다.

이와 같이 사용자의 관심 정보를 분석하기 위한 다양한 방법들이 연구되고 있다. 가장 보편화된 방법은 사용자가 사이트 방문 초기에 명시적으로 표현한 개인 정보나 관심 정보를 이용하는 것이다[4].

국외에서도 사용자의 클릭 히스토리를 이용하여 개인화 검색시스템에 반영하여 단어의 중의성 문제 해결을 시도하고 있다. 또한 질의어-질의어의 재조합을 통하여 사용자의 관심 분야를 검색하는 방법도 제공되고 있다. 즉 사용자가 “Windows"라는 질의어를 입력할 때 "Windows XP"와 "House windows" 등으로 질의어-질의어를 재구성하여 검색 결과를 다양하게 처리하는 방법도 제공되고 있다[5]. 2007년 Ahu Sirg 등은 온톨로지 기반의 사용자 프로파일 정보와 사용자 행위 정보를 이용하여 사용자에게 적합한 문서를 재순위화 시키는 방법을 이용한 개인화 검색서비스를 제안 하였다[6][7].

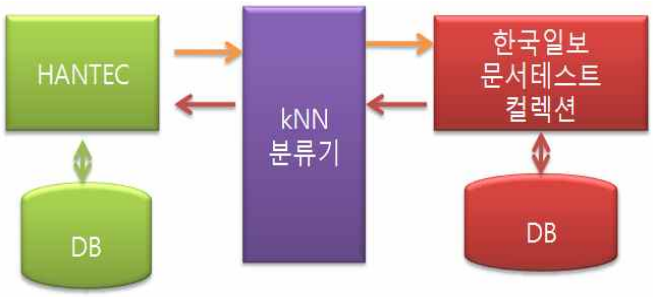
ODP(Open Directory Project) 택소노미를 이용하여 다양한 분야 및 웹기반 개인화 검색시스템에 적용한 것도 있다[8][9]. 구글의 PageRank 알고리즘을 확장한 Personalized Page Rank기술을 적용한 웹기반 개인화 검색시스템도 있다[10].

이러한 문제점들을 해결하기 위해서 최근 개인화 검색시스템에 대한 관심도가 점점 높아지고 있고 활발한 연

구들이 진행되고 있다. 하지만 실제 개인화 검색시스템을 개발하고 시스템의 성능을 평가를 위해서 전문가가 직접 검색 결과를 평가를 하는 방식을 대부분 사용하고 있다. 본 논문에서는 한글 테스트 컬렉션을 이용하여 주제별 분류기반의 개인화 검색시스템을 평가하는 방법을 연구하였다.

2. 평가 모델

본 논문에서는 주제별 분류기반의 개인화 검색시스템을 평가하기 위해서 한글 테스트 컬렉션의 문서들을 kNN 분류기를 이용하여 주제별로 문서들을 분류를 하였다.



<그림 1> HANTEC 주제 분류 시스템

<그림 1]>과 같이 HANTEC를 주제 분류하기 위해서 한국일보 문서 테스트 컬렉션을 이용하였다. 기존에 HANTEC의 문서들을 한국일보-40075의 실험문서집합의 분류체계로 분류를 하였다.

한국일보-40075 실험문서집합은 한국일보가 제공한 1998~1999년의 2년간 신문기사를 바탕으로 40,075개의 각 문서별로 3단계 분류체계의 말단 범주를 부여하여 구축하였다. 한국일보-20000 실험문서집합은 한국일보-40075 집합의 기사 중 20,000건을 별도로 추출하여 분류체계를 보다 현실적으로 수정하였으며, 3단계 분류체계의 모든 노드에 기사를 할당하여 구축한 계층적 분류체계의 문서범주화용 실험문서집합이다[1].

<표 2> 한국일보-40075 실험문서집합의 2003범주당 문서수 (일부)

범주당 문서수	/대분류/중분류/소분류
61	/건강과 의학/건강/영양 식품 식사
35	/건강과 의학/건강/체력단련
41	/건강과 의학/의약학/성인병
12	/건강과 의학/의약학/수의학
60	/건강과 의학/의약학/질병(암)
228	/건강과 의학/의약학/질병(암외의질병)
46	/건강과 의학/의약학/치의학
40	/건강과 의학/의약학/한의학 전통의학
144	/경제/가계 물가/가계 물가
227	/경제/국가/수입
548	/경제/국가/수출
348	/경제/국가/재정 경기전망
95	/경제/금융/보험(생명)
...	...

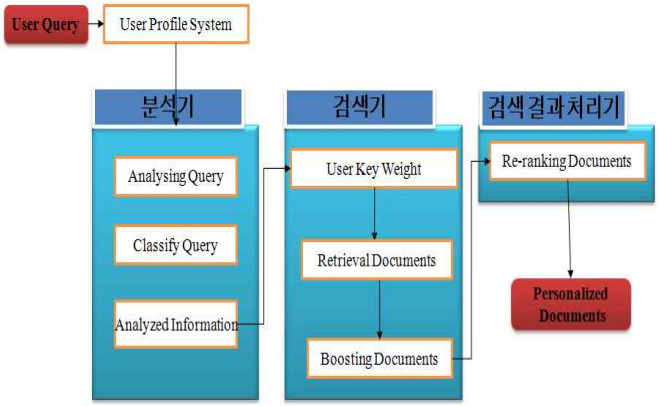
<표 2>에서는 각 소분류 범주에 할당된 문서수를 일부 보여주고 있다. 2003분류체계의 첫 번째 특징은 모든 문서에는 소분류 단위에서 범주가 부여되어 있다. 위의 예에서 볼 수 있듯이 모든 범주는 3단계로 표현된다. 한국일보-40075 실험문서집합에서 대분류의 수는 9개, 중분류는 32개, 소분류의 수는 120개이다. 두 번째 특징은 모든 문서가 단일 범주를 가진다는 것이다. 한국일보-40075 실험문서집합은 40,075개의 문서에 1:1로 범주가 부여되어 있어서 총 부여된 범주의 수는 문서의 수와 동일한 40,075개 이다[1].

<표 3> 한국일보-40075/2003분류체계의 대/중/소분류별 통계

대분류	문서수	중분류범주수	중분류범주별 평균 문서수	소분류범주수	소분류범주별 평균 문서수
/건강과 의학	523	2	261.5	8	65.4
/경제	7300	5	1460.0	19	384.2
/과학	794	2	397.0	8	99.3
/교육	680	4	170.0	4	170.0
/문화와종교	3457	4	864.3	20	172.9
/사회	5273	2	2636.5	8	659.1
/산업	4890	5	978.0	18	271.7
/여가생활	614	2	307.0	5	122.8
/정치	16544	6	2757.3	30	551.5
계	40075	32	1252.3	120	334.0

<표 3>과 같이 한국일보-40075 실험문서집합의 경우 "/정치" 대분류에 속하는 범주를 가진 문서가 전체 문서집합의 41.3%에 달할 정도로 매우 많다는 것이 특징이다

[1].



<그림 2> 분류기반 개인화 검색시스템

<그림 2>의 개인화 검색시스템은 사용자의 질의어를 이용하여 범주화된 주제 분류에 따라 개인의 성향 정보를 분석하고 이것을 기반으로 일반 검색 결과에 대해서 적합한 문서들을 사용자의 성향 정보에 맞게 재순위화를 처리하는 시스템이다[2]. 위와 같은 주제별 분류기반의 개인화 검색시스템을 평가하기 위해서 HANTEC를 주제별 분류로 분류를 처리한다. 분류된 정보를 가지고 개인화 검색시스템을 평가한다.

3. 실험 방법 및 평가

3.1 실험 데이터 셋

본 연구의 실험을 위해서 HANTEC v2.0를 사용하였다. 시스템의 서버 사양은 리눅스 Redhat 4.1.2, 메모리 12G, 2CPU 인텔 Xeon 1.6GHz를 사용하였다.
HANTEC의 DB의 섹션 구성은 <표 4>과 같다.

<표 4> HANTEC의 섹션 구성

섹션 이름	설 명
DOCID	문서식별자
CAT	분류
CAT1	대분류
CAT2	중분류
CAT3	소분류
TITILE	제목
CONTENT	본문

본 논문에서는 <표 4>과 같이 대분류, 중분류, 소분류로 나누어 하나의 섹션 처리하였다. 이것은 실제 주제별 개인화 검색시스템에서 대분류, 중분류나 소분류를 선택하여 주제 범위 별로 개인화 검색 시스템을 테스트를 수행하였다.

<표 5> HANTEC 테이블당 문서 수

테이블 이름	문서 건수	비고
KED94	39,474	
KWDI	110	
SATURN	12,000	
WWW	18,000	
HKIB94	22,000	
KRIST	10,000	
TREND	17,929	
전체	119,513	

<표 5>와 같이 HANTEC 전체 문서 건수는 119,513건의 문서이다.

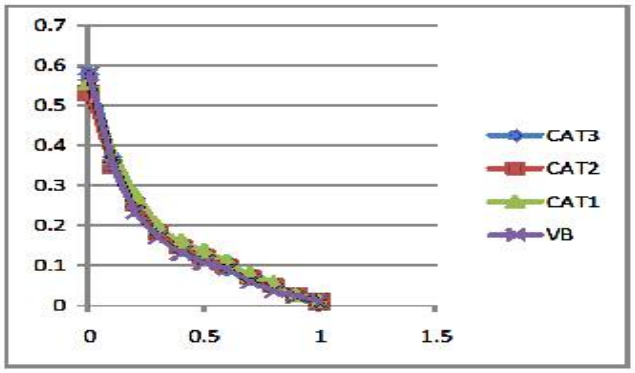
본 논문에서 실험을 위해서 50개의 질의어 셋을 다시 분류기로 분류를 하였다. 분류되어 나오는 값들을 개인의 성향 정보로 보았다. 그리고 이것을 이용하여 벡터모델로 검색을 수행한 것과 개인화 검색 결과 값을 비교하였다.

3.2 실험 결과

본 연구에서는 벡터모델을 이용하여 일반 검색시스템의 결과와 개인화 검색시스템 간의 정확률과 재현율을 측정하였다. 각 50개의 질의어를 대분류(CAT1), 중분류(CAT2)와 소분류별(CAT3)로 11-수준 평균 정확률(11-point interpolated average precision)을 측정하였다. 그 결과는 <표 6>과 <그림 3>으로 나타났다.

<표 6> 11-point interpolated average precision

Recall	CAT3	CAT2	CAT1	VB
0	0.5771	0.5315	0.558	0.5788
0.1	0.37	0.3457	0.3723	0.3474
0.2	0.247	0.2545	0.2756	0.2326
0.3	0.1806	0.1826	0.1963	0.1708
0.4	0.1467	0.1452	0.1588	0.1308
0.5	0.1196	0.121	0.1348	0.1062
0.6	0.0901	0.0971	0.109	0.0912
0.7	0.0666	0.0693	0.0756	0.0609
0.8	0.0492	0.0503	0.0566	0.0355
0.9	0.0252	0.0271	0.0272	0.0225
1	0.0069	0.0101	0.0105	0.0091



<그림 3> 11-point interpolated average precision

일반 검색 모델(VB)의 정확도 평균값은 0.1444, 대분류(CAT1)의 정확도 평균값은 0.1638, 중분류(CAT2) 정확도 평균값은 0.1509이며 소분류(CAT3)의 정확도 평균값은 0.1509로 나타났다. 그 결과 개인화 검색시스템에서는 대분류를 사용하는 것이 중분류나 소분류보다 높게 측정이 되었다.

4. 결 론

본 연구에서는 주제 분류를 이용한 개인화 검색 시스템의 평가를 위해서 한글 테스트 컬렉션을 한국일보-40075의 실험문서집합의 분류체계로 분류를 하였다. 그 결과 기존의 일반 검색시스템보다는 약간 높게 나타났다. 그러고 개인화 검색 시스템에서 사용하는 개인의 성향 정보를 기준을 대분류, 중분류와 소분류 중에 대분류로 하는 것이 더 좋은 결과가 나타났다. 이것은 검색 결과 문서들 중에서 개인에게 적합한 문서들을 판단할 때 좁은 주제 범위로 사용하는 것보다 넓은 주제 범위를 사용하는 것이 더 좋은 것으로 나타났다.

본 연구에서는 실험적으로 한글 테스트 컬렉션을 한국일보-40075의 실험문서집합의 분류체계로 분류를 하였지만 향후 연구 과제로는 한국일보-20000 분류체계를 적용하여 실험을 수행하여 상호 결과를 비교 분석이 수행할 것이며 한글 테스트 컬렉션보다 TREC으로 그 성능을 분석할 필요가 있다.

5. 참고문헌

[1] 한국과학기술정보연구원, “한국일보-20000/한국일보-40075 문서범주화 실험문서집합” http://www.kristalinfo.com/TestCollections/readme_hkib.tml [sited 2010.4.12]

[2] 김광영, 심강섭, 곽승진, “분류와 사용자 질의어 정보에 기반한 개인화 검색시스템,” 한국문헌정보학회지, 제43권, 제3호(2009), pp.163-180.

[3] Jinsuk Kim, Ho-Seop Choe and Beom-Jong You, “HKIB-20000 & HKIB-40075: Hangul Benchmark Collections for Text Categorization Research“, Journal of Computing Science and Engineering, Vol. 3, No. 3, September 2009, pp.165-180.

[4] G. Linden, B. Smith and J. York, "Amazon.com Recommendations Item-to-Item Collaborative Filtering,", IEEE Internet Computing(2003), pp.76-80.

[5] Filip Radlinski, Susan Dumais, "Improving Personalized Wes Search using Result Diversification,", Proceedings of the 29th annual international ACM SIGIR(2006), pp.691-692.

[6] Ahu Sieg, Bamshad Mobasher and Robin Burke, "Web Search Personalization with Ontological User Profiles,", Proceedings of the sixteenth ACM conference on Conference on information and knowledge management(2007), pp.525-534.

[7] J. Trajkova and S. Gauch, “Improving ontology-based user profiles,”, Proceedings of the Recherched' Information Assistear Ordinateur, RIAO(2004), pp.380 - 389.

[8] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschutter, “Using odp metadata to personalize search,”, Proceedings of the 28th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval,SIGIR(2005), pp.178 - 185.

[9] C. Ziegler, K. Simon, and G. Lausen, “Automatic computation of semantic proximity using taxonomic knowledge,”, Proceedings of the15th ACM International Conference on Information and Knowledge Management, CIKM(2006), pp.465 - 474.

[10] G. Jeh and J. Widom, “Scaling personalized web Search,”, Proceedings of the12th international conference on World Wide Web(2003), pp.271 - 279.