개인화된 웹 검색을 위한 선호 기준 분석

이수정†

요 약

웹 문서 수의 급증으로 인해 인터넷을 검색할 때마다 발생하는 정보의 과부하 문제가 심각하게 부각되었다. 웹 검색 결과를 개선하기 위하여 개발된 기존의 알고리즘들은 주로 사용자의 질의어 및 선호어와 문서의 링크수를 이용하였다. 본 연구에서는 실험을 통하여 이 두가지 요소들을 이용한 검색 결과의 성능을 알아보고 이들 요소들 외에 선호하는 웹문서의 선택 기준을 조사 분석하였다. 실험 결과 질의어 및 선호어를 이용한 개인화된 검색 결과는 현 검색 엔진에 비해 최대 약 1.7배의 성능 향상을 가져 왔으며, 링크수를 이용한 검색 결과는 최대 약 1.3배의 향상을 보였다. 사용자가 웹문서를 선호하는 기준은 문서 내용이 최우선이었으나, 가독성과 문서가 포함한 이미지도 큰 비중을 차지하였다. 따라서 질의어 및 선호어 개수 이외에 각 사용자의 성향에 부합하는 객관적 데이터를 추가적으로 활용한다면 웹 검색 개인화 알고리즘의 성능이 크게 향상될 수 있을 것이다.

주제어 : 웹 검색 개인화, 정보 필터링, 협력 필터링, 사용자 프로필, 추천 시스템

Analysis of Preference Criteria for Personalized Web Search

Soojung Lee†

ABSTRACT

With rapid increase in the number of web documents, the problem of information overload in Internet search is growing seriously. In order to improve web search results, previous research studies employed user queries/preferred words and the number of links in the web documents. In this study, performance of the search results exploiting these two criteria is examined and other preference criteria for web documents are analyzed. Experimental results show that personalized web search results employing queries and preferred words yield up to 1.7 times better performance over the current search engine and that the search results using the number of links gives up to 1.3 times better performance. Although it is found that the first of the user's preference criteria for web documents is the contents of the document, readability and images in the document are also given a large weight. Therefore, performance of web search personalization algorithms will be greatly improved if they incorporate objective data reflecting each user's characteristics in addition to the number of queries and preferred words.

Keywords : Web Search Personalization, Information Filtering, Collaborative Filtering, User Profile, Recommendation System

† 정 회 원: 경인교육대학교 컴퓨터교육과 교수
논문접수: 2009년 11월 3일, 심사완료: 2009년 12월 11일
* 본 논문은 2007년 한국학술진흥재단의 학술연구비에 의하여 지원되었음(No. KRF-2007-313-D00679)

1. 서 론

웹 문서 수가 급증함에 따라 인터넷 검색 시 발생하는 정보의 과부하 문제가 심각하게 부각되었다. 대부분의 검색 엔진은 사용자가 누구이건 간에 상관없이 질의어에 대해 동일한 검색 결과를 제공한다. 따라서 검색 작업에 상당한 시간과 노력을 투자하지만 사용자는 자신이 원하는 정보를 얻지 못할 수 있다. 이 문제를 경감시키기 위해 사용자의 선호도에 부합하는 웹 환경을 조성하여 주는 개인화 작업이 주목받고 있다[1]. 현재 웹 개인화의 가장 유명한 예는 추천 시스템으로서 주로 고객이 구매하기 원하는 상품을 찾을 수 있도록 도움을 제공한다[2]. 추천 시스템의 예는 Amazon.com[3]과 MovieLens[4] 등이다.

일반적으로 웹 개인화는 특정 사용자의 요구에 맞도록 웹 사이트를 적응시켜 나가는 과정이다. 웹 개인화는 주로 네 가지 범주로 사용되는데, 웹 검색경로 예측, 개인화된 정보 보조, 검색 내용의 개인화, 그리고 검색 결과의 개인화이다[5]. 웹 검색 결과를 개인화하면 사용자의 흥미에 적합한 결과를 얻을 수 있다.

결론적으로 웹상의 정보량이 급증함에 따라 개인 성향에 부합하는 정보를 찾아주는 개인화 시스템은 매우 중요함에도 불구하고 대부분의 검색 엔진은 사용자 질의어에만 초점을 두어 응답 결과를 산출하고 있다. 최근에 사용자의 선호도나 문서의 링크수에 근거하여 문서 순위를 산출하는 알고리즘을 개발하는 등 앞에서 언급한 문제를 해결하려는 노력이 있어 왔다. 본 연구에서는 이들 근거 요소, 즉, 선호도와 링크수가 웹 검색 결과를 충분히 개선시키는지 알아보기 위하여 실제 사용자 9명을 대상으로 약 30번의 검색 실험을 하게 하여, 검색 결과 웹문서들을 선호와 비선호로 구분하고, 선호 이유를 알아보았다. 실험 결과, 사용자의 선호어 뿐만 아니라, 문서의 가독성, 문서가 포함하는 이미지수, 문서 길이도 선호 문서를 결정하는데 큰 비중을 차지한 반면에, 기존 연구에서 문서 순위를 책정하는 주요 요소인 링크수는 상대적으로 가장 낮은 비중을 차지하였다. 또한 질의어와 선호어를 기준으로 검색 결과 문서들의 순위를 책정하는 것은 기존 검색 엔진의 결과와 비교했을 때 약 1.7배의 성능 향상 효과를 가져왔으며, 링크수를 기준으로 한 순위 책정 방식은 약 1.3배의 향상을 보였다. 따라서, 기존의 웹 검색 알고리즘에서 사용하는 문서 선호 기준들을 보다 확장시킬 필요가 있으며, 확장된 통합 기준에 의거한 새로운 검색 알고리즘의 개발이 필요하다는 결론을 도출할 수 있다.

본 논문의 구성은 다음과 같다. 2절에서 관련기존 연구에 대해 기술하고, 3절에서는 선호어와 링크수를 이용한 방법이 웹검색 결과에 미치는 효과에 대해 실험한 결과를 기술하며, 또한 문서 선호 기준의 실태 조사 결과를 제시한다. 4절에서 논문의 결론을 맺는다.

2. 관련연구

웹 개인화를 실행하는 두 가지 주된 방법은 협력 필터링과 정보 필터링(사용자 프로필)이다[5]. 두 가지 방법은 모두 사용자가 관심을 보일 만한 항목들을 식별하여 정보 과부하를 경감시키고자 하는 시도이다.

협력 필터링은 공통된 흥미를 갖고 있는 여러 다른 사용자들로부터 정보를 획득하여 그들의 의견에 따라 웹사이트를 추천하는 방법이다. 이 방법은 서적, 식품점, 예술과 엔터테인먼트 등 다양한 영역에서 사용되었다[6]. 이 방법의 주요 장점들 중 하나는 추천되는 항목의 내용을 고려하지 않은 채 다른 많은 사람들이 선호한다는 이유 하나만으로 새로운 항목들을 발견할 수 있다는 것이다. 그러나 새로운 문서에 대해서는 축적된 사용자 선호도 정보가 없기 때문에 새로운 문서를 추천할 수 없다는 단점이 있다. 또다른 단점으로 인기 있는 문서를 결정하기 위해 많은 사용자로부터 평가 정보를 필요로 한다는 것이다. 이러한 단점들을 보완하기 위해 손창환과 김기수[7]는 사용자 기반 협력 필터링과 아이템 기반 협력 필터링을 결합하여 하이브리드 협력 필터링 기법과 사례기반 추론 기법을 혼합한 하이브리드 추천 시스템을 제안하였다.

정보 필터링은 내용 분석을 토대로 개인적인 사용자 흥미도 프로필을 구축하는데 초점을 둔다. 프로필 구축은 사용자가 직접 입력하거나 사용자의 행위로부터 간접적으로 학습할 수 있다. 예로써, WebMate[8]는 사용자가 흥미를 보이는

문서로부터 간접적이고 자동적으로 사용자의 선호영역을 학습한다. 또한 Persona시스템[9]은 사용자의 검색경로로부터 관심과 비관심 영역을 분류하여 학습한다. 이 방법들은 협력 필터링의 단점을 극복할 수 있는데, 아직 평가되지 않은 새로운 문서에 대해 그 내용을 살펴봄으로써 사용자의 흥미여부를 예측할 수 있기 때문이다. 그러나 다른 사용자들로부터 정보를 필요로 하지 않지만 프로필에 축적된 사용자의 흥미도 외에 새로운 정보를 발견할 수 없는 단점이 있다.

일반적으로 웹 검색 결과를 개선하기 위해 링크 기반의 순위 계산 기법들이 다양하게 개발되었다. Kleinberg가 개발한 HITS 알고리즘[10]은 검색 결과 문서들과 그들과 근접한 문서들로 구성된 웹 그래프에서 hub와 authority들을 검색이 진행되는 시간 내에 추론해내는 방법을 사용한다. Bharat과 Henzinger[11]은 정확도를 증진시키기 위해 HITS 알고리즘을 확장하였는데 질의어의 주제와 관련된 문서들을 추출해 내어 내용 분석을 가미하였다.

HITS 알고리즘이 질의 시간 동안 실행되어 시간적 부담을 가져오는데 반하여, PageRank 알고리즘[12]은 웹의 링크 구조를 기본으로 한 방법으로서, 웹 상의 모든 문서들에 대해 미리 그 중요도를 계산하여 순위 벡타를 유지함으로써 질의에 대한 응답 결과가 지연되는 것을 다소 예방할 수 있다. 보다 많은 수의 유입 링크를 가진 문서는 그 숫자가 적은 문서보다 더욱 중요할 것이라는 아이디어를 기반으로 한다. HITS와 구별되는 PageRank의 또다른 장점은 전체 웹 그래프를 사용하여 순위가 매겨지기 때문에, 국부적인 링크 스팸의 영향을 덜 받게 된다는 점이다.

위의 방법들은 사용자의 흥미도를 고려하지 않았으므로, 보다 개인화된 검색 결과를 위해 사용자 정보를 이용한 방법들이 개발되었다. Page 등은 PageRank 공식에서 사용되는 E 벡타를 사용자 선호도에 맞게 개선하였다[13]. 사용자 선호도는 즐겨찾기와 홈페이지를 참고하여 발견하였다. Haveliwala[14]는 검색어의 주제를 고려하여 문서 순위를 산출하는 방법을 개발하였다. Open Directory Project[15]의 주요 주제 16 종류를 기반으로 미리 PageRank 벡타를 형성한 후, 질의어가 어느 주제에 적합한 것인지를 확률적으로 계산하여, 각 주제에 해당하는 문서들의 순위를 결정한다. 그러나 이 과정에서 사용자의 선호도는 고려하지 않는다. 또한 계산상의 필요에 의해 단지 16개의 주제로 한정하는 제한점이 있다. Teevan 등[16]은 사용자 프로필을 이용하여 검색 결과 문서의 순위를 산정하는 방법을 제안하였는데, 사용자 컴퓨터 내에 저장된 각종 파일들을 프로필로 간주하여, 사용자의 흥미도를 간접적으로 파악하였다.

사용자 기기에 저장된 로그데이타를 분석하여 개인화 기법에 활용한 방안으로서 이승화 등[17]이 제안한 방법은 로그 데이터로부터 사용자의 관심분야를 추론하고 웹 문서에서 정보 블록만을 식별하여 프로파일에 반영함으로써 사용자 정보가 부족한 서버에서 초기 사용자에게 추천 항목을 제시할 수 있도록 하고 사용자의 선호 변화를 반영할 수 있도록 하였다. 사용자 로컬 시스템의 로그 데이터를 분석한 또다른 연구가 김은수 등[18]에 의해 제시되었는데, 이들의 논문에서는 개인화된 광고 기법을 제안하였다. 즉, 사이트 방문 회수를 기초로 사용자 선호도와 성향 및 분류 카테고리를 결정하고, 각 분류 카테고리의 가중치를 계산하여 가중치가 큰 분류의 광고들을 선정하는 방식이다.

사용자의 검색 의도와 관심사에 보다 근접한 검색 결과를 추출하기 위하여, 박건우와 이상훈[19]은 질의어 사용 빈도수와 이에 따른 순위를 데이터베이스로 구축한 후 질의어에 대한 랭킹 정보를 통해 사용자의 주요 관심사를 파악하고 주요 관심사별 커뮤니티를 형성하여 검색을 수행하는 개인화 검색 시스템을 제안하였다. 이제까지의 거의 모든 시스템이 사용자의 선호도를 고려하였으나, 심상희와 이수정[20]이 제안한 시스템에서는 사용자의 이해수준에 맞는 난이도의 문서들을 우선적으로 검색 결과로 제시하였다.

이상과 같이 기존 연구의 주요 관심사는 개인의 선호도의 정확한 측정법에 있으며, 링크수 계산을 통한 알고리즘 개발을 통해 웹 검색 결과를 개선하고자 하였다. 본 연구에서는 개인의 선호를 정확히 파악하여 적용했을 경우 웹검색 결과가 얼마나 향상되는지 실험을 통해 알아보고, 이 밖에 웹검색 결과문서의 선호기준이 무엇인지 파악하고자 한다.

3. 성능 평가 실험

기존 연구에서 웹 검색 결과 개선을 위하여 사용한 주요 방법인 선호어 적용과 문서 링크수 계산의 성능 개선 효과를 알아보기 위하여 다음과 같이 실험하였다.

3.1 실험 배경

실험은 예비 실험과 본 실험으로 구성하여 다양한 연령대의 9명의 실험자를 통해 진행하였다. 우선 예비 실험에서는 임의의 웹 검색 결과 후 문서를 살펴보아 선호한다고 선택한 경우, 그 이유를 작성하게 하였고, 전체 실험자가 나열한 선호 이유들 중에서 주요한 6개의 공통 항목을 도출하였다. 도출된 항목은 문서가 포함한 이미지, 링크, 문서내용, 문서길이, 가독성과 디자인이다. 이와 같은 예비실험 완료 후, 본 실험에서는 검색엔진에서 총 30번의 검색을 하도록 하였으며, 이 때 질의어 개수는 한 개 이상, 임의의 주제로 하였고, 각 검색 결과 웹문서 중 20개를 대상으로 선호문서를 선택하게 하고, 그 이유를 예비실험에서 도출된 항목들 중에서 한 개 또는 복수개 선택하게 하였다. 또한 각 검색마다 질의어와 관련된 선호단어들을 나열하게 하였다.

실험자가 선호한 문서들이 높은 순위로 검색결과에 제시된다면, 그 검색엔진의 성능은 우수하다고 할 수 있다. 즉, 본 실험의 목적을 성취하기 위하여, 질의어 및 선호어들을 기준으로 문서 점수를 산정하고, 또한 링크수를 기준으로 점수를 산정하여, 높은 점수의 문서들이 선호한 문서들이라면, 질의어 및 선호어 또는 링크수 기준으로 문서 순위를 결정하는 것이 바람직하다 하겠다. 전자의 방법을 Q&P, 링크수 기준인 후자의 방법을 LNK로 표기하자. 각 문서에 대해 이 두 방법으로 점수를 산정하는 것 외에, 질의어만을 기준으로 산정하는 방법(QRY라고 표기), 또한 검색엔진의 결과를 그대로 제시하는 방법(SE라고 표기)과 비교하여 성능을 분석하고자 한다.

문서화일 f에 대하여, q를 각 질의어, Q를 질의어 집합, i를 각 선호어, I를 선호어 집합이라고 할 때 SE를 제외한 각 방법에 따른 f의 점수는 아래 계산식에 의한다.

<표 1> 각 방법의 문서 점수 계산식

방법	계산식
LNK	문서가 포함한 링크의 총개수
QRY	$\sum_{q \in Q \wedge f} TF_{q^*} IDF_q$
Q&P	$\sum_{q \in Q \wedge f} TF_{q^*} IDF_q + 0.5 \sum_{i \in I \wedge f} TF_{i^*} IDF_i$

$$TF_q = \frac{f\text{가 포함한 } q\text{의 개수}}{|f|}$$

$$IDF_q = 1 + \log \frac{\text{전체문서수}}{q\text{를 포함한 문서수}}$$

즉, 질의어 개수가 많고 문서 길이가 짧을수록 TF(Term Frequency) 값은 증가하고, 질의어 q가 희귀할수록 IDF (Inverse Document Frequency) 값이 커짐을 알 수 있다. 또한 질의어는 선호도보다 중요하게 취급되어야 하므로, 동일한 TF*IDF 값이라 할지라도 선호어의 가중치는 질의어 가중치의 절반으로 하였다.

3.2 평가 기준

앞절에서 언급한 LNK, QRY, Q&P 세가지 방법들을 선호도를 고려하지 않은 일반 검색 엔진(SE)의 성능과 비교하였다. 일반 검색 엔진에서는 검색된 결과 문서가 제시되는 순서를 그 문서의 순위로 간주한다.

평가 기준으로는 기존 연구에서 흔히 사용하는 Precision, Recall을 도입하고, 여기에 추가적으로 Rank Rate를 새로이 정의하여 사용한다. 성능 측정에 사용할 문서의 범위를 Domain Size(DN이라 표기)라고 하고, 각 시스템에서 산출한 1위부터 DN 순위까지의 문서들을 기준으로 평가한다. 구체적으로 성능 평가 기준의 계산 방식을 설명하면 다음과 같다. Precision이란 검색 결과 DN 순위 내 문서들 중 실험자가 선호한 문서수의 비율을 나타낸 것이다. Recall이란 Precision과 유사한 개념으로 산출 과정이 거의 같으나 단지 나누는 수(제수)를 DN이 아닌 실험자가 선호한 총문서수로 나눈다. 즉, DN 순위 내의 문서들 중 실험자가 선호한 문서수가 총 선호 문서수 중 얼마만큼의 비율로 나타나는지를 보기 위한 값이다. Precision과 recall의 각 비중을 달리하여 통합 측정하는 방식인 F 기준을 사용하면 평가 대상들의 성능을 한눈에 비교하기 편리하다. 본 논문에서는 precision P와 recall R의 비중을 같게 취급한

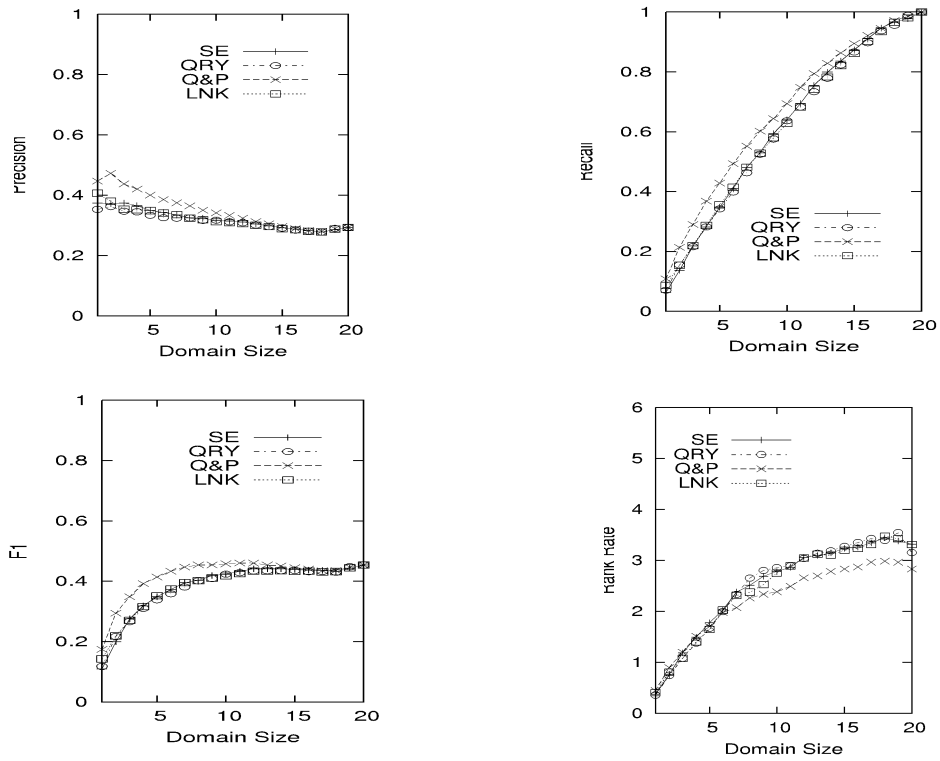
F1을 사용하기로 한다. F1의 계산 방식은 2RP/(R+P)이다[21].

Precision과 recall 값은 Domain Size 내에 선호 문서들이 얼마나 포함되었는지만을 나타내고 얼마나 높은 순위를 차지하는지는 알 수 없다. 이를 해결하기 위한 지표로서 본 연구에서는 Rank Rate를 정의하여 사용한다. 즉, Rank Rate는 실험자가 선택한 문서들이 DN 순위 안에서 얼마나 높은 순위를 차지하는지를 알아보기 위한 것이다. DN 순위 내 문서들 중 실험자가 선호한 문서들의 순위값을 합한 후, 이를 최적의 문서순위합, 즉, 1부터 선호문서수까지의 합으로 나누어 비율로 나타낸다. 예를 들어, DN=10이고 10위 내 선호 문서들의 순위가 2, 5, 7이라면 Rank Rate=(2+5+7)/(1+2+3)=2.33이 된다. Rank Rate 계산식의 분모와 분자가 같으면 최적의 검색 결과이므로 Rank Rate가 1에 근접할수록 우수하다.

3.3 성능 비교

앞에서 언급한 방법들의 성능을 각 실험자별로 domain size에 따라 측정하여 그 평균값을 구한 결과는 <그림 1>과 같다. 네 방법 모두 precision은 domain size가 증가함에 따라 점차적으로 다소 감소하는 경향을 보인다. 이는 선택된 선호문서수의 증가율이 domain size의 증가율을 미치지 못하기 때문이다. 방법들 간에 전반적으로 근소한 성능 차이가 있으나, 작은 domain size에서 Q&P 방법은 다른 세 방법에 비해 최대 약 1.3배의 높은 precision을 보인다. 이는 선호 문서를 선택할 때 링크나 포함된 질의어 수보다 선호어를 실험자가 더욱 비중 있게 고려함을 의미한다. 이와 비슷한 맥락은 recall 성능에서도 나타나는데, 그림 1의 그래프에서 Q&P를 제외한 방법들 간에 거의 동일한 성능값을 보이고 있다. Q&P 방법은 LNK에 비해 최대 약 1.4배, SE에 비해 최대 약 1.7배의 recall값을 나타냈다. Domain size가 커질수록 recall 값이 1에 근접하는 이유는 3.1절에서 언급하였듯이 실험자가 총 20개의 문서들 중에서 선호 문서를 선택하였기 때문이다.

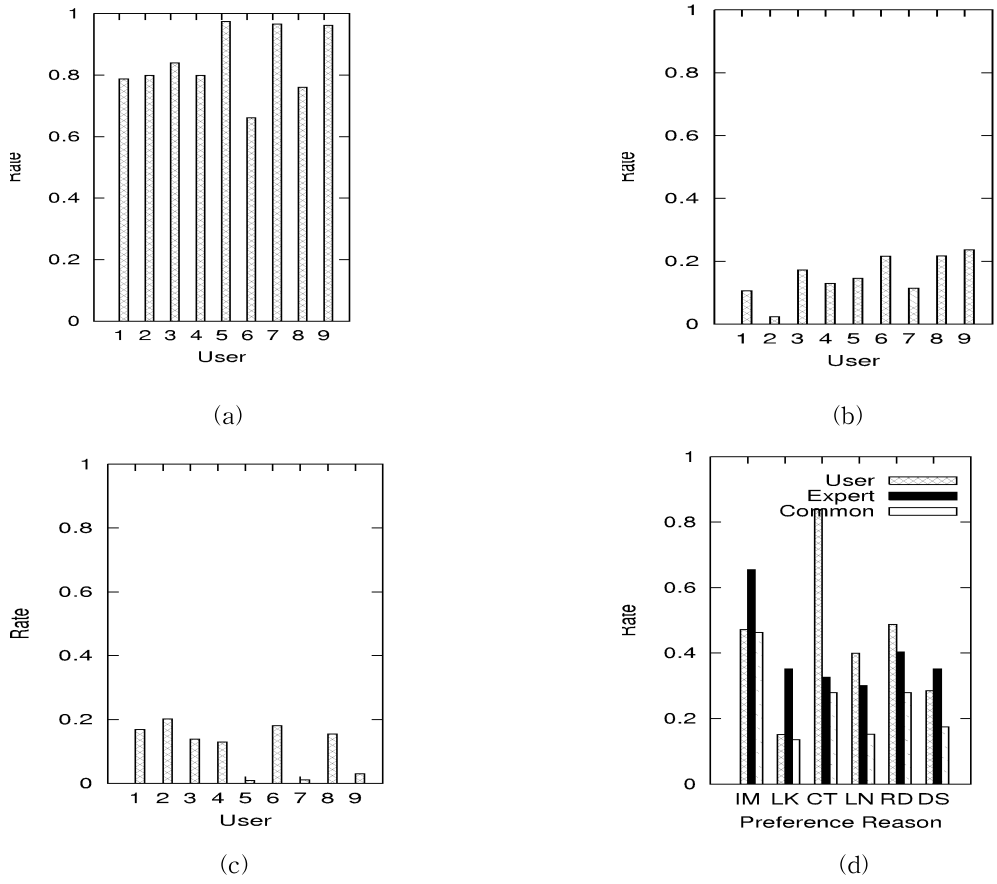
Precision과 recall을 통합 반영한 기준인 F1의 결과를 <그림 1>의 세번째 그래프에서 제시하였다. 앞서의 두 기준과 마찬가지로 Q&P의 성능이 가장 우수하고 나머지 방법들은 유사한 결과를 나타냈다. 구체적으로는 Q&P와 LNK는 SE의 최대 1.7배와 1.3배의 성능 향상을 각기 보였다.



<그림 1> Domain size에 따른 각 방법의 precision, recall, F1과 rank rate.

<그림 1>의 마지막 그래프에서 보듯이 precision과 recall의 경우처럼 rank rate에 있어서 LNK, SE, QRY의 세가지 방법의 성능은 거의 유사하고 Q&P 방법은 이들 방법들을 약간 능가하는 결과치를 보임을 알 수 있다. 이는 Q&P에 의해 문서 순위를 결정하면 실험자의 선호에 맞는 문서들이 다른 방법들에 의한 순위보다 높은 순위를 차지하기 때문이다. 이 결과는 domain size가 작을 때는 드러나지 않지만 커지면 선호 문서수가 많아지므로 차이가 나타나게 된 것이다. 구체적으로, Q&P의 rank rate는 타 방법들에 비해 최대 0.84~0.86배의 범위 내에 존재한다.

이상과 같이 질의어와 선호어를 동시에 고려하는 것이 문서 검색 결과를 개인화하는데 바람직함을 알 수 있었으나, 그 성능 향상 정도는 크지 않았다. 이에 실험자의 선호기준은 무엇인지 살펴보았다. 전체 실험자가 선정한 문서 선호의 공통이유는 6가지로 요약되는데, IM(이미지), LK(링크), CT(내용), LN(문서길이), RD(가독성), DS(디자인)이다. 이들 중 기존 연구의 주요 관심이었던 CT와 LK의 중요성을 파악하기 위해, 전체 선호문서들 중 CT와 LK 각각이 선호 이유(들) 중 하나로 선택된 문서비율을 측정한 결과를 <그림 2>의 (a)와 (b)에 제시하였다. <그림 2>(a)에서 CT의 경우 각 실험자별로 다소 차이를 보였으며, 그 최대 차이값은 약 0.3으로 나타났다. 특히 실험자 6은 약 39%의 선호 문서들을 내용이 아닌 다른 이유 때문에 선택하였음을 알 수 있다. 전체 실험자의 평균값은 약 83%로서 내용이 아닌 다른 이유로 인한 문서 선택률이 17%에 달하여 무시할 수 없는 수치이다. <그림 2>(b)는 링크(LK)가 선호 이유(들) 중 하나로 선택된 선호 문서수의 비율을 보여준다. <그림 2>(a)의 내용(CT)의 경우보다는 훨씬 낮은 비율을 나타내고 있는데, 약 2%~24% 범위의 선호 문서들이 이에 해당하였다. 이러한 결과는 [12][13]에서 제안한 것과 같은 링크 기반의 검색 알고리즘들이 실제 사용자의 선호도에 부합되는 것인지



<그림 2> 실험자별 문서 선호 이유 분포. (a) CT가 선택된 선호문서수 비율. (b) LK가 선택된 선호문서수 비율. (c) CT도 LK도 선택되지 않은 선호 문서수 비율. (d) 각 선호 이유의 선택 비율.

의문을 가져온다. CT나 LK 이외에 다른 이유로 인해 문서를 선호한 비율을 살펴보면 <그림 2>(c)와 같다. 실험자들 간에 차이가 커서 0.8%~20%의 큰 범위를 나타내고 있다. 그러나 6명의 실험자가 15~20%의 선호문서들을 CT나 LK가 아닌 다른 이유 때문에 선호하였으므로, 이러한 결과는 웹 개인화 알고리즘에 반드시 반영되어야 한다고 판단된다.

특정 문서를 선호하는지의 여부는 실험자의 주관이 개입되어 신뢰성이 떨어질 수 있으므로, 각 문서에 대한 객관적인 판단 자료와 비교한 결과를 <그림 2>(c)에 제시하였다. 객관적 판단 기준은 다음과 같다. 우선 IM과 LK는 해당 문서가 관련 이미지나 링크를 포함하고 있으면 선호 이유로 채택하였고, CT는 특정 질의어에 대한 검색 결과 문서들을 Q&P 방법에 따라 정렬한 후 5위 내 문서라면 선호 이유로 채택하였다. 문서 길이는 일정한 범위, 즉, 200~650 단어를 포함하는 문서에 대해 선호 이유로 채택하였다. RD와 DS는 전문가 3인의 의견을 취합하여 2인 이상이 수긍하면 채택하였다. 그림에서 'Common'은 이와 같은 객관적 판단 결과와 실험자의 판단 결과가 일치하는 경우이다. 객관적 판단과 실험자 판단 사이의 차이 정도를 살펴 보면, RD와 DS는 상대적으로 가장 낮고, CT가 가장 높으며 그다음으로 LK인 것으로 확인된다. 따라서, 실험자가 문서 내용을 판단하는 기준과 검색 알고리즘에서 질의어와 선호어 포함 개수로써 판단하는 것과는 상당한 차이가 있음을 알 수 있다. 실험자들이 문서 선호를 판단하는 기준은 CT가 가장 우선적이었으나, 선호 문서들을 객관적인 판단 기준에 의해 살펴보았을 때는 이미지와 가독성이 가장 큰 이유인 것으로 드러났다. 이는 검색 시스템이 내용 뿐만 아니라 문서가 포함한 이미지와 가독성 등을 기준으로 문서 점수를 산정하여 결과를 제시하는 것이 타당함을 말한다.

4. 결론 및 향후 연구 과제

본 논문은 일반적인 웹 검색에서 사용자가 문서를 선호하는 기준을 알아보기 위한 것이다. 9명의 실험자에게 검색 엔진 상에서 임의의 주제로 검색을 실행하게 한 후 선호 문서와 그 이유를 선택하게 하였다. 질의어 및 선호어 기준, 질의어 기준, 링크 기준 등의 방법으로 검색 결과 문서를 정렬한 후 다양한 평가 척도로 각 방법의 성능을 측정하였다. 실험 결과 기존 연구에서 가장 크게 비중을 두었던 질의어 및 선호어 기준의 방법은 검색 엔진에 비해 큰 성능 향상을 이루지 못하였다. 이는 실험자의 문서 선호 이유 중 '내용'이 가장 큰 비중을 차지하였지만, 질의어 및 선호어 기준의 문서 점수 산정 방식이 실험자의 의도를 제대로 반영하지 못하였기 때문이다. '내용' 뿐 아니라 가독성과 문서가 포함한 이미지, 문서길이도 판단 기준으로 높은 순위를 차지한 것으로 밝혀졌다. 따라서 질의어 및 선호어 개수 이외에 객관적 측정이 가능한 데이터를 추가적으로 활용하여 검색 결과를 제시한다면 웹 검색 알고리즘의 성능이 크게 향상될 수 있을 것이다.

본 연구 결과는 전체 실험자의 평균치를 다룬 것이므로, 각 사용자마다 다른 취향과 선호를 반영하여 웹 검색 개인화를 이루기 위해서는 선호 판단 기준의 기록과 분석을 통하여 새로운 알고리즘의 개발이 필요할 것이다. 예를 들어, 데이터 필터링이나 연관 규칙 등의 데이터 마이닝 기법을 본 연구 방법에 적용한다면 보다 성능이 우수한 웹 개인화가 이루어질 것으로 사료된다.

참 고 문 헌

[1] Arotaritei, D. & Mitra, S. (2004). Web mining: a survey in the fuzzy framework. *Fuzzy Sets and Systems*, *148*, 5-19.

[2] Shahabi, C. & Chen, Y.-S. (2003). Web information personalization: challenges and approaches. *3rd International Workshop on Databases in Networked Information Systems*. 5-15.

[3] Linden, G., Smith, B., & York, J. (2003). Amazon.com recommendations : item-to-item collaborative filtering. *IEEE Internet Comput.*, *7*, 76-80.

[4] Miller, B., Albert, I., Lam, S., Konstan, J., & Riedl, J. (2003). MovieLens unplugged: experiences with an occasionally connected recommender system. *Proc. International Conf. on Intelligent User Interfaces*.

[5] Kim, H.-R. & Chan, P.K. (2005). Personalized search results with user interest hierarchies learnt from bookmarks. *7th International Workshop on Knowledge Discovery on the Web*.

[6] Good, N., Schafer, J., Konstan, J., Borchers, J., Sarwar, B., Herlocker, J., & Riedl, J. (1999). Combining collaborative filtering with personal agents for better recommendations. *Conference of the American Association of Artificial Intelligence,* 439-446.

[7] 손창환・김기수 (2005). Web 상에서 개인화된 상품 추천을 위한 Hybrid 추천 시스템에 관한 연구. **한국정보시스템학회 2005년도 춘계학술대회 발표 논문집**, 393-408.

[8] Chen, L. & Sycara, K. (1998). WebMate: a personal agent for browsing and searching. *2nd International Conference on Autonomous Agents,* 132-139.

[9] Tanudjaja, F. & Mui, L. (2002). Persona: a contextualized and personalized web search. *The 35th Annual Hawaii International Conference on System Sciences*.

[10] Kleinberg, J. (1998). Authoritative sources in a hyperlinked environment. *Proc. ACM-SIAM symposium on Discrete Algorithms*.

[11] Bharat, K. & Henzinger, M.R. (1998). Improved algorithms for topic distillation in a hyperlinked environment. *Proc. ACM-SIGIR*.

[12] Brin, S., Motwani, R., Page, L., & Winograd, T. (1998). What can you do with a web in your pocket. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*.

[13] Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The pagerank citation ranking: bringing order to the web. *Tech. Rep. Standford Univ. Database Group*.

[14] Haveliwala, T.H. (2002). Topic-sensitive pagerank. *Proc. 11th Intl. World Wide Web Conference*.

[15] Open directory project. http://dmoz.org/.

[16] Teevan, J., Dumais, S.T., & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *Proc. SIGIR*.

[17] 이승화・최형기・이은석 (2009). 사용자 기기에서 이용한 웹 데이터 분석을 통한 사용자 취향 분석 방법. **정보과학회논문지, 15**(3), 189-199.

[18] 김은수・송강수・이원돈 (2003). 웹 마이닝을 이용한 개인 광고기법에 관한 연구. **한국컴퓨터정보학회 논문지, 8**(4), 92-103.

[19] 박건우・이상훈 (2009). 질의어 패턴 자동분석을 통한 커뮤니티 기반 개인화 검색. **정보과학회논문지, 36**(4), 321-326.

[20] 심상희・이수정 (2009). 사용자의 이해수준에 따른 효율적인 웹문서 검색. **정보과학회논문지, 15**(1), 38-46.

[21] Kuflik, T., Boger, Z., & Shoval, P. (2006). Filtering search results using an optimal set of terms identified by an artificial neural network. *Information Processing & Management, 42*(2), 469-483.

이 수 정

1985 이화여자대학교 과학교육과 (이학사)
1990 미국 Texas A&M대학교 컴퓨터공학과 (석사)
1994 미국 Texas A&M 대학교 컴퓨터공학과 (박사)
1994～1998 삼성전자 통신개발실 선임연구원
1998～현재 경인교육대학교 컴퓨터교육과 교수
관심분야: 컴퓨터교육, 웹마이닝
E-Mail: sjlee@gin.ac.kr