분류와 사용자 질의어 정보에 기반한 개인화 검색 시스템

A Personalized Retrieval System Based on Classification and User Query

김 광 영(Kwang-Young Kim)*
심 강 섭(Kang-Seop Shim)**
곽 승 진(Seung-Jin Kwak)***

목 차

초 록

본 논문은 사용자가 검색에 사용한 질의어를 기반으로 개인의 성향정보를 분석하고자 한다. 이를 위하여 사용자가 검색을 하기 위해서 입력한 질의어를 문서분류기를 이용하여 범주를 부여한다. 본 연구에서는 각 레코드에 미리 부여된 DDC 분류코드를 분류정보로 활용하였다. 이러한 방식을 사용하여 사용자의 질의어를 기반으로 개인의 특징을 분석한다. 분석된 개인의 성향정보를 검색 결과에 반영하고 개인의 의도에 맞는 문서를 재순위화시키는 개인화 검색시스템을 개발하였다. 또한 개인의 성향정보를 이용하여 단어의 중의성 문제를 해결할 수 있었다. 본 논문에서는 한국과학기술정보연구원이 운영 중인 과학기술학회마을 데이터베이스를 이용하여 개인화와 단어중의성 해소에 관한 실험을 수행하였다. 실험과 사용자 평가를 통해서 개인화 검색 및 단어중의성 해소 성능을 제시하였다.

ABSTRACT

In this paper, we describe a developmental system for establishing personal information tendency based on user queries. For each query, the system classified it based on the category information using a kNN classifier. As category information, we used DDC field which is already assigned to each record in the database. The system accumulates category information for all user queries and the user's personalized feature for the target database. We then developed a personalized retrieval system reflecting the personalized feature to produce search result. Our system re-ranks the result documents by adding more weights to the documents for which categories match with the user's personalized feature. By using user's tendency information, the ambiguity problem of the word could be solved. In this paper, we conducted experiments for personalized search and word sense disambiguation (WSD) on a collection of Korean journal articles of science and technology arena. Our experimental result and user's evaluation show that the performance of the personalized search system and WSD is proved to be useful for actual field services.

키워드: 분류, 개인화, 개인화 검색 시스템, 의미 검색
Classification, Personalization, Personalized Retrieval System

---

* 한국과학기술정보연구원 정보기술연구실 선임연구원(kykim@kisti.re.kr)
** 한국과학기술정보연구원 정보기술연구실 연구원(goodsoul@nate.com)
*** 충남대학교 사회과학대학 문헌정보학과 조교수(sjkwak@cnu.ac.kr)
논문접수일자: 2009년 8월 17일 최초심사일자: 2009년 8월 24일 게재확정일자: 2009년 9월 16일
한국문헌정보학회지, 43(3): 163-180, 2009. [DOI:10.4275/KSLIS.2009.43.3.163]

1. 서 론

정보와 전자출판 기술의 발달로 웹 문서, 전자 자료 및 DB들의 양은 기하급수적으로 증가하고 있다. 이러한 정보환경의 변화에 대응하여 1960년대 이후 도서관 및 정보센터에 컴퓨터가 도입된 이래로 정보 검색 시스템은 목록정보의 전산화, 온라인 목록(OPAC) 등이 구축되므로 정보제공의 시간적, 공간적으로 제한점을 어느 정도 해결 할 수 있었다. 그러나 단순히 검색엔진의 등장만으로도 근본적으로 문제를 해결 할 수 없었기 때문에 전통적인 도서관 및 정보서비스 업체에서 제공하는 선택적인 정보배포(SDI: Selective Dissemination of Information) 서비스가 중요한 대안으로 제시되었다(남궁황 2003).

이러한 맞춤정보 서비스는 실제 단순한 사용자의 프로파일 정보에 기반을 두고 프로파일과 일치되는 모든 정보를 제공함으로써 여전히 정보 과잉의 한계점을 극복할 수 없었다. 기존의 정보 검색 시스템들은 시스템 위주로 운영이 되었기 때문이다. 그러나 최근에는 이용자가 웹과 상호작용하는 과정에서 발생하는 일련의 정보탐색과정에 초점을 맞추는 연구가 시도되고 있다(김성진 2006). 또한 정보시스템의 이용자만족지수를 측정할 수 있는 모델을 제시하는 연구도 있었다(김희섭 2004). 이용자의 정보요구 및 성향이 다양해지고 검색 서비스에 대한 기대 수준이 점점 높아지면서 웹 포털 검색뿐만 아니라 의학, 공학 등 전문 검색 서비스에도 개인화와 관련된 다양한 기술들을 접목하여 사용하고 있으며 관련된 연구도 진행되고 있다.

일반적으로 개인화(Personalization)라는 용어는 이용자 정보요구에 부합되는 콘텐츠를 제공한다는 의미로 광범위하게 사용된다(Shahabi, Cyrus and Yi-Shin Chen 2003). Riecken은 "의미 있는 1대1 관계를 형성하여 이용자의 서비스에 대한 충성도(service royalty)를 제공하는 것"으로 규정하였는데, 이용자의 충성도 부분은 특히 포털 서비스에서 가장 중심적으로 추구하는 개념이다.

개인화 검색의 유형은 개인이 적극적으로 프로파일의 사항을 입력하면, 이를 이용하여 기본 프로파일을 작성하는 방법과 서비스 이용형태를 기반으로 하는 방식으로 클릭한 문서, 클릭 수 등의 정보를 이용하여 프로파일을 작성하는 방법과 이용자의 사회화 프로파일(Social Profile)기반으로 하는 방식이 있다(이소영, 정영미 2006).

개인의 어떤 내용을 개인화하여 검색 결과에 반영하는지에 따라 링크정보를 이용하거나, 질의어를 확장하거나, 결과를 재순위화하거나, 메타 검색, 혹은 도메인별 검색 등이 있다. Jeh 와 Widom(2003)은 이용자가 즐겨 찾는 페이지에서 링크되거나, 해당 페이지가 링크한 페이지에 더 많은 가중치를 두어 검색랭킹에 반영하는 형태를 연구하였다.

또한 개인화된 메타검색 엔진들도 개발되었는데, Inquirus2는 이용자가 원하는 주제 분야를 선정하고, 검색엔진을 선택할 때 해당 주제의 내용을 이용하였다(Glover et a1. 2000).

그 흐름을 보면 가장 간단한 유형으로 이용자의 속성 정보에 근거한 개인화이다(Bonnet, Monica 2001). 이용자가 입력한 프로파일에 기반을 두어 맞춤형 검색을 제공하였고 그 다음으로는 이용자의 행동 정보에 근거하는 개인화로서 검색 이력(history) 위주의 다양한 기

능을 제공하는 서비스이다. 2004년 10월 야후가 'My Search'를 시범 서비스로 시작하였으며 검색 결과를 저장하고 편집할 수 있는 기능 위주로 만들어진 시스템이다. 그 다음으로는 이용자의 자산 데이터를 대상으로 하는 개인화이다. 단순하게 개인 컴퓨터의 데이터를 효율적으로 찾아주는 기존의 소규모 솔루션의 데스크 탑 검색 시스템이다. 그리고 2003년부터 폭발적인 인기를 얻고 있는 사회 연결망 개념을 도입한 서비스이다. 기본적인 'My Search'의 개념 모형에 자신의 정보를 다른 사람과 공유하는 것이다. 이 모델의 문제점은 활발한 공유가 가능한 네트워크를 구축하기가 너무 어렵다는 것이다(이소영, 정영미 2006).

대형 웹 포털이나 전문 검색 시스템에서 이러한 개인화 검색 서비스를 제공하기에 현실적으로 많은 어려운 점들을 가지고 있다. 현재 대형 웹 포털 사이트들도 이러한 문제점들을 극복하기 위해서 개인화된 다양한 서비스들을 제공하고 있다. 대표적으로 Google이나 네이버에서 개인화된 일정관리, 포토 앨범, 가계부 등을 제공하고 있다.

개인이 직접 관심 주제를 입력하는 것이 가장 바람직하나 적극적인 사용자가 아닌 이상 이러한 작업들을 번거롭게 생각한다. 또한 로그 데이터를 이용해 사용자의 관심 주제를 추출할 경우에 로그 데이터가 가지는 단편성으로 이용자의 의도나 해석과 같은 것을 정량적으로 측정하기 힘들다. 또한 이용자의 이용형태를 이용하는 방식은 클릭한 사이트, 문서, 클릭 수 등을 피드백 처리하는 과정이 복잡하여 시스템의 많은 부하를 가지고 온다(이소영, 정영미 2006). 개인화 서비스를 위해서 다양한 방법들이 연구되고 있지만 개인의 성향 정보는 계속 변화될 수 있으며 정확하게 판단하는 것이 어렵다. 그러나 사용자들은 검색을 하기 위해서는 반드시 검색 질의어는 입력을 해야만 한다.

본 연구의 목적은 정보 검색의 정확성과 효율성을 높이기 위하여 분류와 사용자 질의어 정보를 이용한 개인화 검색 시스템 개발에 있다. 즉 사용자가 입력하는 질의어만을 이용하여 보다 정확한 사용자 성향 정보를 분석함으로써 사용자를 번거롭게 하는 주제 분야 선택과 이용 형태 정보를 사용하지 않는다. 또한 사용자 성향 정보를 이용하여 단어의 중의성 문제를 해결하여 의미 검색도 가능하게 하고자 한다.

연구목적 달성을 위하여 다음과 같은 연구 내용과 방법으로 연구를 진행하였다. 첫째 사용자의 검색 질의어 중심으로 사용자의 정보를 관리 및 분석하는 시스템을 설계를 한다. 둘째 사용자가 입력한 질의어를 서지 DB에 범주화된 DDC 분류 값으로 분류기를 사용하여 분류를 한다. 셋째 분류된 DDC 정보와 사용자의 질의어 정보를 이용하여 자동으로 사용자의 성향을 분석한다. 넷째 분석된 사용자 성향 정보를 이용하여 검색된 결과를 사용자의 성향에 맞는 문서들을 상위로 재순위화 시킬 수 있는 개인화 검색 시스템을 개발 하였다. 그리고 현재 한국과학기술정보 연구원에서 서비스 중인 과학기술 학회마을[1]에서 제공하는 논문서지 DB 80만 건을 적재하여 실험하고 사용자가 직접 평가하였다.

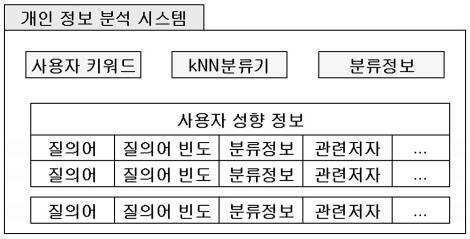
1) 과학기술 학회마을. 〈http://society.kisti.re.kr〉.

2. 분류와 사용자 질의어 기반의 시스템

2.1 질의어 기반의 개인 정보 분석 시스템

본 논문에서는 사용자가 직접 주제 분류를 선택하지 않는다. 직접 시스템이 사용자가 입력한 질의어 기반으로 개인 정보를 자동으로 구축한다. 본 연구에서는 실험적으로 서지 DB에 기술된 DDC 분류를 선택하여 사용하였다. 사용자가 입력한 질의어를 범주화된 서지 DB의 DDC 분류 값으로 분류한다. 즉 사용자가 검색을 하기 위해서 입력한 질의어를 분류기를 이용하여 DDC 값으로 분류를 한다. 그 결과는 3.2절의 〈표 2〉와 같다. 분류기는 kNN[2] 분류기를 이용하여 분류한다. 이렇게 분류된 DDC 정보와 사용자 질의어 정보는 사용자 프로파일에 자동으로 구축된다. 개인 정보를 관리하기 위해서 질의어, 질의 빈도, DDC 분류 값 등을 개인 프로파일에 자동으로 저장하는 시스템을 개발하였다. 본 연구에서는 실험적으로 DDC 분류를 사용하였지만 KDC, UDC, DDC에 상관이 없으면 다른 분류 체계나 개별 시스템의 고유 분류 정보를 이용하여도 된다.

개인 성향 정보 관리를 위해서는 아래의 〈그림 1〉과 같이 사용자가 입력한 질의 정보, 질의어의 빈도 및 질의어를 기반으로 분류된 DDC 정보를 이용하여 사용자의 정보를 분석한다. 또한 개인 정보 분석 시스템에서는 분석된 질의어 기반의 개인 정보(이전 질의어, 질의어 빈도, DDC 분류 정보)와 사용자가 검색을 하기위해서 입력한 질의어를 검색 시스템에 넘겨준다.



〈그림 1〉 질의어 기반의 개인 정보 분석 시스템

검색 시스템에서는 사용자 성향 정보와 검색 질의어를 기반으로 검색을 수행하며 사용자 성향 정보를 이용하여 사용자가 원하는 정보들을 상위에 랭킹 시키는 방법을 이용한다.

개인화 검색 시스템은 개인의 관심분야에 맞는 맞춤 정보를 제공하기 위해서는 많은 정보들이 필요하며, 이런 정보들을 기반으로 개인의 성향이나 관심 분야를 결정하고 결정된 정보를 이용하여 사용자에게 필요로 하는 자료, 문서 등을 제공하는 것이다.

개인 프로파일 정보, 서비스 이용 형태, 사회화 프로파일 정보, 사용자의 질의어 정보 중에서도 가장 중요한 것은 사용자가 입력하는 질의어 정보이다. 그러나 사용자 질의가 갖는 명확한 의도가 검색 시스템에 반영되지 못하는 중요한 이유는 질의어가 갖는 의미적 중의성(Semantic ambiguity) 문제가 있다. 부적절한

2) K-Neatest Neighbors: 문서의 자동분류에서는 새로운 입력문서에 범주를 할당하기 위해 K개의 유사문서로부터 범주별 문서의 분류빈도나 유사도를 이용한다. 본 논문에서 직접 구축한 kNN분류기를 사용하였으며 문서의 자질은 범주가 주어진 문서 내에 출현하는 단어를 추출하고 자질 선별은 주요 단어를 선정하여 문서의 자질로 표현한다. 문서의 분류는 범주-자질 정보와 문서의 자질을 비교하여 범주 부여를 한다.

색인어 및 탐색어의 선정을 비롯하여 동형이의어나 다의어와 같은 중의성을 가진 단어들을 질의어나 색인어로 사용할 경우 사용자가 원하지 않는 검색 결과로 제시할 확률이 매우 높다(윤성희 2007).

일반 웹 검색 환경에서의 사용자들은 다양한 계층 및 연령층으로 구성된다. 트랜잭션 로그 분석을 이용한 연구의 결과를 통해 볼 때에는 '웹 검색에 있어서 사용자의 검색 방식의 단순함'이 가장 두드러진 특징으로 나타나고 있다(윤성희 2007).

일반 웹 검색 엔진에서 사용자들은 정보검색을 정보나 지식을 찾는 것으로 주로 이용한다고 생각하고 있으나, 실제 검색은 미디어성 및 사이트 검색을 더 많이 하고 있는 것으로 분석되었다. 즉, 로그 분석 결과 79.24%가 미디어와 사이트 검색에 치중하고 있으나, 이용자의 53.9%가 정보나 지식을 찾기 위해서 검색 포털을 이용한다고 인식하고 있다(이소영, 조영환 2004).

웹 검색엔진의 탐색관련 특성에 관한 연구에서는 질의어의 수가 1~3개인 질의가 전체 질의의 84%에 달하는 것으로 분석되었으며, 하나의 질의어가 평균 2.2개의 질의어로 구성되었음을 실험을 통해 보이기도 하였다. 이와 같은 사용자의 질의가 단일 어휘이거나 단순한 명사구의 형태를 갖는 많은 경우에는 사용자의 질의 의도는 더욱 판별하기 어렵게 된다(윤성희 2007).

개인 특성을 판단하는 것 중에서 가장 중요한 것은 사용자 질의어들이다. 그러나 사용자의 질의어는 1~2개정도로 짧고 단순한 단어들로 구성되며 또한 중의성을 가진 단어들이 많다. 본 논문에서도 이와 같은 중의성 문제점을 사용자 성향 정보를 이용하여 처리 할 수 있도록 하였다. 본 연구에서는 개인 정보 분석 시스템에서 분석한 사용자의 성향 정보를 이용하여 사용자가 검색을 하기 위해서 입력한 중의성을 가진 단어들에 대해서도 사용자의 성향 정보에 따라 그 결과를 다르게 제공함으로써 단어의 중의성 문제를 하결할 수가 있었다. 예를 들면 아래의 〈표 1〉과 같이 "바이러스"라는 질의어는 다양한 분야에서 사용되고 있다.

〈표 1〉 바이러스 단어의 중의성

바이러스	동물, 식물, 세균 따위의 살아 있는 세포 안에서만 증식이 가능한 미생물
	컴퓨터를 비정상적으로 동작하게 만드는 프로그램

"바이러스"란 단어를 개인화를 하지 않은 일반 검색 시스템에서의 결과는 아래의 〈그림 2〉와 같이 다양한 분야에서 나타난다. 이 경우는 DDC 분류 정보에서는 순수과학, 응용과학, 컴퓨터 분야 등의 다양한 분야에서 "바이러스" 질의어에 대한 검색 결과를 보여 주고 있다. 〈그림 2〉와 같은 사례는 대부분 일반 검색 시스템에서 쉽게 볼 수 있다.

만약 의학 분야에 관심이 많은 사용자가 의학 관련 분야의 "바이러스" 정보를 검색할 경우 〈그림 2〉와 같이 여러 분야에 나타나는 수천 건이 넘는 검색 결과에서 대해서 정확한 정보를 찾는 것은 어렵다. 그러나 분석된 사용자 성향 정보를 사용할 때는 그 결과가 다르게 나타난다. 즉 개인 성향 정보 분석 결과 의학(DDC:610) 분야에 관심이 높은 사용자에게는 검색된 문서들 중에 의학 분야의 "바이러스" 정보를 상위로 랭킹 시킬 수가 있다.

〈그림 3〉에서 "변의학자"[3]가 질의어 기반의 개인 정보를 이용하여 "바이러스" 질의어를 검색한 결과 의학 분야(DDC:610)에 대한 검색 결과들이 상위에 랭킹 되는 것을 볼 수 가 있다. 즉 검색 결과가 사용자 성향 정보에 따라 적합 문서들이 상위로 분포되는 것을 알 수 있다.

〈그림 2〉 개인화를 하지 않은 일반 검색 결과 화면

〈그림 3〉 "변의학자"의 질의어 기반의 개인 정보 중심 검색 화면

3) 의학 분야에 관심을 가지고 있는 가상의 인물.

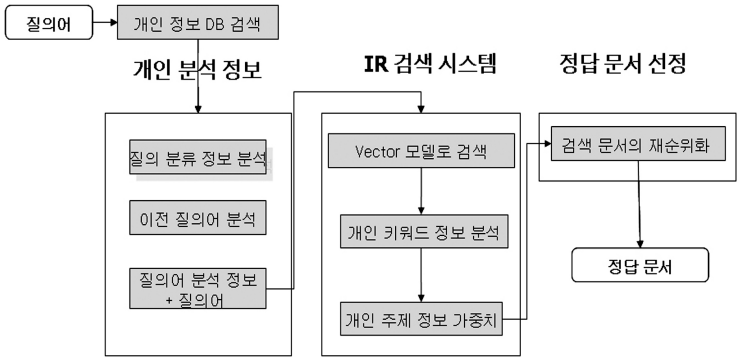
이와 같이 사용자 질의어 정보를 이용하여 범주화된 서지 DB의 DDC분류 정보를 이용하여 개인의 관심 주제를 판단하고 이전의 사용자 질의어를 기반으로 가중치를 다시 재순위화 할 때 사용자가 원하는 정보를 상위 문서로 랭킹 시킬 수가 있었다. 또한 일반 질의어나 중의성을 가진 단어들도 사용자 성향 정보를 이용하여 분별할 수가 있다.

2.2 질의어 기반의 개인 정보를 이용한 검색 시스템

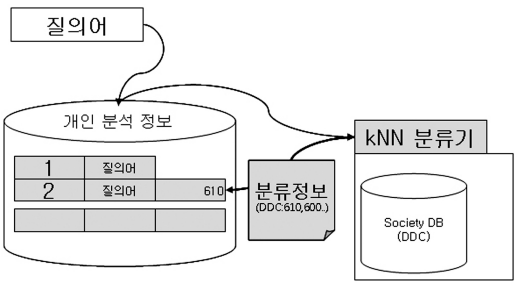
시스템의 구성은 〈그림 4〉와 같다. 〈그림 4〉와 같이 사용자 성향을 분석하는 개인 분석 정보 시스템과 분석된 정보를 이용하여 검색을 처리하는 정보 검색 시스템으로 구성이 된다. 사용자 성향 정보를 분석한 것을 검색 결과에 반영하여 사용자 특성에 맞는 결과를 제공할 수 있도록 시스템 설계 및 개발하였다.

〈그림 5〉는 사용자가 검색을 위해서 입력한 질의어가 이미 분석되었는지 여부를 개인 정보 분석 DB를 검색한다. 만약 새로운 질의어일 때는 사용자의 질의어를 kNN분류기를 호출하여 사용자의 질의어에 대해서 서지 DB에 범주화된 문서들을 DDC값으로 분류한다.

예를 들면 사용자가 "암"이라는 질의어로 검색을 하였을 때 분류기가 서지 DB를 검색하여 "암"이라는 색인어가 나타나는 문서들에 대해



〈그림 4〉 질의어 기반의 개인화 정보 검색 시스템



〈그림 5〉 사용자 질의어를 이용한 분류도

서 DDC 분류 값을 기준으로 분류를 한다. 그 결과 의학(610)에 52%, 응용과학(600) 15% 등으로 나타날 것이다. 분류기에서 분류한 값들 중에서 가장 확률이 높은 값 한 개를 선택을 하게 된다. 분류된 DDC 정보와 사용자 질의어는 개인 분석 정보 DB에 저장 한다.

그 다음 단계로 분석된 사용자 성향 정보를 개인화 검색 시스템을 수행 한다. 검색 결과들은 우선 Vector Space 모델로 문서를 랭킹 시킨다. 그리고 이전 질의어 정보와 DDC 분류 정보로 중심으로 다시 가중치를 계산한다. 그리고 최종 가중치를 중심으로 문서들을 재순위화를 하여 개인화된 검색 결과를 제공한다.

3. 실험 및 결과

3.1 실험 설계

본 논문에서는 과학기술 학회마을의 서지 DB에 기술된 DDC, ACC, UDC 등의 분류 중에 실험적으로 DDC 분류를 이용하여 실험을 하였다. 사용자가 검색을 하기 위해서 입력한 질의어를 DDC 분류 정보 값으로 분류기를 이용하여 분류를 한다. 3.2절의 〈표 2〉와 같이 분류된 DDC값과 사용자 질의어 중에서 빈도(TF)가 높은 것을 선택하여 사용자 성향을 판단하여 개인의 관심 주제를 선정하게 된다.

실험 DB는 KISTI에서 운영하는 과학기술 학회마을 논문서지 DB 80만 건의 문서를 대상으로 하였다. 개인화 검색 시스템의 서버 사양은 리눅스 Redhat 4.1.2, 메모리 12G, 2CPU 인텔 Xeon 1.6GHz를 사용하였다.

본 논문에서 실험 결과를 평가하기 위해서 관련 분야 3명의 평가자가 직접 검색 결과를 평가하였다. 평가 방법으로는 개인별 특정에 맞는 정답 후보 집합을 만들어서 정확도를 평가해야 하나, 본 실험에서는 직접 평가자들이 검색하여 적합하다고 생각하는 문서가 상위에 얼마나 분포하는지 여부를 조사하기 위해서 역순위평균(MRR) 방식을 이용하였다.

역순위평균은 일반적으로 질의/응답 시스템의 평가를 위해서 사용되는 방법으로 정답 문서가 1번째의 순위에 나타나면 1 점, 2번째에 나타나면 1/2 점, N번째의 순위에 나타났으면 1/N으로 점수를 부여하는 방식이다.

본 실험에서는 개인화 검색 결과와 일반 검색 결과를 1~50위까지의 문서 순위 중에서 적합한 문서가 나타나는 순위에 따라 단순하게 역순위 값으로 점수를 부여하고 1~10위, 1~20위, 1~30위, 1~40, 1~50위까지의 구간별 역순위 값의 평균을 구했다. 즉 구간별 평가자가 적합하다고 평가한 문서들이 상위로 어떻게 분포되는지 여부를 실험하였다. 질의어 셋 구성은 〈표 3〉에서 나타나는 중의성을 가진 단어들을 중심으로 한 단어로 구성된 질의어 셋 10~15개로 구성하여 가상의 "변의학자", "요리왕" 및 "사서"에 대해서 평가자가 직접 그 결과를 개인화 검색과 일반 검색을 비교 평가를 하였다.

3.2 질의어 기반의 개인 정보를 판단하는 방법

실험 방법으로는 분류 정보에 따라 이전 질의어의 빈도가 가장 높은 것을 가지고 질의어 기반의 개인 정보를 판단하는 방법을 실험한다. 〈표 2〉는 사용자가 입력한 질의어를 DDC분류

로 분류된 것을 나타낸다. 빈도(TF)는 사용자가 입력한 질의어의 빈도를 나타낸다. 〈표 2〉에서 질의어 빈도(TF)가 가장 높은 DDC값 3개(610, 600, 005)를 선택하여 사용자 성향 정보를 판단하여 개인화 검색 결과에 반영하였다. 〈표 2〉에서 개인의 성향 정보는 의학(610)에 가장 관심이 높고, 그 다음은 응용과학(600), 마지막으로 프로그래밍(005)에 관심이 있는 것으로 나타나고 있다.

〈표 2〉와 같이 DDC 분류 정보 값과 질의어 빈도(TF)를 기반으로 개별 가중치를 계산하여 같은 DDC 분류 정보 값으로 병합하여 최댓값을 선정하여 개인 성향 정보를 분별하는데 사용하였다. 본 실험에서는 실험적으로 최대값들 중에서 3개를 선정하여 사용자 성향 정보를 개인화 검색 결과에 반영을 하였다.

〈식 1〉 $$DDC_i = \frac{tf_i}{\sum_{i=1}^{n} tf_i}$$

〈식 1〉은 각 DDC별로 빈도에 따른 가중치를 계산하는 수식이다. 위의 〈식 1〉를 이용하여 개별 질의어들의 가중치를 계산 한다. 사용자 질의어 대한 벡터 공간 모델의 가중치를 〈식 2〉와 같이 구한다. 각 검색 질의어 별로 문서에 대한 가중치를 계산한다. 구한 결과에 대해서 질의어 기반의 개인 정보에 대해서 α값을 더해서 전체 가중치를 〈식 3〉과 계산한다.

〈식 2〉 $$w_{t,d} = tf_t \log \frac{|D|}{|t \in d|}$$

〈식 3〉 $$Weight_{total}(d) = \frac{(W(d) + \alpha)}{2}$$

$$0 < \alpha \leqq 1$$

전체 가중치는 기본 벡터 모델의 가중치 0~1.0 값과 질의어 기반의 개인 정보 가중치를 합한 0~1.0 값을 2로 나누어 전체 1.0의 가중치를 가지도록 하였다. 알파 값은 〈식 1〉에서 같은 DDC별로 병합한 값으로 전체 1.0을 넘을 수가 없다.

평가 방법은 상위 1~10위, 1~20위, 1~30위, 1~40위, 1~50위 별로 검색한 결과에 대해서 사람이 직접 판단하는 방식을 사용하였다. 사람이 직접 검색한 문서에 대해서 맞는다고 생각하는 것을 〈식 4〉와 같이 역순위평균(Mean Reciprocal Rank) 방식을 사용하여 평가를 하였다.

〈표 2〉 DDC기준으로 분류된 사용자 질의어

질의어	DDC	빈도(TF)	가중치
암	610	19	0.283
항암화학요법	610	13	0.194
키	005	6	0.089
바이러스	600	5	0.074
방사선치료	610	5	0.074
마이크로	600	5	0.074
분자 구조	600	5	0.074
Molecular control	600	4	0.059
	…	…	…

〈식 4〉 $MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}$

사용자가 입력한 질의어들 중 빈도가 높은 순서로 10개 정도를 추출하여 기본 벡터 모델의 가중치와 DDC의 분포에 따른 가중치를 질의어의 빈도를 이용하여 추가하는 방식을 이용하였다. 예를 들면 〈표 2〉와 같이 상위 높은 사용자 질의어가 있다고 가정을 하다면 같은 분류 정보로 묶었을 때 DDC:610의 가중치는 0.641, DDC:600의 가중치는 0.253이고 DDC:005의 가중치는 0.104를 기준으로 질의어 기반의 개인 정보에 따른 가중치를 산정하게 된다.

〈표 3〉은 허정의 연구(허정, 옥철영 2006; 허정, 서희철, 장명길 2006)에서 제시한 중의성 단어들과 인터넷에서 찾은 단어들이다.

〈표 3〉 중의성을 가진 질의어 리스트

키	신장	머리	신체
	열쇠		수량
	컴퓨터 자판	바이러스	동식물에 기생하는 미생물
새	조류		컴퓨터 바이러스
	광석 속의 알갱이	벌	사람에게 주는 고통
	새(관형사)		곤충
여	여자		단위
	여당	돌	광물질
	감탄사		어린아이가 태어난 한 해가되는 날
말	언어	내	개천
	동물		일정한 범위의 안
김	식물(해초)		냄새
	수증기	기관	몸
	잡풀		조직
풀	식물		장치
	쌀/밀가루 전분질에서 빼낸 끈끈한 물질	기사	신문이나 잡지
	기운		운전기사
	수영장		기술업무 담당자
밤	밤나무열매		말을 탄 기사
	낮과밤	기입	수첩이나 문서에 적음
눈	안구		카드 목록 작성
	얼음 결정체(기상)	수서	서명
	눈금		물살이
배	나룻배, 군함		책을 수집
	신체	LC	미국 의회
	과일		약어로 다양하게 사용됨
손	신체	판	출판
	손님		장면
	후손		널빤지

3.3 질의어 기반의 개인 정보를 이용한 평가

3.3.1 "변의학"자 평가

"변의학"자가 일반 검색을 한 경우와 자신의 "의학" 성향 정보를 가진 질의어 기반의 개인 정보를 이용한 경우에 대해서 아래의 질의어 "키"를 입력했을 때를 그 검색 결과에 대해서 성능을 평가 해 본 경우이다.

예를 들면 〈표 4〉와 같이 "키"라는 단어는 다양한 분야에서 사용된다. 의학 분야에서는 신장(몸높이)의 의미로 많이 사용될 것이고 컴퓨터 분야는 문제 해결 실마리/컴퓨터 자판 등으로 사용될 것이다.

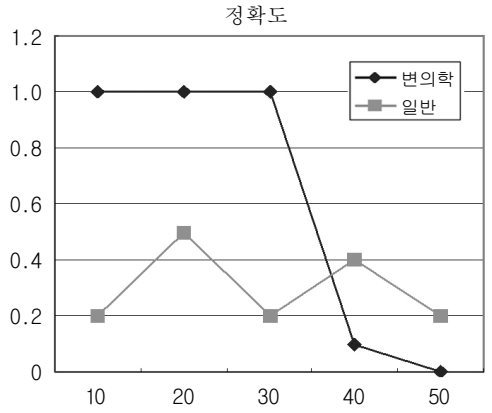
〈표 4〉 "키" 단어의 중의성

키	신장(몸높이)
	열 쇠
	컴퓨터 자판
	…

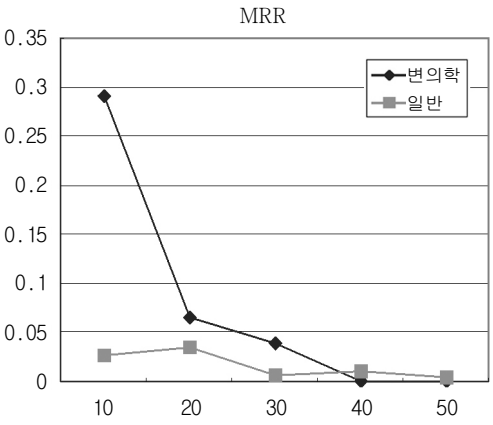
〈그림 6〉은 "키"에 대한 질의어에서 평가자가 상위 10위, 20위, 30위, 40위, 50위까지 구간별 문서에 대해서 검색 결과가 적합하다고 판단한 것에 대해서 MRR 평균을 기준으로 표시한 것이다.

〈그림 7〉은 "키"에 대해서 평가자가 1~10위, 1~20위, 1~30위, 1~40위, 1~50위 까지의 문서에 대해서 검색 결과가 적합하다고 판단한 것에 대해서 역순위평균값(MRR)을 기준으로 표시한 것이다. 〈그림 7〉을 살펴보면 "변의학"자의 질의어 기반의 개인 정보를 이용하여 검색을 수행한 결과들이 상위 문서 1~20위 까지 모두 나타나는 반면 일반 검색으로 했을 경우에는 검색된 문서들이 전체적으로 분포되어지는 것을 볼 수가 있었다.

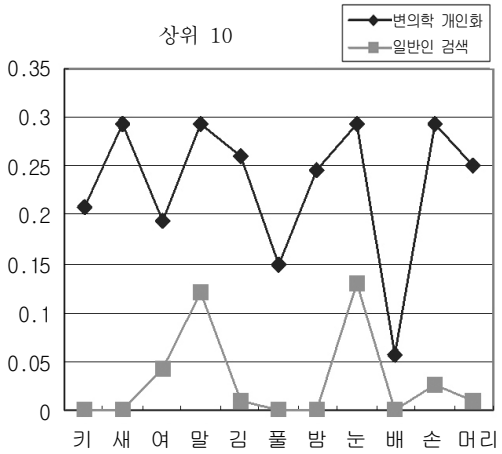
〈그림 8〉은 "변의학"자에 대해서 질의어 기반의 개인 정보를 이용하여 몇 개의 질의어들을 가지고 그 검색 결과를 직접 평가한 것으로 상위 1~10위 까지 MRR 평균값들의 분포도를 나타낸 것이다. 상위 1-10위까지 "변의학"자에 대해서 개인화 처리를 했을 때는 MRR 평균 0.23이고 일반 검색으로 했을 경우에는 0.03으로 나타났다.



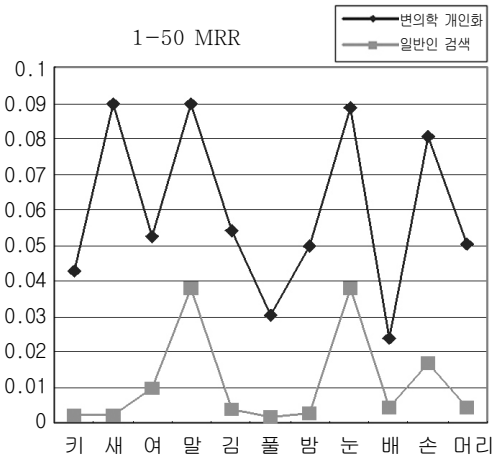
〈그림 6〉 "키" 질의어에 대해서 정확도 기준으로 평가



〈그림 7〉 "키" 질의어에 대해서 전체 구간별 MRR 기준으로 평가



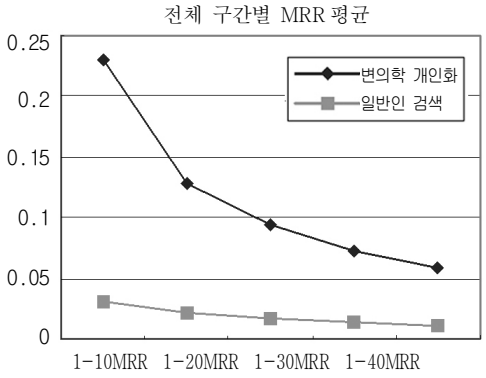
〈그림 8〉 질의어에 대해서 1-10 MRR 분포



〈그림 9〉 질의어들에 대해서 1-50 MRR 분포

〈그림 9〉와 같이 상위 1-50위까지 "변의학"자에 대해서 개인화 처리를 했을 때는 MRR 평균 0.059이고 일반 검색으로 했을 경우에는 0.011로 나타났다.

〈그림 10〉과 〈표 5〉는 "변의학"자의 질의어 기반의 개인 정보를 이용한 질의어들에 대해서 전체 구간별 MRR평균 분포도를 나타낸 것이다. 〈그림 10〉에서 나타난 것과 같이 "변의학"의 질의어 기반의 개인 정보를 이용하여 검색한 경우가 원하는 문서들이 전체적으로 상위에 분포되는 것을 알 수 있다. 반면 일반 검색으로 한 경우에는 정답 문서들이 전체적으로 나타나는 것을 볼 수가 있다. "변의학"자에 대해서 전체 MRR 평균값은 0.116이며 일반 검색에 대해서 전체 MRR 평균값은 0.018로 나타났다. 9.8배정도로 향상된 것을 알 수 있다.



〈그림 10〉 질의어들의 구간별 전체 MRR 평균 분포

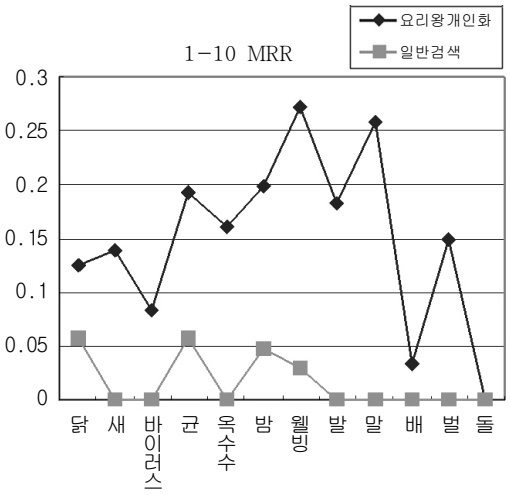
〈표 5〉 질의어들의 구간별 전체 MRR 평균 분포

구 분	1-10MRR	1-20MRR	1-30MRR	1-40MRR	1-50MRR
변의학 개인화	0.230	0.127	0.093	0.072	0.059
일반인검색	0.030	0.021	0.016	0.013	0.011

3.3.2 "요리왕" 평가

"요리왕[4]"이 일반 검색을 한 경우와 자신의 "음식이나 요리" 성향 정보를 가진 질의어 기반의 개인 정보를 이용한 경우에 대해서 아래의 질의어들을 입력했을 때를 그 검색 결과에 대해서 성능을 평가 해 본 경우이다.

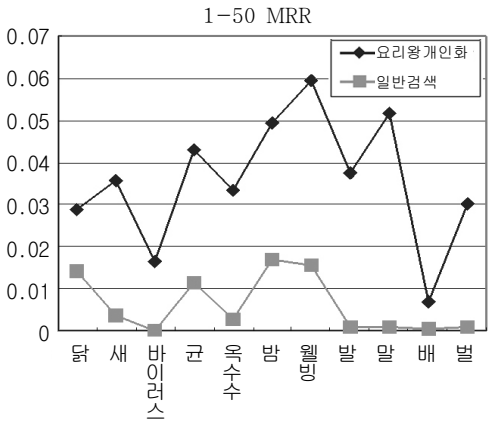
〈그림 11〉은 상위 1-10위 문서 중에서 평균 MRR 값들을 구한 것이다. 위의 그림에서 나타난 결과를 보면 개인화 검색을 한 경우에 보다 좋은 문서들이 상위에 나타나는 것을 볼 수가 있다.



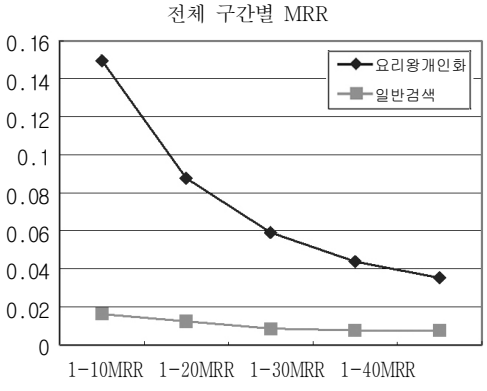
〈그림 11〉 "요리왕" 개인화 1-10 MRR 분포

〈그림 12〉는 상위 1-50위 문서 중에서 MRR 값들을 구한 것이다. 〈그림 12〉에서 나타난 결과를 보면 개인화 검색을 한 경우에 보다 좋은 문서들이 상위에 나타나는 것을 볼 수가 있다. 특히 "닭(닭요리관련)", "옥수수(닭요리관련)", "웰빙(음식관련)" 등의 질의어는 중의성을 가지고 있지 않는 단어들이지만 의미적 검색으로 요리에 관련된 검색 결과들이 나오기를 기대하고 평가를 하였다.

〈그림 13〉은 "요리왕"의 개인화 정보를 이용한 질의어들에 대해서 전체 구간별 MRR평균 분포도를 나타낸 것이다. 〈그림 13〉에서 나타난 것과 같이 "요리왕"의 질의어 기반의 개인 정보를 이용하여 검색한 경우가 원하는 문서들이 전체적으로 상위에 분포되는 것을 알 수 있다. 반면 일반 검색으로 한 경우에는 정답 문서들이 전체적으로 나타나는 것을 볼 수가



〈그림 12〉 "요리왕" 개인화 1-50 MRR 구간별 분포



〈그림 13〉 "요리왕" 질의어에 대한 구간별 전체 MRR 평균 분포

4) 요리 분야에 관심을 가지고 있는 가상의 인물.

〈표 6〉 "요리왕" 질의어들의 구간별 전체 MRR 평균 분포

구 분	1-10MRR	1-20MRR	1-30MRR	1-40MRR	1-50MRR
요리왕개인화	0.149	0.087	0.058	0.044	0.035
일반검색	0.016	0.012	0.008	0.007	0.007

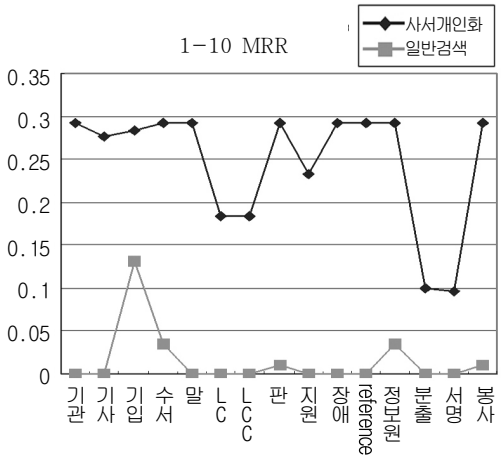
있다. "요리왕"자에 대해서 전체 MRR 평균값은 0.075이며 일반 검색에 대해서 전체 MRR 평균값은 0.01로 나타났다. 6.5배 향상 된 것을 알 수 있다.

3.3.3 "사서" 평가

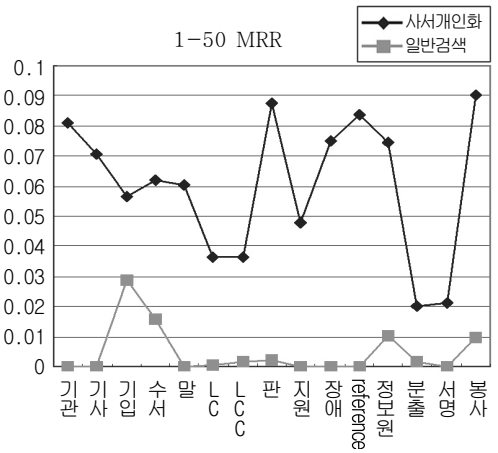
"사서[5]"가 일반 검색을 한 경우와 자신의 "도서관의 사서" 성향 정보를 가진 질의어 기반의 개인 정보를 이용한 경우에 대해서 아래의 질의어들을 입력했을 때를 그 검색 결과에 대해서 성능을 평가 해 본 경우이다. 질의어 중에 "LC"는 도서관에서 주로 미의회도서관을 의미하는 단어로 자주 사용된다. 하지만 일반 웹이나 검색 사이트에서 검색을 했을 때 토익의 LC나 다양한 축약 형태로 검색 결과가 나타난다. 사서 성향을 가진 사용자는 "reference"나 "봉사"라는 단어는 주로 "참고 봉사"를 생각할 것이다.

〈그림 14〉는 상위 1-10위 문서 중에서 MRR 값들을 구한 것이다. 〈그림 14〉에서 나타난 결과를 보면 개인화 검색을 한 경우에 보다 좋은 문서들이 상위에 나타나는 것을 볼 수가 있다.

〈그림 15〉는 상위 1-50위 문서 중에서 MRR 값들을 구한 것이다. 〈그림 15〉에서 나타난 결과를 보면 개인화 검색을 한 경우에 보다 좋은 문서들이 상위에 나타나는 것을 볼 수가 있다.



〈그림 14〉 "사서" 질의어에 1-10 MRR 분포

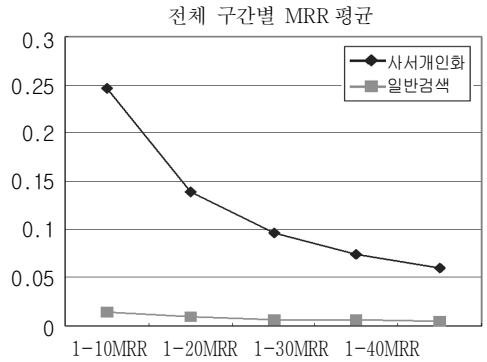


〈그림 15〉 "사서" 질의어에 1-50 MRR 구간별 분포

5) 요리 분야에 관심을 가지고 있는 가상의 인물.

〈표 7〉 사서의 질의어들의 구간별 전체 MRR 평균 분포

구 분	1-10MRR	1-20MRR	1-30MRR	1-40MRR	1-50MRR
사서개인화	0.246	0.138	0.096	0.074	0.060
일반검색	0.014	0.009	0.006	0.005	0.004



〈그림 16〉 "사서" 질의어에 전체 MRR 평균 분포

〈그림 16〉은 "사서"의 개인화 정보를 이용한 질의어들에 대해서 전체 구간별 MRR 평균 분포도를 나타낸 것이다. 〈그림 16〉에서 나타난 것과 같이 "사서"의 질의어 기반의 개인 정보를 이용하여 검색한 경우가 원하는 문서들이 전체적으로 상위에 분포되는 것을 알 수 있다. 반면 일반 검색으로 한 경우에는 정답 문서들이 전체적으로 나타나는 것을 볼 수가 있다. "사서"에 대해서 전체 MRR 평균값은 0.123이며 일반 검색에 대해서 전체 MRR 평균값은 0.008로 나타났다. 11.5배 향상 된 것을 알 수 있다.

4. 결론 및 제언

일반적인 개인화 검색 시스템에서는 개인의 성향 정보를 분석하기 위해서 직접 주제 분야를 선택하고 이용 형태 정보 등을 관리하여 사용자 성향 정보를 분석하고 있다. 이러한 다양한 정보를 분석 및 관리하기 위해서는 많은 양의 데이터를 수집 및 분석해야하는 어려운 점들이 있다. 또한 이렇게 분석된 다양한 성향 정보를 이용하여 개인에 적합한 문서를 랭킹 시키기에는 너무나 많은 변수들을 고려되어야한다. 어떤 변수를 개인화 특성에 반영하는 가에 따라 그 검색 결과는 다르다. 또한 상황에 따라 사용자의 성향이 급변해야하는 경우도 있다. 이러한 많은 변수들을 고려하기 위해서 그 상황에 맞은 많은 모델들을 고려해야만 한다. 대부분의 사용자들은 단순하게 질의어만 넣어서 검색을 하고 있는 반면 적극적인 사용자는 자신의 프로파일의 사항을 입력하고 맞춤형 정보를 찾기 위해서 적극적인 활동을 하고 있다. 하지만 대부분의 사용자들은 간단하게 몇 개의 질의어로 검색을 하고 좋은 결과를 기대하고 있다.

본 논문에서 사용자의 질의어를 서지 DB에 있는 범주화된 분류 정보로 분류하여 사용자의 성향을 판단할 수가 있었다. 서지 DB에 있는 분류 정보는 사람이 직접 분류한 자료들임으로 기계적으로 분류한 자료들보다 정확도가 높을 것이다. 이런 분류 정보와 질의어 정보 들을 이용한 개인화 검색 시스템은 일반 검색 시스템으로 검색 한 경우에 보다 적합한 문서들을 상위로 분포되는 것을 평가 결과로 볼 수 있었다.

평가 결과 가상의 인물 "변의학"자의 경우에는 개인화한 결과는 전체 MRR 평균값은 0.116이고 일반 검색으로 한 경우 전체 MRR 평균값

은 0.018로 약 9.8배 더 향상 된 것을 알 수 있다. 가상의 인물 "요리왕"의 경우에는 개인화 한 결과는 전체 MRR 평균값은 0.075이고 일반 검색으로 한 경우 전체 MRR 평균값은 0.01로 약 6.5배 더 향상 된 것을 알 수 있다. 가상의 인물 "사서"의 경우에는 개인화한 결과는 전체 MRR 평균값은 0.123이고 일반 검색으로 한 경우 전체 MRR 평균값은 0.008로 약 11.5배 더 향상 된 것을 알 수 있다. 위의 실험 결과와 같이 개인화 시스템을 적용했을 때 최소 6.5에서 최대 11.5배 더 성능이 향상되는 것을 볼 수가 있다. 이것은 개인화 검색 시스템에서 사용자의 검색 결과 만족도가 높게 평가되고 있음을 알 수 있었다.

또한 사용자가 질의어를 입력하여 검색할 때마다 사용자 성향 정보가 분류되어 반영이 되기 때문에 개인의 특성을 즉시 반영할 수 있었다. 그리고 사용자 성향 정보를 이용하여 중의성 문제도 해결할 수가 있었다. 일반적으로 중의성은 실제 중의성을 가지는 단어, 색인기에서 잘못 색인된 단어, 외래어, 약어 등등에서 많이 나타나는 것으로 보인다. 의학에 관심이 있는 사용자와 컴퓨터에 관심이 많은 사용자에게 "바이러스", "키" 등의 질의어는 다른 의미적인 정보를 포함하고 있다. 의학에 관심이 있는 사용자에게는 생물적인 바이러스로 분류된 문서들을 상위로 컴퓨터에 관련된 사용자에게는 컴퓨터 바이러스로 분류된 문서들을 상위로 검색 결과를 제공함으로써 의미 검색도 가능하였다.

향후 개인화 검색 시스템의 관련된 연구로는 질의어 기반의 개인 정보 분석 방법 및 가중치를 산정하는 방법에 대해서 더 많은 연구가 필요하다. 이는 가중치 계산 방법에 따라 검색 결과에 많은 변화가 올 수가 있기 때문이다. 그리고 서지DB에서 관련 저자들을 자동으로 군집화하는 방법을 이용하여 개인화 서비스에 이용될 수 있는 모델 개발과 비슷한 성향을 가진 사용자 자동 추천 모델 등이 연구되어야할 것이다. 또한 서지 DB에 분류된 분류 정보에서 대분류만 사용할 것인지, 중분류까지 사용할 것인지 따라 검색 결과가 달라짐으로 연구가 필요하다. 또한 개인화 검색 시스템에 맞는 개인별 정답 집합을 구축하여 정확도를 측정할 필요가 있다.

참 고 문 헌

[1] 김성진. 2006. 이용자 중심 웹 정보탐색 연구의 실체이론 분석. 『정보관리학회지』, 23(3): 127-146.

[2] 김희섭, 박용재. 2004. 정보시스템의 이용자만족지수 모형 개발 및 측정. 『정보관리학회지』, 21(4): 153-171.

[3] 남궁황. 2003. 학습시스템에 기반한 개인화 정보 서비스에 관한 연구. 『정보관리학회지』, 20(4): 113-134.

[4] 윤성희. 2007. 질의어 의미정보와 사용자 피드백을 이용한 웹 검색엔진의 성능향상. 『한국산학기술

학회논문지』, 8(2): 280-285.
[5] 이소영, 정영미. 2006. 웹 포털 이용자 로그 데이터에 기반한 개인화 검색 서비스 모형의 설계 및 평가. 『정보관리학회지』, 23(4): 179-196.
[6] 이소영, 조영환. 2004. 검색 포탈에서 사용자 질의분석을 통한 검색형태 연구. 『정보과학회지』, 22(4): 47-51.
[7] 허정, 서희철, 장명길. 2006. 상호정보량과 복합명사 의미사전에 기반한 동음이의어 중의성 해소. 『정보과학회논문지』, 33(12): 1073-1088.
[8] 허정, 옥철영, 2000. 사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템. 『정보과학회논문지』, 28(9): 688-698.
[9] Bonnet, Monica. 2001 "Personalization of Web Services: Opportunities and Challenages." *Ariadne Issue 28*. [online]. [cited 2009.08.17] 〈http://www.ariadne.ac.uk/issue28/personalization〉.
[10] Glover, E. J., Lawrence, S., Gordon, M. D., Bormingham, W. P., & Giles, C. L. 2000. "Web Search - Your Way." *Communications of the ACM*, 44(12): 97-102.
[11] Jeh, G., & Widom, J. 2003. "Scaling Personalized Web Search." *In Proceedings of the 12th International World Wide Web Conference*, 271-279.
[12] Riecken, D.. 2000 "Personalized Views of Personalization." *Communication of the ACM*, 43(8): 27-28.
[13] Shahabi, Cyrus, & Chen, Yi-Shin. 2003. "Web Information Personalization: Challenges and Approaches." *In Proceedings of 3rd Workshop on Databases in Networked Information System(DNIS)* 5-15.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Sung-Jin Kim. 2006. "Analyzing Substantive Theories in User Studies of Information Seeking on the Web." *Journal of the Korean Society for information Management*, 23(3): 127-146.
[2] Heesop Kim, & Yong-Jae Park. 2004. "Development and Measurement of User Satisfaction Index Model for Information Systems." *Journal of the Korean Society for information Management*, 21(4): 153-171.
[3] Nam-Goong Hwang. 2003. "A study on the personalization information service based on learning system." *Journal of the Korean Society for information Management*, 20(4): 113-134.
[4] Sung-Hee Yoon. 2007. "Improving Performance of Web Search Engine using Query Word

Senses and User Feedback." *Journal of the Korea Academia-Industrial cooperation Society*, 8(2): 280-285.

[5] Soyoung Lee, & Young-Mee Chung. 2006. "Design and Evaluation of a Personalized Search Service Model Based on Web Portal User Activities." *Journal of the Korean Society for information Management*, 23(4): 179-196.

[6] So-Young Lee, & Young-Hwan Cho. 2004. "Search Behavior Research on Internet Search Service by User Query Abalysis." *Journal of Electrical Engineering and Information Science*, 22(4): 47-51.

[7] Jeong Heo, Hee-Cheol Seo, & Myung-Gil Jang. 2006. "Homonym Disambiguation based on Mutual Information and Sense-Tagged Compound Noun Dictionary." *Journal of KISS*, 33(12): 1073-1088.

[8] Jeong Hur, & Cheol-Young Ock. 2000. "A Homonym Disambiguation System based on Semantic Information Extracted from Dictionary Definitions." *Journal of KISS*, 28(9): 688-698.