DOI : 10.5392/JKCA.2010.10.11.461

개인화 검색 시스템 프레임워크 개발

Development of a Personalized Search System Framework

김광영
한국과학기술정보연구원

Kwang-Young Kim(kykim@kisti.re.kr)

요약

본 논문에서는 다양한 콘텐츠들을 특징을 이용하여 각 콘텐츠에 적합한 특징들을 설계하고 수행할 수 있는 개인화 검색 시스템의 프레임워크를 개발하였다. 이를 이용하여 한국과학기술정보 연구원에서 제공하는 국내학술, 특허, 동향 정보 등 다양한 콘텐츠들을 이용하여 개인화 검색 시스템의 다양한 알고리즘에 적용하여 수행하였다.

■ 중심어 : | 개인화 | 개인화 검색 시스템 | 검색 시스템 | 프레임워크 | 분류기 |

Abstract

In this paper, the framework of a personalized search system designing and can perform features which are suitable for each contents by using various contents a feature was developed. In addition, by using various contents including the domestic paper, patent, trend information, and etc., it applied to the various algorithm of the personalized search system.

■ keyword : | Personalization | Personalized Search System | Retrieval System | Framework | Classifier |

I. 서 론

정보와 전자출판 기술의 발달로 웹 문서, 전자 자료 및 DB들의 다양한 콘텐츠들의 양은 기하급수적으로 증가하고 있다. 정보검색 시스템은 이러한 정보환경의 변화에 1960년대 이후 도서관 및 정보센터에 컴퓨터가 도입된 이래로 목록정보의 전산화, 온라인 열람용 목록(OPAC) 등이 구축됨으로 정보제공의 시간적, 공간적으로 제한점을 어느 정도 해결 할 수 있었다.

그러나 단순히 검색엔진의 등장만으로도 근본적으로 문제를 해결 할 수 없으므로 전통적인 도서관 및 정보 서비스 업체에서 제공하는 선택적인 정보배포(SDI: Selective Dissemination of Information) 서비스가 중요한 대안으로 제시되었다[1]. 이러한 맞춤정보 서비스는 실제 단순한 사용자의 프로파일 정보에 기반을 두고 프로파일과 일치되는 모든 정보를 제공함으로써 여전히 정보 과잉의 한계점을 극복할 수 없었다.

인터넷의 발전과 함께 개인 이용자들의 정보요구 및 성향이 다양해지고 검색 서비스에 대한 기대 수준이 점점 높아져다. 이를 위해서 웹 포탈 검색 및 특히 전문 검색 서비스에도 개인화와 관련된 다양한 기술들을 접목하는 시도들이 증가하고 있다.

웹이나 다양한 전문 포탈 검색들의 다양한 콘텐츠에 적합한 개인화 검색 시스템을 개발하기 위해서는 그 콘텐츠에 맞게 최적화 작업을 수행해야 한다.

본 연구에는 제 2장에서는 개인화 검색 시스템에 관

접수번호 : #100727-002
접수일자 : 2010년 07월 27일
심사완료일 : 2010년 09월 15일
교신저자 : 김광영, e-mail : kykim@kisti.re.kr

련된 연구를 조사 및 분석을 하였고 제 3장에서는 다양한 콘텐츠에 대해서 쉽게 개인화 검색 시스템을 적용할 수 있는 프레임워크 개발에 대해서 기술한다. 그리고 마지막 장에서는 4장에서는 결론 및 향후 연구에 대해서 기술한다.

II. 관련 연구

일반적으로 개인화(Personalization)라는 용어는 이용자 정보요구에 부합되는 콘텐츠를 제공한다는 의미로 광범위하게 사용된다[2]. 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[3].

대부분의 개인화 검색시스템에 관한 주요한 기능들은 개인의 클릭 정보나 개인의 검색 히스토리 정보를 이용하여 개인화 검색에 많이 활용을 하고 있다. 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[4].

또한 개인화된 메타검색 엔진들도 개발되었는데, Inquirus2는 이용자가 원하는 주제 분야를 선정하고, 검색엔진을 선택할 때 해당 주제의 내용을 이용하였다[5].

그 흐름을 보면 가장 간단한 유형으로 이용자의 속성 정보에 근거한 개인화이다[6]. 이용자가 입력한 프로파일에 기반을 두어 맞춤형 검색을 제공하였고 그 다음으로는 이용자의 행동 정보에 근거하는 개인화로서 검색 이력(history) 위주의 다양한 기능을 제공하는 개인화 검색 시스템이다. 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 검색시스템이다. 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[3].

개인화의 중심점을 사용자의 질의어 패턴 분석을 통하여 개인의 성향 정보를 분석한 시스템들도 있다. 즉 시스템은 질의어 랭킹 정보(Query Ranking Information)를 참조하여 웹 사용자의 주요 관심사를 파악한다[7].

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[8].

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[9]. 2007년 Ahu Sirg 등은 온톨로지 기반의 사용자 프로파일 정보와 사용자 행위 정보를 이용하여 사용자에게 적합한 문서를 재순위화 시키는 방법을 이용한 개인화 검색서비스를 제안 하였다[10][11].

ODP(Open Directory Project)[12] 텍소노미를 이용하여 다양한 분야 및 웹기반 개인화 검색시스템에 적용한 것도 있다[13][14]. 구글의 PageRank 알고리즘을 확장한 Personalized Page Rank기술을 적용한 웹기반 개

인화 검색시스템도 있다[15]. 이와 같이 개인화 검색시스템에 대한 관심도가 점점 높아지고 있고 활발한 연구들이 진행되고 있다.

III. 개인화 검색 시스템 프레임워크 개발

다양한 콘텐츠에 대해서 개인화 검색 시스템을 적용하기는 어렵다. 현재 개념기반의 개인화 검색 시스템들 대부분은 오픈 소스인 Open Directory Project(ODP) 개념 구조를 이용하여 개인 성향 정보를 주제별로 분류하여 접근을 하고 있다. 예를 들면 OBIWAN 프로젝트[16]는 초기에는 Magellan 사이트의 4 단계의 주제어로부터 4,417 주제어들을 얻어 개념적 계층구조를 참조하여 사용하였지만 그 후에는 오픈 소스인 ODP 구조를 사용하였다. 하지만 ODP는 사이트들을 개념기반으로 계층적으로 분류를 하며 대부분 웹 검색에 기반을 두고 있다.

Sensus 온톨로지[17][18], 7만 개의 노드의 텍사노미와 야후! 디렉토리의 하위집합을 사용하여 그들의 계층적 개념 구조를 사용하였다. Outride 개인화 검색 시스템[19]은 ODP 디렉터리로부터 단지 1,000 개념들을 사용하였고 폭 넓은 동향(trend) 정보를 잡는데 중점을 두었다. 이들 개인화 검색 시스템들은 대부분 웹 검색을 기반으로 하고 있고 비계층적인 구조로 된 개념들에 대해서 적용하기가 어렵다.

본 연구에서는 계층적인 개념 구조 또는 비계층적인 개념 구조에도 쉽게 개인화 검색 시스템을 적용할 수 있는 유연한 구조로 시스템을 개발하였다. 즉 국내학위논문에 대해서는 KISTI 표준 분류 체계를 이용하였고 특허에 대해서 IPC 분류 체계를 이용하였고 동향정보에 대해서는 자체 디렉토리 분류를 체계를 이용하여 개념적 개인화 검색 시스템을 쉽게 구현할 수가 있었다.

1. 개인화 검색 시스템 프레임워크 구조

[그림 1]은 개인화 검색 시스템 프레임워크 구조를 나타내고 있다. 그 구성을 보면 일반 문서들을 적재하고 색인할 수 있는 검색 시스템, 개인의 성향정보를 기반으로 하여 개인 성향에 맞는 문서들을 재순위화 처리를 담당하는 개인화 시스템, 이들 시스템과 사용자 인터페이스와 연결을 위한 자바 API와 실제 사용자들이 다양한 콘텐츠에 쉽게 사용할 수 있는 인터페이스 구성을 위한 FLEX UI 구조로 되어있다.

그림 1. 개인화 검색 시스템 프레임워크 구조

1.1 검색 시스템 구성

검색 시스템은 다양한 콘텐츠들의 검색을 위해서 색인 DB와 원문 정보 등의 관리를 담당한다. 콘텐츠 유형별로 또는 정형화된 데이터, 반정형 데이터 및 비정형화된 데이터들을 처리할 수 있으며 다양한 콘텐츠별로 검색을 수행할 수 있도록 다양한 색인 타입 등을 제공하는 시스템이다.

1.2 개인화 시스템 구성

개인화 시스템은 검색 시스템에서 제공하는 결과를 개인의 프로파일 정보에 따라 재순위화를 처리를 한다. 이를 위해서 [그림 2]와 같이 개인화 프로파일 정보와 사용자가 자주 보는 문서와 자주 검색하는 키워드 정보를 관리하는 시스템을 필요로 한다.

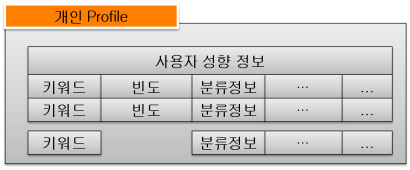

그림 2. 개인화 Profile 정보

개인 프로파일 정보는 기본적으로 사용자가 검색을 하기 위해 입력한 키워드 정보를 중심으로 구성한다. 본 시스템에서는 개인 프로파일을 개념기반에 프로파일을 구성을 한다. 이를 위해서 사용자가 입력하는 키워드 정보를 이용하여 범주화된 각 콘텐츠들 정보들을 중심으로 kNN분류기를 이용하여 분류한 값을 함께 저장을 한다. 이 분류된 값들은 각 콘텐츠의 주제 분류에 값에 따라 다를 수가 있다. 예를 들면 DDC, IPC분류, KISTI 자체 표준 분류(BIST) 등의 정보가 될 수가 있다. 저장된 키워드 정보와 분류 정보를 중심으로 개인화의 성향을 분석하여 검색된 결과에 사용자의 성향 정보를 반영 한다. 각 주제 분류 중에서 대분류, 중분류, 소분류가 있을 경우 본 시스템에서는 대부분 kNN 분류기를 이용하여 중분류를 중심으로 개인의 성향을 유추한다. 그러나 사용하는 콘텐츠에 따라 그 설정을 시스템에서 다르게 설정 할 수가 있다.

사용자 인터페이스 Flex UI에서 개인에게 필요한 개인 키워드 클라우드, 개인 검색 결과의 2차원 결과 값, 개인 주제 정보, 저자 내비게이션 등의 기능들을 호출하여 할 수 있도록 설계되어 있다.

1.3 자바 API 구성

자바 API는 사용자의 UI 정보와 개인화 시스템간의 연결을 담당한다. 직접 검색 시스템을 처리할 수도 있고 개인화 시스템과 연동하여 처리할 수 있는 것이다. 사용자 인터페이스의 다양한 서비스를 처리를 위해서 Flex와 직접적인 연동을 수행한다. 사용자 인터페이스의 다양한 서비스를 제공하기 위해서 직접적으로는 검색 시스템과 개인화 검색 시스템과 연동을 처리한다.

1.4 사용자 인터페이스 Flex 구성

자바 API를 이용하여 개인화에 적합한 인터페이스 정보를 담당하게 된다. Flex UI는 웹 브라우즈에 독립적으로 동작을 한다. 자바 API를 이용하여 개인화 시스템을 위한 시각화를 담당하게 된다. [그림 3]과 같이 사용자의 키워드 클라우드, [그림 4]와 개인 검색 결과의 2차원 구조 표현, 저자 내비게이션 기능 등을 쉽게 구현할 수 있도록 한다. 다양한 콘텐츠에 따라 인터페이스에서는 환경설정에서 필요한 기능을 On/Off처리를 할 수가 있다. [그림 3]은 개인 사용자별로 자신이 자주 검색한 키워드를 중심으로 클라우드 형태로 제공한다. [그림 4]는 개인이 검색한 결과 문서들에 대해서 X축은 가중치 값을 기준으로 표시하고 Y축은 년도 별로 표시하여 가중치가 낮지만 최신 문서를 소팅 처리 없이 한눈에 파악할 수 있고 클릭함으로써 문서의 서지 정보들을 쉽게 볼 수 있는 구조로 구성이 되어 있다.

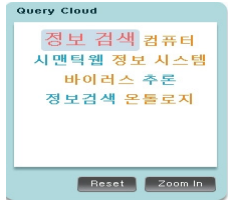

그림 3.개인 키워드 클라우드

그림 4. 개인 검색 결과 2차원

[그림 5]는 개인별로 자주 검색한 키워드를 중심으로 관련된 토픽정보를 클라우드로 보여 준다. 검색 시스템에서 각 콘텐츠별로 필요한 주제 정보를 자동으로 추출하여 미리 구성을 한다. 논문의 경우에는 자동으로 추출하는 주제정보와 논문의 저자 키워드 정보를 중심으로 구성을 한다.

그림 5. 개인 키워드기반 주제 클라우드

또한 사용자의 자주 보는 문서/ 자주 찾는 검색어, 유사한 사용자, 저자 내비게이션(네트워크) 및 일반 검색 기능은 기본적인 개인화 검색 프레임워크에서 제공한다.

2. 다양한 콘텐츠 개인화 검색 시스템 적용

본 시스템은 다양한 콘텐츠를 이용하여 실제 개인화 검색 시스템의 프레임워크에 적용을 하였다.

대상 콘텐츠는 NDSL(www.ndsl.kr)에서 실제 서비스하고 있는 콘텐츠를 이용하였다.

우선 3개의 콘텐츠를 이용하여 개인화 검색 시스템 프레임워크에 적용을 하였다. 첫 번째는 국내학술 저널을 이용하여 KISTI 표준 분류 체계를 이용하여 개념기반의 프로파일을 구성하여 개인화 시스템을 적용하였고, 두 번째는 특허DB는 IPC 분류 체계를 이용하여 개념기반의 프로파일을 구성하여 개인화 시스템을 적용하였고 마지막으로 동향분석 DB는 자체 디렉토리 분류 체계를 이용하여 개념기반의 프로파일을 구성하여 개인화 시스템을 적용하였다.

표 1. 개인화 검색 시스템 적용 콘텐츠

DB명	구성	분류	건수
국내학술 저널	논문	KISTI 표준분류	874,455
특허	국내, 미국특허	IPC	5,768,422
동향분석	글로벌 동향브리핑, 과학기술 정책동향, 분석리포트, 미래정보	디렉토리 분류	135,222

[표 1]과 같이 개인화 검색시스템의 프레임워크에 적용하기 위해서 3가지 콘텐츠별로 각각 시스템을 구성하였다. 각 콘텐츠별로 시스템을 구성할 수 도 있고 전체 하나의 통합 시스템으로 구성을 할 수가 있다. 각 개별적으로 구성할 경우에는 각 콘텐츠별의 주제 분류 정보를 개인화 검색 시스템에 반영함으로서 개별 콘텐츠의 특징을 잘 반영을 할 수가 있었다.

그림 6. 국내학술 저널 개인화 검색 시스템

[그림 6]는 국내학술 저널 80만 건을 이용하여 기본 개인화 검색 시스템의 프레임워크에 적용하였다. 그리고 개별 콘텐츠의 특징을 이용하여 저자 주제어 정보와 공동 저자 내비게이션 등의 추가적인 기능을 할 수가 있다. 국내학술 저널의 경우에는 DDC분류를 중심으로 주제 분류 기반의 개인화 검색 시스템을 적용하였다.

그림 7. 특허 개인화 검색 시스템

[그림 7]은 국내, 미국 특허 5백만 건을 이용하여 기본 개인화 검색 시스템의 프레임워크에 적용하였다. 특허 시스템에서는 특허 분석 맵을 제공하여 출원일/등록일, 발명자, 회사, 기술정보를 중심으로 검색한 문서들에 대해서 간단한 특허 맵을 제공하고 있다. 특허 개인화 검색 시스템에서는 IPC분류 체계를 이용하여 주제 분류 기반의 개인화 검색 시스템을 적용하였다.

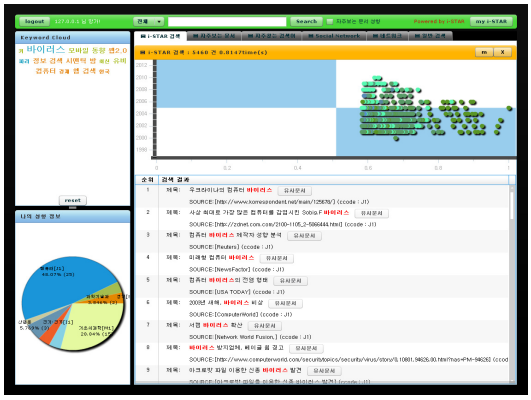

그림 8. 동향 정보 개인화 검색

[그림 8]은 동향분석 13만 건을 이용하여 기본 개인화 검색 시스템의 프레임워크에 적용하였다. 동향분석의 경우에는 디렉터리 서비스에서 사용하는 자체 분류 체계를 이용하였다.

이와 같이 개인화 검색 시스템의 프레임워크를 이용하여 다양한 콘텐츠들을 적용할 수가 있었다.

IV. 결론

본 논문에서는 다양한 콘텐츠들을 특징을 이용하여 각 콘텐츠에 사용하는 주제 분류를 기본으로 구성하여 개인화 검색 시스템의 프레임워크를 쉽게 적용할 수 있는 시스템을 개발하였다. 이를 이용하여 한국과학기술정보 연구원에서 제공하는 국내학술 80만 건, 국내, 미국 특허 300만 건 , 동향 정보 10만 건의 대상 콘텐츠들을 이용하여 개인화 검색 시스템 프레임워크를 적용하고 또한 고유한 콘텐츠들의 특징과 함께 적용될 수 있는 다양한 서비스 및 알고리즘에 적용하여 개발할 수가 있었다. 향후에는 현재 개발된 개인화 검색 시스템의 프레임워크를 서비스 특성에 따라 체계화하여 개인화 검색과 개인화 서비스를 더욱 쉽게 개발 및 적용할 수 있는 시스템을 개발할 필요성이 있다.

참 고 문 헌

[1] 남궁황, “학습시스템에 기반한 개인화 정보 서비스에 관한 연구”, 정보관리학회지, 제20권, 제4호, pp.113-134, 2003.

[2] Shahabi, Cyrus and Yi-Shin Chen, "Web Information Personalization : Challenges and Approaches," In Proceedings of 3rd Workshop on Databases in Networked Information System, pp.5-15, 2003.

[3] 이소영, 정영미, “웹 포털 이용자 로그 데이터에 기반한 개인화 검색서비스 모형의 설계 및 평가” 정보관리학회지, 제23권, 제4호, pp.179-196, 2006.

[4] G. Jeh and J. Widom, "Scaling Personalized Web Search," Proceedings of the 12th International World Wide Web Conference, pp.271-279, 2003.

[5] E. J. Glover, "Web Search - Your Way," Communications of the ACM, Vol.44, No.12, pp.97-102, 2000.

[6] Bonnet, Monica. "Personalization of Web Services: Opportunities and Challenages," Ariadne Issue 28. http://www.ariadne.ac.uk/issue28/personalization

[7] 박건우, 이상훈, “질의어 패턴 자동분석을 통한 커뮤니티 기반 개인화 검색”, 정보과학회논문지, 제36권, 제4호, pp.321-326, 2009.

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

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

[10] 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, pp.525-534, 2007.

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

[12] http://www.dmoz.org

[13] 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, pp.178 - 185, 2005.

[14] 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, pp.465-474, 2006.

[15] G. Jeh and J. Widom, "Scaling personalized web Search," Proceedings of the12th international conference on World Wide Web, pp.271-279, 2003.

[16] Pretschner, "Ontology Based Personalized Search," Proceeding of the 11th IEEE International Conference ICTAI, pp.391-398, 1999.

[17] N. Guarino, C. Masolo, and G. Vetere, "Content-Based Access to the Web," IEEE Intelligent Systems, pp.70-80, 1999.

[18] K. Knight and S. Luk, "Building a Large Knowledge for Machine Translation," Proceedings of AAAI, pp.773-778, 1999.

[19] J. Pitkow and H. Cass, "Personalized search," CACAM, pp.50-55, 2002.

저 자 소 개

김 광 영(Kwang-Young Kim) 정회원

- 2001년 부산대학교 전자계산학과(석사)
- 2001년 ~ 현재 : 한국과학기술정보 연구원 선임연구원

<관심분야> : 정보검색시스템, 개인화, 디지털도서관