기계학습 기반 피드백 과정을 통한 SDI 시스템의 성능향상에 관한 연구*

Machine-Learning Based on Relevance Feedback: A Powerful Engine to Enhance the Performance of SDI System

노 영 희(Young-Hee Noh)**

초 록

정보시대의 도래로 정보량은 기하급수적으로 증가하게 되었고, 이러한 대량의 정보로부터 이용자 개개인에게 적합한 정보를 적시에 제공할 수 있는 방법으로 SDI 서비스가 연구·개발되어 왔지만, 현실적으로 그 활용도는 매우 낮은 것으로 조사되었다. 이에 본 논문에서는 그 원인을 분석하고 SDI 시스템의 성능을 개선시킬 수 있는 적합성 피드백 기반 SDI 시스템을 개발하고자 하였다. 본 연구의 실험을 위해 개발된 실험시스템은 이용자 최소개입 피드백기반 SDI 시스템, 완전자동 피드백기반 SDI 시스템, 그리고 이용자 최대개입 피드백 기반 SDI 시스템이며, 새로 개발된 3개 시스템의 성능 개선정도를 평가하기 위해 네 번째 시스템으로서 전통적인 SDI 서비스에서 사용하고 있는 방법으로 시스템을 개발하였다. 실험결과 이용자 최대개입 피드백 기반 SDI 시스템이 가장 높은 성능을 보여 주었고, 완전자동 피드백 기반, 이용자 최소개입 피드백기반, 전통적 SDI 시스템 순으로 나타났으며, 피드백 기반 시스템들은 피드백이 진행될수록 그 성능이 향상되는 것으로 나타났다.

ABSTRACT

As the Internet facilitates the rapid increase of information availability, the study on SDI service that provides users with relevant document in a timely manner has been developed. However, the practical use of this service has been low. This thesis aims at analyzing the reasons for this and developing relevance feedback based SDI system to improve the performance of the existing SDI system. Experimental systems that are developed for this study are SDI system based on users' minimum intervention feedback, SDI system based on perfect automation feedback, and SDI system based on users' maximum intervention feedback. The fourth system that utilizes the traditional SDI system is also studied to evaluate the level of performance improvement of the newly developed three types of SDI system. As a result of this study, SDI system based on users' maximum intervention feedback showed greatest performance improvement. The next performance improvement happened in order of SDI system based on perfect automation feedback, SDI system based on users' minimum intervention feedback, and the traditional SDI system. Feedback based systems showed greater performance improvement as they went through more feedback processes.

키워드: 기계학습, 적합성 피드백, 최신정보주지, 적합문헌
SDI, Machine Learning, Relevance Feedback

---

* 이 논문은 2004년도 건국대학교 신임교원 연구비 지원에 의한 논문임.
** 건국대학교 문헌정보학과 교수(irs4u@kku.ac.kr)
■ 논문접수일자 : 2004년 11월 17일
■ 게재확정일자 : 2004년 12월 13일

1. 서 론

1. 1 연구의 배경

SDI 서비스에 대한 연구는 인터넷의 등장으로 인한 정보량의 기하급수적 증가로 인해 다시 활기를 찾기 시작했다. 대량의 정보로부터 각 개인에게 적합한 정보를 적시에 제공할 수 있는 방법으로 SDI 서비스 또는 PUSH 서비스 등이 활용되었으며, 특히 도서관 및 정보센터에서는 현재까지 수작업으로 진행되었던 최신정보주지서비스를 자동화하여 왔다. 그러나 그 활용도는 매우 낮은 것으로 조사되었다.

2003년 4월에 시행된 국내 4년제 국공립 및 사립대학도서관의 SDI 서비스의 제공 현황을 조사한 바에 따르면(노영희 2003), SDI 서비스를 제공하고 있는 대학은 19.1%에 지나지 않는 것으로 나타났으며, 그 서비스의 질도 매우 낮은 것으로 분석된 바 있다.

활용도가 낮은 이유로 첫째, SDI 시스템이 제공하는 검색결과의 낮은 만족도이다. 둘째, 부적합문헌이 제공되었을 때 이용자는 SDI 시스템에 재접속하여 정보요구 프로파일을 수정해야 하는데, 이용자는 이러한 작업을 매우 번거로워 하는 경향이 있다.

이에 본 연구에서는 많은 학자들의 SDI 서비스의 유용성 및 활용성의 주장에도 불구하고 여전히 활성화되지 못하고 있는 국내 SDI 서비스의 성능을 개선하고자 하였으며 그 결과 서비스의 활성화에 많은 기여를 할 수 있을 것으로 보인다.

1. 2 연구의 목적

SDI 서비스는 이용자의 정보요구를 프로파일에 등록하여 놓고, 새로운 정보자료에 대한 최신 정보를 수록한 데이터베이스가 만들어지면 등록된 프로파일과 대조하여 적합한 문헌만을 검색하며, 검색된 결과를 프로파일을 작성한 각 이용자에게 제공하는 최신정보주지서비스의 일종이다. 이용자는 자료의 제공 주기 및 제공될 자료의 주제범위 등 다양한 선택옵션을 통해서 관심 주제 분야의 자료를 e-mail을 통해서 또는 서비스 시스템의 개인공지화면을 통해서 신속하게 받아 볼 수 있는 매우 유익한 서비스이다.

그러나 이용자가 시스템에서 정보검색을 한 후 그 결과가 만족스럽지 못할 경우 바로 재검색을 수행할 수 있는 소급탐색과 달리 SDI 시스템에서는 시스템이 제공한 검색결과를 이용자가 받아 본 후 그 제공된 결과가 만족스럽지 못할 경우, 이용자는 SDI 시스템에 다시 접속하여 자신의 관심주제가 표현될 수 있도록 정보요구 프로파일을 수정해야 한다. 이러한 번거로운 작업은 SDI 시스템의 이용률을 낮추는 원인이 될 것이다.

또한, 특정 주제 분야에 대해서 연구를 시작한 연구자는 연구를 진행하다가 연구방향을 수정하거나 연구주제를 세부 주제로 좁혀갈 수 있으며, 이에 따라 새로운 문헌을 탐색하기 위한 탐색문, 즉 SDI 시스템의 정보요구 프로파일을 수정할 필요가 있다. 만약 이용자가 시스템에 재접속하여 시스템의 프로파일을 수정하지 않고도, 정기적으로 제공되는 SDI 시스템의 검색결과 중 적합문헌만을 선정하여 시스템

에 재전송하는 간단한 피드백과정을 이용하여 자동으로 관심주제의 변화를 반영할 수 있다면, SDI 시스템의 효율성 및 이용자의 편의성은 크게 개선될 수 있을 것으로 보인다.

이와 같이 만족스럽지 못한 검색결과의 제공을 지양하고 동일한 관심 주제 분야 내에서의 주제의 세분화 및 연구방향의 변화를 반영하기 위한 방법으로 이용자와 SDI 시스템간의 피드백 과정을 적용할 수 있다. 피드백 과정이 있는 SDI 서비스를 이용할 경우 이용자는 검색결과로 제공된 문헌의 적합성을 판정한 후 그 판정결과를 시스템에 전송함으로써 이용자의 요구를 반영할 수 있다.

즉, 피드백을 통한 정보요구의 변화를 반영하는 가장 간단한 방법으로, 이용자가 시스템이 제공한 검색결과로부터 자신의 요구에 적합한 몇 개의 문헌을 선택, 표시함으로써 시스템이 자동으로 이용자 프로파일을 학습하고 이를 바탕으로 새로운 문헌을 이용자에게 제공할 수 있도록 하는 것이다. 이용자가 제공한 이 자료들은 시스템의 적합문헌집단이 되며, 이 적합문헌집단을 활용하여 이용자의 변화된 관심 주제 분야를 반영할 수 있다. 계속적인 피드백 과정을 통하여 이용자의 관심변화에 따라 적합문헌집단도 달라지며 변화된 적합문헌집단을 기반으로 신착문헌집단으로부터 이용자의 요구에 적합할 문헌을 검색한다면, 이용자의 만족도는 높아질 것으로 보인다.

본 연구에서는 이와 같이 적합성 피드백을 기반으로 한 SDI 시스템의 성능을 개선할 수 있는 방안을 연구하였다. 본 연구의 실험을 위해 개발된 실험시스템은 이용자 최소개입 피드백기반 SDI 시스템, 완전자동 피드백기반 SDI 시스템, 그리고 이용자 최대개입 피드백 기반 SDI 시스템이며, 새로 개발된 3개 시스템의 성능 개선정도를 평가하기 위해 네 번째 시스템으로서 전통적인 SDI서비스에서 사용하고 있는 방법으로 시스템을 개발하였다.

본 논문을 수행함으로써 다음과 같은 부분에 기여하고자 하였다. 첫째, 이용자(연구자)가 자신의 정보요구를 한 번 작성한 후 수정을 위해 시스템에 재접속할 필요성을 줄인다. 둘째, 최소한의 이용자 개입으로 적합문헌집단을 형성하고 이 적합문헌집단을 기반으로 이용자의 관심 주제분야의 변화를 반영할 수 있도록 한다. 셋째, SDI 서비스의 최대 목표로서, 연구자에게 최적의 정보를 적시에 제공함으로써 정보의 활용도를 높이고 연구지원 및 학문발전에 기여할 수 있도록 한다.

2. 이론적 배경

2.1 선행 연구

SDI 관련 연구로 이용자 만족도 조사, DB 통합적 SDI 서비스, SDI의 자동화로 인한 서비스 향상, SDI 유료화가 서비스에 미치는 영향, SDI의 유용성 및 성능평가, 그리고 각종 상용화된 데이터베이스의 SDI 서비스 제공사례 등에 대한 연구 논문이 비교적 많이 발표되었으며, 1970년대 이후 주춤했던 연구가 인터넷의 활성화로 다시 활기를 찾기 시작했다. 특히 이용자 피드백을 통한 SDI 시스템의 성능을 향상시키고 활성화 시키려는 연구 노력이 1990년대 후반에 두드러지게 나타나게 된다.

일반적으로 정보검색 분야 연구에서는 이용

자의 역할에 그다지 많은 관심을 두지 않았었다. 특히 이용자 관심의 정의, 관심의 적절한 표현, 그리고 시스템과 이용자간의 상호작용을 통한 관심사의 표현에는 소홀했었다(Belkin & Croft 1992). 이용자 프로파일에 대한 Myaeng와 Korfhage(1990)의 연구는 이 분야의 몇 개 안되는 중요한 연구노력 중의 하나이다. 그들은 이용자의 정보요구 프로파일을 정보검색시스템에 통합하여 개선된 검색시스템에서 질의와 프로파일을 다양하게 조합하였다. 그러나 그 프로파일들은 이용자가 직접 입력해야 했다.

이용자의 관심변화를 자동으로 반영하기 위해 Rich(1983)는 이용자의 요구정보가 충분하지 않은 상황에서 장기적인 관심사를 표현할 수 있는 이용자모형을 개발하기 위한 방식을 제안했다. 그러나 이 방법은 이용자 모형(user stereotype)을 구축하는데 있어서 지식공학의 실질적인 인간개입을 요구했다. 따라서 질문 갱신목적으로 개발되었던 매우 제한적이고 간접적인 이용자 요구의 표현방법인 적합성 피드백이 사용되었다(Frants et al. 1993; Goker & McCluskey 1991).

이후 인공지능 기반 기법을 적용해서 프로파일을 수집하고 유지하는 문제와 관련된 연구가 진행되었다. Malone(1987) 등은 이용자가 규칙을 사용해서 프로파일을 생성할 수 있게 하는 InfoLens라는 지능형 메시지 공유시스템(intelligent message-sharing system)을 개발했다. 그 규칙은 메시지 유형, 날짜, 그리고 보낸 사람 등과 같은 내용기반 요소에 따른 해당 행위를 기술한다. 그와 같은 명확한 이용자 기반 지식수집방법은 높은 투명성과 최신성을 유지할 수 있게 한다.

이와 유사한 기법을 적용한 InforScope 시스템은 유즈넷 뉴스를 필터링 하기 위해 개발되었다(Fischer and Stevens 1991). InforScope은 적절한 실행을 위해 공통적인 이용패턴(세션의 수, 뉴스그룹 읽기, 기사 내 적합용어의 빈도 등)과 관련된 발견적 규칙을 사용한다. 프로파일을 수정하기 위해 이용자는 프로파일로부터 용어를 제거하거나 추가해야 하며 적절한 규칙 작동기준을 설정해야 한다. 이 시스템은 프로파일 관리를 위해 이용자가 요구를 직접적이고 명확하게 입력할 것을 요구하며 그와 같은 규칙기반 접근방법은 효과적인 프로파일 수정에 많은 단점을 드러냈다.

News-Weeder(Lang 1995)도 유즈넷 필터링 시스템이다. 시스템 이용자의 문헌 평가는 기계학습알고리즘을 위한 학습집단으로 사용되며, 밤에 실행되어 다음 날 이용자 정보요구 프로파일을 생성한다. 이 시스템은 이용자가 문헌 평가만 하도록 제한하기 때문에 이용자의 관여를 줄일 수 있으나 온라인 상태에서 프로파일을 수정할 수 없어서 유용성을 제한하기도 한다.

SIFT(Stanford Information Filtering Tool) 또한 유즈넷 뉴스를 필터링하기 위해 개발되었다(Yan & Garcia-Molina 1995). SIFT는 이용자가 초기 프로파일을 생성하도록 핵심어를 입력할 수 있도록 하고 있다. 이용자의 선택에 따라, 필터는 벡터스페이스 모형 또는 간단한 불리언 모형을 사용하여 표현된다. 벡터스페이스 접근방법이 선택되면, 프로파일 수정에 유동성이 있게 되고, 이용자가 적합성 피드백(관심있는 문헌을 체크함으로써)을 할 수 있도록 하며, 이를 기반으로 프로파일 내 가중치가 수정된다.

NewT(news tailor)는 공통적인 주제 영역을 포함하는 미리 정의된 프로파일 집합을 이용자에게 제공하고, 하나 이상의 프로파일을 선택할 수 있도록 하였다. 또한 프로파일 수정을 위한 적합성 피드백을 제공한다. NewT는 프로파일의 적응성을 향상시키기 위한 유전적 알고리즘(genetic algorithm)을 사용한다(Seth 1994).

Callan(1996)은 InRoute라는 통계적 필터링 시스템을 개발했다. InRoute가 추론망 모형(inference network model)을 기반으로 하고 있지만, 이는 전형적으로 가장 통계적인 필터링시스템이다. 그는 논문에서 질문의 구조화를 위해 MinTerm 색인기법이라는 새로운 프로파일 선정기법을 제안했다. 이 기법은 특히 필터링 속도에 있어서 매우 효과적인 것으로 밝혀졌다.

Mostafa 등(Mosta, Mukopadhyay, Lam, & Palakal 1997)은 정보필터링 환경에서, 이용자 관심의 변화 및 유동적인 문헌의 흐름과 관련된 불확실성은 효과적으로 관리되어야 한다고 주장하였다. 그들은 논문에서 이러한 불확실성에 대처하기 위한 다중 적응기법(multiple adaptation techniques)의 필요성을 강조하고 인공지능을 기반으로 한 SIFTER라는 필터링 시스템을 개발하였다. SIFTER 시스템은 내용과 이용자의 특정 관심사를 기반으로 한 필터를 수행한다. 이용자의 관심사는 선택적인 적합성 피드백으로 최소한의 이용자 개입을 바탕으로 자동으로 학습되고 갱신될 수 있다.

Amati 등(amati, Crestani, & Ubaldini 1997)은 정보필터링과 선택적 정보배포를 위한 학습시스템(learning system)을 제안했다. 전자정보량이 증가하고 있는 오늘날의 상황에서 이들 자료를 선택적으로 선별해서 배포할 필요가 있고 효과적인 정보 필터링 시스템은 이용자가 최소한의 노력으로 정보요구를 기술하고 그 요구에 맞는 적합한 정보를 제공하는 시스템이라고 주장하고 있다. 이 논문에서 그들은 일반화된 확률검색모형을 응용한 정보필터링 학습모형을 제시하고 그 성능을 평가하였다. 이 모형은 "불확실성 샘플링(uncertainty sampling)" 개념을 기반으로 하며, 적합문헌과 부적합문헌에 대한 적합성 피드백과정을 이용하는 기법이다. 그들이 제안한 학습모형은 ProFile(Probabilistic Filtering)이라는 정보필터링시스템의 핵심이 되었다.

Blake(1997)는 그의 논문에서 인트라넷의 개개 이용자에게 유료서비스를 제공하는 개인화된 경고서비스, 'Muscat'에 대해 논의하고 있다. Muscat 경고서비스는 확률검색과 통계적 기법을 통합한 검색기법을 사용하고 있으며, 검색결과를 받은 이용자가 검색된 문헌에 마킹함으로써 초기 탐색문의 탐색어 가중치가 수정될 수 있도록 하고 있다.

Walker 등(Walker, Robertson, Boughanem, Jones, & Spark Jones 1997)은 기울기 값을 산출하고 응용함으로써 Okapi 시스템을 적응형 정보필터링으로 확장하며 이용자 프로파일을 개선하기 위해 논리적 회귀분석 방법을 사용한다.

Oard와 DeClaris(1995)는 필터링이 인터넷 정보량의 폭발적인 증가로 그 중요성이 향상 되었다고 주장하고 이 분야의 최근 동향을 살펴보고 이용자 평가에 대한 기록을 바탕으로 인지적이고 상호작용적인 실험시스템을 개발했다. 상호작용적인 필터링 시스템 모형은

'Gaussian' 이용자 모형이라는 새로운 인지적 필터링 기법을 사용하고 있다. 이 연구에서 이 시스템의 성능이 높은 것으로 분석되어 Cornell SMART 텍스트 검색시스템에 적용하기도 하였다.

Boughanem과 Tmar(2002)은 적응형 필터링 과정을 제안했다. 적응형 필터링은 문헌과 이용자 프로파일을 비교하는 과정에서 이용자 프로파일과 배포기준값을 수정함으로써 시간이 지남에 따라 필터링의 성능이 향상될 수 있도록 하는 방법이다. 이용자 프로파일 수정 방법으로는 검색된 적합문헌에 가중치를 높게 부여하는 것이며, 배포기준값 또한 필터된 적합문헌 분석을 통해 수정된다.

단어의 사용은 주제분야 의존적이며, 한 분야에서 일반적으로 사용되는 단어가 다른 분야에서는 그다지 빈번하게 사용되지 않을 수 있다. Singhal(Singhal, Mitra, & Buckley 1997) 등은 이러한 단어 사용의 속성을 활용해서 문헌 필터링을 개선하고자 하였다. 연구결과 질문영역(query domain)에 있는 문헌으로부터 학습된 'routing query(정보요구 프로파일)'가 질문영역이 사용되지 않는 문헌으로부터 학습된 프로파일보다 8~12% 정도 높은 성능을 보여주었다.

2. 2 SDI 서비스 모형개발

선행 연구를 조사한 결과 1990년대 이후 이용자 피드백을 통한 SDI 시스템의 성능향상과 서비스 개선에 관한 연구가 비교적 활발하게 이루어졌다는 것을 알 수 있다. 그러나 대부분의 연구가 이용자가 판정한 적합문헌과 부적합문헌을 이용하여 프로파일 내 용어를 변경하거나 용어의 가중치를 변경하는 방식이었다. 본 연구에서는 이와 같은 접근 방법과 함께 이용자의 정보요구에 적합한 적합문헌 집단을 형성한 후 이 문헌집단과 신착문헌집단을 비교하는 접근방법을 사용하였다.

이를 위해 본 연구에서 제안하고 있는 SDI 모듈은 크게 3가지이며, 나머지 하나의 모듈은 현재까지 개발되어 서비스되고 있는 모듈로서, 본 연구에서 개발될 시스템의 성능을 평가하기 위해 개발된 것이다. 피드백 기반 SDI 서비스 모듈로 첫째는 이용자 최소개입 피드백기반 SDI 서비스이고, 둘째는 완전자동 피드백 기반 서비스이며, 셋째는 이용자 최대개입 피드백 기반 서비스이다. 나머지 하나는 전통적인 SDI 서비스 방법이다. 이용자 최소개입과 최대개입 서비스의 경우 중간에 이용자가 개입하여 적합문헌을 선정해야 하는데, 본 실험은 실험문헌집단에 이미 선정되어 있는 적합문헌 집단을 사용하고 있다. 그 내용을 구체적으로 설명하면 아래와 같다.

1) 이용자 최소개입 피드백 기반 서비스

이용자의 개입을 최소화하기 위해 처음 한 번만 이용자 피드백을 수행하여 최초의 적합문헌집단을 형성하고, 그 적합 문헌집단과 새로운 문헌집단을 비교하여 검색결과를 이용자에게 제공하는 방법이다.

이용자의 피드백이 발생하기 전에는 적합문헌 집단이 없으므로 이용자의 정보요구 프로파일과 신착 문헌집단을 비교하여 검색결과를 제공해야 한다. 이용자가 시스템에 제공한 검색결과를 평가 한 후 그 중 적합문헌을 선정하여

SDI 시스템에 전송함으로써 최초의 적합문헌 집단이 생성될 수 있다.

최초의 적합문헌 집단을 생성하기 위한 키워드와 신착문헌집단과의 유사도는 코싸인 유사계수 공식을 사용하게 된다. 이 때 이용자가 최초로 입력한 키워드의 가중치는 1이고 신착문헌 집단 내 각 용어의 가중치는 tf*idf 공식을 이용한다.

그 다음 단계부터 시스템은 이용자의 정보요구 프로파일의 키워드가 아닌 적합문헌집단과 신착문헌 집단을 비교하여 그 검색결과를 이용자에게 제공하며 이용자는 다시 적합문헌을 재전송할 필요가 없다(그림 1 참조).

2) 완전자동 피드백 기반 서비스

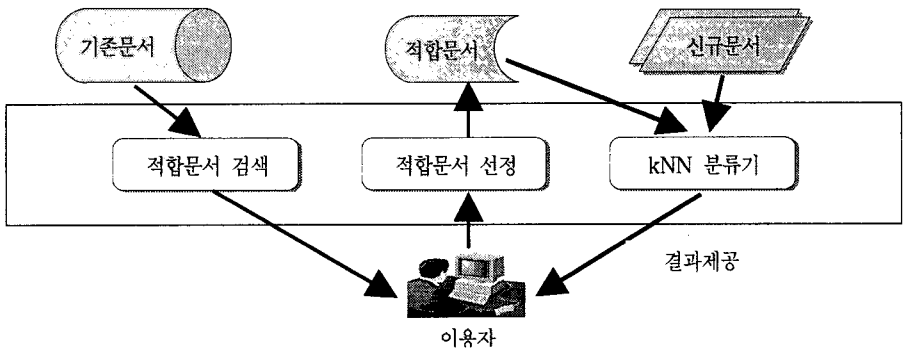
이용자가 SDI 서비스 프로파일을 작성한 후 최초의 SDI 서비스 결과를 제공하기 전에 시스템이 자동으로 소급탐색을 수행한다. 소급탐색은 이용자의 프로파일을 기반으로 소장문헌 집단에 대하여 탐색하고 코사인 유사계수 공식을 기반으로 유사도 값을 산출한다. 유사도 순으로 검색결과를 정렬한 후 상위 순위의 문헌을 적합문헌 집단으로 분류하고, 이후에 제공되는 SDI 서비스는 이 적합문헌집단과 신착문헌집단을 비교하여 적합문헌집단과 유사도가 높은 문헌만을 제공하는 방법이다.

적합문헌집단은 피드백이 진행됨에 따라 계속적으로 수정되어야 하는데, 그 수정방법으로 본 모듈에서는 검색된 문헌과 기존의 적합문헌집단을 통합하여 가중치 순으로 정렬한 후 유사도가 0.3 이상인 문헌 중 상위 10위의 문헌만을 적합문헌집단으로 포함시키도록 하였다.

3) 이용자 최대개입 피드백 기반 서비스

소급탐색을 기본으로 하고 계속적인 피드백 과정을 통해서 적합문헌 집단을 수정해 가면서 이용자에게 적합문헌을 제공하는 모듈이다.

소급탐색을 수행하여 얻은 적합문헌집단은 이용자가 등록해 놓은 이용자의 질문프로파일에 의해 검색된 것일 뿐 이용자의 요구에 적합하다는 판정이 내려진 집단은 아니다. 따라서 이용자가 정기적인 SDI 서비스를 받기 전에 시스템이 소급탐색을 수행해서 형성한 적합문헌 집단을 먼저 본 후 적합문헌을 선정할 수 있는 절차를 삽입한다. 이용자는 소급탐색 기반 검색결과를 직접 평가하고 몇 개의 적합문



〈그림 1〉 이용자 최소개입 피드백 기반 서비스 시스템 구성도

헌을 선정함으로써 이용자의 의견이 반영된 적합문헌집단을 형성할 수 있으며, 따라서 이용자의 개입이 전혀 없는 두 번째 방법보다는 높은 성능을 보여줄 것으로 예측된다.

본 서비스 방법의 SDI 서비스 과정은 〈그림 2〉와 동일하나 단지 이용자가 적합문헌을 선정하는 과정을 〈그림 2〉는 단 한번만 수행하지만 본 시스템에서는 매번 이용자의 요구를 반영하기 때문에 그 과정이 반복된다는 것이 다르다.

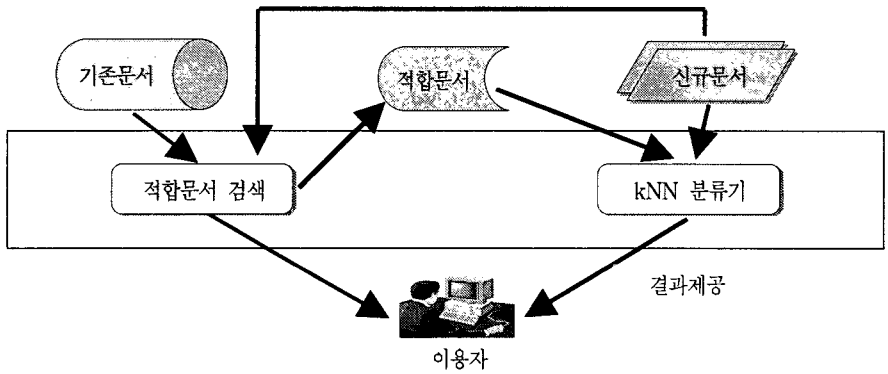
4) 전통적인 SDI 서비스 방법

가장 일반적으로 사용되고 있는 기존의 방법으로서 이용자가 직접 키워드를 수정하게 하는 방법이 있으나 본 실험의 목적은 SDI 질문 프로파일을 한 번 작성한 후 시스템에 재접속하지 않고도 간단한 피드백 과정을 거쳐서 적합문헌 집단을 계속적으로 받아 볼 수 있도록 하는 방법이다. 이 시스템에는 적합문헌집단이 없으며, 실험에서 제공되는 네 번의 SDI 서비스는 모두 이용자가 처음 등록해 놓은 이용자 프로파일만을 기반으로 제공된다.

3. 실험시스템 설계

3. 1 실험집단

본 연구를 위해 사용된 실험문헌집단은 KTSET 1.0이며 실험문헌 집단의 크기는 4,414건이다. KTSET은 정보과학회지, 정보과학회 논문지, 정보관리학회지에 실린 논문들의 초록으로 구성된 한국어 정보검색 실험집단이다. 이 실험문헌집단은 소급탐색을 위한



〈그림 2〉 완전자동 피드백 기반 서비스 시스템 구성도

〈그림 3〉 전통적인 SDI 서비스의 시스템 구성도

문헌집단 1개와 SDI 서비스를 위한 몇 개의 신규문헌집단으로 구분된다.

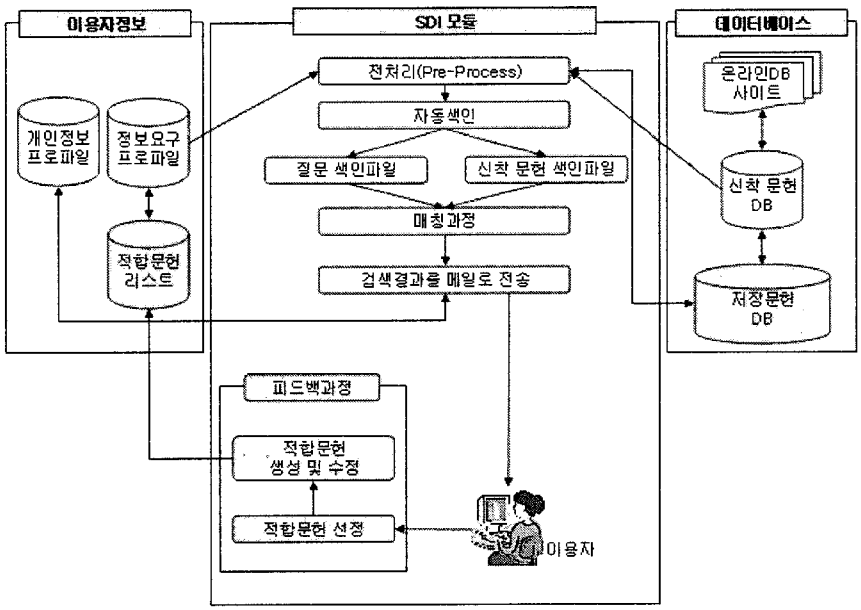
신규문헌집단의 수는 피드백을 몇 단계까지 할 것인가에 따라 달라진다. 본 연구에서는 피드백 과정을 4단계까지 하기로 결정하고 4개의 실험문헌집단을 형성하였다. 이 중에서 하나의 집단은 소급탐색용 실험집단이 되는데, 이는 실험설계 중 소급탐색을 기반으로 한 SDI 탐색을 수행하는 단계가 실험과정에서 수행되기 때문이다. 소급탐색용 집단의 크기는 전체 실험문헌집단 크기의 2분의 1로 잡았다. 또한 5개의 신착문헌집단의 크기는 약 800건씩으로 하였다.

실험을 위한 질문 프로파일의 개수는 30개로 하였다. 질문 프로파일은 자연언어가 아니라 탐색어로 이루어지며 질문프로파일을 이루는 탐색어의 개수는 2-6개로 이루어져 있다.

탐색문과 신착 문헌집단과의 유사도는 코싸인 유사계수 공식을 사용하여 산출하고 이용자에게 문헌을 제공할 때는 유사도순으로 순위를 매겨 제공한다. 이용자는 순위를 참조하여 피드백 과정에서 적합문헌을 선정할 수 있다.

3. 2 실험시스템 설계

본 연구의 실험을 위해 개발된 실험시스템은 2장 2절에서 설명했듯이 크게 4개로 구분된다. 첫째는 이용자 최소개입 피드백기반 SDI 시스템이다. 둘째는 완전자동 피드백 기반 시스템이고, 셋째는 이용자 최대개입 피드백 기반 시스템이며, 네 번째는 전통적인 SDI 서비스 시스템이다. 전체적인 시스템의 구성도는 〈그림 4〉와 같고, 각각의 실험시스템에 대해 구체적으로 살펴보면 아래와 같다.



〈그림 4〉 피드백 기반 SDI 시스템 구성도

1) 정보요구 프로파일 로드(load)

이용자 정보파일에는 개인정보 프로파일과 정보요구 프로파일이 있다. SDI 모듈은 이용자의 정보요구를 처리하기 위해 정보요구 프로파일로부터 이용자의 질문을 가져온다.

2) 데이터베이스로부터 각 문헌정보 로드(load)

데이터베이스에는 도서관에 입수되는 신착문헌 DB와 온라인데이터베이스, 전자저널 및 인터넷 문서 등이 있으며 SDI 시스템은 각 문헌들의 정보를 가져온다.

3) 전처리과정

이 단계에서는 정보자료 및 이용자 정보요구 파일에 대한 자동색인 과정을 거치게 되며, 각 문헌에 출현하는 색인어의 가중치도 산출한다.

4) 매칭과정

전처리 과정을 거친 정보요구 파일 및 문헌정보파일을 비교하는 단계이다. 이용자의 개별 정보요구에 대한 적합문헌을 산출하는 기법은 불리언 검색기법, 매칭함수 기법 등 기타 다른 다양한 검색기법을 적용할 수 있으며, 사용되는 검색기법에 따라 검색결과에 순위를 매겨 이용자에게 제공할 수도 있다.

5) 검색된 결과를 이용자에게 전송한다.

6) 이용자는 제공된 결과의 적합성을 판정하고 적합문헌을 선정해서 SDI 시스템에 전송한다.

7) 이용자 정보요구 프로파일 갱신

이용자가 제공한 적합문헌을 기반으로 적합문헌 집단을 생성하고 피드백 모듈에 의해 이용자 정보요구 프로파일을 갱신한다.

8) 1)~7) 과정 반복

1)에서 7)과정을 반복하지만, 7)에서 생성된 적합문헌 집단과 이용자가 처음 작성했던 정보요구 프로파일을 동시에 로드한다.

3. 3 적합문헌집단 선정 및 갱신

본 연구에서는 이용자 피드백과정을 통해서 SDI 시스템의 성능을 향상시키고자 하였다. 이 때 SDI시스템에는 각 이용자의 정보요구 프로파일에 해당하는 적합문헌집단이 존재해야 한다. 이용자가 정보요구 프로파일을 작성하면서 적합문헌집단을 등록하는 방법이 가장 효율적인 방법이겠지만, 이러한 방법은 이용자에게 매우 번거로운 일이 될 것이다.

따라서, 적합문헌집단은 최소한의 이용자 개입으로 작성되거나 이용자가 전혀 개입하지 않고 작성되는 것이 바람직하다. 적합문헌집단은 다음과 같이 세 가지 방법으로 형성될 수 있다. 첫째, 이용자가 정보요구 프로파일을 작성하면, 그 프로파일을 기반으로 소급탐색을 수행함으로써 검색된 문헌집단을 코싸인 유사계수 공식을 사용하여 순위화한 후 상위 순위의 문헌들을 적합문헌집단으로 분류하는 방법이다.

둘째, 첫 번째 방법으로 작성된 적합문헌집단은 피드백과정이 반복됨에 따라 갱신되어야 할 것이다. 적합문헌집단을 갱신하는 방법은 다시 두 가지로 나뉘며, 이용자로부터 적합문헌집단을 피드백 받는 방법과 시스템이 자동으로 갱신하는 방법이다. 이용자로부터 적합문헌집단을 피드백 받는 방법은 처음에 소급탐색에 의해 작성된 적합문헌집단을 기반으로 이용자에게 SDI 서비스가 제공된 후 이용자로 하여금 제공된 문헌 중 적합문헌들을 선정해서 시스템에 보내도록 하되, 아주 간단히 그 일을

수행할 수 있도록 하는 방법이다. 시스템이 자동으로 갱신하는 방법은 1차적인 SDI 서비스 제공을 위해 신착문헌집단과의 유사도를 비교할 때, 검색된 문헌집단과 적합문헌집단에 속한 각 문헌들을 대상으로 이용자가 처음 제공한 정보요구 프로파일과의 유사도를 산출하고 유사도순으로 정렬하여 특정 순위까지의 문헌을 다시 적합문헌집단으로 재선정하는 과정을 반복하는 방법이다.

셋째, 적합문헌집단이 시스템에 의해서 작성되는 것이 아니라 처음부터 이용자가 작성하는 방법이다. 즉, SDI시스템이 이용자의 정보요구 프로파일에 기반해서 검색된 정보자료를 이용자에게 제공하면, 이용자가 제공된 정보자료 중에서 자신의 요구에 적합한 문헌을 선정해서 시스템에 보냄으로써 작성되는 방법이다.

위에서 설명했듯이 적합문헌집단을 이용자가 직접 제공하는 경우와 적합문헌집단을 시스템이 자동으로 생성하는 경우로 나누어 볼 수 있다. 적합문헌집단을 자동으로 생성하기 위해서는 1차적으로 소급탐색을 수행해야 하며, 소급탐색에서 탐색 및 검색결과의 순위화에 사용한 공식은 유사계수 공식이다. 유사계수 공식에는 다이스 계수, 자카드 계수 등 다양한 공식이 있으나 본 연구에서는 코싸인 유사계수 공식을 사용하였다.

$$W(D_i, Q_j) = \frac{\sum_{i=1}^{t} t_{ik} \times q_{jk}}{\sqrt{\sum_{i=1}^{t} (t_{ik})^2 \times \sum_{i=1}^{t} (q_{jk})^2}} \quad <\text{공식 } 1>$$

공식 1은 소급탐색용 문헌집단내 각 문헌 i와 이용자 정보요구 프로파일내 질문 j간의 유사도를 측정하는 공식이다. n은 데이터베이스 내 문헌의 총수를 나타낸다. t_{ik}는 문헌 i내의 용어 k의 가중치를 나타내며($0 \leq d_{ij} \leq 1$), q_{jk}는 질문 j내의 용어 k의 가중치를 나타낸다. 매칭함수에 의해 산출된 유사도순으로 검색된 결과가 순위화되고 이용자의 요구에 적합할 확률순으로 정렬된다. 이 때 정보요구 프로파일을 이루는 탐색어의 가중치는 1로 하였다.

본 연구에서는 이용자의 정보요구 프로파일당 적합문헌집단의 크기인 k 값을 10으로 하여 실험하였다.

3. 4 적합문헌집단과 신착문헌과의 유사도

적합문헌집단과 신착문헌과의 유사도는 kNN 분류기의 원리를 이용하여 산출할 수 있다. kNN(k-Nearest Neighbor classification)은 몇 년 동안 패턴인식분야에서 집중적으로 연구되어온 통계적 접근방법으로 잘 알려져 있다(Dasarathy 1991). 그 이후 kNN은 다른 기계학습 기반 자동분류 알고리즘과 함께 문헌범주화에 응용되었다(Masand, B., G. Linoff, & D. Waltz 1992; Yang 1994; Iwayama, Makato, & Takenobu Tokunaga 1995).

kNN 알고리즘은 새로이 분류될 입력문헌이 있을 때, 시스템이 학습문헌집단 중에서 k개의 최근접 문헌을 찾아낸다. 그리고 k개의 최근접 문헌들이 할당된 범주정보를 이용하여 후보 범주에 가중치를 부여할 수 있다. 즉, 입력문헌과 각 근접문헌과의 유사도는 이웃문헌이 속한 범주의 가중치가 되는 것이다. 만약 k개의 최근접 문헌 중 여러 개가 하나의 범주에 분류되어 있다면 여러 개의 근접 문헌들의 가

중치가 그 범주에 모두 더해지며, 그 결과로서 나온 가중치의 합은 입력문헌에 대한 그 범주의 유사도로 사용될 수 있는 것이다. 후보 범주의 가중치를 정렬하여 입력문헌을 후보 범주 중 하나에 최종 분류할 수 있다.

본 연구에서는 이러한 kNN알고리즘의 원리를 이용하여 적합문헌집단과 신착문헌과의 유사도를 산출할 수 있도록 하였다. 즉, 신착문헌이 입력되면 이 신착문헌과 적합문헌집단에 속한 문헌들과의 유사도를 산출할 수 있으며, 신착문헌과 적합문헌집단 내 각 문헌과의 유사도의 합의 평균을 구하여 그 유사도 평균이 특정 기준치 이상인 신착문헌들만 해당 프로파일들의 이용자에게 제공하도록 할 수 있다. 신착문헌과 적합문헌집단과의 유사도를 산출하는 공식은 다음과 같으며, Yang(1994)의 kNN 공식을 변형한 공식이다.

아래 공식에서 신착문헌 Nx와 특정 이용자의 정보요구 프로파일에 해당하는 k개의 적합문헌집단 각각의 문헌들간의 유사도를 구하고 모두 합산하여 이용자의 정보요구 프로파일에 대한 적합성 점수, $rel(Pk|Nx)$를 구한다. 적합성 점수의 평균을 구하여 특정 기준치 이상인 신착문헌들만을 이용자에게 전송한다.

$$rel(P_k \mid N_x) \approx \sum_{j=1}^{j=k} sim(N_x, R_j) \times \frac{1}{k} \quad \text{<공식 2>}$$

위 식에서 sim(Nx, Rj)는 공식 3과 같이 다시 쓸 수 있고, 선정된 자질들을 벡터로 표현하기 위해 $\log tf$를 가중치로 사용하였으며, 신착문헌 N_x와 적합문헌집단 R_j의 유사도를 산출하기 위해서 다음과 같이 코싸인 유사계수 공식을 이용한다.

$$W(N_x, R_j) = \frac{\sum_{i=1}^{n} t_{xk} \times t_{jk}}{\sqrt{\sum_{i=1}^{n} (t_{xk})^2 \times \sum_{i=1}^{n} (t_{jk})^2}} \quad \text{〈공식 3〉}$$

위 공식은 신착문헌 x와 적합문헌집단 j간의 유사도를 측정하는 공식이다. txk는 신착문헌 x내의 용어 k의 가중치를 나타내며, t_{jk}는 적합문헌 j내의 용어 k의 가중치를 나타낸다.

3. 5 자질 축소

자질 축소란 특정 문헌을 대표하는 단서어로 그 문헌을 대표하게 할 때, 그 문헌에 출현한 모든 색인어를 사용하는 것이 아니라 문헌빈도(Document Frequency)를 이용하여 문헌빈도 순위 20~30%에 해당하는 색인어만을 자동분류의 자질로 사용하는 것을 말한다.

본 연구에서는 적합문헌집단과 신착문헌의 유사도를 산출할 때 적합문헌집단 내 각 문헌과 신착문헌의 유사도를 산출할 때 사용될 자질을 100%로 하였다. 자질 축소를 하지 않고 불용어 등을 제외한 색인어 전체를 자질로 사용하였다.

이는 몇몇 연구에서 보여 주듯이 자질을 축소함으로써 성능이 크게 개선되는 것이 아니라 단지 처리 속도가 빨라질 뿐이며 성능에는 그다지 영향을 주지 않는 것으로 나타났기 때문이다(Chung & Noh 2003). 이 연구에서는 자질을 축소함으로써 정확률은 조금 올라갔으나 반대로 재현율은 떨어지는 경향을 보였다.

또한, 본 연구에서 자질축소를 하지 않은 두 번째 이유로 실험문헌집단이 초록정보의 길이이기 때문에 자질축소의 의미는 크지 않을 것

으로 판단되었기 때문이다.

3. 6 평가방법

본 연구에서 다양한 방법으로 개발된 피드백 기반 SDI 시스템의 성능을 평가하기 위해 사용한 평가척도는 재현율과 정확률, 그리고 복합척도이다. 이 중 복합척도는 재현율(R)과 정확률(P)을 모두 반영하여 평가하는 방법이다(정영미 1993). β가 1/2이면 재현율에 정확률의 1/2의 중요도를 부여하는 경우이고, β가 1이면 동일한 중요도를, β가 2이면 재현율에 정확률의 2배의 중요도를 부여한 것이다. 복합척도 E값이 적을수록 검색효율은 높은 것이다.

$$E = 1 - \left(\frac{(\beta^2 + 1)PR}{\beta^2 P + R} \right) \quad \langle 공식\ 4 \rangle$$

4. 실험의 결과

본 연구에서 개발된 실험시스템은 이용자 최소개입 피드백기반 시스템, 완전 자동 피드백기반 시스템, 이용자 최대개입 피드백기반 시스템, 그리고 전통적인 SDI시스템이다. 전통적인 SDI 시스템의 경우 코싸인 유사계수 공식을 사용하여 질문과 문헌집단을 비교하며 그 성능을 측정하였다. 그 외 모듈은 1차 검색에서는 질문과 문헌집단을 비교하기 위해 코싸인 유사계수 공식을 사용하며 피드백 과정에서는 적합문헌집단과 신규 문헌집단을 비교하기 위해 kNN 알고리즘을 적용하여 그 성능을 측정하였다.

각 실험시스템의 성능 측정 및 성능비교를 위해, 먼저 피드백을 수행하기 전에 기존 문서를 대상으로 적합문헌집단을 선별하는 과정을 갖게 되는데, 이 때 적합문헌집단을 찾기 위한 적절한 코싸인 유사도 값을 실험을 통하여 발견하고자 하였다. 코싸인 유사도 값을 0.2에서 0.5까지 다양하게 변화시켜 가며 성능을 측정한 결과 0.3에서 비교적 높은 성능을 보여 주었다. 따라서 본 연구에서는 기존 문서를 대상으로 최초의 적합문헌을 생성하는 유사도 값으로 0.3을 선택하고 피드백 과정에서는 유사도 값을 0.3, 0.4, 0.5로 각각 변화시켜 가며 그 성능을 측정하였다.

1) kNN 알고리즘을 이용한 피드백 기법의 성능 비교

최초의 적합문헌 선정시 0.3으로 유사도 값을 주어 적합문헌을 선정했을 때의 재현율과 정확률은 각각 40.3과 53.6이었다. 초기 검색을 기반으로 적합문헌집단과 신착 문헌집단과의 유사도 값을 0.3에서 0.5까지 변화시켜가면서 1차에서 3차까지의 피드백 과정을 거쳐 얻은 검색 결과는 아래와 같다(표 1, 표 2, 표 3 참조).

아래 〈표 1〉에서 〈표 3〉까지 모두 이용자 최소개입과 이용자 최대개입의 경우 1차의 성능은 모두 동일한데, 이는 기존 문서 검색을 기반으로 검색된 결과에 대한 적합문헌 판정은 두 모듈 모두 1차 검색 전 단계까지는 동일하게 발생하기 때문이다.

〈표 1〉에서 〈표 3〉까지의 실험 결과를 통해 피드백 진행에 따른 성능평가와 적합문헌 갱신 방법에 따른 성능평가가 가능하다.

먼저, 피드백 진행에 따른 성능 평가 결과를

보면 〈표 1〉과 같이 유사도 값을 0.3으로 하였을 때, 재현율에 있어서는 이용자 최소개입의 경우가 72.73으로 평균적으로 가장 높은 성능을 보여 주었고, 이용자 최대개입, 완전 자동 피드백 순으로 나타났다. 정확률에 있어서는 이용자 최대개입이 평균 64.6으로 가장 높은 성능을 보여 주었고 이용자 최소개입, 완전 자동 피드백 순으로 나타났다.

유사도 값을 0.4로 하였을 때의 성능을 보면 유사도 값을 0.3으로 했을 때 보다 재현율과 정확률이 평균적으로 하강하는 것으로 나타났다. 평균으로 보았을 때 재현율에 있어서는 이용자 최대개입, 완전자동, 이용자 최소개입 피드백 기법순으로 나타났고, 정확률에 있어서는 이용자 최대개입이 가장 높고 이용자 최소개입, 완전 자동 피드백 기법 순으로 나타났다(표 2 참조).

유사도 값을 0.5로 하였을 때에는 정확률은 크게 증가하였으나 재현율에 있어서 크게 떨어지는 것으로 나타났으며 3개의 성능이 비슷한 결과를 보여 주었다(표 3 참조).

세 가지의 경우에 동일하게 나타나는 현상

〈표 1〉 유사도 값을 0.3으로 하였을 때의 성능

성능 / 회차 / SDI기법	재 현 율				정 확 률			
	1차	2차	3차	평균	1차	2차	3차	평균
이용자 최소개입	58.5	80.8	78.9	72.73	54.7	60.8	50.6	55.3
완전자동	53.8	60.8	70.7	61.77	51.3	45.8	49.0	48.7
이용자 최대개입	58.5	63.9	84.0	68.80	54.7	66.1	73.0	64.6

〈표 2〉 유사도 값을 0.4로 하였을 때의 성능

성능 / 회차 / SDI기법	재 현 율				정 확 률			
	1차	2차	3차	평균	1차	2차	3차	평균
이용자 최소개입	32.4	34.9	47.1	38.1	23.4	42.9	63.6	43.3
완전자동	20.5	41.1	54.0	38.5	11.3	51.7	62.4	41.8
이용자 대 개입	32.4	57.2	70.2	50.6	23.4	67.5	83.9	58.3

〈표 3〉 유사도 값을 0.5로 하였을 때의 성능

성능 / 회차 / SDI기법	재 현 율				정 확 률			
	1차	2차	3차	평균	1차	2차	3차	평균
이용자 최소개입	26.5	23.5	31.0	27.0	64.0	60.0	91.1	71.7
완전자동	18.5	18.2	24.7	20.5	56.0	64.0	88.2	69.4
이용자 최대개입	26.5	26.6	36.4	29.8	64.0	67.0	76.5	69.2

은 첫째, 이용자 최대개입 피드백 서비스는 재현율과 정확률에 있어서 거의 모든 경우에 피드백이 진행 될수록 그 성능이 향상되는 것으로 나타났다. 일반적으로 피드백이 진행될수록 문헌 수가 많아지기 때문에 성능이 떨어질 수 있으나 본 연구에서는 실험문헌집단의 크기가 크지 않고 게다가 적합문헌이 선정되어 있는 실험문헌집단의 결과를 활용하여 피드백을 수행하기 때문에 비교적 높은 성능을 보여 주는 것으로 나타났다.

둘째, 완전 자동 피드백 기법은 거의 모든 경우에 다른 두 기법에 비해 낮은 성능을 보여 주었으나 피드백이 진행될수록 그 성능은 다른 기법과 마찬가지로 조금씩 증가하는 것으로 나타났다.

실험 결과를 요약해 보면, kNN 알고리즘에서 사용하는 유사도 값은 0.3으로 하는 것이 재현율과 정확률에 있어서 비교적 높은 성능을 보여 주었고, 이용자 최대개입 기법이 전체적으로 높은 성능을 보여 주었으며, 대부분의 경우에 피드백이 진행될수록 그 성능이 향상되었다.

2) 전통적인 SDI 기법과 kNN 알고리즘 기반 피드백 기법과의 성능 비교

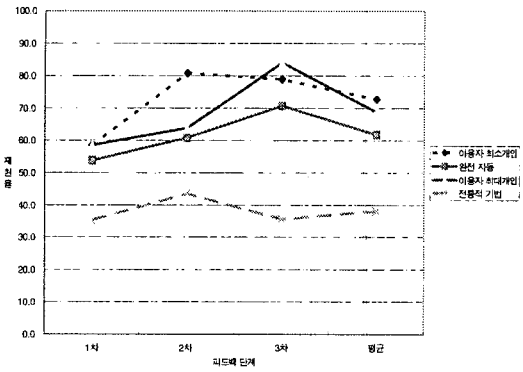
kNN 알고리즘을 사용하여 피드백을 수행한 위의 세 기법을 전통적인 SDI 기법과 그 성능을 비교 하였는데, 앞의 실험에서 비교적 높은 성능을 보여준 유사도 값, 0.3을 사용했을 경우와 비교하였다. 〈표 4〉와 〈표 5〉는 각각의 재현율과 정확률의 성능을 나타낸 것이고 〈그림 5〉와 〈그림 6〉은 이를 그림으로 나타낸 것이다. 표와 그림에서 보듯이 전통적인 기법은 재현율에 있어서 다른 세 기법보다 그 성능이 많이 떨어지는 것으로 나타났으며 정확률에 있어서는 완전자동 피드백 기법보다는 약간 높게 나타나 3순위의 성능을 보여 주었다.

〈표 4〉 SDI 기법들의 재현율 비교

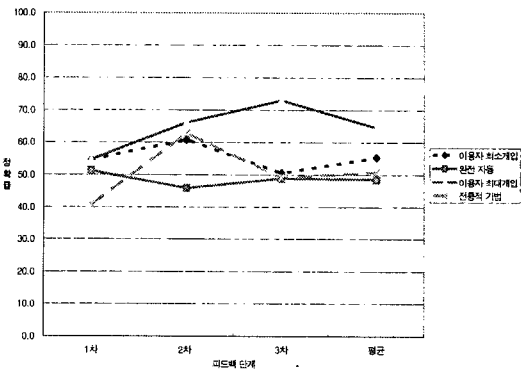
재현율	1차	2차	3차	평균
이용자 최소개입	58.5	80.8	78.9	72.7
완전 자동	53.8	60.8	70.7	61.8
이용자 최대개입	58.5	63.9	84.0	68.8
전통적 기법	35.3	43.5	35.7	38.2

〈표 5〉 SDI 기법들의 정확률 비교

정확률	1차	2차	3차	평균
이용자 최소개입	54.7	60.8	50.6	55.3
완전 자동	51.3	45.8	49.0	48.7
이용자 최대개입	54.7	66.1	73.0	64.6
전통적 기법	40.7	62.8	49.2	50.9



〈그림 5〉 전통적인 SDI 기법과 피드백 기반 SDI 기법과의 재현율 비교



〈그림 6〉 전통적인 SDI 기법과 피드백 기반 SDI 기법과의 정확률 비교

위 실험 결과, 전통적인 기법만을 사용하여 SDI 서비스를 제공할 경우 재현율과 정확률 모두 비교적 낮은 성능을 보여 주는 것으로 나타났고, 특히 재현율에 있어서 큰 성능 차이를 보여 주는 것으로 나타났다. 따라서 전통적인 SDI 서비스를 이용하는 것보다 피드백 기반 SDI 기법을 이용하는 것이 훨씬 이용자 만족도를 높일 수 있고, 그 중에서 이용자 개입이 제공되는 SDI기법을 사용하는 것이 좀 더 높은 만족도를 줄 것으로 나타났다.

마지막으로 재현율(R)과 정확률(P)을 모두 반영하여 평가하는 방법인 복합척도를 사용하여 성능을 비교하여 보았다. β 값을 1로 하였을 때, 즉 재현율과 정확률에 동일한 중요도를 부여하여 성능을 평가한 결과는 〈표 6〉과 같다.

위 표에서 보면, 이용자 최대개입 기법이 가장 높은 성능을 보여주고 이용자 최소개입, 완전 자동, 그리고 전통적 SDI 기법 순으로 나타났다. 특히 이용자 최대개입 기법은 복합척도로 비교할 때에도 피드백이 진행될수록 성능이 높아지는 것으로 나타났다.

5. 결 론

정보시대의 도래로 정보량은 기하급수적으로 증가하게 되었고, 이러한 대량의 정보로부터 이용자 개개인에게 적합한 정보를 적시에 제공할 수 있는 방법으로 SDI 서비스를 연구하고 개발해 왔지만, 현실적으로 그 활용도

〈표 6〉 복합척도에 의한 성능 비교

β=1	1차	2차	3차	평균
이용자 최소개입	56.6	70.8	64.7	64.0
완전 자동	52.5	53.3	59.9	55.2
이용자 최대개입	56.6	65.0	78.5	66.7
전통적 기법	38.0	53.1	42.4	44.5

는 매우 낮은 것으로 조사되었다. 이에 본 논문에서는 그 원인을 분석하고 SDI 시스템의 성능을 개선시킬 수 있는 방안을 모색해 보고자 하였다.

이를 위해 본 연구에서는 이용자가 한번의 정보요구 프로파일 작성 작업을 수행한 후 자신의 정보요구 변화를 반영하는 검색결과를 받아 볼 수 있는 시스템을 적합성 피드백 기반으로 개발하고자 하였다.

본 논문을 수행함으로써 다음과 같은 부분에 기여하고자 하였다. 첫째, 이용자(연구자)가 자신의 정보요구를 한 번 작성한 후 수정을 위해 시스템에 재접속할 필요성을 줄인다. 둘째, 최소한의 이용자 개입으로 적합문헌집단을 형성하고 이 적합문헌집단을 기반으로 이용자의 관심 주제분야의 변화를 반영할 수 있도록 한다. 셋째, SDI 서비스의 최대 목표로서, 연구자에게 최적의 정보를 적시에 제공함으로써 정보의 활용도를 높이고 연구지원 및 학문발전에 기여할 수 있도록 한다.

본 연구에서는 이와 같이 적합성 피드백을 기반으로 한 SDI 시스템의 성능을 개선할 수 있는 방안을 연구하였다. 본 연구의 실험을 위해 개발된 실험시스템은 이용자 최소개입 피드백기반 SDI 시스템, 완전자동 피드백기반 SDI 시스템, 그리고 이용자 최대개입 피드백 기반 SDI 시스템이며, 새로 개발된 3개 시스템의 성능 개선정도를 평가하기 위해 네 번째 시스템으로서 전통적인 SDI 서비스에서 사용하고 있는 방법으로 시스템을 개발하였다.

실험을 통하여 발견한 사실은 첫째, 이용자 최대개입 피드백 서비스는 재현율과 정확률에 있어서 거의 모든 경우에 피드백이 진행될수록 그 성능이 향상되는 것으로 나타났다.

둘째, 완전 자동 피드백 기법은 거의 모든 경우에 다른 두 기법에 비해 낮은 성능을 보여주었으나 피드백 횟수가 많아질수록 그 성능은 다른 기법과 마찬가지로 증가하는 것으로 나타났다.

셋째, kNN 알고리즘에서 사용하는 유사도 값은 0.3으로 하는 것이 재현율과 정확률에 있어서 비교적 높은 성능을 보여 주는 것으로 나타났고, 이용자 최대개입 기법이 전체적으로 높은 성능을 보여 주는 것으로 나타났으며, 대부분의 경우에 피드백이 진행될수록 그 성능이 향상되는 것으로 나타났다.

마지막으로 피드백 기법을 사용한 경우와 피드백이 없는 전통적인 기법만을 사용한 경우를 비교하였는데, 전통적인 기법만을 사용하여 SDI 서비스를 제공할 경우 재현율과 정확률 모두 비교적 낮은 성능을 보여 주는 것으로 나타났고, 특히 재현율에 있어서 큰 성능 차이를 보여 주는 것으로 나타났다. 따라서 전통적인 SDI 서비스를 이용하는 것보다 피드백 기반 SDI 기법을 이용하는 것이 훨씬 이용자 만족도를 높일 수 있고, 그 중에서 이용자 개입이 제공되는 SDI 기법을 사용하는 것이 좀 더 높은 만족도를 줄 것으로 보인다.

참 고 문 헌

노영희. 2003. 국내 대학도서관의 SDI 서비스 제공현황 분석 및 통합형 서비스 시스템 구축 방안에 관한 연구. 『정보관리학회지』, 20(3): 199-223.

정영미. 1993. 『정보검색론』. 개정판. 서울: 구미무역.

Amati, Gianni, Fabio Crestani, and Flavio Ubaldini. 1997. "Learning System for Selective Dissemination of Information." *Proceedings of IJCCAI-97, 15th International Joint Conference on Artificial Intelligence*, (1): 764-769.

Belkin, N. J. and W. B. Croft. 1992. "Information Filtering and Information Retrieval: Two Sides of the Same Coin." *Communications of the ACM*, 35(12): 29-38.

Blake, P. 1997. "Exploring the News." *Information World Review*, 191: 17-18.

Bonifati, Angela, Stefano Ceri, and Stefano Paraboschi. 2001. "Pushing Reactive Services to XML Repositories using Active Rules." *Computer Networks*, 39(5): 633-641.

Boughanem, M, and M. Tmar. 2002. "Incremental Adaptive Filtering: Profile Learning and Threshold Calibration." *ACM*, 640-644.

Callan, J. P., W. B. Croft, and J. Broglio. 1995. "TREC and TIPSTER Experiments with INQUERY." *Information Processing and Management*, 31(3): 327-343.

Callan, Jamie. 1996. "Document Filtering with Inference Networks." *SIGIR' 96*, 262-269.

Chung, Young Mee, Young Hee Noh, 2003. "A Study on Automatic text categorization of internet documents." *Journal of Information Science*, 29(1): 117-126.

Dasarathy, Belur V. 1991. *Nearest Neighbor(NN) Norms: NN Patern Classification Techniuqes. McGraw-Hill Computer Science Series.* Las Alamitos, California: IEEE Computer Society Press.

Fischer, G. and C. Stevens. 1991. "Information Access in Complies, Poorly Structured Information Spaces." In *Proceedings of ACM Special Interest Group on Human Computer Interaction Annual Conference (New Orleans, La., Apr. 27-May 2). ACM, New Youk: 63-70.*

Foltz, P. W. and S. T. Dumais. 1992. "Personalized Information Delivery: An Analysis of Information Filtering Methods." *Communications of the ACM*, 35(12): 51-60.

Frants, V. I., N. I. Kamenoff, and J. Shapiro. 1993. "One Approach to Classification of Users and Automatic Clustering of Documents." *Information Processing and Management*, 29(2): 187-195.

Goker, A. and T. L. McCluskey. 1991. "Toward an Adaptive Information Retrieval System." In *Proceedings of 6th International Symposium (Charlotte, N.C., Oct. 16-19). ISMIS: 348-357.*

Iwayama, Makato, and Takenobu Tokunaga. 1995. "Cluster-based Text Categorization: A Comparison of Category Search Strategies." *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR): 273-281.*

Lang, K. 1995. "NewsWeeder: An adaptive multi-user text filter." *Tech. Rep., School of Computer Science, Carnegie Mellon Univ., Pittsburgh, Pa.*

Manlone, T. W., K. R. Grant, F. A.. Trubak, F.A., Brobst, S.A., and M. D. Cohen. 1987. "Intelligent Information Sharing Systems." *Commun. ACM 30*, 5(May): 390-402.

Masand, B., G. Linoff, and D. Waltz. 1992. "Classifying News Stories Using Memory based Reseonin." *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR): 59-64.*

Myaeng, S. H. and R. R. Korfhage R. 1990. "Integration of User Profiles: Models and Experiments in Information Retrieval," *Information Processing Management.,. 26*(6): 719-738.

Oard, Douglas W. and Nicholas DeClaris. 1995. "Experimental Investigation of High Performance Cognitive and Interactive Text Filtering." In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, Vancouver, Canada: 4398-4403.*

Packer, K. H. and D. Soergel. 1979. "The Importance of SDI for Current Awareness in Fields with Severe Scatter of Information." *Journal of the American Society for Information Science*, 30(3): 125-135.

Resnick, Paul, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. "GroupLens: An Open Architecture for Collaborative Filtering of Netnews." In *R. K. Faruta and C. M. Neuwirth, editors, Proceedings of the Conference on Computer Sup-*

ported Cooperative Work, 175-186.

Rich, E. 1983. "Users are Individuals: Individualizing User Models." *International Journal. Man-Mach. Studies*, 18: 199-214.

Robertson, S. E., S. "Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at TREC-3." In D. K. Harman, editor, *The Third Text Retrieval Conference (TREC-3)*, Gaithersburg, MD, 1995. National Institute of Standards and Technology, Special Publication: 500-225.

Salton, Gerald and J. McGill Michael. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Seth, B. D. 1994. *A Learning Approach to Personalized Information Filtering*. M. S. Thesis, Electrical Engineering and Computer Science Dept., MIT, Cambridge, Mass.

Singhal, Amit, Mandar Mitra, and Chris Buckley. 1997. "Learning Routing Queries in a Query Zone." *SIGIR 97: 25-32.*

Walker, S., S. E. Robertson, M. Boughanem, G. J. F. Jones, and K. Spark Jones. 1997. *Kkapi/Keebow ant TREC-6 automatic and ad hoc, VLC, Routing, Filtering and QSDR, TREC-6.*

Wyle, M. F. and H. P. Frei. 1989. "Retrieving Highly Dynamic Distributed Information." In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval: 108-115.*

Yan T. and H. Garcia-Molina. 1995. "SIFT - A tool for wide-area information dissemination." In *Proceedings of the* 1995 *USENIX Technical Conference: 177-186.*

Yan T. and H. Garcia-Molina. 1994. "Distributed Selective Dissemination of Information." In *Proceedings of the 3rd International Conference on Parallel and Distributed Information Systems (PDIS, Austin, TX, Sept.): 89 - 98.*

Yang, Y. 1994. "Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval." *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR): 11-21.*

Yochum, J. A. 1985. "A High-Speed Text Scanning Algorithm Utilizing Least Frequent Trigraph." In *Proceedings of the IEEE International Symposium on New Directions in Computing: 114-121.*