사용자 로그 분석과 클러스터 내의 문서 유사도를 이용한 동적 추천 시스템

(A Dynamic Recommendation System Using User Log Analysis and Document Similarity in Clusters)

김 진 수 † 김 태 용 †† 최 준 혁 ††† 임 기 욱 †††† 이 정 현 †††††

(Jin-Su Kim) (Tae-Yong Kim) (Jun-Hyeog Choi) (Kee-Wook Rim) (Jung-Hyun Lee)

요 약 웹 문서들은 빠른 생성과 소멸의 특징 때문에, 사용자는 찾고자하는 웹 문서를 신속하고 정확하게 추천해 줄 시스템을 요구하고 있다. 정제되지 않은 웹 데이타에는 사용자들의 축적된 경험들을 포함하는 유용한 정보들을 포함하고 있다. 현재, 이러한 유용한 정보를 마이닝 기법이나 통계학적 측정 방법 등을 가지고 정제하여 추천 시스템을 통해 사용자에게 제공하려는 노력이 시도되고 있다. 기존의 정보 필터링 방식은 사용자들의 프로파일을 반드시 이용해야 하는 문제점을 갖고 있으며, 협력적 필터링 방식은 First Rater 문제와 Sparsity 문제가 있다. 또한 사용자 브라우징 패턴을 이용하는 동적 추천 시스템은 연관성이 없는 웹 문서들을 결과로서 제공한다는 문제점이 있다.

본 논문에서는 웹 문서 형식에 따라 웹 문서 사이의 유사도를 이용하여 웹 문서를 분류하고, 웹 서버에 기록된 로그 파일을 이용하여 사용자 브라우징 순차 패턴 DB를 생성한다. 이렇게 생성된 정보들과 사용자들의 세션 정보를 이용하여, 사용자가 웹 문서에 접근했을 때 현재 웹 문서와 유사도가 높은 상위 N개의 연관 웹 문서 집합을 제공하고, 순차적인 특성을 갖는 웹 문서를 추천 문서로 제공하는 시스템을 제안한다.

키워드 : 추천시스템, 연관 규칙, 웹 마이닝, 순차 패턴, 클러스터링

Abstract Because web documents become creation and disappearance rapidly, users require the recommend system that offers users to browse the web document conveniently and correctly. One largely untapped source of knowledge about large data collections is contained in the cumulative experiences of individuals finding useful information in the collection. Recommendation systems attempt to extract such useful information by capturing and mining one or more measures of the usefulness of the data. The existing Information Filtering system has the shortcoming that it must have user's profile. And Collaborative Filtering system has the shortcoming that users have to rate each web document first and in high-quantity, low-quality environments, users may cover only a tiny percentage of documents available. And dynamic recommendation system using the user browsing pattern also provides users with unrelated web documents.

This paper classifies these web documents using the similarity between the web documents under the web document type and extracts the user browsing sequential pattern DB using the users' session information based on the web server log file. When user approaches the web document, the proposed Dynamic recommendation system recommends Top N-associated web documents set that has high similarity between current web document and other web documents and recommends set that has sequential specificity using the extracted informations and users' session information.

Key words : Recommendation System, Association Rule, Web Mining, Sequential Pattern, Clustering

† 비 회 원 : 인하대학교 전자계산공학과
kjspace@nlsun.inha.ac.kr
†† 비 회 원 : 문경대학 인터넷정보계열
tykim@mkc.ac.kr
††† 종신회원 : 김포대학 컴퓨터계열 소프트웨어전공 교수
jhchoi@kimpo.ac.kr
†††† 종신회원 : 선문대학교 산업공학과 부교수
rim@sunmoon.ac.kr
††††† 종신회원 : 인하대학교 컴퓨터공학부 교수
jhlee@inha.ac.kr
논문접수 : 2002년 2월 20일
심사완료 : 2004년 1월 28일

1. 서 론

정제되지 않은 웹 데이타에는 사용자들의 축적된 경험들을 포함하는 유용한 정보들을 가지고 있다. 이러한 유용한 정보를 마이닝 기법이나 통계학적 측정 방법 등을 가지고 추출하여 추천 시스템들을 통해 사용자에게 제공하려고 시도되고 있다[1].

기존의 정보 필터링(Information Filtering) 방식에 의한 추천 시스템은 항목 컨텐츠의 분석과 사용자들이 입력한 흥미로운 프로파일에 근거하여 사용자들의 특성을 파악하여 추천 서비스를 제공하려 하였다. 그러나 사용자들이 직접 입력한 정보는 매우 왜곡된 것일 수 있으며, 사용자 프로파일은 정적이기 때문에 시간이 지남에 따라 프로파일의 질적인 효과는 감소할 수밖에 없다는 문제점이 있다[2].

이러한 정보 필터링 방식에 의한 추천 시스템의 문제점을 개선하기 위해 사용자들로부터 먼저 웹 문서에 대한 평가를 입력받아 평가된 축적 정보를 다른 사용자들에게 제공하려는 협력적 필터링(Collaborative Filtering) 방식에 의한 추천 시스템이 제안되었다. 이러한 협력적 필터링 방식을 이용하는 추천 시스템에는 FireFly[3], GroupLens[4,5]와 같은 시스템들이 있다. 협력적 필터링 방식의 추천 시스템은 일정한 목적없이 웹 사이트를 방문한 사용자들에게는 미처 생각하지 못한 웹 문서들을 추천 문서로 제공받는다는 장점이 있다. 그러나 새로운 웹 문서가 출현할 때, 사용자들이 일정한 수 이상 평가되기 전까지는 다른 사용자들에게 추천 집합으로 제공되지 못하는 First Rater 문제, 대량의 웹 문서에서 사용자가 찾고자 하는 웹 문서가 상당히 드물 때 발생하는 Sparsity 문제 등의 다양한 문제점이 존재한다.

최근에는 이러한 단점을 보완하기 위해 협력적 필터링 방식에 컨텐츠를 적용하여 First Rater 문제나 Sparsity 문제를 해결하려는 연구가 있지만[5], 웹 문서들 사이의 연관성에 대한 고려가 여전히 미흡하다. 또한, 기존의 협력적 필터링 방식을 이용한 동적 추천 시스템이 사용자들의 입력 자료에 지나치게 의존하고 있는 문제점을 개선하기 위해 웹 마이닝 기법을 이용하여 사용자들의 브라우징 패턴 정보로부터 추천 문서를 제공하려는 연구도 시도되었다[6,7]. 그러나 이러한 추천 시스템은 사용자 브라우징 패턴을 분석하고 결정하기 위해 데이타 마이닝의 연관 규칙 알고리즘을 이용하고 있는데, 이러한 방법은 사용자들의 브라우징 순서를 고려하지 않고 단순히 빈번하게 동시에 발생하는 웹 문서들의 요청에 대해서만 규칙을 생성하기 때문에 시간상의 선후 관계가 존재하는 브라우징 패턴을 정확하게 분석하지 못하고, 웹 문서들의 내용 정보를 무시함으로써 웹 문서들간의 내용적 측면에서의 연관성을 고려하지 않는다. 또한 연관 웹 문서 분류와 브라우징 순차 패턴을 이용한 동적 링킹 시스템(Dynamic Linking System; DLS)[8]에서는 WEBMINER 시스템[9]이 지닌 문제점을 해결하기 위해 웹 마이닝 기법 중 순차 패턴 알고리즘[10]을 이용하여 사용자들의 브라우징 순서에 대한 정보까지 고려하였다. 웹 문서들간의 연관성을 이용하기 위해서는 Association Rule Hypergraph Partitioning(ARHP) 알고리즘[11]을 이용하였지만, 추천 문서로 하이퍼링크 위주의 탐색 페이지와 같은 정보가 없는 단순 링크 기능을 지닌 불필요한 웹 문서까지 추천 문서로 제공하는 문제점이 있다.

본 논문에서 제안하는 동적 추천 시스템(DRS)은 웹 문서의 형식 결정 단계를 거친 후 탐색 페이지를 제외한 나머지 웹 문서들간의 유사도를 측정하여 사용자에게 웹 문서를 추천한다. 그리고 웹 서버 로그 파일에 포함된 유용한 정보들로부터 사용자들의 브라우징 순차 패턴을 생성, 웹 문서의 형식에 따라 연관된 웹 문서뿐만 아니라 사용자들이 자주 지나간 순차적인 특성을 가진 웹 문서를 추천 문서로 제공한다. 이때 추천 웹 문서 집합이 탐색 페이지이면 사용자 브라우징 순차 패턴 DB에서 사용자들이 자주 항해하는 순차적인 웹 문서 중 정보가 들어있는 내용 페이지를 사용자에게 최종 추천 문서로 제공한다.

2. 관련연구

2.1 데이타 마이닝(Data Mining)

데이타 마이닝이란 대량의 실제 데이타로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 데이타 마이닝 기법에는 여러 가지가 있지만 본 논문에서 사용한 기법은 다음과 같다.

2.1.1 연관 규칙(Association Rules)

연관 규칙은 항목들 사이의 강한 의존관계를 나타내는 지식 표현의 하나이다. X와 Y는 항목들(Items)의 집합이라고 할 때, 연관 규칙은 $X \Rightarrow Y$로 표시하며, "X를 포함하고 있는 데이타베이스 내의 트랜잭션(Transaction)은 Y도 함께 포함하고 있다"는 것을 의미한다. 생성된 연관 규칙이 트랜잭션들의 상황을 얼마나 잘 뒷받침해 주는 척도는 지지도(Support)와 신뢰도(Confidence)라는 측정 기준이 있다. 지지도와 신뢰도를 구하는 공식은 다음과 같다.

$$\text{지지도} = \frac{X\text{와 } Y\text{의 모든 항목들을 포함하고 있는 트랜잭션의 수}}{\text{데이타베이스 내의 전체 트랜잭션의 수}} \quad (1)$$

$$\text{신뢰도} = \frac{X\text{와 } Y\text{의 모든 항목을 포함하고 있는 트랜잭션의 수}}{X\text{의 항목을 포함하고 있는 트랜잭션의 수}} \quad (2)$$

전체 트랜잭션 데이타베이스로부터 사전에 정의된 최소 지지도(Minimum Support)를 만족하는 항목들의 집합을 Large 항목집합 또는 Frequent 항목집합이라고 한다.

2.1.2 순차 패턴(Sequential Patterns)

순차 패턴은 한 트랜잭션 안에서 발생하는 항목들간의 연관 규칙에 시간의 변이를 추가한 것이다[2]. 순차 패턴에서는 주어진 트랜잭션 데이타베이스에서 사용자가 정의한 최소 지지도(Minimum Support)를 만족하는 모든 시퀀스들 사이에서 최대 시퀀스를 찾는 것이다. 여기서 시퀀스는 트랜잭션 시간에 따라 정렬된 트랜잭션들의 리스트를 말한다.

2.2 WEBMINER

WEBMINER 시스템의 구조는 크게 두 부분으로 구성된다. 첫 번째 부분은 웹 데이타를 적당한 트랜잭션의 형태로 변환시키는 부분이고, 두 번째 부분은 연관 규칙, 순차 패턴과 같은 데이타 마이닝 기법을 사용하는 부분으로 이루어진다. 이 시스템은 웹 로그로부터 정보를 얻어 연관된 웹 문서나 순차적인 웹 문서를 추천은 해 주지만 웹 문서들 간의 내용적 연관성을 배제함으로써 성능을 떨어뜨린다.

2.3 DLS(Dynamic Linking System)

DLS 시스템은 크게 두 부분으로 구성된다. 첫 번째 부분은 웹 로그로부터 사용자들의 브라우징 순차 패턴을 분석하는 부분이고, 두 번째 부분은 웹 문서들간의 유사도를 연관 규칙을 이용하여 분류하는 부분이다. 전자의 경우는 WEBMINER 시스템에서 제안한 방법과 유사하며, 후자의 경우에 사용된 기법은 각 웹 문서에서 연관된 단어들을 추출하고, 각 웹 문서와 각 단어 클러스터에 존재하는 단어들에 대해 역문헌빈도(TFIDF) 계산을 통해 그 값의 총합이 가장 큰 단어 클러스터에 문서를 할당함으로써 연관 웹 문서들을 분류한다. 그러나 역문헌빈도를 사용하면 연관된 웹 문서들을 분류하는데 처리 시간이 많이 걸린다. 또한 추천 스코어가 크면 사용자에게 추천을 해 주기 때문에 단순 하이퍼링크로만 이루어진 정보가 없는 문서까지 추천해 준다는 단점이 있다.

3. 웹 로그분석과 문서 유사도를 이용한 동적 추천 시스템

본 논문에서 설계한 시스템은 크게 두 부분으로 구성되어 있다. 첫째, 사용자 브라우징 순차 패턴 DB를 구축하는 부분과 둘째, 연관 웹 문서 DB를 생성하는 부분이다. 그리고 두 과정에서 추출된 사용자 브라우징 순차 패턴 DB와 연관 웹 문서 정보를 이용하여 사용자가 웹 문서에 접근했을 때 동적 추천 알고리즘을 이용하여 사용자에게 연관된 웹 문서나 먼저 방문한 다른 사용자들의 순차적 경험이 내포된 웹 문서를 추천 문서로 제공한다.

그림 1은 본 논문에서 설계한 동적 추천 시스템에 대한 전체 구성도이다.

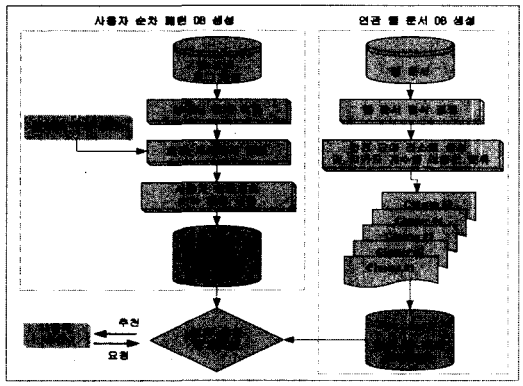

그림 1 동적 추천 시스템 구성도

3.1 사용자 브라우징 순차 패턴 분석

3.1.1 데이타 정제 작업(Data Cleaning)

사용자 순차 패턴 정보를 추출하기 위해 가공되지 않은 원시 데이타인 웹 서버의 로그 파일에서 불필요한 정보들(이미지, 에러 등)을 제거하는 데이타 정제 작업을 수행해야 한다.

본 논문에서는 사용자가 요청한 파일의 확장자가 "*.htm", "*.html"인 파일을 제외한 모든 기록들을 제거하고, 사용자의 *IP* 주소, 요청 시각, 요청 URL 필드만을 남기고 나머지 데이타는 제거한다. 그리고 정제된 로그 파일로부터 세션 트랜잭션 결정을 위한 항목 $I=\{IP, TIME, URL\}$을 생성한다. 여기서 *IP*는 사용자가 웹 서버에 접근했을 때 사용된 컴퓨터의 *IP* 주소이고, *TIME*은 사용자가 웹 문서를 요청한 시간, *URL*은 사용자가 웹 서버에 요청한 웹 문서의 *URL*이다.

3.1.2 사이트 구성 정보 복원 작업

사용자들이 사용하는 브라우저의 로컬 캐쉬와 프록시 서버의 캐쉬 사용으로 인해 요청되지 않은 기록을 복원하기 위한 사이트 구성 정보가 필요한데, 이를 Path Completion이라 한다[13]. 본 논문에서는 제약 사항으로 각 사용자는 프록시 서버를 사용하지 않고, 고정 할당 IP 주소를 사용하고 있다고 가정한다.

3.1.3 세션 트랜잭션 결정

사용자 브라우징 순차 패턴을 추출하기 위해서는 요청된 웹 문서들을 항목으로 하는 세션 트랜잭션이 결정

되어야 한다. 본 논문에서는 한 사용자가 한 번의 방문 동안 요청한 웹 문서들에 대응하는 항목 I들의 집합으로 세션 트랜잭션을 구성한다. 세션 트랜잭션을 결정하기 위한 방법에는 Reference Length Module, Maximal Forward Reference Module, Time Window Module 등이 있다[9,14].

본 논문에서는 대부분의 상용 제품에서 사용되는 Time Window Module을 이용하고, 윈도우 한계값은 30분을 사용한다[9]. 세션 트랜잭션 ST는 다음과 같이 정의된다[8].

$ST_k = [\{IP, TIME_1, URL_1\}, \ldots,$

$\{IP, TIME_i, URL_i\}, \ldots, \{IP, TIME_n, URL_n\}]$

여기서, 임의의 k 세션 트랜잭션(ST_k)에 포함된 IP는 모두 동일하고, $TIME_{i+1} - TIME_i$ 는 30분 미만이어야 한다.

3.1.4 사용자 브라우징 순차 패턴 생성

세션 트랜잭션이 결정되면 세션 트랜잭션들 중에서 URL들만을 항목으로 하는 URL 트랜잭션(UT)을 다음과 같이 구성한다.

$$UT_k = \{ URL_1, URL_2, \ldots, URL_n \}$$

URL 트랜잭션을 대상으로 AprioriAll 알고리즘[10]을 이용하여 사용자 브라우징 순차 패턴을 생성하며, 생성된 Large 항목집합을 Large 시퀀스라고 부른다. 이때 Large 시퀀스 자체가 얻고자 하는 순차 패턴이다. 위의 방법을 통해 얻어진 Large 시퀀스는 사전에 정의된 최소 지지도를 만족하는 순차 패턴을 의미하는데, 이 순차 패턴에 포함된 URL들은 "최소 지지도를 만족하면서 사용자들이 한번의 방문(세션)동안 순차적으로 방문한 웹 문서들"이라는 의미를 지닌다.

3.2 웹 문서 형식 결정

웹 문서는 형식에 따라 주요 페이지(Head page), 내용 페이지(Content page), 탐색 페이지(Navigation page), 참조 페이지(Look-up page), 개인 페이지(Personal page) 등으로 분류할 수 있다[13,15]. 주요 페이지는 사용자가 웹 사이트를 방문하였을 때의 첫 번째 페이지이고, 내용 페이지는 웹 사이트가 제공하는 정보의 내용이 포함되어 있는 페이지이고, 탐색 페이지는 웹 문서 내의 하이퍼링크를 통해 내용 페이지로 안내하는 페이지이다. 그리고 참조 페이지는 정의와 약어 표현을 위한 페이지이고, 개인 페이지는 각 개인들의 특성을 지닌 정보가 들어있는 페이지이다. 본 논문에서는 웹 문서 형식을 탐색 페이지와 내용 페이지만으로 분류한다[16].

웹 문서의 형식 결정 기준은 문서 내의 단어 수와 링크 수, 단어들간의 유사도를 측정하여 판별하며, 웹 문서 형식 결정 기준은 표 1을 따른다.

표 1 웹 문서 형식 결정 기준

	단어 수 (*Doc_N*)	링크 수 (*Link*)	단어들 간의 유사도 (*Word_Sim*)
내용 페이지	$Doc\_N \geq \alpha$	$Link < \beta$	$Word\_Sim \geq \gamma$
탐색 페이지	$Doc\_N < \alpha$	$Link \geq \beta$	$Word\_Sim < \gamma$

표 1에서 단어들간의 유사도를 측정하기 위해서는 단어들의 연관성을 정량적으로 나타내는 상호 정보량[17]을 이용한다.

단어들간의 유사도를 측정하기 위해 본 논문에서는 단어들의 연관성을 정량적으로 나타내는 상호 정보 수식을 사용한다. 크기가 N인 말뭉치에서 단어 x와 y가 출현할 횟수를 각각 $f(x)$, $f(y)$라하고, x와 y가 한 문장에서 함께 출현한 빈도 수를 $f(x,y)$라고 했을 때 단어 x와 y의 상호 정보량은 다음과 같이 정의할 수 있다[21].

$$MI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} \approx \log_2 \frac{f(x,y)}{f(x)f(y)} \quad (3)$$

웹 문서에 출현하는 단어와 연결된 하이퍼링크의 수도 중요하지만 본 논문에서는 단어들 간의 유사도에 가중치를 높게 부여하여 내용 페이지와 탐색 페이지를 분류하였다. 즉 하이퍼링크의 수는 웹 문서의 내용에 따라 많을 수도 있다. 그렇다고 탐색 페이지로 분류하는 것이 아니라 단어들 간의 유사도를 측정하여 한계값 이상이면 정보가 있는 내용 페이지로 분류한다.

3.3 연관 웹 문서 분류

연관 웹 문서 분류를 수행하기 위해 웹 문서 형식이 내용 페이지인 웹 문서들을 대상으로 본 연구실에서 개발한 형태소 분석기를 사용하여 추출된 명사들로부터 불용어를 제거하고 각 문서에서 출현하는 단어들을 이용하여 동일한 식별자를 갖는 하나의 트랜잭션을 구성한다. 그리고 연관 규칙 알고리즘을 이용하여 단어들간의 연관 규칙을 생성하고, 생성된 연관 규칙의 신뢰도를 가중치로 사용하여 ARHP 알고리즘을 적용, 여러 웹 문서에서 동시에 출현하는 명사들의 리스트를 군집화한다. 그리고 각 클러스터 내에 포함된 단어 리스트를 가지고 벡터 모델, $V_{Cluster_i} = (w_{i1}, w_{i2}, \ldots, w_{ij}, \ldots, w_{in})$로 표현한다. 본 논문에서는 웹 문서의 특징을 추출하기 위해 역문헌빈도를 이용하는 것이 아니라, 처리 속도와 정확도를 높이기 위해 웹 문서에서 추출된 모든 명사들을 이용한다. 따라서 확률 모델보다는 벡터 모델이 적합하다.

벡터 모델에서 w_{ij}의 값은 이진수로 표현되며, 단어가 $Cluster_i$에 포함되면 1로 표현하고 아니면 0으로 표현한다. 분류될 웹 문서를 위와 같이 벡터 모델로 변환시킨

후 유사도를 측정한다. 유사도 측정은 식 (4)의 자카드 계수[12]를 이용한다. 자카드 계수를 이용하는 이유는 유사도 측정 공식 중 자카드 계수가 문서 클러스터링에 사용되는 가장 보편적인 함수이기 때문이다.

$$Sim(V_i, V_j) = \frac{\sum_{k=1}^{n}(w_{ik} \cdot w_{jk})}{\sum_{k=1}^{n} w_{ik} + \sum_{k=1}^{n} w_{jk} - \sum_{k=1}^{n} w_{ik} \cdot w_{jk}} \quad (4)$$

여기서 n은 $Cluster_i$에 포함된 단어의 수이고, w_{ik}는 벡터 V_i의 k번째 단어의 값이다[18].

식 (4)의 $Sim(V_i, V_j)$ 공식은 두 번 사용되는데, 첫 째는 클러스터 내에 포함된 단어 리스트들을 가지고 전체 웹 문서들과의 유사도를 측정하여 가장 큰 값을 갖는 문서를 각 클러스터에 분류할 경우에, 그리고 두 번 째로는 사용자들이 웹 문서에 접근했을 때 각 클러스터 내에 포함되어 있는 웹 문서들과 현재 접근중인 웹 문서 사이의 유사도를 측정하여 추천할 집합을 찾을 경우에 사용한다. 이때, 첫 번째 경우는 $V_i = (1,1,\ldots,1)$과 웹 문서 V_j의 벡터를 이용하며, $Sim(V_i, V_j)$ 값들 중 가장 큰 클러스터에 웹 문서를 할당하여 연관 웹 문서를 분류한다. 그리고 두 번째 경우는 현재 접근중인 웹 문서 V_i와 클러스터내에 포함된 n개의 웹 문서들 사이의 $Sim(V_i, V_j)$를 측정하여 상위 N개의 추천 집합을 생성한다.

3.4 동적 추천 알고리즘

동적 추천 알고리즘은 사용자 브라우징 순차 패턴으로부터 현재 사용자의 세션을 갖는 추천 집합과 사용자가 방문하고 있는 현재 웹 문서와 가장 연관된 웹 문서를 제공할 추천 집합을 생성하는 알고리즘으로 본 논문에서 설계한 동적 추천 알고리즘은 그림 4와 같다.

그림 4에서 사용자 세션 중 last_url의 웹 문서 형식이 탐색 페이지인지 내용 페이지인가에 따라 추천 집합을 생성한다. 사용자의 last_url의 웹 문서 형식이 내용 페이지이면, 현재 세션을 포함하면서 1이 더 큰 large 시퀀스 집합 중에서 추천 집합을 생성한다. 만약 last_url의 웹 문서 형식이 탐색 페이지이면, 사용자 브라우징 순차 패턴 DB에서 last_url을 포함하면서 최소 지지도 이상을 가진 large 시퀀스 집합에서 추천 집합을 생성한다. 이때 동적 추천 알고리즘에서 사용하는 신뢰도($session \Rightarrow url$)는 다음 식 (5)와 같다.

$$\text{신뢰도}(session \Rightarrow url) = \frac{|\ session \cap url\ |}{|\ session\ |} \quad (5)$$

신뢰도($session \Rightarrow url$)는 추천 집합의 순위 결정을 위한 가중치로 사용되며, 최소 신뢰도를 만족하는 url들만을 추천 집합에 포함시킨다. 이때 추천될 url의 웹 문서 형식이 탐색 페이지이면 Seq 함수로부터 반환된 웹 문서를 포함시킨다. Seq 함수는 url을 포함하는 순차 패턴 DB에서 사용자의 세션을 가지고 사용자들이 자주 항해하는 순차적인 특성을 가진 웹 문서를 반환하는 함수이다. 예를 들어, A, C문서는 내용 페이지이고, B문서는 탐색 페이지라고 가정할 때, 만약 A→B→C로 항해하는 사용자가 많다고 가정하면, 동적 추천 시스템은 사용자가 A문서를 방문했을 때 B문서를 추천하는 것이 아니라 C문서를 추천 집합에 포함시켜 제공하는 것이다. 이것은 제안한 알고리즘에 의해 B문서가 웹 문서 형식이 탐색 페이지이기 때문에 C문서를 제공하는 것이다. 또한 last_url을 가진 클러스터내의 웹 문서들 사이의 유사도가 높은 상위 N개의 웹 문서들을 추천 집합에 포함시킨다.

```
Input : cur_session // 현재 사용자 세션
        last_url    // 사용자가 가장 최근에 요청한 URL
        δ           // 최소 지지도
        α           // 최소 신뢰도
        ξ           // 유사도 한계값
Output : Recommend1 // 순차 패턴에 포함된 추천 문서 집합
         Recommend2 // 연관 문서에 포함된 추천 문서 집합

Algorithm Recommend
begin
  Recommend1 <- ∅, Recommend2 <- ∅
  if last_url.type = content_page then
    begin
/* Items은 cur_session을 포함하는 size가 |cur_session|+1인 large
  sequence 집합 */
      for each Items do
        begin
          if 지지도(Items) >= δ then
            begin
              confidence = 신뢰도(session⇒url)
              if confidence >= α then
                url.score <- confidence
                if url.type = navigation_page then
                  url = Seq(url)
                Recommend1 = Recommend1 + url
              enddo
        enddo
      for each url do
        if (value = Sim(last_url, url) >= ξ then
          Recommend2 = Recommend2 + url
    enddo
  else if last_url.type = navigation_page then
    while (url = Seq(last_url)) >= δ do
      begin
        if (url.confidence >= α ) then
          begin
            url.score <- url.confidence
            if url.type = navigation_page then url = Seq(url)
            Recommend1 = Recommend1 + url
          enddo
      enddo
  output(Recommend1+Recommend2)
enddo.
```

그림 4 동적 추천 알고리즘

4. 실험 및 결과

본 논문의 실험 환경으로는 Windows NT 서버 4.0, 시스템의 구현을 위해 Microsoft Visual C++ 6.0을 사용하였다. 실험 데이타로는 인하대학교 대학원 홈페이지를 서비스하는 웹 서버의 Common Logfile Format (CLF) 로그 파일 가운데, 2001년 1월 18일부터 2월 2일까지의 기록된 정보를 이용하였고 대학원 웹 문서 188개 중 연관 웹 문서 분류를 위해 웹 문서 형식이 내용 페이지인 169개의 웹 문서들을 이용하였다. 연관 웹 문서 분류에서 제거된 탐색 페이지들은 [menu.html],

[main.htm] 등과 같이 범용 브라우저에 기본적으로 보여 지는 부분 혹은 아무런 의미가 없는 단순 하이퍼링크로 구성된 문서들이 대부분이다.

사용자 브라우징 순차 패턴을 생성하기 위한 측정 기준으로 지지도의 최소 지지도 한계값은 30%로 설정하였다. 웹 서버 로그 파일로부터 순차 패턴 알고리즘[10]을 이용하여 최소 지지도 30%를 만족하는 사용자 브라우징 순차 패턴을 생성하였으며, 이때 추출된 순차 패턴의 수는 481개이다.

표 2 사용자 브라우징 순차 패턴 Database의 일부분

	지지도 (%)
[/~grad/]→[/~grad/menu.htm]	51.78
[/~grad/]→[/~grad/sugang/2001-1iljung.hwp]	50.23
[/~grad/]→[/~grad/sugang/main.htm]	47.46
[/~grad/]→[/~grad/sugang/main.htm]→[/~grad/menu.htm]	27.52
[/~grad/]→[/~grad/notice/2001-1enter/2001-1_grad.html]	24.85
[/~grad/]→[/~grad/labs/main.htm]→[/~grad/labs/eng.htm]	23.85
[/~grad/sugang/index.html]→[/~grad/sugang/main.htm]	20.34
...	...

표 2는 추출된 사용자 브라우징 순차 패턴에 대해 지지도를 기준으로 정렬한 데이타베이스의 일부분을 나타낸다. 표 2의 왼쪽 필드는 사용자 브라우징 순차 패턴들이고, 오른쪽 필드는 각 순차 패턴에 대한 지지도를 나타낸다.

웹 문서 형식을 결정하기 위한 원본 웹 문서는 다음 그림 5이고, 그림 6은 웹 문서로부터 추출한 형태소 분석 결과이며 그림 7은 추출된 명사들과 빈도수를 나타낸다.

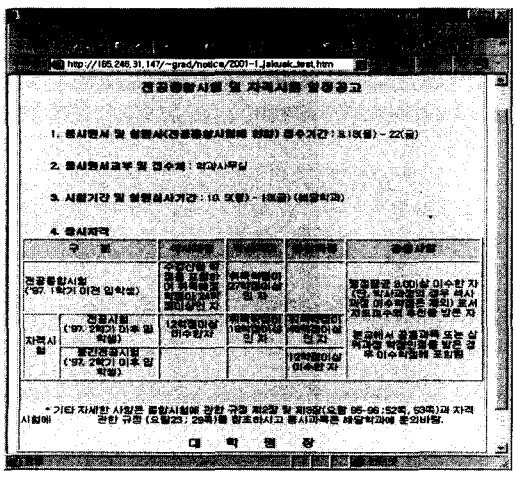

그림 5 웹 문서 원본 예

```
[전공종합시험 ((전공 종합 시험) (N N N))]
[및 ((및) (AD))]
[자격시험 ((자격 시험) (N N))]
[일정공고 ((일정 공고) (N N))]
[응시원서 ((응시 원 서) (N N PCA)) ((응시 원서) (N N))]
[및 ((및) (AD))]
[접수처 ((접수 처) (N PCA))]
[전공종합시험에 ((전공 종합 시험 에) (N N N PCA))]
[한함 ((한 함) (N N)) ((한하 ㅁ) (V EN))]

[자세한 ((자 세 하 ㄴ) (V EE VA ED)) ((자 세 하 ㄴ) (V EE AA ED)) ((자 세한) (N N)) ((자세 한) (N N)) ((자세하 ㄴ) (A ED))]
[사항은 ((사항 은) (N PS))]
[종합시험에 ((종합 시험 에) (N N PCA))]
[관한 ((관하 ㄴ) (V ED)) ((관한) (AD))]
[규정 ((규정) (N))]
[자격시험에 ((자격 시험 에) (N N PCA))]
[관한 ((관하 ㄴ) (V ED)) ((관한) (AD))]
[규정 ((규정) (N))]
[요람 ((요람) (N))]
[쭉 ((쭉) (N))]
[응시과목은 ((응시 과목 은) (N N PS))]
[해당학과에 ((해당 학과 에) (N N PCA))]
[문의바람 ((문의 바람) (N N))]
```

그림 6 원문의 형태소 분석 결과

경우	2
공고	1
공동	1
공과	1
과목	2
교규	1
기간	2
기타	1
문의	1
바람	1
방법	2
본사	1
사무	2
사항	1
상석	2
수시	1
시험	10
신청	1
신청	1
예정	2
요람	2
응시	4
이상	5
이수	5
이후	1
인정	2
입학	1
전공	3
접수	5
접수처	1
제외	1
종합	4
주간	1
지도	2
처	1
취득	1
평균	1
학과	3
학기	3
정도	14
해당	2
자격	4

총 124 단어

그림 7 추출된 명사 및 빈도

표 3 단어들 간의 상호정보량(일부)

	학점	전공	종합	시험	취득
자격	0.346	1.896	-0.381	0.073	1.520
학점		1.180	-0.345	-0.477	1.256
전공			0.619	1.295	-0.254
종합				1.635	0.082
시험					1.535

표 3은 웹 문서에 나타난 단어들 간의 상호정보량을 보여주는 예이다. 이때 0보다 큰 값 일수록 단어들 간의 관계가 강하다. 웹 문서 형식 결정을 위한 한계값으로 문서내의 단어 수와 링크 수는 각각 100개와 25개로 설정하였고, 단어들 간의 연관도는 상호 정보량의 평균값을 이용하였으며, 이때 한계값은 0.15로 설정하였다.

웹 문서 형식이 내용 페이지인 169개의 웹 문서를 대상으로 문서에 빈번히 출현하는 단어들의 집합으로부터 최소 지지도 30%, 최소 신뢰도 50%를 만족하는 단어들의 쌍을 연관 규칙 알고리즘[19]을 이용하여 추출한다. 이때 ARHP 알고리즘에 적용하기 위해 연관 단어들의 신뢰도를 가중치로 하여 5개의 단어 클러스터를 생성하였다. 각 단어 클러스터에 포함된 단어들을 이용하여 169개의 웹 문서를 분류하며, 이때 각 클러스터에 포함된 단어들을 벡터로 표기하고, 각 문서들과의 유사도를 측정하기 위해 자카드 계수를 이용하여, 유사도의 크기가 가장 큰 클러스터에 웹 문서를 할당한다.

표 4는 분류된 연관 웹 문서의 벡터 표현과 해당 클러스터 내에 포함된 웹 문서 수를 나타낸다. 분류된 연관 웹 문서 정보는 사용자가 웹 서버에 접근했을 때, 현재 웹 문서와 클러스터 내의 문서들과의 유사도를 측정하여 방문하고 있는 웹 문서와 가장 유사한 문서를 찾아내어 추천 집합으로 사용자에게 제공한다. 이렇게 생

표 4 분류된 연관 웹 문서의 벡터 표현과 해당 클러스터의 웹 문서 수

클러스터링 된 연관 웹 문서의 벡터 표현	문서수
[개설][입학][자격][전공][전형][제출][과목][교수][구술][학과][학문][합격][특별][학부][학술지]... [~grad/sugang/2001-1orientation.html] (1,1,1,1,0,1,1,0,0,0,1,1,1,1,0,1,0,1,0, ... ,0,1,1) [/~grad/sugang/2001-1iljung.hwp] (1,1,0,1,1,1,1,0,0,1,0,1,0,1,1,1,1,1,0,1,0, ... ,0,1,0)...	73
[정보][정의][정책][제도][계산][서류][설명][교육][공동][공통][과학][동일][문학][발표][지원]... [~grad/about/cooperation.htm] (0,0,1,1,1,1,1,1,1,1,1,1,0,0,0,1,1,1, ... ,0,1,1) [~grad/notice/2001-1jaguekhab.html] (1,1,1,0,1,1,0,1,0,1,0,1,1,0,0,0,0,1,1,0, ... ,1,1,1)...	26
[경제][관리][기관][흐름][모델][목표][문화][물리][반응][분류][사례][발달][운영][과학][정론]... [~grad/labs/sci.htm] (0,1,0,0,0,0,1,0,0,1,1,0,1,0,0,1,0,0,0,0, ... ,0,0,1) [~grad/course/degree/sangsan.htm] (0,1,1,0,1,0,0,1,1,1,0,1,0,0,1,0,0,0,0,1, ... ,1,1,0)...	13
[강의][개념][개발][개요][검토][제반][고찰][과정][소개][동향][교과][관계][습득][원리][주제]... [~grad/course/degree/bio.htm] (1,1,1,1,0,1,1,1,0,1,1,1,1,1,0,1,1,0,1,1, ... ,1,1,1) [~grad/course/degree/교육-1.htm] (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1, ... ,1,1,1)...	45
[공정][공학][관점][세미나][기술][시스템][목적][제어][발전][효율][컴퓨터][자료][수행][설계][에너지]... [~grad/labs/eng.htm] (1,1,1,0,1,0,1,0,1,0,1,0,0,0,1,1,1,1, ... ,1,0,0) [~grad/about/student.htm] (0,0,1,0,0,1,0,1,0,1,0,0,0,0,0,1,1,0,0, ... ,0,1,0)...	12

표 5 [수강 신청] 문서 방문시의 추천 문서 예

WEBMINER	DLS	DRS
[/~grad/]	[/~grad/]	[/~grad/sugang/2001-1iljung.hwp]
[/~grad/menu.htm]	[/~grad/sugang/2001-1iljung.hwp]	[/~grad/sugang/2001-1orientation.html]
[/~grad/sugang/2001-1iljung.hwp]	[/~grad/menu.htm]	[/~grad/sugang/2001-1gangjwa.html]
[/~grad/sugang/2001-1orientation.html]	[/~grad/sugang/2001-1orientation.html]	[/~grad/notice/2001-1enter/2000-2_deonglok.htm]
[/~grad/sugang/2001-1gangjwa.html]	[/~grad/sugang/2001-1gangjwa.html]	[/~grad/course/degree.htm]
...	...	...

표 6 [대학원 추가 모집] 문서 방문시의 추천 문서 예

WEBMINER	DLS	DRS
[/~grad/]	[/~grad/]	[/~grad/notice/2001-1enter/2001-1_1grad.html]
[/~grad/menu.htm]	[/~grad/notice/2001-1enter/2001-1_1grad.html]	[/~grad/notice/2001-1enter/2001-1_3grad_chuga.html]
[/~grad/notice/2001-1enter/2001-1_1grad.html]	[/~grad/menu.htm]	[/~grad/notice/2001-1enter/2001-1_6grad.html]
[/~grad/notice/2001-1enter/2001-1_3grad_chuga.html]	[/~grad/notice/2001-1enter/2001-1_3grad_chuga.html]	[/~grad/notice/2001-1enter/2001-1_4grad.html]
[/~grad/notice/index.html]	[/~grad/notice/2001-1enter/2001-1_6grad.html]	[/~grad/notice/2001-1enter/2001-1_7grad_chuga.html]
...	...	...

성된 순차 패턴 DB와 연관 웹 문서 정보는 사용자가 웹 페이지에 접근했을 때의 입력 자료로 사용된다.

표 5와 6은 사용자가 [수강 신청]과 [대학원 추가모집] 웹 문서를 방문했을 때의 추천된 웹 문서들 중 일부를 보여주는 예이다. 비교하는 추천 시스템들은 WEBMINER, DLS, 그리고 본 논문에서 제안한 DRS이다.

표 7은 표 5와 6을 포함한 10개의 웹 문서를 접근했을 때 각 추천 시스템에서 제공되는 상위 문서 4, 8, 12개의 문서를 대상으로 정보검색에서 널리 사용되는 정확도, 재현율, F-measure[20]를 계산한 평균값을 나타낸다. 여기서, 정확도와 재현율은 약간 변형하였으며, F-measure는 정확도나 재현율을 결합하여 같은 가중치를 부여함으로써, 질적인 판단을 높이기 위해 이용하는데, 측정 방법은 식 (6)과 같다.

$$\begin{aligned} \text{정확도} &= \frac{\text{검색된 적합 문서 수}}{\text{상위 } N \text{ 문서 수}} \\ \text{재현율} &= \frac{\text{검색된 적합 문서 수}}{\text{테스트 문서 총 수}} \\ F-measure &= \frac{2 \times \text{정확도} \times \text{재현율}}{(\text{정확도} + \text{재현율})} \end{aligned} \quad (6)$$

여기서 적합 문서는 전문가가 평가한 문서이고, 테스트 문서는 실험에 사용된 웹 문서들을 말한다.

표 7 Top-N 성능 비교표

	정확도(%)			재현율(%)			F-measure		
	WEBMINER	DLS	DRS	WEBMINER	DLS	DRS	WEBMINER	DLS	DRS
Top-4	42.500	55.000	85.000	1.006	1.302	2.012	0.01965	0.02543	0.03931
Top-8	45.000	52.500	83.750	2.130	2.485	3.964	0.04068	0.04746	0.07571
Top-12	45.000	61.667	77.500	3.195	4.379	5.503	0.05967	0.08177	0.10276

표 8은 표 7에서 사용한 10개의 웹 문서들을 나타낸다. 이때 사용된 10개의 웹 문서는 표 4에 나타난 5개의 클러스터에서 사용자들이 가장 빈번하게 접근한 문서들 중에서 2개씩을 선택하였다.

표 8 Top-N 성능 비교에 사용된 웹 문서들

웹 문서 제목	웹 문서
[수강 신청]	[~grad/sugang/main1.htm]
[대학원 추가 모집]	[~grad/notice/2001-1enter/2001-1_grad.html]
[영어 강좌 안내]	[~grad/sugang/2001-1gangjwa.html]
[대학원 학사제도 안내]	[~grad/notice/2001-1enter/2001-1_7grad.html]
[장학 제도]	[~grad/about/encourage.htm]
[신입생 오리엔테이션 안내]	[~grad/sugang/2001-1orientation.html]
[학위 과정]	[~grad/course/degree.htm]
[교류/협동]	[~grad/about/cooperation.htm]
[교과 과정]	[~grad/course/curri.htm]
[생산 기술 연구원]	[~grad/course/degree/sangsan.htm]

표 9는 각 추천 시스템에서의 추천된 문서들에 대한 정확도와 재현율을 나타낸다. 본 논문에서 제안한 방법이 평균 WEBMINER 시스템보다 7.32%, DLS 시스템보다 2.32% 더 높은 성능을 나타내고 있다.

표 9 세 시스템의 성능 비교표

재현율(%)	WEBMINER	DLS	DRS
0	66.70	68.90	77.50
10	44.00	55.10	64.00
20	39.60	50.80	46.90
30	32.30	41.30	44.00
40	33.66	41.80	41.35
50	23.30	30.33	33.20
60	17.70	22.40	26.13
70	8.60	8.60	11.65
80	2.00	2.12	3.40
90	0.40	1.70	0.40
100	0.14	0.39	0.39

그림 8은 표 9의 세 시스템을 재현율에 따른 정확도의 변화를 그래프로 보인 것이다.

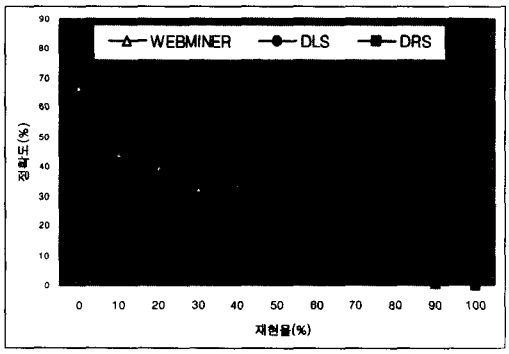

그림 8 세 추천시스템의 추천 문서 성능 비교

5. 결 론

본 논문에서는 기존 추천 시스템의 문제점을 개선하기 위해, 사용자 브라우징 패턴 분석 시 데이타 마이닝 알고리즘의 순차 패턴을 사용하여, 사용자들의 웹 사이트 방문에 대한 순차적 특성을 고려하였다. 이때 사용자의 현재 세션과 가장 유사한 사용자 브라우징 순차 패턴 정보를 추천 문서로 제공하였다. 또한 탐색 페이지와 같이 정보가 없고 단순히 내용 페이지로의 하이퍼링크만을 제공하는 웹 문서를 추천 문서에서 제외시킴으로써 사용자들이 보다 편리하고, 정확하게 사이트를 브라우징할 수 있는 시스템을 설계하였다. 이때, 사용자들에게는 이전 사용자들의 사전 경험과 연관된 웹 문서들을 이용하여 연관문서를 추천하기 때문에 시스템의 성능을 높이고 사용자가 원하는 웹 문서를 빠르게 제공할 수 있다.

향후, 각 사용자의 개별화를 통해 문서를 군집화하고 사용자가 웹 문서를 방문하였을 때 비슷한 관심을 갖고 있는 사용자들로부터의 정보를 얻어 제공하는 연구가 필요할 것으로 판단된다. 또한 제공된 추천 집합이 사용자들에게 얼마나 적용되었는지 로그 파일의 재분석을 통해 추천 알고리즘이나 전처리 부분을 재조정할 필요가 있다. 그리고 사용자 브라우징 순차 패턴 DB 상에 포함되지 않은 예상치 못한 상황이 발생하였을 경우를 대비하여 사용자 브라우징 순차 패턴 집합의 기계 학습을 통해 추천 집합을 사용자에게 제공해야 할 필요가 있을 것으로 판단된다.

참 고 문 헌

[1] Stephen C. Gates, Charu C. Aggarwal, "Recommender System: Knowlege from Mining User Experiences," IBM Research Report, 1999.

[2] B. Mobasher, et al., "Automatic Personalization on Web Usage Mining," Technical Report TR99-010, Department of Computer Science, Depaul University, 1999.

[3] U. Shardanand and P. Maes, "Social information filtering : algorithms for automating 'word mouth'," Proc. of ACM CHI Conference, 1995.

[4] J. Konstan, et al., "GroupLens: applying collaborative filtering to usenet news," Communications of the ACM (40) 3, 1997.

[5] Sarwar, B. et al., "Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System," Proc. ACM CSCW 98, pp. 345-354, 1998.

[6] Tak Woon Yan, et al., "From user access patterns to dynamic hypertext linking," Computer Networks an ISDN Systems 28, pp.1007-1014, 1996.

[7] J. Srivastava, R. Cooley, M. Deshpande, P-T. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations, (1) 2, 2000.

[8] 박영규, 연관 웹 문서 분류와 브라우징 순차 패턴을 이용한 동적 링킹 시스템, 인하대학교 대학원 공학 석사 학위 논문, 2000.

[9] B. Mobasher, N. Jain, E. Han, and J. Srivastava. "Web mining: Pattern discovery from world wide web transactions," Technical Report TR 96-050, University of Minnesota, Dept. of Computer Science, Minneapolis, 1996.

[10] R. Agrawal and R. Srikant, "Mining Sequential Patterns," Proc. of the Int'l Conference on Data Engineering (ICDE), Taipei, Taiwan, March 1995.

[11] G. Karypis, R. Aggarwal, V. Kumar, and S. Shekhar, "Multilevel hypergraph partitioning: Application in VLSI domain," In Proceeings ACM/IEEE Design Automation Conference, 1997.
[12] 정영미, 정보검색론, 구미무역 출판부, 1993.
[13] R. Cooley, et al., "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, Vol. 1-1, 1999.
[14] J. S. Park, et al., "Using a Hash-Based Method with Transaction Trimming for Mining Association Rules," TKDE 9(5), pp. 813-825, 1997.
[15] P. Pirolli, J. Pitkow, and R. Rao. "Silk from a sow's ear: Extracting usable structures from the Web," In Proc. of 1996 Conference on Human Factors in Computing Systems(CHI-96), Vancouver, British Columbia, Canada, 1996.
[16] 김진수, 김태용, 이정현, "웹 문서 형식과 클러스터 내의 문서 유사도를 이용한 동적 추천 시스템", 제28회 한국정보과학회 춘계학술발표 논문집(B), pp. 274-276, 2001.
[17] 전미선, 박세영, "상호 정보를 이용한 어의 모호성 해소에 관한 연구", 제6회 한글 및 한국어 정보처리 학술발표 논문집, pp. 369-373, 1994.
[18] T. Tokunaga and M. Iwayama, "Text categorization based on weighted inverse document frequency," IPSJ SIG Report. NL100 (5), 1994.
[19] R. Agrawal and R. Srikant, "Fast Algorithm for Mining Association Rules," Proc. of the 20th VLDB Conference, pp. 487-499, 1994.
[20] Yang, Y., and Liu, X. "A Re-examination of Text Categorization Methods," In Proceedings of ACM SIGIR'99 conference, pp. 42-49, 1999.

김 진 수
1998년 인천대학교 전자계산공학과(공학사). 2001년 인하대학교 전자계산공학과(공학석사). 2001년~현재 인하대학교 전자계산공학과 박사과정. 2002년~현재 김포대학 컴퓨터계열 겸임교수. 관심분야는 웹 마이닝, 데이타마이닝, 기계학습, 정보검색, 자연어처리, 웹 마이닝, 정보검색

김 태 용
1992년 인천대학교 전자계산공학과(공학사). 1995년 인하대학교 전자계산공학과(공학사). 2000년 인하대학교 전자계산공학과 박사과정 수료. 1995년~1998년 ㈜현대정보기술 정보기술연구소 선임연구원 재직. 1998년~현재 문경대학 웹마스터과 교수. 관심분야는 웹마이닝, 텍스트마이닝, 정보검색, 자연어처리

최 준 혁
1990년 경기대학교 전자계산학과(이학사) 1995년 인하대학교 전자계산공학과(공학석사). 2000년 인하대학교 전자계산공학과 박사(공학박사). 1997년~현재 김포대학 컴퓨터계열 조교수. 2003년~현재 특허청 외부 심사 자문위원. 관심분야는 정보검색, 데이타마이닝, 신경망, 유전자 알고리즘

임 기 욱
1977년 인하대학교 공과대학 전자공학과 졸업. 1987년 한양대학교 전자계산학 석사. 1994년 인하대학교 전자계산학 박사 1977년~1983년 한국전자기술연구소 선임연구원. 1983년~1988년 한국전자통신연구소 시스템소프트웨어 연구실장. 1988년~1989년 미 캘리포니아주립대학(Irvine)방문 연구원. 1989년~1997년 한국전자통신연구원 시스템연구부장 주전산기(타이컴) Ⅲ,Ⅳ개발 사업책임자. 1997년~2000년 정보통신연구진흥원 정보기술 전문의원. 2000년~현재 선문대학교 교수. 관심분야는 실시간 데이타 베이스시스템, 운영체제, 컴퓨터구조

이 정 현
1977년 인하대학교 전자공학과 졸업. 1980년 인하대학교 대학원 전자공학과(공학석사). 1988년 인하대학교 대학원 전자공학과(공학박사). 1979년~1981년 한국전자기술연구소 시스템 연구원. 1984년~1989년 경기대학교 전자계산학과 교수 1989년~현재 인하대학교 컴퓨터공학부 교수. 관심분야는 연어처리, HCI, 정보검색, 음성인식, 음성합성, 컴퓨터구조