

# 무선 환경에서 사용자 검색 성향을 반영한 웹 방송 정보 재구성 기법

김 원 철\* · 이 수 철\*\* · 황 인 준\*\*\* · 변 광 준\*\*\*\*

## 요 약

최근 다양한 무선 단말기의 보급과 네트워크 기술의 발전으로 인하여 무선 단말기를 이용한 인터넷 접속이 보편화되고 있다. 특히 디지털 방송의 도입에 따른 다양한 방송 프로그램과 방송사 뉴스 서비스는 무선 단말기의 제한된 환경에서의 이용률이 높다. 그러나 대부분의 방송사 웹 페이지들은 한 페이지에 많은 내용으로 인한 세분화된 섹션을 담고 있기 때문에 제한된 화면과 입력장치를 가진 무선 단말기를 이용하여 사용자가 원하는 부분에 접근하기까지 반복적인 스크롤링을 해야 하는 불편함이 있다. 이러한 문제를 해결하기 위해 본 논문에서는 방송사 웹 페이지 내에서 실시간으로 사용자가 선호하는 방송사 웹 페이지의 섹션을 추출하고, 무선 환경에 적합하도록 각 섹션의 순서를 재구성하여 무선 단말기에 제공해 주는 기법을 제안한다. 제안된 기법을 통해 사용자는 무선 단말기의 단점을 극복함과 동시에 방송사 웹에서 선호하는 섹션의 맞춤형 방송사 웹 서비스를 제공받을 수 있다.

## Reconstructing Web Broadcasting Information based on User Retrieval Pattern

Woncheol Kim\* · Soocheol Lee\*\* · Eenjun Hwang\*\*\* · Kwang Jun Byeon\*\*\*\*

## ABSTRACT

Today the fastest growing communities of web users are mobile visitors who browse web page with wireless PDAs and cellular phones. However, most web pages are optimized exclusively for desktop clients on the broadband network and are inconvenient to users with small screen mobile devices. They display only a few lines of text and cannot run client-side programs or scripts due to lack of system resource. Even worse, their connections are usually slow to support most of the data-intensive applications. In this paper, we propose a pageslet scheme that makes it feasible to browse ordinary web pages on small screen mobile devices. It extracts broadcasting sections of user preference from broadcasting web pages and automatically reorganizes the extracted sections for convenient browsing on mobile devices.

키워드 : 재구성(Reconstruction), 무선 단말기(Mobile Device), 맞춤형 서비스(Personalized Services)

### 1. 서 론

최근 디지털 기술의 급속한 발전으로 디지털 방송의 도입은 방송환경에 커다란 변화를 가져오고 있다. 즉, 가용채널과 부가 방송 서비스의 폭발적인 증가로 인하여, 다양한 사용자의 요구에 부응할 수 있는 시청자 중심의 방송 서비스 시대가 열릴 것으로 예상된다. 이러한 변화에 발맞추어, 무선 단말기의 대중화와 무선 인터넷 기술의 발전으로 사용자들은 무선 단말기를 이용하여 언제 어디서나 자신이 원하는 방송 정보를 얻고자 한다.

이와 같은 환경 변화로 인해, 신문이나 책자에 활자로 출

력된 프로그램 가이드 대신 전자적인 정보를 통해 방송 프로그램을 조회할 수 있는 서비스인 전자 프로그램 가이드(EPG: Electronic Program Guide)가 출현하였다. 신문이나 잡지를 찾지 않아도 EPG는 방송정보의 조회가 가능하고, 방송 스케줄이 변경되더라도 수정이 불가능한 인쇄매체의 정보와는 달리 즉각적인 수정이 가능하다는 장점을 가지고 있다.

현재 디지털 방송을 위해 제공되고 있는 대부분의 전자 프로그램 가이드 서비스들은 디지털 TV 화면이나 각 방송사 웹 사이트 및 무선 단말기 등을 통해서 서비스되고 있다. 그러나, 방송 채널의 수가 증가하고 사용자가 시청할 수 있는 프로그램이 다양해 지면서, 시청자는 많은 방송 프로그램들 가운데 자신이 원하는 프로그램을 찾기위해 부가적인 노력이 필요하게 되었다.

또한, 실질적으로 사용자 입장에서는 원하는 방송 프로그램 정보를 찾기 위해서 수백개의 채널에 달하는 방송 프로

\* 본 연구는 대학 IT 연구센터 육성·지원사업의 연구결과로 수행되었음.  
† 정 회 원 : 아주대학교 대학원 정보통신  
\*\* 준 회 원 : 아주대학교 대학원 정보통신  
\*\*\* 중신회원 : 고려대학교 전자공학부 교수  
\*\*\*\* 중신회원 : 아주대학교 정보 및 컴퓨터 공학부 교수  
논문접수 : 2004년 6월 4일, 심사완료 : 2004년 7월 27일

그램에 대한 정보를 일일이 검색하는 것은 매우 힘든 일이며, 무선 단말기를 이용하여 각 방송사의 웹 사이트 정보를 일일이 찾아보는 것도 힘들다. 사용자가 원하는 방송에 대한 키워드를 직접 입력하는 방식 또한 일시적인 사용자의 기억에만 의존함으로써 폭넓은 검색이 어렵다.

이러한 문제를 해결하기 위해 기존의 연구들은 방송 메타 데이터를 이용한 프로그램에 정보의 단순 검색뿐만 아니라 이용자가 미리 작성한 사용자 프로파일(Profile)을 이용하여 시청자의 취향에 맞는 프로그램을 제공하고자 노력하였다.

그러나 이러한 연구는 사용자가 프로파일 작성에 대하여 무지하거나, 프로파일 작성을 번거롭게 생각하는 경우 제대로 활용되지 못하는 단점이 있다. 또한 프로파일을 작성하였다고 하더라도 작성시점 이후의 사용자 취향변화에 대해 사용자가 적절히 반응하지 않음으로 인해 불필요한 정보를 제공받게 된다.

본 논문에서는 사용자의 웹 검색 성향을 통하여 사용자가 관심을 가지는 방송 정보들에 대해 추출한다. 웹 사이트 내에서 사용자의 의미있는 접근 성향을 추출하기 위해서는 웹 검색시 사용하는 링크의 앵커 텍스트를 이용하며, 앵커 텍스트는 링크에 대한 가장 중요한 정보를 포함하고 있다.

사용자가 무선 단말기를 이용하여 방송 정보 및 방송 웹 페이지를 요청할 때, 파악된 선호도를 바탕으로 웹 페이지를 재구성해 줌으로써 무선 단말기의 단점을 극복할 수 있으며 방송에서 선호하는 프로그램의 맞춤형 서비스를 제공할 수 있다.

## 2. 관련 연구

최근 대부분의 방송이 아날로그 방송에서 디지털 TV 방송으로 전환되면서 점차적으로 채널의 신설, 추가, 변경, 삭제가 빈번해지고 각 채널을 소개해야 할 필요성이 증가하였다. 사용자는 선택의 폭이 늘어난 만큼 자신의 취향에 맞는 프로그램을 선택하는 시간과 노력을 더 들여야만 한다. 게다가 이제는 인터넷 방송국까지 생겨나면서 수 많은 프로그램 정보를 한꺼번에 모두 얻기 힘들게 되었다. 이에 따라, 방송 프로그램에 대한 정보를 보여주는 EPG(Electronic Program Guide)의 중요성이 부각되었다. 예를 들어, 사용자가 원하는 프로그램을 찾기 위해 수많은 채널의 정보를 일일이 브라우징 하는 것은 사용자에게 매우 번거로운 일이기 때문이다.

이러한 문제를 해결하기 위해 MPEG-7 UsageHistory DS (Description Schemes)나 MPEG-21을 이용하여 사용자의 정보(프로그램 시청정보, 시간대별 선호장르)를 저장하고, 사용자 정보를 동적으로 추출하여 일치하는 EPG 데이터를 우선적으로 표시함으로써 EPG를 효율적으로 제공하는 방법들을 제시하는 연구들이 활발히 진행되고 있다.

이와 더불어 컴퓨터 기술의 급속한 발전 및 웹의 확산으로 개인이 얻을 수 있는 정보의 양이 증가함에 따라, 필요한

관련 정보를 탐색하는 것과 다량의 정보로부터 지식을 추출한다는 것이 어렵게 되어 사용자에게 대한 학습 과정 및 맞춤형 서비스 요구가 증가하게 되었다. 이러한 문제들을 해소하기 위해 웹으로부터 정보를 얻을 수 있는 방법이 필요하게 되었으며, 또한 이를 통해 얻은 정보를 이용하여 웹 사용자들의 성향을 식별할 수 있는 웹 마이닝 방법이 필요하게 되었다.

웹 마이닝(Web Mining)[6-9]이란 웹 환경에서 적용되는 데이터 마이닝 기술로써 웹으로부터 얻어지는 방대한 양의 정보로부터 유용한 정보를 찾아내어 분석하는 것을 말한다. 즉, 웹 환경의 여러 정보로부터 수집된 방대한 양의 데이터를 처리하기 위해 데이터 마이닝 기법을 연결하여 사용자들의 특성을 나타내는 패턴을 찾아내는 기법이라고 할 수 있다. 웹 마이닝은 다음과 같이 크게 3가지 영역으로 분류될 수 있다.

웹 콘텐츠 마이닝(Web Content Mining)은 온라인(On-line) 상에서 이용 가능한 정보를 자동적으로 찾아주는 데이터 마이닝 기법이다. 즉, 웹 기반의 구조화되지 않은 데이터로부터 비슷한 형태를 찾기 위한 것으로 관계형 데이터베이스를 활용하기 위한 데이터 마이닝 기법이다. 웹 콘텐츠 마이닝은 자료의 콘텐츠를 직접 찾아주는 웹 페이지 콘텐츠 마이닝(Web Page Content Mining)과 탐색 엔진(Search Engines)과 같은 도구를 사용하여 콘텐츠 탐색을 향상시키는 질의 결과 마이닝(Search Result Mining)으로 구분할 수 있다. 웹 구조 마이닝(Web Structure Mining)은 웹 환경에서 참조한 페이지와 참조된 페이지 사이의 관계와 구조에 대한 정보 및 웹 사이트나 웹 페이지에 대한 요약된 구조를 생성시키는 데이터 마이닝 기법이다. 웹 용법 마이닝(Web Usage Mining)은 웹 서버로부터 사용자들의 접속 유형과 사용자의 패턴을 찾는 데이터 마이닝 기법으로 웹 로그 마이닝(Web Log Mining)이라고도 한다. 즉, 사용자들이 검색했던 기록을 남기는 웹 서버 로그로부터 사용자들의 접속 유형을 발견하는 것을 목적으로 하는 데이터 마이닝 기법이다. 웹 용법 마이닝은 접속 성향과 유형을 이해하기 위해 웹 로그를 분석하는 것과 사용자 개인을 위해 웹 사이트를 조정하는 것으로 구분할 수 있다.

웹 마이닝 과정은 크게 네 단계로 구분할 수 있는데, 첫째가 웹에서 얻어지는 정보를 수집하는 단계(Web Data Gathering)이다. 예를 들면, 웹 용법 마이닝의 경우에 서버의 로그 파일이 데이터 수집 단계의 정보 소스가 될 수 있다. 두 번째 단계는 전처리 과정(Preprocessing)이다. 정보 수집 단계의 데이터들은 데이터 내에 마이닝과 연관되지 않거나 중복되는 불필요한 데이터들이 다수 존재하는데, 이러한 정보들을 제거하는 작업을 전처리 과정이라 하며 로그파일 정제(log file cleaning), 사용자 및 작업 식별(user and transaction identification) 등의 작업을 포함한다. 세 번째 단계에서는 정제된 자료 및 정보를 이용하여 통계적인 기법 및 데이터 마이닝 기법을 이용하여 특정한 패턴을 찾아내는 작업

을 수행한다. 마지막으로 자료로부터 찾아낸 패턴을 이용하여 적절한 마케팅 전략을 수립하고 개인 대 개인 마케팅 및 개인화 등에 적용하는 단계이다.

웹 마이닝 기법을 이용하여 사용자의 사용자 접근성향을 분석한 후 사용자에게 서비스를 제공해 주는 연구들을 살펴보면 다음과 같다.

WebWatcher[10]는 Carnegie Mellon 대학에서 수행중인 과제인데, 사용자가 웹 페이지를 방문할 때 페이지 내의 링크 중에서 사용자가 관심을 가지고 있다고 판단되는 링크를 추출하여 사용자에게 추천해 주는 일종의 에이전트 기술이다. Letizia[11]는 웹에서 사용자의 검색 패턴을 추적하여 사용자의 최종 목적을 추론하는 에이전트이다. 사용자가 페이지를 읽는 동안 Letizia는 추론한 것을 바탕으로 검색을 수행하게 되고 연관된 페이지가 발견되면 사용자에게 연관된 페이지를 추천해 준다. Letizia의 목적은 사용자가 미래에 검색할 페이지를 예상하고 사용자를 대신해서 추천을 하는 것이다. 이러한 에이전트 시스템 이외에 진행되고 있는 또 다른 연구로는 웹 상에서 사용자의 로그파일에서 사용자의 검색 패턴을 분석하는 것이다. Chen[12]은 웹 페이지 간의 traversal path graph라고 불리는 방향 그래프를 이용하여 사용자의 검색 패턴을 분석하였다. Yan[13]은 사용자들의 접근 빈도를 통해서 사용자들을 분류였다. 이러한 연구는 전자상거래 등에서 사용자 별로 온라인 카탈로그를 설계하는 것과 같은 응용 분야에 유용하다. 그 이외에도 서버의 웹 로그를 분석하고 사용자의 행동 패턴[14, 15]을 파악한 후, 사용자의 패턴에 따라 적응형 웹 사이트[16, 17]를 제공해주는 연구들도 활발히 진행되고 있다.

### 3. 사용자 웹 검색성향 추출 기법

최근 늘어나는 사용자들의 요구에 대처하기 위해, 웹은 페이지 내에 많은 정보를 포함하고 있다. 다른 주제를 가지는 여러 부분들이 존재함에도 각 페이지를 하나의 의미를 갖는 개체로 간주하고 분석을 하면, 사용자의 패턴이나 선호사항 분석시 여러 가지 판단 오류가 발생할 수 있다. 본 논문에서는 페이지내의 주제와 기능성에 따라 페이지의 각 부분을 분리하고, 나누어진 부분에 방송내의 각 섹션에 해당하는 의미를 부여하는 작업을 한다. 이렇게 나뉘어진 각 부분을 **페이지렛(Pagelet)**이라 한다. 방송사 웹 페이지로부터 의미 단위의 **페이지렛**을 추출하기 위해 **페이지렛**은 방송사 웹의 하나의 주제나 기능을 충분히 표현할 수 있어야 한다. 본 논문에서는 **페이지렛**을 다음과 같이 정의한다.

정의 : **페이지렛**은 방송사 웹 페이지내에 주제나 기능성을 표현하는 의미적인 단위이고, 페이지렛내에는 같은 주제나 같은 기능성을 가진 다른 의미구역이 포함되지 않는다.

본 논문에서는 사용자가 선호하는 방송사 웹 페이지에서

각 섹션들을 추출하기 위해서 세 단계로 진행된다. i) 방송사 웹 페이지 내에서 대표 키워드 추출, ii) 웹 페이지 내에서 사용자의 검색 패턴분석을 통한 사용자 검색 순서 결정, iii) **페이지렛** 추출.

#### 3.1 방송사 웹 페이지로부터 키워드 추출

웹 페이지내의 각 **페이지렛**이 방송사 웹의 어느 섹션에 속하는가를 결정하기 위해서 각 섹션들을 대표하는 키워드들과 **페이지렛** 내의 키워드들을 추출하고 추출된 키워드들을 비교하여 방송사 웹의 섹션을 파악한다.

각 방송사 사이트에서 키워드를 추출할 때 문서내의 모든 키워드들을 추출하면 너무나 많은 양의 키워드가 추출되고, 또한 많은 키워드들로 인하여 방송사 웹의 각 해당 섹션에 대한 경계가 모호해 질 것이다. 본 논문에서는 신문 사이트의 페이지에서 키워드를 추출할 경우 우선적으로 각 방송사 웹 분야의 페이지에서 해당 페이지 내에서 중요한 문장들을 먼저 추출한 후에 추출된 문장에서 키워드들을 추출한다. 따라서 방송사 웹의 각 분야에서 보다 효과적으로 키워드들을 추출하기 위해서 웹 문서의 각 섹션에서 중요한 문장들을 추출한 후에 추출된 문장에서 키워드들을 추출한다. 일반적으로 중요한 문장을 추출하기 위해 Lam-Adesina and Jones[18] 방법이 주로 사용된다. 중요한 문장들을 추출하기 위해 각 문장들에 대한 랭킹(Ranking) 방법을 이용하는데, 예를 들면 문서 내에서 문장의 위치, 전체 문서 내에서 단어의 빈도 수를 측정, 문장 내에 특정 단어나 절의 존재 여부, 다른 문장과 단어, 절과의 관계의 여부 등이 포함된다. 또한 문서로부터 중요한 문장(Sentence)을 선택하기 위하여 본 논문에서는 Luhn's Keyword cluster[19] 기술과 HTML문서 내에서 <Title> 태그내의 키워드의 빈도 수, 키워드의 가중치 등을 사용한다.

본 논문에서는 중요한 문장을 선택하기 위해 첫 번째로 TF/IDF[20]의 가중치 값과 Luhn의 Keyword cluster 기법을 변형하였다. 이 방법은 문장내의 어떤 키워드들이 임계값보다 크면 중요한 문장이라고 판단한다. 문장들의 가중치 값을 구하는 공식은 다음과 같다.

$$SW1 = \frac{SK^2}{TK}$$

where SW 1 = the sentence weight score

SK = the number of significant keywords

TK = the total number of keywords

일반적으로 웹 문서내의 <Title> 태그는 그 문서의 가장 중요한 정보를 포함하고 있으므로 문장이 <Title> 태그 내의 키워드들을 얼마만큼 포함하는 가를 살펴보면 다음과 같다.

$$SW2 = \frac{NTT}{TNT}$$

where  $SW_2$  = the sentence weight score

$NTT$  = the number of title terms found in a sentence

$TNT$  = the total number of terms in a title

보다 정확한 측정을 위해서 우리는 문장내의 키워드들간의 차이를 측정하고 중요한 문장을 선택한다. 일반적으로 웹 문서 저자는 문장 내에서 중요성을 표현하기 위해 키워드들을 강조한다. 예를 들면 페이지 내에서 키워드에 대해 색이나 <b> 태그를 이용해 굵은 글씨체를 사용하여 강조한다. 본 논문에서는 강조된 키워드들을 찾아서 중요한 문장을 선택할 때 이 키워드들에 대해 높은 가중치를 이용하여 중요한 문장에 적용한다. 따라서 강조된 키워드들을 이용한 중요성 평가의 수식은 다음과 같다.

$$SW_3 = \frac{MW^2}{TW}$$

where  $SW_3$  = the sentence weight score

$MW$  = the number of modified words

$TW$  = the total number of words

문장의 중요성에 대한 평가는 위의 3가지 계산을 통해 이루어진다.

$$SSW = SW_1 + SW_2 + SW_3$$

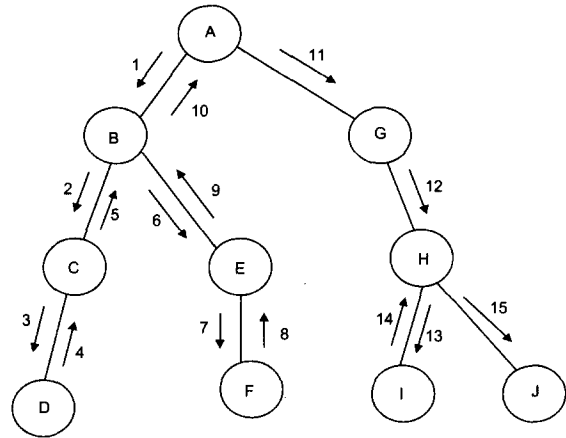
where  $SSW$  = Sentence Significant Weight Score

중요성을 평가한 후 비교될 키워드들을 추출하여 각 색션을 대표하는 대표 키워드로 정의한다.

### 3.2 방송사 웹 색션 검색 순서 추출

순회 패턴(Traversal pattern)은 웹 환경에서 사용자의 검색 패턴을 찾기 위해 제안되었다[22]. 본 논문에서는 순회 패턴 알고리즘을 이용하여 사용자의 웹 사이트 검색 순서와 방송사 웹 검색 순서를 분석한다. 웹 사이트의 페이지들이 서로 연결되어 있다고 가정할 때, 사용자는 자신이 원하는 정보를 얻기 위해 웹 사이트의 링크를 따라 앞이나 뒤로 페이지를 검색하는 경우가 일반적이다. 그 결과로 어떤 페이지의 링크는 그 내용보다는 위치 때문에 방문될 가능성이 있다. 예를 들면 웹 환경에서 웹 사이트를 하나의 트리구조 형식으로 생각해 볼 경우, 형제노드 위치에 있는 페이지를 방문하기 위해서 사용자는 새로운 URL을 사용하는 대신에 보통 역방향 아이콘을 누르고 다시 형제 노드 위치의 페이지를 방문한다. 결과적으로 웹 사이트 내에서 사용자의 기록을 가지고 있는 로그 화일에서 의미있는 사용자 접근 패턴을 추출하기 위해서는 이러한 역방향 검색의 영향을 고려하여 관심있는 실제 접근 패턴을 발견하도록 해야 한다. 이런 관점에서 역방향에 대한 참조는 주로 내용을 검색하기 보다는 검색의 편리를 위해 이루어진다고 가정한다. 그러므

로, 역방향 참조는 같은 사용자 접근에 의해 이전에 방문한 객체를 재방문하는 것을 의미한다. 역방향 참조가 일어나면, 순방향 참조 경로는 종결된다. 이렇게 얻어지는 순방향 참조 경로를 최대 순방향 참조(maximal forward reference)라고 한다.



(그림 1) 순회 패턴(Traversal patterns)의 예

예를 들어, (그림 1)에서 사용자가 웹 페이지 검색시 A, B, C, D, C, B, E, F, E, B, A, G, H, I, H, J의 순회패턴을 가진다고 가정하자. 사용자가 웹 페이지 검색시 최대 순방향 참조를 추출한 후 각 최대 순방향 참조가 어떤 뉴스 색션에 해당 하는가를 결정하는 알고리즘이 그림 2에 나타나 있다. (그림 2)의 알고리즘을 적용하여 순방향 참조 집합을 추출하면 ABCD, ABEF, AGHI, AGHJ와 같다. 즉 사용자가 웹을 검색함에 있어서 역방향 링크를 다시 검색할 경우 순방향 참조가 종료되고 하나의 최대 순방향 참조를 만들게 된다.

(그림 2)의 알고리즘을 통해 순방향 참조 집합을 구할 때 각 노드에 해당하는 링크 앵커의 키워드들을 추출한다. 링크 앵커는 링크내 <a> 태그 사이에 존재하는 문자로, 예를 들면 <a href = "http://www.newskorea.com"> korea news homepage</a> 같은 태그가 존재할 경우 "korea news homepage"을 링크 앵커 텍스트라 한다. 링크 앵커 텍스트는 링크에 대한 가장 중요한 정보들을 포함하고 있다. 본 논문은 링크 앵커 텍스트의 키워드들을 링크 앵커 키워드라고 정의 하며, 링크에 대해 가장 잘 설명해주는 링크 앵커 키워드들을 이용하여 방송사 웹의 색션을 구분한다.

순방향 참조에서, 각 링크에서 결정된 방송사 웹의 색션들을 비교하고 분석하여 최대 순방향 경로의 방송사 웹 색션을 최종 결정한다. 예를 들면 사용자가 (A, B, C, D)의 순서로 링크를 방문했을 경우 각 링크에서 결정된 방송사 웹의 색션들이 (news, movie, news, news)라고 하면, 최대 순방향 경로가 news에 관련된 것이라 판단할 수 있다. 본 논문은 추출된 각 순방향 참조 집합에 대하여 반복적으로 링크 앵커 키워드들을 추출한 후에 방송사 웹 색션을 판단한다.

그러나 순방향 참조 경로의 링크를 통해 각 링크의 방송사 웹 섹션을 판단할 때 사용자가 다음링크를 선택하기 전까지 방송사 웹의 섹션을 판단하지 못할 경우, 이 링크는 방송사 웹의 콘텐츠를 위한 것이 아니라 다음 페이지로 가기 위한 클릭 행동이라고 간주한다. 따라서 본 논문에서 각 작업은 사용자의 참조가 언제 일어날지 모르고, 각 작업이 데드라인이 중요하지 않는 비주기적인 작업이라 가정한다.

```

end if
append B to the end of string Y // step4
fetch anchor text and store it to K
if (F is equal to 0)
    F = 1
end if
set i = i + 1 // step5
}
make Ns's path to section path's graph
    
```

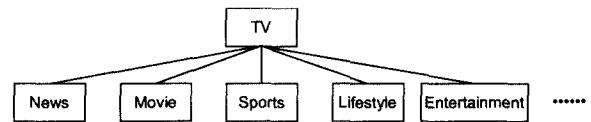
(그림 2) 사용자 선호사항 추출 알고리즘

```

FindNewsPath(L)
make traversal sequence such as {(s1, d1), (s2, d2)... (sn, dn)}
set i = 1 and string y to null for initialization, where string y is
used to store the current forward reference path. // setp1
set the flag F = 1 to indicate a forward traversal.

while(the sequence is not completed scanned)
{
    Let A = si and B = di //step2
    if (A is equal to null)
        Write out the current string Y to the database Df
        Set string Y = B
        Goto step5
    end if
    if (B is equal to some reference in string y) //step3
    if (F is equal to 1)
        write out string Y to database Df
        using keyword matching between K and decide
        K's news section
        append decided news section to the end of string Ns
    end if
    discard all the references after the j-th one in string Y
    F = 0;
    goto step5
}
    
```

(그림 3)은 방송 콘텐츠의 장르를 나타낸다. 사용자의 검색 성향을 알기 위해 순방향 참조가 완료된 후에 각 링크의 링크 앵커 키워드들과 앞 절에서 미리 추출한 방송 각 섹션에 대표 키워드들을 비교하여 방송장르를 결정한다.



(그림 3) 방송 콘텐츠의 장르 구분

(그림 3)과 같이 웹 사이트에서 사용자의 링크정보와 사용자의 정보를 이용하여 방송사 웹 사이트에서 사용자의 일반적인 접근 순서가 N → L → E → S → M → N → S와 같음을 알 수 있다.

사용자가 검색을 재시작 할 경우, 위와 같은 과정을 반복적으로 수행하여 접근 순서에 대한 성향을 분석하고 사용자의 최근 성향이 점차 어떻게 변하고 있는지를 분석한다. 사

Link information	Link anchor keywords	
~/newsArticle.jhtml?type=politicsNews&storyID=2844205	Congress, Try ,Solve ,Medicare Problems	News section
~/newsArticle.jhtml?type=politicsNews&storyID=2844971	Israel, Palestinians, Peace, Steps	
~/newsArticle.jhtml?type=politicsNews&storyID=2840293	Iraqi, Weapons, Reason, War-Wolfowitz	
...	...	News (N)
~/_tscsls/markets/marketstory/10090182.html	lifestyle, babies, Adoption,	
~/financeNewsArticle.jhtml?type= storyID=2845238	style, life, Food, Wine Festival	
~/financeNewsArticle.jhtml?type=News&storyID=2845323	Boys, girls, children, Television	Lifestyle (L)
...	...	
~/newsArticle.jhtml?type=entertainmentNews&storyID=2844570	Thanks, Memory, Americans, Rock	
~/newsArticle.jhtml?type=entertainmentNews&storyID=2842373	Idol, Fox, Pop, Albums, Charts	Entertainment (E)
...	...	
~/newsArticle.jhtml?type=sports&storyID=2843665	star, Premier, League, club	
~/newsArticle.jhtml?type=sports&storyID=2837894	world, game, baseball	Sports (S)
...	...	
~/newsArticle.jhtml?type=internetNews&storyID=2842114	Matrix, Reloaded, radio, television,	
~/newsArticle.jhtml?type=internetNews&storyID=2839697	X-Men, United, Hulk, Movie	Movie (M)
...	...	
~/newsArticle.jhtml?type=politicsNews&storyID=2840773	Bush, G8, Iraq, Debate	
~/newsArticle.jhtml?type=politicsNews&storyID=2843578	Anti-War, Polish, Mayor, Bush Visit	News (N)
...	...	
~/newsArticle.jhtml?type=golfNews&storyID=2820132	Leading, prize-money, winners, U.S., PGA, tour,	
~/newsArticle.jhtml?type=baseballNews&storyID=2820183	Language, Police, baseball, homerun	Sports (S)
...	...	

(그림 4) 웹에서 사용자의 검색 성향

용자가 재구성된 정보를 요청할 경우 분석된 접근 성향 순서를 이용하여 문서를 재구성한다.

이러한 과정은 기존의 웹 로그 화일을 분석해서 사용자의 성향을 파악하는 방법과 달리 많은 전처리작업이 불필요하고 사용자의 접근 성향의 변화를 쉽게 파악할 수 있다.

### 3.3 자동 페이지렛 추출 기술

본 논문에서는 HTML 엘리먼트내의 링크의 존재와 페이지렛 내의 키워드들에 대한 가중치를 부여함으로 각 페이지 내의 섹션 부분을 구분한다. 페이지렛 추출에 대한 알고리즘은 (그림 5)와 같다. 사용자가 웹 페이지를 검색할 때, 제안한 시스템은 웹 문서를 분석한 후 각 HTML 요소들을 추출하고 파스트리를 만들며, 파스트리의 각 노드를 큐에 저장한다. 다음 단계는 큐로부터 노드 엘리먼트를 추출한다. 지식 노드의 링크가 임의의 값보다 크면 독립적인 의미를 포함하고 있다고 판단하고 그 노드를 페이지렛으로 결정한 후 페이지렛 큐에 저장한다. 그러나 임의의 수보다 링크를 적게 가지고 있을 경우 그 노드는 부모노드의 주제에 종속된다고 본다.

```

NewsletPartition(p)
Begin procedure
  parse HTML page
  construct Parse tree
  Queue Parse Tree
  while( Queue is not empty)
    if (top element in Queue has a child with at least k links)
      push all the children of top element to Queue
    else
      declare it as a newslet
      if ( user click is included in this newslet)
        return newslet
      end if
    end if
  end if
End procedure
    
```

(그림 5) 페이지렛 추출 알고리즘

(그림 5)의 알고리즘을 통해 페이지렛을 추출한 후, 페이지렛이 어떤 방송사 웹 섹션에 포함되는가를 판단해야 한다. 일반적으로 방송사 웹 정보는 시간에 따라 분류가 되고 고정된 날짜의 방송사 웹 페이지는 여러 개의 하위 목록으로 연결되고 하위 목록들은 또 다른 하위 페이지로 반복적으로 연결되어 있다. 예를 들어, 방송사 웹 사이트는 정치, 경제, 스포츠 등의 여러가지 섹션으로 나누어져 있으며, 스포츠는 다시 야구, 축구, 농구 등의 하위 목록으로 연결되어 있다. 따라서 페이지렛의 섹션을 판단하기 위해 기존의 방송사 웹 페이지에서 추출한 각 섹션의 대표 키워드들과 비교하여 섹션의 유사도를 측정해야 한다. 본 논문에서는 기존연구에서 주로 사용하는 Luhn's Keyword cluster기술과 TF/IDF를 사용해서 키워드 가중치를 계산하여 페이지렛의 섹션을 판단한다.

사용자 선호 섹션의 랭킹을 계산하기 위해서 한 방송사 웹 페이지 내에서 사용자 자신이 선호하는 기사를 검색하는 시간을 Sessiontime이라 하고 또한 사용자가 검색한 각 분야의 검색 시간을 Branchtime이라고 하면 방송사 웹의 각 섹션에 대한 사용자의 선호도는 다음과 같다.

$$UI_{field} = \frac{BT_{field}}{ST}, 0 \leq UI_{field} \leq 1$$

where  $BT_{field}$  = Branchtime for news branch  
 $ST$  = Sessiontime

방송사 웹 페이지내에서 사용자의 각 섹션에 대한 선호도를 파악하고 최근까지 누적된 정보들을 이용하여 방송사 웹의 각 섹션에 대한 전체적인 사용자 선호도를 추출해야 한다. 전체 선호사항 추출 계산은 다음과 같다.

$$TUI = \sum_{i=0}^s \frac{UI_i \times func(date)}{TS - ST_i}$$

where  $TUI$  = Total User Interests per each branch

$func(date)$  = function of date

$TS$  = Total Sessiontime

$ST$  = Sessiontime

$S$  = count of each UI

누적된 정보들 중 최근에 사용자가 더 많은 선호도를 가진 방송사 웹 섹션이 더 높은 값을 갖게 하기 위해  $func(date)$ 을 이용한다. 이 함수는 날짜 정보를 이용하여 최근 날짜가 수치적으로 더 높은 값을 가지도록 한다.

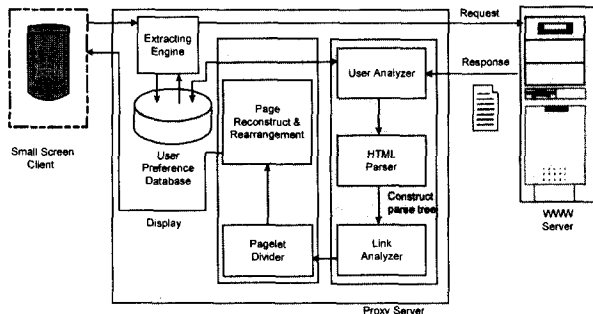
## 4. 방송사 웹 콘텐츠 재구성 기법의 전체 구조

사용자의 검색 성향을 분석하는 방법은 사용자의 정보를 분석하는 위치에 따라 다음과 같이 세 가지로 나눌 수 있다. 첫째, 서버측의 분석 방법은 웹 사이트내의 사이트 방문자들에 대한 기록을 가지고 있는 웹 서버의 로그 파일을 이용하는 것이다. 웹 서버 로그에 기록된 데이터의 대부분은 여러 사용자들이 웹 사이트를 방문한 기록들이다. 이러한 웹 로그는 웹 서버에 따라 다른 형태를 가지게 된다. 그리고 웹 서버 로그 내에는 사용자의 검색 행동이외에 복잡한 여러 가지 추가 정보를 가지고 있다. 이러한 문제로 사용자의 행동을 분석하고 의미 있는 정보를 추출하기 위해서는 여러 단계의 정제과정을 거쳐야 한다. 또한 웹 환경 내에서 다양한 형태의 웹 페이지에 대한 캐싱이 존재하여 캐쉬된 페이지는 서버로그에 기록되지 않기 때문에 로그 파일의 기록이 웹 사용자들에 대한 모든 행동을 완전하게 보여준다고 보기는 어렵다. 또한 POST 메소드를 통해서 전달되는 중요한 정보는 서버의 로그 내에 기록되지 않게 되어 정확한 사용자에 대한 정보를 추출하기 어렵다.

둘째, 클라이언트 측의 분석 방법은 원격(remote) 에이전트(즉 자바 스크립트, 자바 애플릿 혹은 클라이언트 측 애플리케이션)에 의해 사용된다. 이런 클라이언트 측의 분석은 캐싱이나 세션 아이디등의 문제를 쉽게 해결하기 때문에 서버 측의 분석보다 효율적이지만 사용자의 동의를 얻어야 하기 때문에 어려움이 있다. 또한 원격 에이전트가 클라이언트 측에서 복잡한 연산을 수행할 경우 사용자의 행동(action)에 대한 적절한 수행하지 못할 수도 있다.

셋째, 프록시는 웹 서버와 클라이언트 브라우저간의 캐쉬 등 이들간의 중재자 역할을 한다. 프록시 레벨의 분석은 사용자가 이전에 방문했을 경우 웹 페이지 로딩과 네트워크 트래픽 감소 등의 효과를 가져오고 프록시 캐싱에서의 성능은 앞으로 사용자가 사용할 페이지를 얼마만큼 정확하게 예측하느냐에 따라 달라진다.

선호도 기반의 방송사 웹 전달 기법은 캐쉬 문제와 모듈 삽입이 용이한 클라이언트와 서버 사이의 프록시 서버에 위치하게 된다. (그림 6)은 방송 웹 메신저 시스템의 전체적인 구조를 나타낸다. 무선 단말기의 사용자가 웹 페이지를 요청하였을 때 처리 순서는 다음과 같다.



(그림 6) 방송 웹 메신저 시스템 전체구조

- ① 무선 단말기 사용자가 웹 페이지를 요청하였을 때 추출 엔진은 무선상의 사용자의 선호 사항을 파악하고 데이터베이스에 저장하거나 저장된 정보를 추출한다.
- ② 프록시 서버가 웹 서버로부터 응답 메시지를 받을 때, 사용자를 파악하고 사용자 선호사항 데이터베이스로부터 저장되었던 정보를 추출한다. 그런 후 사용자 정보와 HTML 문서를 HTML 파서에 전달한다.
- ③ HTML 파서 모듈이 문서를 전달 받았을 때, 파서는 문서내의 배너와 그림을 제거하고, 문서를 각 노드로 분리하고 파스 트리를 생성한다. 페이지렛을 만들기 위하여 Link analyzer 모듈이 파싱된 트리를 생성한 후 링크 정보를 추출한다.
- ④ Pagelet Divider 모듈은 방송사 웹 페이지내에서 방송사 웹의 단위를 파악하기 위해 파싱된 트리 노드의 링크수를 분석하여 페이지렛을 만든다. Page Reconstruct and Rearrangement 모듈은 추출된 사용자의 선호하는 방송사 웹 섹션을 바탕으로 페이지렛의 순서

를 재배치한다.

위와 같은 처리순서를 이용한 방송사 웹 메신저 시스템은 사용자가 더 많은 선호사항을 가지고 있는 방송사 웹의 섹션을 무선 단말기 화면에 보다 위에 보여주고, 불필요한 배너와 그림을 제거함으로써 무선 단말기의 작은 화면에 모든 정보를 표현하려고 할 때 제약사항을 완화시켰다.

무선 단말기를 이용하여 EPG를 원할 때 사용자에게 EPG 정보를 그대로 전달해 주면 사용자가 원하는 프로그램을 검색하기 위해서 많은 스크롤링이 요구되며, 원하는 프로그램을 찾기가 쉽지 않다.

따라서 본 논문은 웹에서 사용자의 성향을 분석후 사용자가 먼저 접근하는 방송 섹션에 해당하는 EPG 정보를 나중에 접근하는 방송섹션보다 위에 보여줌으로 작은 화면의 제약을 해결하였다.

## 5. 실험 및 분석

### 5.1 실험 환경

본 논문은 웹 기능을 가지는 PDA와 동일한 환경을 제공하기 위해 모바일 에뮬레이터를 이용하여 구현하였고 이를 이용하여 성능 평가를 하였다. 방송사 웹 페이지로부터 키워드를 추출하고, 방송사 섹션에 관한 정보를 데이터베이스에 저장하는 것은 데스크톱환경에서 자바 J2SE1.4.1 Swing API 이용하였고, Windows XP 환경에서 개발되었다. 추출된 사용자의 선호사항을 이용하여 모바일 에뮬레이터에 재구성하는 것은 임베디드 비주얼베이직 3.0을 이용하여 구현하였다. 링크에 연결되어있는 링크 앵커 키워드를 추출하기 위해 Swing이 제공하는 DTD(Document Type Definition) 기반의 파서를 이용하였다.

### 5.2 성능 평가

대부분의 웹 페이지들이 데스크톱에 최적화되어 있어 무선 단말기를 이용하여 웹 페이지를 검색할 경우 콘텐츠의 전체 구조를 이해하기 위해 많은 스크롤 작업이 필요하고 낮은 대역폭으로 인해 페이지의 내용을 가져오기까지 많은 시간이 걸린다. 따라서 본 논문에서 제안하는 시스템 성능을 평가하기 위해 사용자가 방송사 웹 페이지를 검색하기 위한 스크롤 수와 단말기를 통해 전달받게 되는 페이지 사이즈를 측정하였다.

<표 1>과 <표 2>는 사용자 선호사항을 기준으로 방송사 웹 페이지를 재구성한 경우의 이점을 보여준다. 방송사 웹에 포함되는 일반적인 6개의 섹션 News, Lifestyle, Drama, Sports, Entertainment, Movie를 각각 P, T, B, S, E, M이라고 표현하였다. 실험에서는 모바일 화면에 제공되는 전형적인 방송사 웹 페이지의 순서가 P, B, T, S, E, M 이라고 가정한다. 재구성된 방송사 웹 섹션이 이점을 갖는지 판단하기 위해 위/아래로의 스크롤링의 수를 계산한다. 주어진 순

서의 방송사 웹 섹션을 읽기 위해 사용자 스크롤의 총합을 비교한다.

〈표 1〉 재구성하지 않은 경우 스크롤의 수

User	Display sequence without reorganization	User sequence	The number of scrolling for each sequence	Total number of scrolling
A	N, D, L, S, E, M	N, D, E, L, M, N, E	0, 0, 2, 0, 1, 3, 2	8
B		M, D, E, N, D, M, D	3, 2, 1, 2, 0, 3, 2	13
C		M, E, D, N, S, L, S	3, 0, 2, 1, 1, 0, 0	7
D		M, L, S, M, E, L, M	3, 1, 0, 1, 0, 1, 1	7
E		E, M, S, L, N, D, N	2, 1, 0, 1, 2, 0, 0	6
F		N, D, S, E, L, M, L	0, 0, 0, 1, 1, 0, 1	3

예를 들면, 무선 단말기 화면에 단지 3개의 섹션만 보인다고 가정하고, 사용자 A의 방송사 웹 사이트 검색에서 패턴이 일반적으로 N, D, E, L, M, N와 같은 경우, 재구성하지 않은 모바일 단말기를 이용할 경우 N, D, L, S, E, M의 순서로 나열된 화면을 통해 방송사 웹을 검색해야 한다.

모바일 단말기의 화면을 통해 단지 3개의 섹션을 보여주기 때문에 처음 모바일 단말기를 통해 N, D, L의 섹션을 볼 수 있다. 처음 사용자가 관심을 가지는 방송사 웹 섹션이 N이므로 사용자는 스크롤을 하지 않아도 된다. 사용자가 일반적으로 이후에 검색을 하는 방송사 웹 섹션은 D이므로 역시 스크롤을 하지 않아도 된다. D를 검색한 후 사용자가 검색하는 방송사 웹 섹션은 E이므로 사용자는 위로 2번의 스크롤을 통해 L, S, E를 화면상에 보게 되고 E의 방송사 웹 섹션을 검색한다. E의 방송사 웹 섹션을 검색후, 사용자가 일반적으로 검색하는 방송사 웹 섹션은 L이므로 사용자는 스크롤을 하지 않아도 필요한 방송사 웹 섹션을 검색할 수 있다. 방송사 웹 섹션을 검색한 후 M을 보기 위해 사용자는 위로 1번의 스크롤 작업을 해야 하고 모바일 스크린상에 S, E, M을 보이게 하여 방송사 웹 섹션을 검색한다. 마지막으로 N와 E 섹션을 보기위해 5번의 스크롤 작업을 해야 한다.

〈표 2〉 재구성한 경우 스크롤의 수

User	Display sequence with reorganization	User sequence	The number of scrolling for each sequence	Total number of scrolling
A	N, D, E, L, M, S	N, D, E, L, M, N, E	0, 0, 0, 1, 1, 2, 0	4
B	M, D, E, N, S, L	M, D, E, N, D, M, D	0, 0, 0, 1, 0, 1, 0	2
C	M, E, D, N, S, L	M, E, D, N, S, L, S	0, 0, 0, 1, 1, 1, 0	3
D	M, L, S, E, N, D	M, L, S, M, E, L, M	0, 0, 0, 0, 1, 0, 0	1
E	E, M, S, L, N, D	E, M, S, L, N, D, N	0, 0, 0, 1, 1, 1, 0	3
F	N, D, S, E, L, M	N, D, S, E, L, M, L	0, 0, 0, 0, 1, 1, 0	2

그러나 <표 2>와 같이, 방송사 웹 섹션을 재구성한 경우 사용자 A는 단지 4번의 스크롤 작업을 필요로 한다; 예를 들면, 무선 단말기 화면에 단지 3개의 섹션만 보인다고 가정

하고, 사용자 A의 방송사 웹 사이트에 대한 검색패턴이 N, D, E, L, M, N와 같은 경우, 재구성을 한 모바일 단말기를 이용할 경우 N, D, E, L, M, S의 순서로 나열된 화면을 통해 방송사 웹을 검색해야 한다.

모바일 단말기의 화면을 통해 단지 3개의 섹션을 보여주기 때문에 처음 모바일 단말기를 통해 N, D, E의 섹션을 볼 수 있다. 처음 사용자가 관심을 가지는 방송사 웹 섹션이 N이므로 사용자는 스크롤을 하지 않아도 된다. 사용자가 일반적으로 이후에 검색을 하는 이후의 방송사 웹 섹션은 D와 E이므로 역시 스크롤을 하지 않아도 된다. E를 검색한후 사용자가 검색하는 방송사 웹 섹션은 L이므로 사용자는 위로 1번의 스크롤을 통해 D, E, L의 화면이 보이게 하고 L의 방송사 웹 섹션을 검색한다. 방송사 웹 섹션을 검색후 M을 보기위해 사용자는 위로 1번의 스크롤 작업을 해야 하고 모바일 스크린상에 E, L, M의 화면을 보이게 하여 방송사 웹 섹션을 검색한다. 마지막으로 N와 E의 섹션을 보기위해 2번의 스크롤 작업을 해야한다. 즉, L를 검색하기 위해 1번, M을 검색하기 위해 1번 그리고 N를 검색하기 위해 2번이 필요하다.

지금까지 방송사 웹 섹션을 재구성한 경우가 재구성하지 않은 경우보다 더 적은 양의 스크롤을 필요로 한다는 것을 알 수 있다. 결과적으로 사용자가 선호하는 섹션으로 재구성하는 방법이 작은 화면을 가진 무선 단말기에 편리하다는 것을 알 수 있다.

본 논문에서 제안된 시스템의 파일 크기와 화면 크기를 측정하기 위해, 여러 방송사 웹 페이지에서 문서를 재구성한 후에 문서의 크기와 화면상의 크기를 비교했다. 표 3은 실험에 사용된 방송사 웹 페이지들의 리스트를 보여준다.

〈표 3〉 실험에 사용된 방송사 웹 페이지

Site	URL
CNN	http://www.cnn.com/
BBC	http://news.bbc.co.uk/
MSNBC	http://www.msnbc.com/news/NEWS_Front.asp?Odm = C = y
ABCNews.com	http://abcnews.go.com/
Google news	http://news.google.com/
USATODAY	http://www.usatoday.com/
CBS SportsLine	http://cbs.sportsline.com/
CNET news.com	http://news.com.com/2001-12-0.html?tag=hd_ts
TechWeb	http://www.techweb.com/

<표 4>는 방송사 웹 페이지의 파일 크기와 페이지를 재구성한 후에 크기, 화면상의 감소된 크기를 비교하고있다. 여기서 우리의 페이지렛 단위의 사용이 파일과 화면의 크기를 줄이는 것을 볼 수 있다. 평균 파일 크기와 화면 감소가 평균 72% 이상이다. 파일 크기의 감소는 무선 환경의 제한된 대역폭에 방송사 웹 콘텐츠를 전송할 경우 계약을 완화시키고, 화면 크기의 감소는 작은 화면의 단말기의 제한과



적은 스크롤링으로 사용자가 보고 싶어하는 내용을 볼 수 있다.

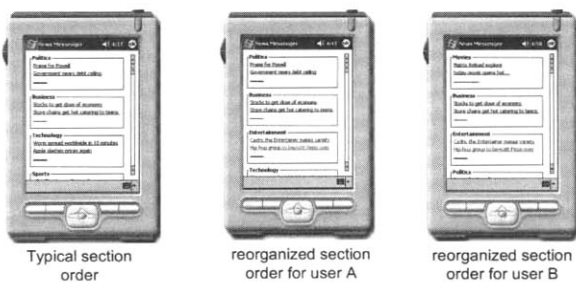
〈표 4〉 재구성한 후 페이지 비교

Page	Original page size	Newsletter size	Display Saving
CNN	122.4Kb	25.5Kb	34.5
BBC	110.6Kb	23.2Kb	33.2
MSNBC	123.4Kb	25.6Kb	31.6
ABCNews.com	115.8Kb	22.5Kb	32.5
Google news	118.2Kb	30.7Kb	32.7
USATODAY	128.1Kb	28.1Kb	33.1
CBS SportsLine	117.4Kb	27.2Kb	38.4
CNET news.com	114.3Kb	29.7Kb	36.2
TechWeb	126.6Kb	27.0Kb	33.4

결과에서 보는 것과 같이 방송사 웹 페이지내에서 사용자가 선호섹션을 무선 단말기의 작은 화면에 차례로 재구성함으로써 검색이 용이하고 스크롤링이 감소되는 것을 알 수 있다. 또한 데스크톱에 최적화되어 있는 불필요한 그림과 팝업 창 등을 제거함으로써 로딩시 많은 시간을 줄일 수 있었다. 결과적으로 본 논문에서 제안한 **페이지렛**을 이용함으로써 무선 단말기 사용자는 웹 페이지에 대한 가독성뿐만 아니라 적은 스크롤링으로 편리한 사용이 가능해졌다.

5.3 결 과

(그림 7)은 사용자가 정치, 경제, 오락등의 순으로 방송사 웹 섹션을 선호할 때, 방송사 웹 페이지를 요청했을 경우의 화면과 경제, 오락, 기술등의 순서로 방송사 웹 섹션을 선호하는 사용자가 방송사 웹 페이지를 요청했을 경우 재구성되지 않은 경우와 사용자의 선호 사항을 통해 재구성된 경우의 화면을 보여준다.



(그림 7) 방송사 웹 페이지의 재구성

6. 결론 및 향후연구 방향

무선 단말기 사용자의 인터넷 접속이 증가하면서 점차 웹 페이지는 유선 사용자뿐 아니라 무선 사용자의 요구를 만족시켜야 할 필요성이 증가하였다. 이러한 필요성으로 인해 현존하는 웹 페이지들이 각각의 기기에 맞게 표현되어야 하

지만, 무선 단말기는 데스크톱에 비해 제한된 인터페이스와 화면 크기를 가지므로 사용자에게 사용의 불편함을 가져온다. 따라서 본 논문에서 무선 단말기 사용자의 요구를 만족시키고 편리한 인터페이스를 제공하기 위하여 실시간 선호사항 분석 기반의 방송사 웹 페이지 전달기법을 제안하였다. 특히 무선 단말기에 맞는 페이지의 별도 준비없이 현존하는 웹 페이지를 통하여 맞춤형 서비스를 제공하게 되었다. 실시간 선호사항을 추출함으로써 방송사 웹의 각 섹션을 사용자의 검색 성향을 이용하여 재구성함으로써 무선 단말기의 작은 화면에서 스크롤링을 줄였고, 사용된 **페이지렛**은 본래 페이지와 비교하여 작은 파일 사이즈로 변형시켰다. 따라서 무선이란 환경의 낮은 대역폭의 제한을 완화시켰고 작은 화면에서 가독성을 증가시켰다.

향후 계획으로는 사용자의 선호사항을 분석한 후에 무선 환경에서 방송사 웹 사이트만이 아닌 일반적인 사이트에도 적용시키는 것이다.

참 고 문 헌

- [1] T. W. Bickmore, B. N. Schilit, "Digester : Device-independent Access to the World Wide Web," Proceedings of the 6th international World Wide Web Conference, Santa Clara, CA, 1997
- [2] B. B. Bederson, J. D. Hollan, "Pda++ : A Zooming Graphical Interface for Exploring Alternate Interface Physics," ACM Symposium on user Interface Software and Technology, 1994.
- [3] S. Bjrk, L. E. Holmquist, J. Redstrm, I. Bretan, R. Danielsson, J. Karlgren and K. Franzn, "WEST : A Web Browser for small Terminals," ACM Symposium on User Interface Software and Technology, 1999.
- [4] Buyukkokten, H. Garcia-Molina and A. Paepcke, T. Winograd, "Power Browser : Efficient Web Browsing for PDAs," Proceedings of CHI'2000, ACM Press, Amsterdam, 2000.
- [5] Wap Forum, White paper(Wireless Internet Today Overview), <http://www.wapforum.org/>, June, 2000.
- [6] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data," In : ACM SIGKDD 2000.
- [7] Baldonado Lizhen Liu, Junjie Chen and Hantao Song, "The research of Web mining," Intelligent Control and Automation, Proceedings of the 4th World Congress on, Vol.3, pp.2333-2337, 2002.
- [8] F. Massegli, P. Poncelet and R. Cicchetti, "WebTool : An Integrated Framework for Data Mining (PS)," Proceedings of the 10th International Conference on Database and Expert Systems Applications (DEXA'99), Florence, Italy,

LNCS, Vol.1677, pp.892-901, September, 1999.

[9] F. Massegli, P. Poncelet and R. Cicchetti, "WebTool : An Integrated Framework for Data Mining (PS)," Proceedings of the 10th International Conference on Database and Expert Systems Applications (DEXA'99), Florence, Italy, LNCS, Vol.1677, pp.892-901, September, 1999.

[10] Armstrong et al., "WebWatcher : A Learning Apprentice for the World Wide Web. Working Notes of the AAAI Spring Symp : Information Gathering from Heterogeneous, Distributed Environments," AAAI Press, pp.6-12, 1995.

[11] Henry Lieberman, "Letizia : An Agent That Assists Web Browsing," In Proc. of the 1995 International Joint Conference on Artificial Intelligence, Montreal, Canada, 1995.

[12] M. S. Chen, J. S. Park and P. S. Yu, "Data mining for Path Traversal Patterns in a Web Environment," Proceedings of the 16th International Conference, May, 1996.

[13] T. W. Yan et al., "From User Access Patterns to Dynamic Hypertext Linking," Proceeding of the Fifth International World-Wide Web Conference, May, 1996.

[14] Wenwu Lou Hongjun Lu Guimei Liu Qiang Yang "Restoring Meaningful Episodes in a Proxy Log," Computer Networks, Amsterdam, Netherlands, 1999.

[15] Feng Tao, Fionn Murtagh, "Towards Knowledge Discovery from WWW Log Data," The International Conference on Information Technology : Coding and Computing (ITCC '00), Las Vegas, Nevada, March, 2000.

[15] Mike Perkowitz and Oren Etzioni, "Adaptive web sites : Automatically cluster mining," In Fifteenth National Conference on Artificial Intelligence, Madison, WI, 1998.

[16] Mike Perkowitz and Oren Etzioni, "Adaptive web sites : Conceptual cluster mining," In Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999.

[17] Adenike M. Lam-Adesina, Gareth J. F. Jones, "Applying summarization techniques for term selection in relevance feedback," In proceedings of the 24th annual International ACM SIGIR Conference, ACM press, pp.1-9.

[18] H. P. Luhn, "The automatic creation of literature abstracts. IBM Journal of Research and Development," IBM journal of research & development, Vol.2, No.2, pp.159-165, April.

[19] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, "Modern Information Retrieval," Addison Wesley.

[20] D. Song and E. Hwang "PageMap : Summarizing Web Pages for Small Display Devices," Proc. of Int'l Conference on Internet Computing, Las Vegas, pp.506-512, June, 2002.

[21] Ming-Syan Chen, Jong Soo Park, Philip S. Yu "Efficient

Data Mining for Path Traversal Patterns," IEEE Transactions on Knowledge and Data Engineering, Vol.10, No.2, pp.209-221.



**김 원 철**

e-mail : wc323@ajou.ac.kr  
 2002년 아주대학교 컴퓨터 공학과(학사)  
 2002년~2004년 아주대학교 정보통신  
 전문대학원(석사)  
 2004년~현재 대한생명 전산팀  
 관심분야 : 데이터베이스, XML 응용



**이 수 철**

e-mail : juin@ajou.ac.kr  
 1998년 한남대학교 컴퓨터 공학과(학사)  
 1998년~2000년 아주대학교 정보통신  
 전문대학원(석사)  
 2000년~2002년 아주대학교 정보통신  
 전문대학원 박사수료  
 2003년~현재 아주대학교 정보통신전문대학원 박사과정  
 관심분야 : 데이터베이스, 멀티미디어 시스템, 정보 통합, XML  
 응용, 유비쿼터스



**황 인 준**

e-mail : ehwang@ajou.ac.kr  
 1988년 서울대학교 컴퓨터공학과(학사)  
 1990년 서울대학교 컴퓨터공학과(석사)  
 1998년 Univ. of Maryland at College  
 Park 전산학과(박사)  
 1998년 Hughes Research Lab. 연구 교수  
 1998년~1999년 Bowie State Univ., Assistant Professor  
 1999년~2002년 아주대학교 정보통신전문대학원 조교수  
 2003년~2004년 아주대학교 정보통신전문대학원 부교수  
 2004년~현재 고려대학교 전자공학부 조교수  
 관심분야 : 데이터베이스, 멀티미디어 시스템, 정보 통합, 전자상  
 거래, XML 응용, 유비쿼터스



**변 광 준**

e-mail : byeon@ajou.ac.kr  
 1985년 서울대학교 컴퓨터 공학 학사  
 1987년 Pennsylvania of State University  
 전산학 석사  
 1993년 University of Southern California  
 전산학 박사  
 1994년~현재 아주대학교 정보 및 컴퓨터공학부 조교수, 부교수  
 관심분야 : 멀티미디어 데이터베이스, 분산 데이터베이스, 분산  
 객체 컴퓨팅