2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

포린 페이지 시스템: 웹 컨텐츠 추출 및 통합을 통한 메타 브라우징 서비스의 설계 및 구현

박남훈, 이원석
연세대학교 컴퓨터과학과
e-mail: { zyonix, leewo}@amadeus.yonsei.ac.kr

Foreign Page System: Design and Implementation of Meta-Browsing Service by Web-Contents Extraction and Composing

Namhun Park, Wonsuk Lee
Dept. of Computer Science, Yonsei University

요약

본 연구는 웹 컨텐츠 통합 서비스에 관한 것으로 메타 브라우저, 중계 웹 서버, 포린 페이지 저작기, 포린 페이지 저장기로 구성한다. 메타 브라우저를 통해 사용자가 웹 사이트를 탐색하면서 웹 컨텐츠를 선택하며, 포린 페이지 저작기를 통해 각 사이트의 컨텐츠들로 포린페이지를 저작한다. 중계 웹 서버에서는 포린 페이지에 사용된 컨텐츠를 주기적으로 모니터링하여 컨텐츠 변화 감지시에 해당 컨텐츠로 구성된 포린페이지도 자동으로 갱신한다. 컨텐츠 추출을 위해 웹 문서로 태그 트리를 구성하고, 그룹 시간 관계를 정의하여 포린 페이지 재생 모델을 제시했으며, 동기화를 위해 종료 제한 시간을 예측한다. 컨텐츠 변화 탐지 및 자동 갱신을 위해 컨텐츠 태그 트리와 웹 문서의 태그 트리간 차이값을 구하여 컨텐츠 변화 감지 방법을 제시한다.

1. 서론

웹은 이질적인 정보들로 구성된 광범위한 네트웍이다. 많은 사람들이 웹 사이트를 통해서 새로운 정보를 얻지만, 수많은 웹 사이트 내에 존재하는 정보들은 빠른 속도로 변화한다.

웹은 하이퍼 링크(Hyperlink)에 의해 연결되어 있어서, 현재의 브라우징 서비스는 하이퍼 링크에 의해 이뤄진다. 이와 달리 메타 브라우징 서비스는 현재 다양한 연구[1,2,3]가 진행중인 새로운 브라우징 기술의 하나로서 이들 연구에서 제시하는 메타 브라우징 해법은 기존의 하이퍼 링크에 의한 검색이나 포탈 서비스와 다르게 사용자가 원하는 컨텐츠를 웹에서 선택, 편집하여 "개인 포털"을 구축하는 서비스이다. 웹에 대한 검색없이 개인 포털을 통해 컨텐츠를 한 눈에 볼 수 있는 기능을 제공하며 컨텐츠가 갱신되면 개인 포털의 컨텐츠도 함께 갱신 된다.

본 논문에서는 메타 브라우징 서비스의 구현을 위해 포린 페이지 시스템을 설계, 구축하였다. 포린 페이지 시스템은 크게 중계 브라우저, 중계 웹 서버, 포린 페이지 저작기, 포린 페이지 저장기로 구성되어 있다.

2장에서 포린 페이지 설계 내용에 대하여 기술하고, 3장에서는 2장의 설계 내용을 바탕으로 포린 페이지 시스템을 구현한다. 마지막으로 4장에서는 논문의 결론을 기술한다.

2. 포린 페이지 시스템 설계

2.1 포린 페이지

메타 브라우징 솔루션에 의한 기존의 웹 페이지를 ***개인 포털***이라 하며, 이는 개인이 필요한 웹 컨텐츠들이나 링크 정보를 모아서 만든 맞춤형 포털 서비스이다. 이와달리 ***포린 페이지***는 개인 포털 기능뿐만 아니라 웹 객체 모델[4]에 따라 컨텐츠와 저작요소가 각각의 속성과 이벤트, 시간 정보를 가지고 재생된다. 따라서 포린 페이지란, 저작 요소가 시간에따라 사용자 입력이나 이벤트, 위치 정보를 가지고 재생되는 개인 포털의 집합으로 정의한다.

2.2 컨텐츠 추출을 위한 웹 문서 모델링

웹 문서를 컨텐츠 추출을 위해 태그 트리로 구조화 하며 태그 트리를 구성하기 위해서 웹 문서를 '오류없는 웹 문서'의 형태로 변환한다.

2.2.1 오류없는 웹 문서

웹 문서는 텍스트와 태그로 이루어진다. 태그는 '<','>' 사이에 태그 이름과 속성리스트로 표시되며 일부태그를 제외한 나머지 태그는 쌍으로 문서에 나타난다. 태그 쌍에서 '/'로 시작하지 않는 태그를 시작 태그라 명하며 반대로 '/'로 시작하는 태그를 끝 태그라 정의한다. 그리고 컨텐츠 추출을 위해 다루는 웹 문서는 다음과 같은 특징을 가진다.

1. 문서내에 태그가 아닌 '<','>' 부호를 가지지 않는다. 문서내의 텍스트에서 '<','>'부호는 < 와 >로 표시한다.
2. 모든 태그는 쌍으로 구성되어 있다. 모든 시작 태그에는 쌍이되는 끝 태그가 뒤따른다.
3. 태그 내의 속성 값은 인용 마크내에 포함된다. (e.g. <a href="www.yahoo.co.kr">
4. 끝 태그 없이 쓰여지는 태그(<img>,<hr>,
)는 끝 태그가 바로 뒤따르도록 수정한다.
5. 태그의 쌍들은 다른 태그를 포함하여 중첩시킨다.

일반 웹 문서를 오류없는 웹 문서로 변환하는 과정을 웹 문서 정규화라 정의한다.

2.2.2 웹 문서의 태그 트리 생성

웹 문서 정규화를 거쳐서 내부 노드는 태그로, 리프노드는
태그 내부의 문자열, 숫자나 그외의 데이터로 구성된 태그 트리를 구성한다. 태그 트리는 [그림 1]과 같이 생성된다.

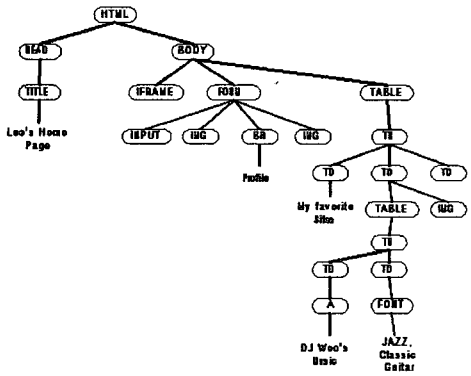

그림 1. 웹 문서로부터 구성한 태그 트리

T=(V,E)를 웹 문서의 태그 트리라 한다. $V = V_T \cup V_C$ V_T 는 태그 노드의 유한집합이며 V_C 는 리프 노드(컨텐츠 노드)의 유한집합이다. $E \subset (V \times V)$ 는 방향성을 지닌 노드간의 연결선이다. 본 연구에서 사용한 태그 트리의 세부개념을 설명하면 다음과 같다.

부모노드 : 노드 n의 부모노드의 정의는 다음과 같다. Parent(n)={w|w∈V,(w,n)∈E}. 루트 노드는 트리내에 부모노드가 존재하지 않는 노드로 정의한다.
자식노드 : 노드 n의 자식노드는 다음과 같이 정의한다. Children(n)={w|w∈V,(n,w)∈E}. 노드간 (u,w)∈E 선분이 존재하면 w는 n의 자식노드라 정의한다.
팬아웃 : 트리내의 노드 n에서 팬아웃 n은 다음과 같다. Fanout(n)=||Children(n)|| , $n \in V_I$ 이며 $n \in V_C$ 이면 Fanout(n)=0이다.
노드크기 : $n \in V$이며 $n \in V_C$ 일 때 NodeSize(n)은 태그 내의 내용(스트링, 텍스트나 그외의 태그 내부의 데이터)의 바이트 크기로 정의한다. 반대로, $n \in V_I$ 인 경우 NodeSize(n)은 Children(n)에 포함되는 모든 리프노드 크기의 합으로 정의한다. $NodeSize(n)=\sum_{v_i \in Children(n)}(NodeSize(v_i))$
태그수 : $n \in V$이며 $n \in V_C$ 일 때 TagCount(n)=0이다. 반대로, $n \in V_I$ 인 경우 $TagCount(n)=\sum_{v_i \in Children(n)}(TagCount(v_i))$. 태그수는 하위노드의 개수를 의미한다.

2.3 포린 페이지 재생 모델

포린 페이지는 웹 객체 모델[4]을 적용하여 태그를 포함하는 컨텐츠와 웹 페이지의 구성요소들을 객체로 정의하여 속성값과 각 객체의 메소드를 재정의할 수 있다. 본 연구에서는 포린 페이지내의 객체들을 객체의 속성 정보와 포린 페이지 재생을 위한 시간에 따른 재생 메소드로 정의한다.

2.3.1 포린 페이지 시간속성 재생 모델

시간 속성 표현을 위한 방법으로 시간 간격에 대한 표기를 주로 사용한다. 시간 상수는 시간축에서 길이가 0인 순간으로 정의하며, 시간 간격은 두 시간 상수로 나타낼 수 있는 시간 상수 사이의 기간(길이)으로 정의한다. 포린 페이지 내의 객체 간의 시간 관계는 객체의 시간 상수와 시간 상수간의 시간 간격으로 표현할 수 있는데, 두 객체간 시간 관계는 before, meets, overlaps, during, starts, finishes, equals로 정의한다[5].

2.3.2 복합 객체간 그룹 시간 관계

많은 객체를 포함하는 웹 문서 재생에 있어 시간 재생 모델은 단일 시간 연산에 의한 동기화 문제를 가진다[6]. 다수의 객체에 대한 동기화의 방법으로 본 연구에선 그룹 시간 관계를 사용하였다. Allen Relation[5]을 다수 객체에 적용한 그룹 시간 관계는 그림 2와 같이 표현된다.

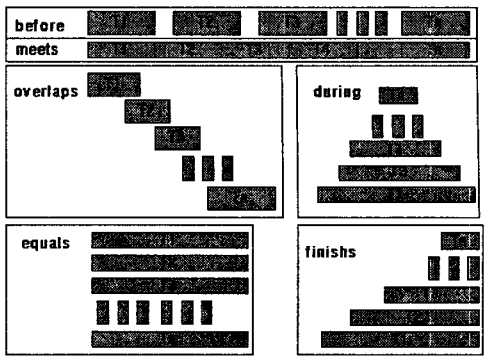

그림 2. 그룹 시간 관계

2.3.3 복합 객체간 그룹 시간 관계에서의 종료제한시간

웹 문서는 포함구조를 지닌 태그들로 구성되어 있으며 각 태그들은 저작속성을 지닌 객체로 구성되어 있다. 다수의 객체가 실시간 재생되는 시스템에서는 동기화를 위해 종료 제한 시간 측정이 필요하다[6]. 이를 위해 그룹 시간 관계에서 각 객체의 정확한 종료제한시간을 [식 1]과 같이 계산한다.

[식 1] $\pi_k = c + \sum_{i=1}^{k-1} \gamma_\delta^i, (1 < k \le n)$

π_n : n 번째 객체의 시작 시간 상수
γ_δ^i : i 번째 객체의 시작상수에서 i+1번째 객체의 시작 상수 간의 시간간격
γtr^n : n번째 객체의 종료제한시간
$\pi_k=c$, (k=1)
c는 시간 상수이다.

증명. $\pi_1=c$, $\gamma_\delta^1=\pi_2-\pi_1$이므로, $\pi_2 = c + \gamma_\delta^1$ 이다.
임의의 수 m에 대해서 $\pi_m = c + \sum_{i=1}^{m-1} \gamma_\delta^i$ 이 성립한다면,
m + 1에 대해서도 $\pi_{m+1} = \pi_m + \gamma_\delta^m$ 이다.
위의 식으로부터

$$\pi_{m+1} = \pi_m + \gamma_\delta^m = c + \sum_{i=1}^{m-1} \gamma_\delta^i + \gamma_\delta^m = c + \sum_{i=1}^{m} \gamma_\delta^i$$

이 된다.

위의 수식을 본 연구에서 사용한 그룹 시간 관계에 적용하여 [표 1] 같이 종료 제한 시간을 구한다.

[표 1] 그룹 시간 관계에 따른 종료 제한 시간

Relation	$\gamma^i, (1 \le i < n)$	γtr^n
before	$<\gamma^i,$	$\sum_{i=1}^{n-1} \gamma_\delta^i + \gamma^n$
meets	$\gamma^i,$	$\sum_{i=1}^{n} \gamma^i, \sum_{i=1}^{n} \gamma_\delta^i$
overlaps	$<\gamma^{i+1}+\gamma^i,$	$\sum_{i=1}^{n} \gamma_\delta^i + \gamma^n$
during	$>\gamma^{i+1}+\gamma^i,$	γ^1
starts	$<\gamma^{i+1}, (\gamma^i,=0)$	γ^n
finishes	$\gamma^{i+1}+\gamma^i,$	γ^1
equals	$\gamma^{i+1}, (\gamma^i,=0)$	$\gamma^i, (1 \le i < n)$

3. 포린 페이지 구현

3.1 포린 페이지 시스템 구성

메타 브라우징 서비스를 구현하기 위해 본 시스템은 [그림 3]과 같이 컨텐츠 추출 및 모니터를 위한 중계 브라우저, 중계 웹 서버 그리고 포린 페이지의 저작 및 재생을 위한 포린 페이지 저작기, 포린 페이지 재생기, 포린 페이지 저장기로 구성한다. 중계 브라우저는 중계 웹 서버를 거쳐서 실제의 웹 사이트, 즉 원격 웹 서버에 접속할 수 있는 도구이며 웹 브라우저 기능외에 웹 페이지 상의 컨텐츠 범위를 지정하여 추출할 수 있는 기능이 있다.

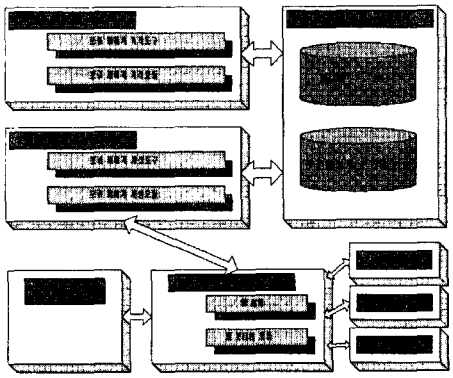

그림 3. 포린 페이지 시스템 구성도

중계 웹 서버는 중계 브라우저와 같이 작동하여 원격 웹 서버에 접속하기 위한 사용자 정보(로그인 정보, 브라우저 내부 변수)와 원격 웹 서버내의 미디어 파일, 웹 문서 파일을 저장하여 중계 브라우저로 전달한다. 컨텐츠 모니터를 위해 컨텐츠가 있는 웹 페이지에 접근하기 위한 사용자 정보를 저장하여 컨텐츠와 컨텐츠 내의 미디어 정보들을 사용자가 포린페이지 저작에 사용할 수 있다.

포린 페이지 저작기는 컨텐츠들을 이용하거나 컨텐츠를 생성하여 새로운 웹 페이지인 포린 페이지를 저작하는 기능을 한다. 포린 페이지 저작 도구는 웹 상에서 브라우저를 통해 사용자가 이용하는 도구이며 포린 페이지 저작 모듈은 포린 페이지 저작도구를 이용하여 저작한 정보를 웹 객체 모델로 변환하여 포린 페이지 저장기로 전달하는 기능을 한다. 포린 페이지 저작기는 웹 시나리오 모델[7]과 같은 방법을 사용하여 구현한다.

포린 페이지 저장기는 컨텐츠의 정보를 포린 페이지 저작모듈을 통해 웹 객체 모델 정보로 받아서 저장한다. 미디어 리소스 관리를 위한 미디어 DB를 분리하였다. 포린 페이지 저장기는 웹 시나리오 모델[7]과 같은 방법을 사용하여 구현한다.

포린 페이지 재생기는 포린 페이지 재생도구와 포린 페이지 재생모듈로 구성되어 있다. 포린 페이지 재생모듈에서는 포린 페이지 저장기 내의 정보를 웹 객체 모델로 생성하여 동적으로 웹 객체 모델의 집합인 웹 페이지를 생성하며 사용자는 브라우저로 웹 페이지 재생도구에 사용하여 포린 페이지를 볼 수 있다. 포린 페이지 재생기는 웹 시나리오 모델[7]과 같은 방법을 사용하여 구현한다.

3.2 중계 브라우저와 중계 웹 서버

중계 브라우저의 사용 화면은 [그림 4]와 같다. 중계 브라우저는 브라우저 상에서 이용할수 있도록 구현하여 URL 입력창과 입력창 하단의 탐색화면으로 나뉘어져 있다. 스크린 스크랩핑(Screen Scraping) 기술을 사용하여 원하는 영역을 마우스로 드래그(drag)하여 '가져오기' 버튼을 클릭하면 선택한 부분이 하나의 컨텐츠로 생성되어 저장된다.

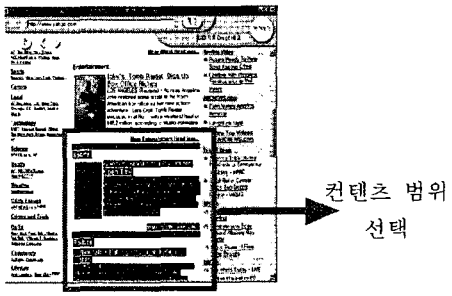

그림 4. 중계 브라우저를 이용한 탐색 화면

[그림 5]는 사용자 탐색과 추출시 내부 작업을 순차적으로 표시하였다. 컨텐츠 추출을 위해 사이트에 접속할 때 사용자 정보를 저장한다. 중계 웹 서버를 거쳐서 웹 사이트를 탐색하게 하여 중계 웹 서버에서 이를 저장하도록 하였으며, 세부적인 과정은 다음과 같다.

첫째, 사용자가 원하는 웹 사이트를 중계 웹 서버가 탐색하여 웹 문서 정규화 작업 후 중계 브라우저를 통해 제공한다.

둘째, 웹 문서 내의 컨텐츠에 공통적으로 적용되는 스타일 시트와 스크립트 언어 부분을 추출하여 컨텐츠와 함께 저장한다. 포린 페이지 저작시에 스타일 시트와 스크립트 언어 부분도 컨텐츠와 함께 추가되어 추출한 컨텐츠가 동일한 속성과 메소드를 갖는다.

셋째, 사용자의 로그인 정보와 웹 사이트 탐색에 쓰이는 내부 변수들을 저장한다. 포린 페이지 내의 컨텐츠를 자동 갱신하기 위해서 모니터 에이전트가 일정주기로 컨텐츠가 있던 웹 사이트에 접근하여 갱신 여부를 확인하는데, 로그인과 같은 절차가 필요한 웹 사이트는 미리 저장된 정보를 사용하여 접근한다.

넷째, [그림 4]와 같이 마우스로 영역을 선택하여 컨텐츠를 추출한다. 선택한 영역내의 텍스트가 포함된 부분을 찾아 독립적인 웹 객체 모델[4]이 되도록 텍스트 앞과 끝에 HTML 태그를 추가한다. 선택한 영역의 웹 객체 정보를 중계 웹 서버에 저장한다.

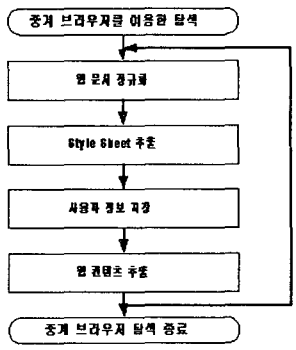

그림 5. 컨텐츠 추출과정

3.3 웹 컨텐츠 모니터링 모듈

웹 컨텐츠 모니터링 모듈은 설정된 시간에 따라 포린 페이지를 구성하는 컨텐츠가 있었던 각각의 원격지 웹 서버상의 변화를 감지하고 컨텐츠 변화시에 포린 페이지 내의 컨텐츠를 자동으로 갱신한다. 중계 웹 서버에서는 사용자가 선택한 컨텐츠가 존재하는 원격지 웹 서버의 URL과 컨텐츠에 접근하기 위해 필요한 사용자 정보와 브라우저 변수를 저장하고 있으며 웹 컨텐츠 모니터링 모듈에서는 필요에 따라 복수개의 모니터 에이전트를 작동시킨다. 모니

터 에이전트는 할당받은 컨텐츠의 정보로부터 원격지 웹 서버상의 컨텐츠와 비교하게 되며 이때 과정은 [그림 6]과 같다.

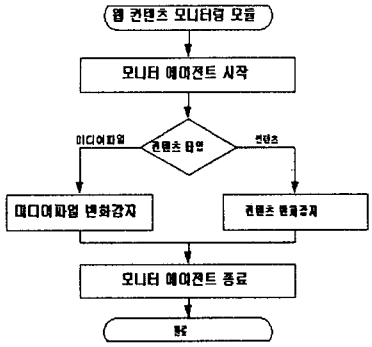

그림 6. 웹 컨텐츠 모니터링 과정

웹 컨텐츠 모니터링 모듈은 설정된 시간에 따라 복수개의 모니터링 에이전트에 순차적으로 컨텐츠를 분배한다. 에이전트는 컨텐츠의 URL과 컨텐츠 내용, 컨텐츠에 접근하기위해 필요한 사용자 인증정보와 브라우저 내부변수를 읽는다.

에이전트는 미디어파일과 텍스트와 태그를 포함하는 컨텐츠를 구분하여 변화감지를 한다. 미디어파일에 대해서는 해당 미디어파일에 접근하여 과거에 미디어파일의 CRC (Cyclic Redundancy Check)값과 현재의 CRC값을 비교한다. CRC값이 같으면 미디어파일은 변화되지 않은 것이며, CRC값이 다르거나 파일이 존재하지 않으면 변화된 것으로 감지한다.

미디어파일외의 컨텐츠에 대해서는 해당 웹 문서에 접근한 후 웹 페이지의 CRC값을 이전 CRC값과 비교한다. CRC값이 동일한 경우 컨텐츠는 변화되지 않은 것이며, CRC값이 다른 경우 웹 페이지 내에서 갱신된 이전의 컨텐츠를 찾는데, 태그 트리 구조를 사용한다.

컨텐츠의 태그 트리 구조와 갱신된 웹 문서의 태그 트리에서 루트 노드부터 선택하며 가장 유사한 서브 트리를 찾는다. 트리간의 매핑(Mapping) 방법은 [그림 7]과 같다.

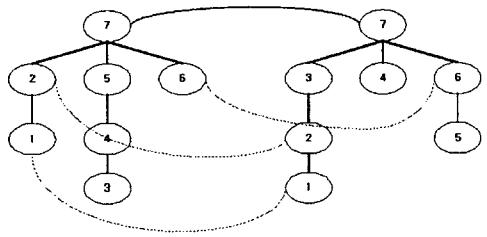

그림 7. 트리 비교도

i_k, j_k는 트리 T_1, T_2, 내의 k번째 노드로 정의한다.(postorder 순으로 노드 번호를 정하였다.)

1. 1대 1매핑 : $i_1 = i_2$ 일 때 $j_1 = j_2$ 일 때 1대 1 매핑이 성립한다.
2. 형제 노드 매핑 : i_1 이 i_2 의 좌측 형제 노드로 존재한다면 j_1 도 j_2의 좌측 형제 노드로 존재해야한다.
3. 부모 노드 매핑 : i_1 이 i_2의 부모 노드라면 j_1도 j_2의 부모노드로 존재해야 한다.

위의 매핑 방법으로 컨텐츠 트리와 웹 문서 내의 각 서브 트리들을 비교하였으며 트리간의 차이값를 구하기 위해 노드 연산 비용에 대한 정의는 다음과 같다.

a, b, c : 비교하려는 노드
γ(a->b) : 노드 a에서 노드 b로 modify 연산에 대한 비용
$\gamma(\Lambda \rightarrow a)$: 노드 a를 insert 연산에 대한 비용
$\gamma(a \rightarrow \Lambda)$: 노드 a를 delete 연산에 대한 비용
$\gamma(a \rightarrow b) \geq 0$; $\gamma(a \rightarrow a) = 0$
$\gamma(a \rightarrow b) = \gamma(b \rightarrow a)$
$\gamma(a \rightarrow c) \leq \gamma(a \rightarrow b) + \gamma(b \rightarrow c)$

위의 정의를 이용하여 트리내의 노드간 연산 비용을 구할 수 있는데 트리 간 차이값은 매핑을 이루는 각 노드에 대한 연산 비용의 합으로 정의하며 [식 2]와 같이 구할 수 있다.

트리 1과 트리 2간의 매핑을 M이라 할 때

[식 2]

$$\gamma(M) = \sum_{(i,j)\in M} \gamma(T_1[i] \rightarrow T_2[j]) + \sum_{i=I} \gamma(T_1[i] \rightarrow \Lambda) + \sum_{j=J} \gamma(\Lambda \rightarrow T_2[j])$$

식 2로부터 컨텐츠를 구성하는 서브 트리와 웹 페이지의 태그 트리내의 각 서브 트리를 비교하여 가장 유사한 컨텐츠를 포함하는 서브 트리를 구한다.

4.결론

지금까지 사용자가 웹 페이지 상에서 자신이 원하는 부분만을 추출하여 각각 사이트의 컨텐츠를 자신의 웹 페이지로 통합할 수 있는 포린페이지 시스템에 관해 살펴보았다. 기존의 브라우징 기술은 링크 기능에 의한 것이며 기존 웹 서비스는 한 사이트내에서 컨텐츠로 맞춤형 서비스를 제공하는 것이었다. 원하는 콘텐츠를 인터넷에서 끌어 모아 편집할 수 있는 메타 브라우징 기술은 현재 인터넷 기업에서 개발중이며 미국에서 개인 포탈이란 이름으로 차세대 브라우징 기능으로 인식되고 있다. 본 논문에서는 포린 페이지 재생 모델로 그룹 시간 관계를 정의하고, 동기화를 위해 그룹 시간 연산과 제한 종료 시간을 구했다. 웹 컨텐츠 추출 모듈을 설계하여 컨텐츠 태그 트리와 웹 문서 서브 트리간 차이값으로 컨텐츠 변화를 감지하는 컨텐츠 모니터 에이전트를 제시하여 포린 페이지 시스템을 구현했다. 포린 페이지 시스템은 개인 포털을 포함하는 메타 브라우징 서비스의 해법으로 앞으로 활용성이 넓을 것으로 기대한다.

[참고문헌]

[1] URL: http://www.looloo.net 루루커뮤니케이션.
[2] URL: http://www.korpage.com 코페이지.
[3] URL: http://www.openbiz2000.co.kr 오픈비즈.
[4] Frank Manola, "Technologies for a Web Object Model", January / February 1999 IEEE Internet Computing http://computer.org/internet.
[5] J.F.Allen, "Maintaining knowledge about temporal intervals," Commun.ACM, vol.26, pp.832-843, Nov. 1983
[6] Thomas D. C. Little and Arif Ghafoor, " Interval-Based Conceptual Models for Time-Dependent Multimedia Data" IEEE transactions on knowledge and data engineering, VOL 5, NO. 4, August 1993 551-563
[7] 임희경 이원석, "웹 기반 인터액티브 멀티미디어 시나리오 모델 설계 및 구현", 2000년 가을 한국멀티미디어 학회.