HTML 문서의 시각적 분석을 이용한 사용자 프로파일 생성

곽 주 현[†] · 이 창 훈[††]

요 약

본 연구는 사용자가 선호하는 웹 문서를 찾아내기 위한 웹 에이전트의 성능을 개선하는 방법을 제안하였다 웹 에이전트에서는 사용자의 선호도를 추출하기 위해서 한 문서에서 사용된 모든 단어의 중요도가 동일한 것으로 가정하는 키워드 추출 알고리즘(TFIDF 알고리즘)을 사용하고 있다. 그러나 HTML 등의 웹 문서는 일반문서와는 달리 단어의 중요성에 따라 글자의 크기 및 강조효과 등의 시각적인 차이를 둔다.본 연구에서는 둔단의 시각적인 중요도에 따라 단어의 가중치에 차등을 두는 방법을 사용하여 웹 에이전트의 성능을 향상시키도록 하였다. 또한 성능향상을 위해 하나의 문서를 여러 개의 문단으로 나누어 중요도를 차별화하도록 하는 방법을 제안하였다 또한 이를 위해 각 문단마다 프로파일을 생성하고, 이 프로파일들을 통합하는 방법을 제시하였다. 제안된 알고리즘은 기존의 알고리즘인 TFIDF와 선호하는 웹 문서를 찾는 성능을 비교하여 그 효과를 실험하였다

User Profile Generation using Visual Differences of HTML Documents

Ju-Hyun Kwak[†] · Chang-Hoon Lee[††]

ABSTRACT

In this study, I've suggested how to improve the function of web-agents to find out the web-document users prefer. Web-agents employ TFIDF, which considers all the words used in a document as equal in importance to find out users' preferences Web-documents like HTML, however, make visual differences by using different sizes of letters and highlighting them based on importance of words. In this study, I've attempted to improve the functions of the web-agents by differentiating the weight of each word in accordance with the visual importance of each paragraph. To enhance functions, I've suggested how to make a profile from each paragraph to be consolidated later. As to suggested algorithms, I've tested their effects by comparing the established TFIDF algorithm with the function which helps users find documents they prefer.

1. 서 론

웹에서 일반적으로 사용자는 검색엔진을 제공하는 사이트에 접속하여 키워드를 입력 함으로서 자신이 원하는 웹 문서를 찾는다. 그러나 이들 검색엔진들이 제공하는 문서는 사용자가 원하는 정확한 문서가 아니기 때문에 사용자는 이들 중 자신이 원하는 문서를 다시 직접 선별해야 된다. 이러한 검색엔진의 불편함을 해결하기 위해 사용자가 원하는 정보에 대한 패턴을 학습하여 더욱 정확하게 원하는 문서를 검색해 주는 프로그램이 개발되었다. 이렇게 사용자를 대신하여 원하는 문서를 검색해 주는 프로그램을 웹 에이전트라고 한다.

† 준 회 원 건국대학교 대학원 컴퓨터정보통신공학과
†† 종신회원 건국대학교 컴퓨터정보통신공학과 교수
논문접수 2000년 2월 25일, 심사완료 : 2000년 3월 24일

웹 에이전트는 사용자가 만족한 문서를 기반으로 사용자의 선호도를 정량적으로 기록한 후 사용자가 필요로 할 가능성이 높은 문서를 미리 검색하여 사용자에게 제시한다. 또한 사용자의 평가를 학습하여 사용자의 선호도에 대한 기록을 갱신한다.이때 이러한 사용자의 선호도에 대한 기록은 에이전트가 검색하는 문서의 정확도에 영향을 주게 되는 매우 중요한 과정이다

WebSailor, WebAce, SiteSeer등이 웹 에이전트의 예들 이며, 이들이 사용자의 선호도를 학습 할 때 일반적으로 사용하는 방법은 TFIDF(Term Frequency Inverted Document Frequency)이다. TFIDF는 사용자가 선호하는 문서들 중에 비교적 높은 빈도로 출현했던 단어를 이용하여 사용자의 취향을 반영하게 된다. 이때 지나치게 일반적인 단어가 선정되는 것을 막기 위해 많은 문서에 무분별하게 출현하는 단어는 배재 시키도록 하고 있다. TFIDF에 사용되는 단어는 모두 균등한 중요성을 부여받게 되므로서 단어의 출현횟수에 의해서만 단어의 가중치가 결정된다.

그러나, 웹 문서가 글자의 크기나 모양이 서로 다른 시각적인 측면을 가지고 있음을 고려한다면, 웹 문서 분석에 단어의 출현횟수만을 고려하는 TFIDF 방식만을 적용 한다면 웹 문서의 시각적인 측면은 무시되게 된다. 그러므로 웹 문서를 효율적으로 분석할 수 있는 새로운 방법이 필요하다.

웹 문서에서 중요한 내용이나 그 문서를 대표할 수 있는 단어의 경우 일반적으로 큰 글꼴이나 굵은 형태의 글꼴로서 강조되고 있다. 그러므로 이러한 정보를 이용하면 보다 더 정확하게 그 문서를 대표하는 단어를 추출할 수 있다. 또 웹 문서에 등장한 각 문단을 중심으로 시각적 특징을 계산하고 이를 기반으로 문단의 관계를 정의하여 문서를 분석할 수 있다 이러한 문단간의 관계는 문단의 중요성을 계산하는 데에 있어 상대적인 가중치의 차이를 부여하게 한다. 본 연구는 웹 문서 분석에서 내용만을 고려하는 기존의 TFIDF 방법과는 달리, 웹 문서의 시각적 정보도 고려하는 분석 방법에 대한 연구이다.

2. HTML 문서의 시각적 분석

웹 문서에서 가장 널리 사용되는 HTML은 HYPER TEXT 형식의 문서로서 일반적인 문서에 비해 사용자에게 보다 더 많은 표현의 영역을 제공한다 이를 통해 검색자가 원하는 내용을 보다 더 효율적이고 알아보기 쉽게 표현할 수가 있다. 예를 들어 큰 글씨로 또는 강조된 형식의 글씨로 표현할 경우 이를 읽는 사람은 그 부분을 더 강하게 인식하게 된다.

그러므로 같은 문장이라도 테그를 분석 함으로서 검색자에 대한 영향력을 분석이 가능하다. 만일 이 문서의 제작자가 이 문서를 검색자가 보다 쉽게 이해하도록 이러한 테그를 사용했을 경우 검색자에게 시각적 영향은 이 문서의 내용상의 중요도와 비례하게 된다. 시각적으로 강한 영향력을 미치는 태그는 다음과 같다.

〈표 1〉 시각적인 강조에 영향을 미치는 태그

태 그	효 과
<H1>~<H6>	글씨의 크기에 관계되며 함께 강조효과도 추가된다.
<font size = ?>	글씨의 크기를 결정하며 ?부분의 수치에 따라 크기가 결정됨
<B>	글씨를 굵게 한다. 단 <H?>의 테그가 쓰일 경우 자동적으로 이 효과가 추가된다.

이는 문서의 각 부분들의 검색자에게 미치는 시각적 효과를 고려할 경우 이 부분이 전체 문서에서 가지는 역할과 영향력이 분석이 가능함을 의미한다 이러한 분석은 단어의 출현 횟수 외에도 그 단어를 표현하기 위해서 사용한 여러 태그등에 의해서 단어의 중요도가 차별화 됨을 의미한다.

본 연구에선 HTML 문서를 이러한 부분들로 분할 후 각 부분간의 관계를 정의하여 HTML 문서를 트리 형태로 구조화 시키는 방법과 이를 통해 개선된 프로파일의 생성방안을 제시한다. 또한 이를 현재 사용되고 있는 HTML 문서에 실험함으로서 기존의 TFIDF 방법에 의해 생성된 프로파일과의 비교를 수행하였다.

2.1 문서의 분할

이러한 구조적 분석을 위해서 가장 처음에 해야 하는 일은 문서를 각 부분별로 구분하는 것이다. 이는 원칙적으로 의미상으로 연결되어 있는 부분 별로 구분하는 것이 이상적이다. 그러나 실제로 현재의 기술로는 정확한 의미파악이 어렵다. 이를 해결하기 위해 본 연구에선 작성자가 의미적 구분을 기준으로 문서를 시각적으로 분할함을 가정함으로서 시각적인 분할을 기준으로 문서의 부분을 분할하였다.

이때 시각적으로 연결되어 쓰여진 문장을 '시각적 문단'으로서 규정하며 이를 기준으로 문서를 분할한다.

시각적 문단을 구분하기 위해서 가장 쉽게 사용할 수 있는 방법이 <P> 태그에 의한 구분이다. <P> 태그는 주로 한 문단을 묶어서 표현하기 위해서 사용되는 태그로서 이 태그 안의 문장은 하나의 문단으로 처리되어서 표현되며 다른 부분과 확연하게 시각적으로 구분된다.

그러나 시각적으로 연결된 모든 문장이 <P> 태그에 의해서 표현되는 것은 아니다. 그러므로 본 연구에선 '시각적 문단'을 다음과 같이 정의한다.

- **시각적 문단**. 하나의 시각적 문단은 줄바꿈 없이 이어서 쓰여진 부분으로서 하이퍼 링크를(A태그)를 제외한 다른 HTML 객체(그림, 표, 스크립트)등을 포함하지 못하며 위의 객체에 의해 시각적으로 구분될 경우 각각 다른 문단으로 간주한다.

2.2 시각적 문단의 가시도

HTML 문서를 각각의 부분(시각적 문단)으로 분할 후에는 이들간의 시각적 연결관계를 분석한다. 이러한 시각적 연결 관계는 사람의 눈에 비치는 문서상에서의 각 부분간의 연결성을 의미한다. 예를 들어 어떤 문서상에 대 제목, 중 제목, 본문 의 형태로 구성되어 있다고 가정하자. 본 연구에서는 이러한 문서에서 각 문단이 시각적으로 영향을 미치는 정도를 그 문단의 '가시도'로 정의하였다. 일반적으로 대 제목은 높은 가시도를 지니도록 표현된다 이는 글씨의 크기나 굵기등에 의한 강조가 많이 쓰이곤 한다. 중 제목은 그보다 더 낮은 가시도를 지니도록 표현되나 본문보다는 높은 가시도를 지닌다 본문의 경우 가장 낮은 가시도를 지니게 된다.

그러므로 이러한 시각적 연결관계를 규정짓기 위해서는 각 문단의 가시도에 대한 고려가 필요하다. 일반적으로 문단의 가시도는 글꼴의 크기와 강조여부에 비례한다. 본 연구에서는 이러한 각 문단의 가시도를 다음과 같이 정의하였다.

$$\text{문단의 가시도}(Ve) = \text{font size} + \text{bold effect} \qquad (1)$$

font size는 HTML 문서에서 글씨의 크기를 정의하는 tag인 <font>에서 사용되는 글자의 크기의 단위를 기준으로 하였다. HTML 문서에서 표현할 수 있는 font의 크기는 1에서 7까지이다. Bold effect는 글자의 강조(bold) 여부를 의미하며 강조된 글자의 경우 1 아니면 0의 값을 지닌다.

<font> 이외의 글자의 크기 지정방법인 <H1>~<H6>의 경우 이에 대응되는 글꼴의 크기에 강조가 된 효과가 나타난다. 이러한 태그에 소속된 문단의 가시도는 다음과 표와 같다.

〈표 2〉 H1~H6의 가시도

태 그	가시도
H1	7
H2	6
H3	5
H4	4
H5	3
H6	2

3. HTML의 구조화 방법

전 단계에서 시각적 문단으로 분류된 문서는 시각적인 차이를 이용해 가중치의 차등을 준다. 이때 유의할 점은 이러한 시각적인 차이는 상대적이라는 것이다.

어떤 두 문서중 한 문서는 소제목을 위해서 글씨의 크기를 9로 설정하고 본문의 내용은 5정도로 큰 차이를 두고 나머지 한 문서는 8과 6정도의 작은 차이를 둔다고 가정하자. 이때 문서의 시각적 강조도를 절대적으로 평가한다면 9와 5로 둔 문서의 경우 본문에 비해서 제목이 매우 높은 강조를 얻게 된다. 그러나 8과 6의 경우는 거의 차등이 없게 된다.

그러나 이러한 표현은 문서의 제작자의 작성 스타일에 따르는 경우가 많다. 그러나 이들 두 문단이 본문과 소제목의 관계임에는 변함이 없다. 그러므로 본 연구에선 이 둘의 문서에서의 영향력의 차이를 이용해 두 문단이 종속적임을 보임으로서 이를 반영하고자 한다. HTML 구조화 과정에서는 이러한 종속성을 표현하기 위해 각 문단들의 관계를 트리형태로 구성한다. 이는 트리 구조상의 계층성을 이용하여 프로파일을 하부구조부터 통합해가면서 자연스럽게 그 중요도가 반영되게 하기 위한 것이다.

3.1 시각적 문단사이의 관계

가시도가 규정된 경우 각 시각적 문단과 문단의 사이에 관계를 판단한다. 이때 두 인접한 문단의 관계는 다음의 세가지로 나뉜다 그 문단의 각각의 관계에 따라 종속, 동등, 역관계로 규정하였는데 종속관계는 선

행 문단의 가시도가 후행문단보다 큰 경우이며 동등관계 두 문단의 가시도가 동일할 경우 두 문단은 유사한 수준의 시각적 영향력을 행사함을 의미한다. 역관계는 하행 문단이 상행문단보다 가시도가 높을 경우는 종속관계와는 다르다. 이는 어떤 문단을 대표하는 문단은 그 앞에 위치함을 고려해야 하기 때문이다. 이 경우는 상행문단보다 한 수준 높이 있는 문단과의 관계를 구한다.

3.2 HTML 트리 구조의 생성

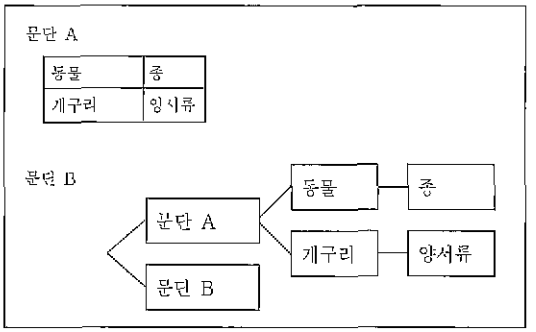
위의 각 문단의 관계를 이용하여 종속관계의 두 시각적 문단을 부모 노드(parent node)와 자식 노드(child node)로 간주하고 동등관계를 형제 노드로서 간주할 경우 각각의 시각적 문단들을 노드로 하는 트리 형태의 표현이 가능하다. 역관계의 경우 종속 및 동등관계를 발견할 때까지 상행문단의 부모 노드를 거슬러 비교해간다. 아래 그림에는 이러한 관계를 이용해 트리를 구성하는 알고리즘을 제시하였다.

```
algorithm   HTML의 시각적 구조트리 생성

최 상위 노드(root node) 생성,
loop  모든 문단에 대해서
  if ( 처음 문단일 경우 )
      첫 문단으로 노드를 생성하고 최 상위 노드의 자식노드로 연결,
  end of if
  else
        비교노드 = 상행문단의 노드;
        while( 비교노드와의 관계 = 역관계)      /* 최상위 노드에 대해서는
              비교노드 = 비교노드의 부모노드,   무조건 종속관계가 성립 */
        end of while
        if 비교노드와의 관계 = 종속관계
              현재 문단을 비교노드의 자식 노드로 인길;
        end of if
        if 비교노드와의 관계 = 동등관계
              현재 문단을 비교노드의 부모노드의 자식노드로 연결,
        end of if
  end of else
end of loop
만일 최상위 노드의 자식노드가 하나라면 최상위 노드를 제거하고 자식노드를
최상위 노드로 등록

end of algorithm
```

(그림 1) HTML 구조화 알고리즘

3.3 HTML 내의 테이블의 처리

HTML 내부에서 테이블은 일반 문서에서의 표와는 달리 문단의 횡 배치를 위해서 사용되기도 한다. 그러므로 테이블의 한칸은 하나의 문단과 같이 생각해서 처리한다 또한 하나의 행에 존재하는 칸들은 모두 종속적으로 연결된다 또한 테이블은 어떠한 경우에도 상위 문단에 대해 종속적이며 테이블의 다음 문단은 테이블과의 비교대신 테이블 상위의 문단과의 비교를 수행한다. 아래 그림에는 테이블처리의 예가 나와있다.



(그림 2) 테이블의 문단화

위의 예제는 테이블의 종속관계를 가로로 구성함을 보인다. 테이블의 우상단의 '종'이라는 단어는 좌상단의 '동물'에게 종속되게 된다. 또한 위의 예제는 실제로는 동물 - 개구리, 종 - 양서류의 관계가 이루어지는 것이 적합하나 자연이 사전을 사용하지 않음으로서 잘못된 관계가 생성될 수도 있음을 보여준다. 이러한 테이블의 처리는 더 많은 연구가 필요함을 알 수 있다.

4. 전체 프로파일의 생성

분할되어 구조화된 HTML문서는 '시각적 문단'을 하나의 노드로 하는 트리형태로 표현된다. 이는 문서에서 높은 가시성을 지닌 노드가 상위노드에 위치하는 형태로 이루어져 있으며 문서에서의 문단의 배치를 고려하여 순서가 배치된다. 이때 각각의 노드(문단)마다 프로파일을 생성 함으로서 트리의 형태로 구조화된 HTML문서의 효율적인 분석이 가능하다.

4.1 문단 단위의 프르파일 생성

모든 단어에 대해 가중치를 갖는 프로파일을 구성하려면 많은 계산량이 요구된다. 또한 단어의 비교를 위해서는 단어의 복수형이나 소유격, 동사의 변화등을 고려하여야 한다. 기존의 정보 검색 즉, TFIDF에서는 이러한 문제의 해결을 위해서 주로 Stop List와 Stemmer 알고리즘을 사용한다. 본 연구에서 역시 위의 두 알고리즘을 이용하여 불필요한 단어를 제거하고 이근

만을 추출한 문단단위의 프로파일을 생성한다. Stemmer 알고리즘은 접두사 또는 어미 등 어근에 붙어있는 부분을 제거하는 알고리즘으로서 단어를 어근으로 분리하므로 같은 단어이지만 접미사나 어미의 변화에 의해 다른 단어로 인식되는 것을 막을 수 있다. 또한 stop list 제거 알고리즘은 문서에서 많이 나오는 일반적인 단어는 키워드로 사용하는데 부적절하므로 주요 키워드 후보에서 제거하는 알고리즘이다.

Welcome to my music page. Here you will find info and samples of various kinds of music. Make this your start page. turn on some music and then surf the web! Note : Midi's are best heard with a 16-bit or higher sound card and WinGroove installed drivers

Mus	1
Pag	2
Find	1
Info	1
Vari	1
Web	1
Midi	2
WinGroove	1
install	1
Driv	1

(그림 3) 소단위 프로파일 생성

4.2 전체 프로파일의 생성

전체 프로파일의 통합은 아래서부터 자식 노드의 프로파일을 고려하여 각각의 프로파일을 수정해 감으로서 수행된다 이러한 수정이 최 상위 노드(Root)의 프로파일에 대해서도 수행된 경우 수정된 Root의 프로파일을 그 문서를 대표하는 프로파일로 간주한다. 역으로 하부 문단의 경우 상위 프로파일에 영향을 주는 형태로서 그 문서에 대해 자신의 특징을 반영한다.

이때 각 문단에서의 단어의 발생 빈도는 그 문서에 영향을 주되 한 문단에 집중적으로 나온 횟수보다 그 문서에 골고루 퍼진 단어에 대해 더 높은 영향력을 가정하였다. 그러므로 하나의 단어 t에 대한 하위 n번째 프로파일 영향력은 다음과 같은 식으로 정의하였다.

$$ACpt = ma(1 - ca^{CWpt}) \quad (2)$$

ACpt(affection by term of child profile) : 하위 프로파일의 영향력

ma(maximum affection) : 최대 영향치 상수

cr(convergency rate) : 수렴 속도 상수

CWpt(weight of term in child profile) : 하위 프로파일 p에서의 단어 t의 가중치

영향력을 행사할 하위 프로파일의 개수가 여러개라고 할 경우 n개의 하위 프로파일에 대한 영향력의 총합은 각각의 영향력의 총합으로 계산하였다.

$$ACt = \sum_{p=1}^{n} ACpt \quad (3)$$

ACt(affection by term of child profile) : 총 하위 프로파일들의 영향력

*ma*는 최대 영향치를 의미한다 즉 빈도수가 높을수록 최대 영향치에 수렴하게 된다. *cr*은 수렴 속도로서 0부터 1사이의 수치를 지닌다. 수치가 낮을수록 빠르게 최대 영향치에 수렴한다. 만일 이 속도가 0.5일 경우 한번 나온 단어는 영향치의 50% 두번 나온 단어는 영향치의 75%를 갖게 된다. 본 연구에서는 0.7정도의 수렴속도를 지니도록 하였다. 이는 빈돗수가 증가함에 따라 최대치의 30%, 51%, 66%, …의 비율로 증가한다.

하위 프로 파일의 영향력(*ACt*)가 계산되면 이 영향력을 상위 프로파일에 더함으로서 상위 프로파일을 변경한다.

이 식은 결과적으로 한 문단에 단어가 여러 번 나오는 것보다 2~3문단에 걸쳐서 등장하는 단어가 더 높은 가중치를 얻게되는 효과가 있다.

5. 실 험

위의 알고리즘의 성능을 테스트 하기 위한 실험은 실제 문서에 대한 그 효용성을 검증하기 위해 Yahoo 홈페이지에서 분류된 문서를 기준으로 수행되었다. 이 문서들은 3월 3일부터 3월 10일 사이에 문학, 공원, 3차원 그래픽이라는 3가지 카테고리에서 순서대로 수집되었다.

5.1 문서에 있어서 글자의 유형이 미치는 효과

알고리즘의 실험을 수행하기 이전에 수집된 데이터에 대한 분석이 먼저 수행되었다. 이는 실제 사용되고 있는 문서들 중에서 시각적인 글자의 수식이 얼마나 널리 사용되고 있으며 중요한 부분이 이러한 효과에 의해서 강조되고 있는가의 여부를 검사함으로서 본 알고리즘이 제안하는 문자의 시각적 유형의 이용에 대한

정당성에 대한 실험이었다.

〈표 3〉 시각적 표현의 사용비율

카테고리	하나이상의 글자 유형을 가진 문서	실제 다른 글자에 비해 강조된 형태의 제목이 존재하는 문서	평균적인 글자의 유형수
문학	50	41	3.38
공원	50	45	4.11
3차원 그래픽	49	23	5.21

위의 실험결과 실제 거의 모든 HTML 문서는 하나 이상의 글자의 유형을 가지고 있으며 또한 약 70% 가량의 문서가 이러한 유형을 통해서 제목을 구별하는 것이 가능함을 나타내고 있다. 위의 결과는 분야에 대한 편차가 있는데 3차원 그래픽에 관련된 컴퓨터 관련 홈페이지들은 다른 분야에 비해 제목을 그림파일을 이용해 표시한 경우가 많았기 때문이었다. 이를 개선하기 위해 이미지를 문자인식 하는 방법도 고려하였으나 이를 위해서는 에이전트가 이미지 파일을 받아서 처리해야 하므로 텍스트 데이터에 비해 많은 네트워크 상의 부하가 발생하기 때문에 제외하였다.

.5.2 대/소 제목의 판별능력

칫 실험은 본 알고리즘의 구조화 부분에 대한 집중적인 실험을 수행하였다. 이 실험의 목적은 더 강조된 내용(인위적인 판단에 의한 대제목/소제목 등)이 생성된 HTML 구조에서 어느 정도 상위에 위치하는 가에 관한 실험이 수행되었다. 이 실험은 위의 문서들 중에서 적절한 전체 제목 및 부분 제목을 포함한 문서 30개씩에 대해서 수행되었다. 우선 대제목이 최상위 노드에 위치하는 지의 여부를 검사하였으며 부분 제목과 그 부분제목이 포괄하는 내용이 그 부-자 노드로 연결되는 가를 검사하였다

〈표 4〉 제목의 종속적 연결 실험

카테고리	최상위 노드에 위치된 전체 제목/ 전체 제목의 총 수	적절히 분석된 부분제목/ 총 부분제목의 수
문학	24/30	70/74
공원	28/30	62/67
3차원 그래픽	22/30	41/46

5.3 상위/하위 구조의 중요도 측정

실제 부분 프로파일의 생성과 통합과정을 위해 최대 영향치 상수와 수렴 속도 상수의 적절한 값이 요구되었다 이는 구조화된 샘플상에서 하위 단어와 상위 단어의 중요도를 분석함으로서 수행되었다

이를 위해서 소 제목에 매우 분명한 하나의 주제어를 판명함과 동시에 본문부분과의 연결성을 가지고 있는 부분을 추출하였다 3가지의 분야에 대해 각각 15씩을 선정하였다. 이때 소 제목은 상위, 본문은 하위노드가 된다. 이때 하위노드에서 평균적인 주제 단어의 가중치를 계산하였다.

〈표 5〉 제목 및 본문의 주요 단어 포함 비율

카테고리	제목부분의 주제어 등장 평균 횟수	본문 부분의 주제어 등장 평균 횟수	본문부분 등장/ 제목부분 등장의 평균 치
문학	1.13	3.7	3.45
공원	1.07	2.56	2.48
3차원 그래픽	1.0	3.43	3.43
평균	1.07	3.23	3.12

위의 데이터는 평균적으로 소제목과 본문상에서 중요한 주제어가 등장하는 비율이 대략 3정도임을 보여준다.

다음의 실험은 문학 관련 문서 15개에 대해서(문학 관련 문서는 주제가 명확한 종류의 문서가 많은 관계로 이를 선정하였다)에서 가장 중요한 키워드를 2~3개 이내에서 추린 후에 각 구조의 상/하위별로 이들의 분포비율을 측정하였다.

〈표 6〉 레벨별 주제어 등장횟수

레벨	주제어 등장 횟수	주제어 등장횟수/전체 단어수 평균
1	1.32	0.143
2	3.21	0.044
3	4.38	0.012

이를 보면 상위 노드에 단어가 주제어일 비율 하위 노드에 비해 3배가까이 높음을 알 수 있다. 특히 1레벨(대제목 레벨)일 경우 유난히 높으며 2~3레벨의 사이에는 약 2배가량의 차이가 있음을 알 수 있었다. 그러므로 본 실험시 하위 레벨의 영향력 상수를 0.33으로 가정하였다. 수렴 속도상수는 임의적으로 0.5정도로 가정하여 실험이 수행되었다.

5.4 생성된 프로파일의 성능

마지막으로 위의 구조 분석에 의해 생성된 통합 프로파일과 기존의 TFIDF에 의해서 생성된 프로파일의 성능을 비교하였다. 대상 문서는 전체의 문서중 에서 위의 2번째 실험에 사용된 30개의 문서 중 적절히 구조적인 분석이 이루어진 15문서에 대해서 수행되었다. 평가의 기준은 title tag를 가진 문서들을 대상으로 생성된 프로

파일에서 상위 5위안에 들어가는 단어들 중 title tag에 소속된 단이의 수를 비교함으로서 평가하였다.

〈표 7〉 TF/IDF와의 비교

카테고리	구조적 분석에 의한 프로파일의 적중률	TF/IDF를 이용한 프로파일의 적중률
문학	35/75	21/75
공원	40/75	29/75
3차원 그레픽	38/75	31/75

위의 실험결과에서 문서의 글꼴 및 구조를 분석하는 프로파일 생성방식은 그 방법을 사용하지 않은 TFIDF에 비해 비교적 높은 성능을 보임을 알 수 있다. 특히 문학쪽에 관련된 문서에 대해서는 특히 높은 비율의 성능 향상이 있었다. 이는 위의 5.1부터 5.3까지의 실험 및 분석결과에서 알 수 있듯이 시각적인 효과의 사용정도와 그 문서가 얼마나 구조적인 형태를 취하는가가 본 알고리즘의 성능에 영향을 줄 수 있음을 의미한다.

6. 결 론

본 알고리즘은 HTML문서가 일정한 형식에 의해 글자의 유형을 사용한다는 가정하에 이를 이용하여 중요한 부분을 구분하고 이를 고려하여 프로파일 내에서의 중요 단어의 가중치를 향상시키는 방법에 관한 것이다. 이를 위해서 우선 유형분석의 기준이 되는 시각적 문단을 정의하였다. 그 후에 시각적 문단에 대해 글자의 유형을 분석함으로서 이 문단의 중요도를 계산하여 시각적 문단의 배열을 트리 구조로 변환한다. 시각적 문단에 대해서 각각의 프로파일이 생성되며 이는 트리 구조를 통해 통합된다. 최상위 노드에서 통합된 프로파일은 이 문서를 대표하는 프로파일로서 간주된다.

이러한 기법은 실험의 결과 본 논문과 유사한 대제목-중제목-소제목의 형식을 가진 문서에 대해서 효율적인 프로파일 생성을 보여주었다. 그러나, 본 논문처럼 완전히 형식화된 문서가 아닌 경우라도 부분적으로는 이러한 관계가 존재하였으며 이 경우에 대해서도 부분적인 향상이 나타났다.

참 고 문 헌

[1] Ju-hyun kwak, chang-hoon lee, "Advanced User Profile Agent Using Structure Analysis of HTML Document," 책임연구, Proc pp.319-323 IC-AI'99, CSREA Press, 1999.

[2] 박영식, 곽주현, 이창훈, "프로파일 생성을 위한 TFIDF 개선방안 연구", 정보처리학회 추계 학술대회, 1998.

[3] Gerard Salton and Michael J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983.

[4] Salton, G., and Buckley, "Term weighting approaches in automatic text retrieval," Technical Report 87-881, Cornell University. Department of Computer Science 1987.

[5] Gerard Salton and Chris Buckly, "Improving Retrieval Performance by Relevance Feedback," 1990.

[6] William B. Fakes, Ricardo Baeza-Yates, "Information Retrieval. Data Structures & Algorithms." ch.7, ch.8. Prentice-Hall, 1992

[7] Marko Balabanovic and Yoav Shoham, "Learning Information Retrieval Agents : Experiments with Automated Web Browsing," AAAI Spring Symposium on Information Gathering, Stanford.

곽 주 현

e-mail : jhkwak@cse.konkuk.ac.kr

1995년 건국대학교 컴퓨터공학과 졸업(학사)

1997년 건국대학교 대학원 컴퓨터공학과 졸업(석사)

1997년 건국대학교 대학원 컴퓨터 공학과 입학(박사과정)

관심분야 : agent, 정보검색

이 창 훈

e-mail : chlee@kkucc.konkuk.ac.kr

1975년 연세대학교 수학과 졸업(학사)

1977년 한국과학기술원 전산과학과 졸업(석사)

1993년 한국과학기술원 전산과학과 졸업(박사과정)

관심분야 : agent, linux, 전자상거래