Proceedings of KFIS Fall Conference, 2003. 5. 3

퍼지 이론을 적용한 웹 페이지 화면추출

Screen Extraction of Web Page Using Fuzzy Theory

김성택 · 손창식 · 정환묵
대구가톨릭대학교 컴퓨터정보통신공학부
Sung-Teak Kim, Chang-Sik Son, Hwan-Mook Chung
Faculty of Computer and Information Communication Engineering,
Catholic University of Daegu
E-mail : zergprotoss@hotmail.com

요 약

본 논문에서는 사용자에게 보다 적합하고 유용한 정보를 제공할 수 있도록 퍼지 α-cut을 이용하여 웹 페이지를 부분적으로 표시할 수 있는 방법을 제안한다. 또한, 사용자의 취향과 패턴을 분석하여, 블록화 되어있는 태그들을 중요도에 따라 나타낼 수 있도록 함으로서 사용자의 정보검색에 소요되는 시간과 비용을 절감할 수 있을 것으로 기대된다.

1. 서론

정보통신 기술의 발달과 인터넷 인프라의 급속한 발전으로 인하여 인터넷이 현대사회에 있어서 정보의 전달 및 취득에 있어서 중요한 정보원으로 발돋움하고 있다.

인터넷 서비스를 제공하는 웹사이트에 있어서 표현되는 웹 페이지의 경우 일반 데스크탑용으로 작성되어서 사용자에게 전달되며, 이때 사용자에게 보여지는 웹 페이지는 사용자가 사용하는 컴퓨터의 종류나 화면의 크기, 또한 웹 페이지가 유용한 정보로 구성되어 있는가? 에 대해서 확신 없이 모든 사용자에게 동일한 단 방향성 정보를 보여 주고 있는 실정이다.

이때 사용자는 표시되는 웹 페이지에서 유용한 정보를 얻기까지 많은 시간이 들어가고, 상대적으로 불필요한 부분까지 표시되어 목표로 한 정보를 얻을 때까지 많은 노력을 요한다. 이런 문제점을 제거하기 위해서는 사용자가 자주 이용하는 정보, 즉 볼 가능성이 높은 정보를 중심으로 표시하고, 사용자의 요구가 적은 정보는 생략하는 것이 적합하다.

웹 페이지에는 수많은 하이퍼링크들과 하이퍼링크 된 이미지들이 존재한다. 이런 하이퍼링크 된 텍스트와 이미지를 가지는 태그(Tag)에 대한 정보들의 클릭빈도를 분석하고, 퍼지 α-cut을 이용하여 태그에 클릭빈도의 상관관계를 파악하여 웹 페이지를 부분표시 할 수 있는 방법을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 일반적인 퍼지 α-cut을 설명하고, 3장에서는 웹 페이지의 HTML을 태그로 분할하고 속성값 계산 방법을 설명한다.

2. 퍼지 α-cut

퍼지집합에 포함된 원소들 중에서 일정한 가능성 (소속함수 값) 이상 포함된 원소들로만 구성된 보통집합을 만들 수 있다. 이것을 α-수준(α-cut)집합이라고 부르는데, 소속함수의 값이 α 이상인 원소들로 이루어진다.

$$A_\alpha = \{ x \in X \mid \mu_A(x) \geq \alpha \}$$

이때 α는 임의로 선택할 수 있다. 이와 같이 만들어진 α-수준집합은 보통집합(crisp set)이 된다[1].

3. 웹 페이지의 부분 표시

3.1 대상 Tag의 선택 및 HTML 분할

해당 웹 페이지에 대하여 대상이 되는 태그를 선택하고 그 태그들을 분할한다. 하이퍼링크가 되어 있는 대상 태그들에 대하여 사용자가 관심이 있고, 유용한 정보라고 생각된다면, 링크 된 부분을 클릭 하여 웹 페이지의 정보를 얻을 것이고, 관심이 없고 불필요하다고 생각된다면, 단지 읽기만 한다던가, 지나쳐 버릴 것이다.

본 논문에서 사용하는 분할 대상은 <A> 태그로 한다.

<A> 태그의 클릭 수를 측정하기 위해서 하이퍼링크 된 텍스트와 이미지에 대하여 카운터를 부여한다. 카운터는 데이터베이스를 이용하는 방법과 웹 로그 분석을 통하여 계산하는 방법이 있는데, 여기서는 웹 로그 분석을 통하여 카운터를 계산하였다[2].

카운터를 가지는 태그는 다음과 같다.

<A Href ~> **Text** </A> <A Href ~> **<img src=~>** </A>

표 3.1 카운터를 가지는 태그

<A href="yahoo.co.kr" target=blank> **야후로 가기**</A> <A href="naver.co.kr" target=blank> **네이버로 가기**</A> <A href="joins.com" target=blank> **중앙일보 가기**</A> <A href="6.htm" target=blank> **<IMG src="noname3.files/tgedu.jpg" Width="130">** </A>

표 3.2 카운터를 부여할 태그들의 예

3.2 속성값 계산

각 태그들에 대하여 클릭 수를 바탕으로 속성값을 다음 식으로 계산한다.

속성값=각 태그의 카운터수 / 카운터수 총합 ..... (1)

위의 표 3.2에 대한 카운터를 구한 결과는 아래 표 3.3과 같다.

이를 각각의 대상 태그에 적용시키면 아래와 같은 속성값을 얻을 수 있다.

블록 번호1 10/24 = 0.4166
블록 번호2 6/24 = 0.25
블록 번호3 4/24 = 0.16
블록 번호4 4/24 = 0.16

위의 속성값을 적용시키면 다음과 같이 나타낼 수 있다.

블록번호	대상태그	속성값
1	<A href="yahoo.co.kr" target=blank> 야후로 가기 </A>	10/24
2	<A href="naver.co.kr" target=blank> 네이버로 가기 </A>	6/24
3	<A href="joins.com" target=_blank> 중앙일보 가기</A>	4/24
4	<A href="6.htm" target=blank> <IMG src="noname3.files/ tgedu.jpg" width=130> </A>	4/24

표 3.3 속성값이 부여된 블록번호

위에서 계산된 속성값을 바탕으로 "반드시 필요한 부분"을 $\alpha_{0.2}$ 로 잡는다면 블록번호는 1,2만 포함하게 되며 퍼지 α-cut 값을 α_0으로 잡는다면 블록번호는 1, 2, 3, 4를 포함한다.

즉, 블록번호 1, 2번의 태그는 "반드시 필요한 부분"에 포함되며 이에 비해 α_0는 웹 페이지를 사용자의 선호도와 취향에 관계없이 $\alpha_{0.2}$보다 상대적으로 불필요한 부분을 포함한 모든 부분

을 보여주게 된다.

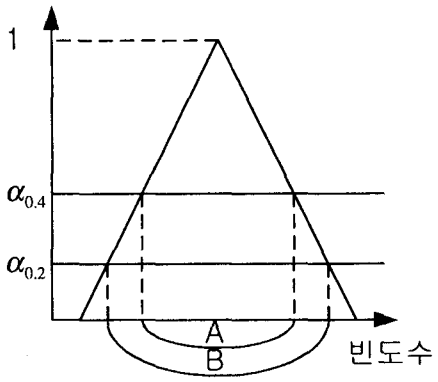


그림 3.1 속성값을 부여한 α-cut

위 그림에서 A구간은 속성값이 가장 높은 블록번호 1을 의미하며, B구간은 속성값이 그 다음으로 높은 블록번호 1, 2를 의미한다.

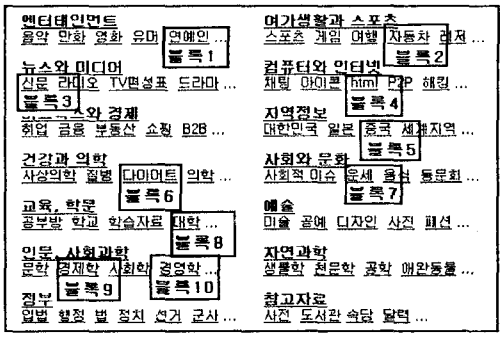
4. 시뮬레이션

① 원 화면

아래 그림에서 보는 것과 같이 웹 페이지는 수많은 하이퍼링크 된 텍스트들과 이미지들로 구성되어 있다.

② 대상 태그의 선정

대상되는 태그로는 아래 그림과 같이 임의로 10가지 태그를 선정하였다.



③ HTML 분할

위에서 선택한 10개의 대상태그를 분할하여 블록번호를 부여하였다.

블록번호	대상 태그
1	<A href=/Entertainment/Entertainers/ entertainers.htm"> 연예인</A>
2	<A href=/Recreation/Automotive/ car.htm> 자동차 </A>
3	<A href=/News_and_Media/ Newspapers/newspaper.htm>신문 </A>
4	<A href=/Computers_and_Internet/ Internet/World_Wide_Web/page_creation /HTML/html.htm> html </A>
5	<A href=/Business_and_Economy/ Shopping_and_Services/ Shopping_Mall/Shopping_Mall.htm>쇼핑
6	<A href=/Regional/Countries/China/> 중국
7	<A href=/Health/Weight_Issues/Weight_Loss /Weight_Loss.htm>다이어트
8	<A href=/Education/Higher_Education/ Education.htm> 대학
9	<A href=/Humanities__Social_Science/ Economics/Economics.htm>경제학
10	<A href=/Humanities__Social_Science/ Management_Science/ Management_Science.htm>경영학

④ 카운터 분석을 위한 코드 삽입

```
<script language='javascript'>
var U_URL = 'http://'+'gtcc1.ace카운터.com' ; var
A_CODE = 'AM2A2912954038' ;
var ACE_dim='',ACE_tz = 20 , ACE_ja = 'Unknown'
,ACE_bn = 'Unknown' ,ACE_bv = 'Unknown' ,ACE_sc
= 'Unknown' ,ACE_sv = 10 ,ACE_ref = '',ACE_arg= '',
ACE_av = '', ACE_ck = 'n', ACE_je = 'n',ACE_ul =
'Unknown',ACE_ua = 'Unknown',ACE_UA =
'Unknown',ACE_url = '',ACE_os= 'Unknown'
ACE_tz = Math.floor((new
Date()).getTimezoneOffset()/60) + 29 ;
if( ACE_tz > 24 ) ACE_tz = ACE_tz - 24 ; ACE_je =
ACE_ck = (navigator.cookieEnabled==true)?'1':'0';
ACE_bn=navigator.appName;
```

⑤ 사용자의 카운터 분석

블록화 된 태그들의 카운터수는 다음과 같다.

블록번호	Text	카운터 수
1	연예인	432
2	자동차	324
3	신문	216
4	html	109
5	쇼핑	108
6	중국	107
7	다이어트	42
8	대학	28
9	경제학	0
10	경영학	0
카운터 합		1366

⑥ 퍼지 α-cut의 적용

위에서 분석한 카운터를 사용하여 아래와 같은 속성값을 구할 수 있다.

블록번호	Text	해당카운터/총카운터	속성값
1	연예인	432/1366	0.3162
2	자동차	324/1366	0.2871
3	신문	216/1366	0.1581
4	HTML	109/1366	0.0797
5	쇼핑	108/1366	0.0790
6	중국	107/1366	0.0783
7	다이어트	42/1366	0.0307
8	대학	28/1366	0.0204
9	경제학	0/1366	0
10	경영학	0/1366	0

위의 속성값을 기준으로 하였을 때, 사용자에 의해서 클릭 된 빈도수가 가장 높은 부분에서 가장 낮은 빈도 수를 가지는 부분으로 나타내어 진다.

$\alpha_{0.08}$ = {0.1581, 0.2871, 0.3162}

α_{0} = {0, 0.0204, 0.0307, 0.0783, 0.790, 0.0797
0.1581,0.2871, 0.3162 }

5. 결론

기존 웹 페이지 표시의 경우, 사용자 취향이나 정보에 대한 접근패턴에 상관없이 단방향적으로 보여주는 것이 대부분이었다. 또한 사용자들의 하드웨어 환경과 개인적 선호도에 따라 개개인의 적합한 웹 페이지를 표시할 필요가 있다.

따라서 본 논문에서는 사용자에게 보다 적합하고 유용한 정보를 제공할 수 있도록 퍼지 α-cut을 이용하여 웹 페이지를 부분적으로 표시할 수 있는 방법을 제안하였다. 제안된 방법은 사용자의 패턴을 분석하고, 블록화 되어있는 태그들을 중요도(속성값)에 따라 나타낼 수 있도록 함으로서 사용자의 정보검색에 소요되는 시간과 비용을 절감할 수 있을 것으로 기대된다.

6. 참고문헌

[1] 이광형 · 오길록,[퍼지이론 및 응용 I권 이론], 홍릉과학출판사, 1991.

[2] AmazingSoft사, www.acecount.com