개인화된 뉴스 서비스를 위한 뉴스 메신저

김원철° 황인준
아주대학교 정보통신전문대학원 정보통신공학과
{wc323°, ehwang}@ajou.ac.kr

News messenger for personalized news services

Woncheol Kim° Eenjun Hwang
The Graduate School of Information and Communication, Ajou University

요 약

최근 다양한 무선 단말기의 보급과 네트워크 기술의 발전으로 인하여 무선 단말기를 이용한 인터넷 접속이 증가하고 있다. 특히 텍스트 기반의 뉴스 컨텐츠들은 무선 단말기의 제한된 환경에서도 비교적 손쉽게 접근이 가능하기 때문에 이용률이 높다. 그러나 대부분의 뉴스 페이지들은 한 페이지에 세분화된 섹션과 많은 내용을 담고 있기 때문에 제한된 화면과 입력장치를 가진 무선 단말기를 이용하여 사용자가 원하는 부분에 접근하기까지 반복적인 스크롤링을 해야하는 불편한 점이있다. 이런 문제점을 해결하기 위해 본 논문에서는 웹의 뉴스 페이지내에서 사용자가 선호하는 뉴스의 섹션을 추출하고, 무선 환경에 적합하도록 뉴스의 각 섹션의 순서를 재구성하여 제공하는 기법을 제안한다. 본 논문에서 제안된 기법을 통해 사용자는 무선 단말기의 각종 단점을 극복함과 동시에 뉴스에서 선호하는 섹션의 맞춤형 뉴스 서비스를 제공받을 수 있다.

1. 서 론

최근 무선 단말기의 대중화와 무선 인터넷 기술의 발전으로 인해 무선 인터넷 접속이 크게 증가하고 있다. 특히, 뉴스 컨텐츠는 대부분 텍스트 위주로 구성되어있기 때문에, 무선 단말기의 낮은 대역폭과 같은 제한된 환경에서도 비교적 쉽게 접근이 가능하여 이용률이 높다. 그러나 대부분의 뉴스 페이지 구성은 데스크톱의 넓은 화면에 최적화되어 있기때문에 상대적으로 작은 화면과 제한된 인터페이스를 가진 무선 단말기를 통한 원활한 검색이 어렵다. 뉴스 페이지내 기사 검색을 위해서는 많은 스크롤링이 요구되며, 주제별 섹션으로 세분화 되어있는 구조의 전체적인 이해가 어렵다.

이러한 단점을 극복하기 위해 뉴스 페이지나 일반적인 웹 페이지를 무선 단말기에 표시하기 위한 대표적인 연구들이 있다. Digestor[1]는 페이지의 구조 변형과 문장의 제거를 이용하여 요약하였고, Pda++[2]은 화면에 줌 형식의 인터페이스를 제공함으로써 작은 화면을 가진 무선 단말기에 웹 컨텐츠를 제공해주었다. WEST[3]는 각 페이지를 포커스(focus)와 문맥(context)방식을 이용해서 작은 화면에 카드 형식의 인터페이스를 사용하였다. Power Browser[4]는 웹 페이지를 STUs(semantic textual units)로 나누어서 요약한 다음 무선 단말기에 3 단계로 나누어서 보여주었다.

그러나 이러한 연구들은 일반적으로 웹 페이지의 일정 부분의 내용을 제거하거나 요약하여 제공함으로 본래 페이지가 가지고 있는 의미 손실에 대해 고려하지 않았고, 웹 페이지에서 사용자의 선호 사항이 다름에도 불구하고 모든 사용자에게 동일한 재구성 방법을 이용함으로 맞춤형 서비스를 제공해 주지 못했다. 본 논문에서는 사용자가 선호하는 뉴스 섹션을 추출하여 무선 단말기의 작은 화면을 통해 제공할 수 있는 기법을 제안한다. 뉴스의 각 섹션은 과거 사용자가 데스크톱을 사용했을 때 사용자의 로그를 이용하고 사용자가 선호하는 섹션을 추출하며, 추출된 섹션은 접근 빈도에 따라 재구성된다. 제안된 기법을 이용하여 무선 단말기의 화면 크기를 극복함과 동시에 사용자는 선호 분야에 대한 맞춤형 뉴스 서비스를 제공받을 수 있다.

본 논문의 구성은 다음과 같다. 2장은 뉴스 페이지내에서 사용자의 선호사항을 추출하기 위해 뉴스 페이지내에서 키워드를 추출하는 과정을 보여준다. 3장은 뉴스 페이지에 존재하는 여러 섹션을 추출하는 기술을 보여주고 4장은 추출된 사용자의 선호사항을 통하여 페이지를 재구성하는 방법에 대해 설명한다. 5장은 본 논문에서 구현한 시스템의 실험 결과를 보여주고 6장에서는 결론 및 향후 계획에 대하여 논의 한다.

2. 뉴스 페이지내의 키워드 추출

최근 웹 페이지는 페이지 내에 가능한 많은 정보를 포함하고 있다. 많은 의미 부분의 존재에도 불구하고 각 페이지를 하나의 주제로 보고 분석을 하게 되면 사용자의 패턴이나 선호사항 분석시 여러 가지 판단상의 오류가 발생할 수 있다. 본 논문은 기존 방식과는 다르게 페이지내의 주제와 기능성에 따라 페이지의 각 부분을 의미단위로 분리하고 나누어진 부분에 뉴스내의 각 섹션에 해당하는 의미를 부여하는 작업을 한다. 그리고 본 논문은 이렇게 나뉘어진 각 부분을 뉴스렛이라 한다. 뉴스 페이지로부터 의미단위인 뉴스렛을 추출하기 위해 뉴스렛은 뉴스의 하나의 주제나 기능을 충분히 표현할 수 있어야 한다. 본 논문에서 우리는 뉴스렛을 다음과 같이 정의한다.

> *정의: 뉴스렛은 뉴스 페이지내에 주제나 기능성을 표현하는 의미적인 단위이고, 뉴스렛내에는 같은 주제나 같은 기능성을 가진 다른 의미구역이 포함되지 않는다.*

본 논문은 정보추출과 사용자가 선호하는 사항을 보다 정확하게 처리하기위해 뉴스렛을 이용한다. 페이지내의 각 의미부분인 뉴스렛이 뉴스의 어느 부분에 속하는가를 결정하기 위해서 뉴스렛 내의 키워드들을 추출하고 추출된 키워드들을 비교하여 뉴스의 분야를 파악한다. 따라서 뉴스의 각 분야에서 보

다 효과적으로 키워드들을 추출하기 위해서 각 문서에서 중요한 문장들을 추출한 후에 추출된 문장에서 키워드들을 추출한다. 본 논문은 키워드 추출을 위해 일반적으로 이용되는 Luhn's Keyword cluster [5]를 이용하고 보다 정확한 측정을 위해 HTML문서내의 <Title>태그내의 키워드 추출방법과 변형 키워드 가중치 방법을 이용한다. 본 논문은 중요한 문장을 선택하기 위해 첫번째로 TF/IDF[6]의 가중치 값과 Luhn의 연구의 변형을 이용한다. 이 방법은 문장내의 어떤 키워드들이 임계값보다 크면 중요한 문장이라고 판단한다. 문장들의 가중치 값을 구하는 공식은 다음과 같다.

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

where SW1 = 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 본 논문에서는 문장이 <Title> 태그 내의 키워드들을 얼마만큼 포함되는 가를 살펴본다. <Title>태그내의 키워드 빈도수를 이용한 중요성 평가 계산은 다음과 같이 한다.

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

where SW2 = 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 웹문서 저자는 문장 내에서 중요성을 표현하기 위해 키워드들을 강조한다. 본 논문에서는 강조된 키워드들을 찾아서 중요한 문장을 선택할 때 이 키워드들에 대해 높은 누적 값을 이용하여 중요한 문장에 적용한다. 따라서 강조된 키워드들을 이용한 중요성 평가의 계산은 다음과 같다.

$$SW3 = \frac{MW^2}{TW}$$

where SW3 = the sentence weight score
MW = the number of modified words
TW = the total number of words

문장의 중요성에 대한 평가는 위의 3가지 계산에 의해 이루어 진다.

3. 자동 뉴스렛 추출 기술

본 논문은 HTML 엘리먼트 내의 링크의 존재와 뉴스렛 내의 키워드들에 대한 가중치를 부여함으로 각 페이지 내의 주제 부분을 추출한다. 뉴스렛 추출에 대한 알고리즘은 그림 1과 같다. 사용자가 뉴스 페이지내에서 클릭한 부분의 뉴스렛을 추출한 후, 뉴스렛이 뉴스의 어느 분야에 속하는가를 파악하고 사용자가 선호하는 뉴스섹션을 랭킹해야한다. 랭킹을 적용하기 위해서 한 뉴스 페이지 내에서 사용자 자신이 선호하는 기사를 검색하는 시간을 SessionTime이라고 정의를 한다.

```
NewsletPartition(p)
  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
```

그림 1. 뉴스렛 추출 알고리즘

또한 사용자가 검색한 각 분야의 검색시간을 BranchTime이라고 하면 뉴스의 각 섹션에 대한 사용자의 선호사항은 다음과 같다.

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

where BT_{field} = BranchTime for news branch
ST = SessionTime

뉴스 페이지내에서 사용자의 각 섹션에 대한 선호사항을 파악하고 최근까지 누적된 정보들을 이용하여 뉴스의 각 섹션에 대한 전체적인 사용자의 선호사항을 추출해야한다. 전체 선호사항 추출 계산은 다음과 같다.

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

where TUI = Total User Interests per each branch
func(date) = function of date
TS = Total SessionTime
ST = SessionTime
S = count of each UI

누적된 정보들중에 최근에 사용자가 더 많은 선호사항을 가진 뉴스섹션에 더 높은 값을 갖게 하기위해 *func(date)*을 이용한다. 이 함수는 날짜 정보를 이용하여 최근 날짜가 수치적으로 더 높은 값을 가지도록 한다.

4. 웹 페이지 재구성

뉴스 메신저 시스템은 클라이언트와 서버 사이의 프록시 서버에 위치하게 된다. 그림 2는 뉴스 메신저 시스템의 전체적인 구조를 나타낸다. 무선 단말기의 사용자가 웹 페이지를 요청하였을 때 처리 순서는 다음과 같다.

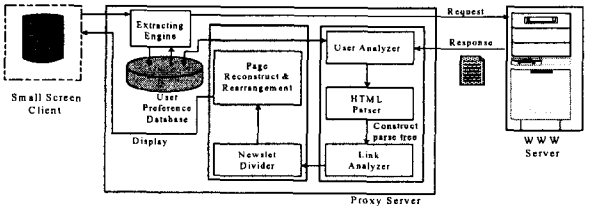

그림 2. 뉴스 메신저 시스템 전체구조

1. 무선 단말기 사용자가 웹 페이지를 요청하였을때 추출엔진은 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 HTML문서를 HTML 파서에 전달한다.
3. HTML 파서 모듈이 문서를 전달 받았을 때, 파서는 문서 내의 배너와 그림을 제거하고, 문서를 각 노드로 분리하고 파스 트리를 생성한다. 뉴스렛을 만들기 위하여 Link analyzer 모듈이 파스 트리를 생성후 링크 정보를 추출한다.
4. Newslet Divider 모듈은 뉴스 페이지내에서 뉴스의 단위를 파악하기 위해 파스 트리의 노드의 링크수를 분석하여 뉴스렛을 만든다.
5. Page Reconstruct and Rearrangement 모듈은 추출된 사용자의 선호하는 뉴스 섹션을 바탕으로 뉴스렛의 순서를 재배치한다.

위와 같은 처리순서를 이용한 뉴스 메신저 시스템은 사용자가 더 많은 선호사항을 가지고 있는 뉴스의 분야를 무선 단말기 화면에 보다 위에 보여주고 불필요한 배너와 그림을 제거함으로 무선 단말기의 작은 화면에 모든 정보를 표현하려고 할 때 제약사항을 완화시켰다.

5. 실험 및 분석

우리는 뉴스 메신저 시스템을 모바일 에뮬레이터를 이용하여 구현하였다. 뉴스 페이지로부터 키워드를 추출하고 사용자가 뉴스 페이지내에서 클릭한 곳의 뉴스렛을 추출하고 사용자의 선호하는 뉴스섹션에 관한 정보를 데이터베이스에 저장하는 것은 데스크톱환경에서 클라이언트측에서 자바 API를 이용하여 구현하였고, 추출된 사용자의 선호사항을 이용하여 모바일 에뮬레이터에 보여주는 것은 임베디드 비주얼베이직 3.0을 이용하여 구현하였다.

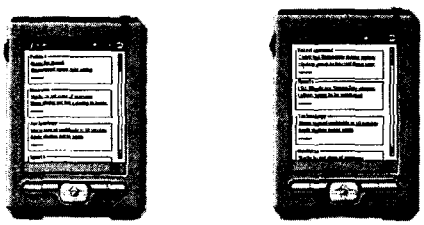

그림 3. 뉴스 페이지의 재구성

그림 3의 왼쪽 그림은 정치, 경제, 과학, 스포츠, 오락의 차례로 뉴스섹션을 선호하는 사용자가 뉴스 페이지를 요청했을 경우 화면이다. 그림 3의 오른쪽 그림은 오락, 스포츠, 경제, 정치, 과학의 순서로 뉴스섹션을 선호하는 사용자가 뉴스 페이지를 요청했을 경우 화면이다.

표 1. 실험에 사용된 뉴스 페이지

Site	URL
CNN	http://www.cnn.com/
BBC	http://news.bbc.co.uk/
MSNBC	http://www.msnbc.com
ABCNews	http://abcnews.go.com/

본 시스템의 효율성을 측정하기 위해 우리는 여러 뉴스 페이지에서 문서를 재구성한 후에 문서의 크기와 화면상의 크기를 비교 했다. 표1은 실험에 사용된 뉴스 페이지들의 리스트를 보여준다.

표 2. 재구성한 후 페이지 비교

Site	Original page size (A)	Newslet size (B)	((A-B)/A) x 100	Display Saving (%)
CNN	122.4 Kb	14.5 Kb	88 %	85%
BBC	110.6 Kb	13.2 Kb	87 %	84%
MSNBC	123.4 Kb	11.6 Kb	90 %	87%
ABCNews	115.8 Kb	12.5 Kb	89 %	86%

표2는 원래 뉴스 페이지의 파일 크기와 페이지를 재구성한 후에 크기, 화면상의 감소된 크기를 비교하였다. 여기서 우리의 뉴스렛 단위의 사용이 파일과 화면의 크기를 줄이는 것을 볼 수 있다. 평균 파일 크기와 화면 감소가 평균 80% 이상이다. 파일 크기의 감소는 무선 환경의 제한된 대역폭에 뉴스 컨텐츠를 전송할 경우 제약을 완화 시키고, 화면 크기의 감소는 작은 화면의 단말기의 제한과 적은 스크롤링으로 사용자가 보고 싶어하는 내용을 볼 수 있다. 결과적으로 본 논문에서 제안한 뉴스렛을 이용함으로 무선 단말기 사용자는 웹 페이지에 대한 가독성뿐 아니라 적은 스크롤링으로 사용시 보다 편리해졌다.

6. 결론 및 향후 연구 과제

무선 단말기 사용자의 인터넷 접속이 증가하면서 점차 웹 페이지는 유선사용자뿐 아니라 무선 사용자의 요구를 만족시켜야할 필요성이 증가하였다. 본 논문에서 무선 단말기 사용자의 요구를 만족시키기 위하여 뉴스 메신저를 제안하였다. 뉴스 메신저에서 사용된 뉴스렛은 본래 페이지와 비교하여 작은 파일 사이즈로 변형 시켰고 이는 무선이란 환경의 낮은 대역폭의 제한을 완화시켰고 작은 화면에서 가독성을 증가시켰다.

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

7. 참고 문헌

[1] T. W. Bickmore, B. N. Schilit, " Digestor: 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 Sysmposium 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] De Bra and R.D.J. Post. "Information retrieval in the World-Wide Web: making client-based searching feasible." Proceedings of the First International World-Wide Web Conference. 1994.

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