동적 데이터 추출을 통한 뉴스 클리핑 시스템

전호철°, 신성혁
이소프팅 개발사업본부

News Clipping System Through Dynamic Data Extraction

HoChul Jeon, SungHyuk Shin
BTOBIZ Research & Development Center.

요 약

인터넷의 빠른 보급으로 많은 양의 정보가 유통되기 시작했다. 그러나 사용자들은 필요한 정보들을 취사 선택하기엔 정보들은 양이 너무 방대하다. 각종 사이트에 있는 뉴스들을 실시간으로 사용자들에게 필요한 정보를 제공할 수 있는 뉴스 클리핑은 이러한 사용자들의 요구를 충족할 수 있다. 하지만 기존의 뉴스 클리핑 시스템은 각 사이트에 접근 후, 수동적인 분석을 통해 뉴스 정보 및 뉴스 기사의 위치를 파악하고 이를 추출하도록 하는 시스템들이다. 본 논문에서 제안하고자 하는 시스템은 사이트의 구조를 파악하고, 뉴스 기사들을 동적으로 추출함으로써 기존 시스템의 단점을 극복하고, 내용 기반의 뉴스기사 검색이 가능하도록 한다.

1. 서론

인터넷이 빠르게 보급되면서 많은 정보들을 접하게 되었다. 21세기는 정보를 지배하는 자가 세계를 지배한다고 했다. 사용자가 수많은 정보들 중에 필요한 정보를 빠르게 실시간으로 입수 한다는 것이 그만큼 중요하다고 할 수 있다. 하지만 현재 인터넷 상에서 유통되고 있는 방대한 양의 정보들은 일반 사용자들로 하여금 필요한 정보를 쉽게 접근하게 하지 못하고 있다.

현재 사용자들이 쉽게 정보를 얻기 위해 각 언론사들의 뉴스를 이용하고 있다. 하지만 정치, 경제, 문화, 사회, 스포츠 그리고 연예 등 다양하고 수 많은 언론사 사이트는 그 양이 너무 방대하기 때문에 사용자가 필요한 정보들을 실시간으로 취사 선택 하는 것이 쉬운 일이 아니다.

또한 각종 정부 기관이나 다양한 업체에서 제공하는 정보들도 그 양이 많아 필요한 정보를 얻기는 쉽지 않다.

이러한 불편함을 해결하기 위하여 뉴스의 정보를 사용자가 요구하는 필요한 정보를 실시간으로 취사 선택 하기 위하여 뉴스를 클리핑하여 제공한다면 사용자들의 부담을 한층 덜 수 있다. 하지만 현재의 뉴스 클리핑 시스템은 수동적으로 뉴스의 위치를 파악하여 제공한다. 따라서 본 논문은 실시간으로 필요한 뉴스 정보들을 동적으로 추출하여 사용자들에게 제공하는 뉴스클리핑 시스템(News Clipping System)을 제안하고자 한다.

2. 관련연구

기존의 뉴스 클리핑 시스템은 각 사이트에 접근

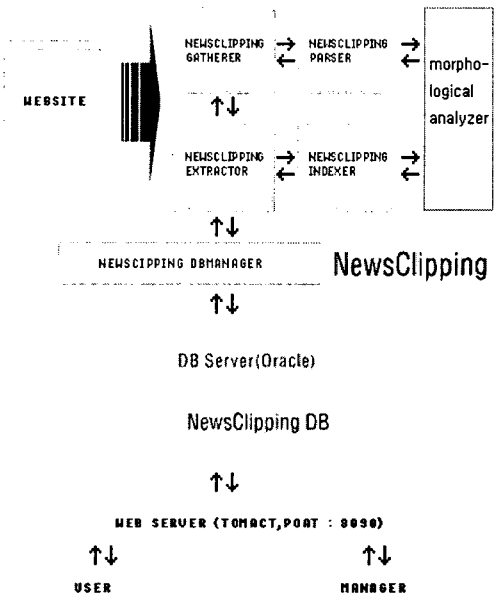
후, 수동적인 분석을 통해 뉴스 정보 및 뉴스 기사의 위치를 파악하고 이를 추출하도록 하는 시스템들이다.

이는 기 분석된 사이트들의 구조가 변경되는 경우 재분석하는 단점이 있으며, 뉴스 기사를 추출하고자 하는 사이트의 추가가 반드시 분석 과정을 통해 이루어 지도록 된다. 본 논문에서 제안하고자 하는 시스템은 사이트의 구조를 파악하고, 뉴스 기사들을 동적으로 추출함으로써 기존 시스템의 단점을 극복하고, 내용 기반의 뉴스기사 검색이 가능하도록 한다.

3. 뉴스 클리핑 시스템

3.1 시스템 구성

본 연구에서 구축하고자 하는 뉴스 클리핑 시스템은 웹에서 정보들을 모아서 필터링 한 후 사용자에게 제공하여 텍스트를 이용할 수 있게 한다.



[그림 1] 뉴스 클리핑 시스템 구성도

본 논문에서는 정보를 필터링 하는 부분 중 데이터를 동적으로 추출하는 방법에 대하여 제안하고자 한다.

3.2 뉴스 기사 위치의 기본 특성

먼저 뉴스 클리핑 시스템에서 먼저 정보를 추출하기 위해서는 뉴스 기사의 위치에 대한 기본 특성을 파악해야 한다.[1]

모든 뉴스 기사는 뉴스의 제목과 근접해 있다.
모든 뉴스 기사는 <TABLE></TABLE> 사이에 있다.
대부분의 뉴스 기사는 전체 문자의 길이가 길다.

위와 같은 특성들은 뉴스 기사 위치의 기본 특성이 된다.[2]

3.2 뉴스 기사 위치의 기본 특성을 이용한 동적 추출

뉴스 사이트에서 링크에 사용된 제목과 실제 뉴스 기사의 제목이 다른 경우는 그 위치를 정확히 판단하는 것이 중요하다. 이를 위해 링크에 사용된 뉴스 기사의 제목을 색인어 단위로 추출하며, 추출된 색인어들은 발생 순서에 따라 가장 많은 색인어가 일치 하는 Table Tag 가 실제 뉴스 기사의 제목으로 결정된다.

$$sum\left(\frac{1}{terms}, match\_count\right)$$

수식 1

terms : 링크에 사용된 제목의 추출된 색인어 수, match_count : 각 TABLE Tag에서 추출된 색인어가 일치 하는 수

이를 식으로 표현하면 아래의 식으로 표현된다.

$$W_{title} : \sum_{1}^{match\_count} \frac{1}{title\_terms}$$

수식 2

$$W_{rank} = \left(\frac{1}{rank}\right) * 0.5$$

수식 3

W_{rank} : 뉴스의 제목과의 거리에 대한 가중치 값
rank : 근사거리에 따른 순위 값
W_{rank} 의 값이 0.1 이상인 TABLE Tag 에 대해서만 이후의 계산 식을 적용한다.
0.1의 값은 실험을 통해 얻어진 수치이다.

$$W_{le\_rank} : \left(\frac{1}{le\_rank}\right) * 0.5$$

수식 4

W_{le_rank} : 각 Table Tag 의 문자 길이에 대한 가중치 값

$$W = W_{le\_rank} + W_{rank}$$

위의 기본 공식만을 사용하는 경우 각 뉴스 기사에 대한 독자의 의견을 포함하는 TABLE Tag의 문자 길이가 실제 뉴스 기사의 길이 보다 긴 경우 가중치 값이 실제 뉴스 기사보다 커지게 된다.
예를 들어 스포츠 경기 결과에 대한 짤막한 기사

이를 위해 독자의 의견 수렴에 사용되는 특정 단어들을 배제 토록 할 필요가 있다.

이러한 단어들을 포함하는 TABLE에 대한 가중치를 줄임으로써 상대적으로 실제 뉴스 기사를 포함하는 TABLE Tag의 가중치 값을 높이도록 한다.

$$W_{keyword} = -\sum_{1}^{match} \frac{1}{keyword}$$

수식 5

$W_{keyword}$: 특정 단어의 포함에 대한 가중치 값
Keyword : 특정 단어의 수
Match : 각 TABLE Tag에서 포함하는 특정 단어의 수

$$W = W_{le\_rank} + W_{rank} + W_{keyword}$$

수식 6

$$W = \left(\frac{1}{le\_rank}\right) * 0.5 + \left(\frac{1}{rank}\right) * 0.5 - \sum_{1}^{match} \frac{1}{keyword}$$

수식 7

계산된 각 TABLE Tag에 대한 가중치 값에 따라 실제 뉴스 기사를 포함하는 TABLE Tag가 결정 된다.

마지막으로 각 사이트에서 뉴스 기사의 동적 추출을 위해 정확한 뉴스 기사의 제목을 추출하고 추출된 뉴스 기사와의 거리에 따른 각 TABLE Tag에 대한 가중치 적용 및 각 TABLE Tag가 포함하는 문자의 길이에 따른 가중치 적용, 특정 단어에 대한 각 TABLE Tag에 대한 가중치 적용을 통해 수행토록 한다.

본 시스템을 통해 뉴스 기사의 추출을 원하는 사이트의 추가가 비교적 쉽게 이루어 지며, 추가하고자 하는 사이트의 수에 제한이 없어 진다.
또한 동적으로 뉴스 기사를 추출함으로써 추출하고

자 하는 사이트의 구조 변경에 능동적으로 대처 가능하다.

반면 뉴스 기사의 제목 및 그 위치에 의존적이므로 이미지만으로 링크가 이루어진 뉴스기사의 추출은 불가능 하고 뉴스 기사라는 제한된 데이터만을 고려 했기 때문에 정보 수집에 적용하기에 부적합할 수 있다. 또한 본 시스템은 지역성(localization)을 고려하지 않았기 때문에 국내 사이트에 제한적이다.

4. 구현 및 고찰

본 시스템은 윈도우 2000 또는 리눅스 7.0이상, JAVA JDK1.4 이상, 웹서버로는 Tomcat 4.0 이상 그리고 오라클 8.15이상에서 동작된다.

5. 결론

각 사이트에서 뉴스 기사의 동적 추출을 위해 정확한 뉴스 기사의 제목을 추출하고 추출된 뉴스 기사와의 거리에 따른 각 TABLE Tag에 대한 가중치 적용 및 각 TABLE Tag가 포함하는 문자의 길이에 따른 가중치 적용, 특정 단어에 대한 각 TABLE Tag에 대한 가중치 적용을 통해 수행토록 한다. 향후 뉴스 기사의 제목에 의존적인 단점을 극복하도록 수정 되어야 하며, 이를 위해 TABLE Tag의 위치에 따른 가중치 적용을 수정할 필요가 있다. 또한 국외 사이트에 적용 가능토록 이에 대한 연구가 필요하다. 텍스트 기반의 미디어 뉴스에 대한 서비스를 제공할 예정이다.

[참고문헌]

[1] David King and Daniel O'Leary, "Intelligent Executive Information Systems", IEEE Expert, December 1996, pp. 30-35

[2] Thomas Gschwind and Manfred Hauswirth: "A Cache Architecture for Modernizing the Usenet Infrastructure", Thirty-second Annual Hawaii International Conference on System Sciences-Volume 8, January 1999, pp. 8011