韓國 컴퓨터情報學會 論文誌
第12卷 第5號, 2007. 11.

2007-12-5-1-1

웹 뉴스의 기사 추출과 요약

한 광 록*, 선 복 근**, 유 형 선***

Text Extraction and Summarization from Web News

Kwang-Rok Han *, Bok-Keun Sun **, Hyoung-Sun Yoo ***

요 약

뉴스 콘텐츠 등 웹을 통해 제공되는 많은 정보들은 불필요한 클러터를 많이 포함하고 있다. 이러한 클러터들은 문서의 요약, 추출, 검색과 같은 자동화된 정보처리 시스템의 구축을 어렵게 한다. 본 논문에서는 웹 뉴스 콘텐츠를 추출하고 이를 요약하는 시스템을 구축하고자 한다. 추출 시스템은 HTML로 된 뉴스 콘텐츠를 입력받아 DOM 트리와 유사한 요소 트리를 구축하며, 이 요소 트리에서 HTML 태그의 하이퍼링크 속성을 갖는 클러터를 제외하면서 본문을 추출한다. 추출 시스템을 통해 추출된 본문은 요약시스템으로 전달되어 핵심 문장이 추출된다. 요약 시스템은 공기관계 그래프를 이용하여 구성한다. 본 논문에서 구현한 시스템을 통해 추출된 요약 문장은 SMS와 같은 메시지 서비스를 통하여 PDA이나 모바일 폰 등에 전송될 수 있을 것으로 기대된다.

Abstract

Many types of information provided through the web including news contents contain unnecessary clutters. These clutters make it difficult to build automated information processing systems such as the summarization, extraction and retrieval of documents. We propose a system that extracts and summarizes news contents from the web. The extraction system receives news contents in HTML as input and builds an element tree similar to DOM tree, and extracts texts while removing clutters with the hyperlink attribute in the HTML tag from the element tree. Texts extracted through the extraction system are transferred to the summarization system, which extracts key sentences from the texts. We implement the summarization system using co-occurrence relation graph. The summarized sentences of this paper are expected to be transmissible to PDA or cellular phone by message services such as SMS.

▸ Keyword : 웹뉴스(Web News), 추출(Extraction), 요약(Summarization), 공기관계그래프(Co-occurrence Relation Graph)

• 제1저자 : 한광록
• 접수일 : 2007. 8.6, 심사일 : 2007. 10.1, 심사완료일 : 2007. 10.11.
* 호서대학교 컴퓨터공학과 교수, ** 호서대학교 프로그램 전담강사
*** 순천향대학교 국어국문학과 부교수

Ⅰ. 서 론

대부분의 웹 뉴스 콘텐츠는 뉴스 이외의 많은 데이터를 포함하고 있다. 광고와 팝업, 이미지등 기사와 관련이 없는 부분들이 이에 해당한다. 이러한 뉴스 이외의 정보들은 문서의 분류 및 검색 그리고 추출 등과 같은 여러 응용분야에서 웹 뉴스를 활용하는데 방해물인 클러터로 작용한다. 또한 PDA나 모바일 폰 등과 같이 사용 환경에 제약이 있는 장치에서 뉴스 데이터에 접근하고자 할 경우에는 클러터와 같은 방해물은 반드시 제거해야 한다.

이러한 문제점으로 인해 데이터에 기초한 정보검색과 어플리케이션을 위한 데이터 자원으로서 웹 문서를 사용하기 위하여 다양한 관점에서 연구가 이루어지고 있다. 클러터를 제거하거나 콘텐츠를 보다 읽기 쉽게 하려는 대부분의 방법들은 폰트 크기를 크게 하거나 이미지를 제거하거나 또는 자바 스크립트를 사용하지 못하게 하는 등 홈페이지 고유의 룩앤필(look-and-feel) 기능을 제거 하는 것들이다. 이와 같은 방법은 만일 소프트웨어가 제어하지 못하도록 프로그램된 레이아웃을 만나게 되면 부정확한 결과를 생성하게 된다[1]. 참고문헌[1-4]에서는 HTML의 구조를 분석하여 클러터를 제거하거나 데이터를 가공한다. 이와 같은 시스템은 룩앤필에 기초한 방법보다 효과적이지만 시스템이 매우 복잡해 질 수 있다. 참고문헌[2]의 시스템을 예로 들면, 문서의 구조를 알아내기 위해서 xpath를 처리하기 위한 전처리 시스템과 역전파 알고리즘을 사용한 학습 및 추론 시스템이 필요하며, 다양한 휴리스틱을 적용하기 위한 시스템도 필요하게 된다.

본 논문에서는 직접 구현한 HTML 파서를 이용해 DOM 트리와 유사한 요소 트리를 구축한 후 HTML 태그의 하이퍼링크 속성을 갖는 노드를 제거해 나가면서 웹 뉴스 데이터를 추출하고, 추출된 데이터를 요약하는 시스템에 대해 논하고자 한다. 또한 추출한 결과를 요약하기 위하여 공기관계 그래프(co-occurrence relation graph)를 이용하는 방법을 적용하였다[5]. 개발된 시스템은 SMS 서비스와 PDA 그리고 모바일 폰 등의 응용프로그램 등에 다양하게 적용해 볼 수 있다.

Ⅱ. 관련 연구

정보검색 및 추출과 어플리케이션을 위한 데이터의 자원으로서 웹 문서를 사용하기 위하여 다양한 관점에서 많은 연구가 이루어지고 있다. 자연어처리와 문법 규칙, HTML 트리구조 처리, HTML 테이블 처리 등 다양한 기술을 기반으로 웹 문서 처리에 관한 연구가 이루어진다.

콜럼비아 대학의 NLP 연구 그룹에서는 단어의 수를 계산하여 웹 페이지의 가장 큰 텍스트 몸체(text body)를 검출하고 그것을 콘텐츠로 분류한다[6]. 이 방법은 간단한 웹 페이지에서는 잘 동작한다. 하지만 이 알고리즘은 특히 불규칙한 광고나 이미지 배치를 포함하는 다중 몸체로 구성된 문서들을 처리하는 데는 불필요한 정보나 부정확한 결과들을 생성한다. 참고문헌[6]에서는 구조분석과 문맥해석을 하여 요약하는 또다른 기법을 제안하고 있다. HTML 문서의 구조가 먼저 분석되고, 이 HTML 문서를 적절하게 보다 작은 부분들로 분해한다. 그리고 나서 이 각각의 부분들의 콘텐츠가 추출되고 요약된다. 그러나 이 방법은 이직 구현단계이다. 더욱이 이 논문은 콘텐츠 추출을 위한 선행 요구조건들만을 나열하고 있고 그것을 구현하기 위한 방법은 제안하지 않고 있다[7].

참고문헌[2]에서는 백프로파게이션 네트워크를 구성하여 웹 문서의 구조를 학습한 후, 이를 활용하여 새로운 웹 문서의 구조를 추론해 내는 시스템을 구성하였다. 시스템은 먼저 xpath에 ID를 부여하는 과정을 통해 웹 문서를 백프로파게이션 네트워크의 입력값으로 변환한다. 백프로파게이션 네트워크의 학습시스템은 정해진 에러율 이하의 값이 나올 때까지 반복하여 학습을 수행 후, 문서를 네트워크에 통과시키면 시스템은 웹 문서의 구조를 추론하고 그 구조에 적합한 정보를 추출해 낸다. 그러나 이 시스템은 학습시스템을 구축하기 전에 xpath를 처리하기 위한 전처리 과정에 많은 계산 비용이 소모되며, 특정 사이트의 구조가 변경되면 네트워크의 학습을 다시 시켜야 된다는 단점이 있다.

참고문헌[1]의 시스템은 openXML[8]을 HTML 파서로 사용하여 HTML을 DOM 트리의 자료구조로 변환 한 후 다양한 필터를 적용하여 콘텐츠를 추출한다. 이 시스템은 본 논문에서 제안한 시스템과 유사한 형태로 자료를 변환하고 콘텐츠를 추출한다.

본 논문은 직접 제작한 HTML 파서를 이용하여 DOM 트리와 유사한 형태의 요소 트리를 구축한 후 이 트리에서 하이퍼링크 속성을 가진 요소의 노드를 삭제하는 과정을 통해 콘텐츠를 추출한다. 표준 HTML 파서와 DOM 트리는 표준화 되어 있기 때문에 다양한 분야에 쉽게 적용이 가능하다. 그러나 제공되는 API가 상위 시스템의 요구사항을 모두 만족 시킬 수 없기 때문에 특정 시스템의 하위 시스템으로 사용할 경우 적용의 유연성이 떨어지게 된다.

III. 웹뉴스 기사의 추출과 요약 시스템

본 논문에서는 추출과 요약을 수행하는 두개의 시스템으로 구성되어 있다. 추출 시스템은 입력으로 들어온 웹 콘텐츠의 클러터를 제거하여 뉴스 데이터만을 추출하며, 요약 시스템은 추출 시스템으로부터 추출된 뉴스 데이터의 핵심 문장을 추출해 요약한다. 그림 1은 전체 시스템의 구조와 문서의 처리 흐름을 보여준다.

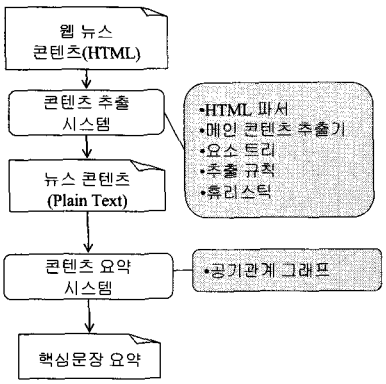

그림 1. 전체 시스템 구조
Fig 1. Total System construction

3.1 태그정보 기반의 텍스트 추출

대부분의 웹 뉴스 콘텐츠는 뉴스 이외의 많은 데이터를 포함하고 있다. 기사와 관련이 없는 광고와 팝업 그리고 이미지 등이 여기에 해당한다.

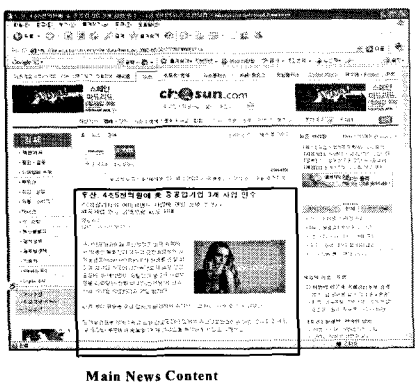

그림 2. Chosun.com 뉴스기사의 예
Fig 2. An example of chosun.com news

그림 2는 조선일보 인터넷 뉴스의 메인 화면이다. 그림에 나타난 바와 같이 중앙의 메인 뉴스 콘텐츠를 제외한 모든 것들은 클러터이다. 조선일보뿐만 아니라 거의 모든 인터넷 뉴스 사이트는 위와 유사한 구조를 가지고 있다. 이러한 유사한 구조를 가진 사이트에 사용된 HTML 태그의 속성을 파악해 보면 대부분의 클러터는 하이퍼링크 속성을 가지고 있다. 표 1은 뉴스 사이트들의 클러터 정보를 요약한 것이다.

표1. 웹 뉴스 HTML에 포함된 클러터 정보
Table 1. Cluster information in Web news HTML

Site Name	Page View	Avg. of Link Clutter	Avg. of Non-Link Clutter
www.cnn.com	40	148	8
www.chosun.com	40	92	7
news.yahoo.com	40	122	11
www.khan.co.kr	40	92	10
www.latimes.com	40	103	6

표 1의 링크 클러터는 대부분 저작권(copyright), 현재 시간, 단순한 헤드를 포함한다. 이와 같이 간단한 텍스트를 제외한 모든 클러터는 하이퍼링크의 속성을 가지고 있다.

본 논문의 태그정보 기반의 뉴스 텍스트 추출 시스템은 HTML 파서를 통해 만들어진 요소 트리를 기반으로 위와 같은 클러터 들을 그림 3과 같이 제거하면서 본문을 추출한다.

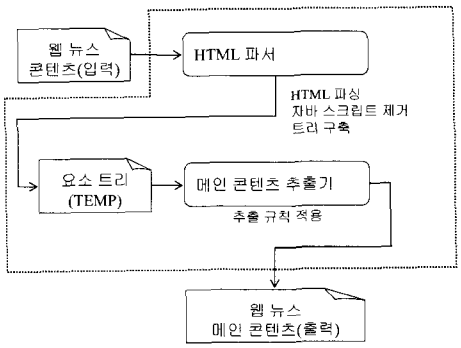

그림 3. 태그정보 기반의 뉴스 텍스트 추출 시스템
Fig 3. News extraction system based on tag information

그림 3에서 웹 뉴스 콘텐츠는 HTML 파서에 의해 요소 트리로 만들어진다. 이 과정에서 스크립트는 버린다. 경험적으로 거의 모든 스크립트는 뉴스 콘텐츠와는 상관이 없다. 다음 단계로 메인 콘텐츠 추출기에서 요소 트리의 모든 노드를 탐색하면서 추출규칙을 적용하게 되면 웹 뉴스의 메인

콘텐츠가 추출 된다. 이때 적용되는 알고리즘은 그림 4와 같으며, 추출규칙은 아래와 같이 요약 할 수 있다.

(1) 텍스트 노드들을 제외한 모든 노들은 무시한다.
(2) 하이퍼링크 속성을 갖는 텍스트 노드는 무시한다.
(3) 휴리스틱 규칙을 적용한다.
(4) 텍스트를 추출한다.

표 1에서 살펴본 바와 같이 평균 7.6% 정도의 클러터는 링크를 가지고 있지 않다. 이러한 클러터는 대부분 위에서 언급한 저작권, 현재 시간, 단순 헤더 등의 짧은 텍스트들이다. 또 다른 문제는 기사 본문 내에 키워드 링크를 추가해 놓은 부분이다. 키워드 링크는 단순 하이퍼링크를 사용한 경우와 스크립트를 사용한 경우로 나뉘게 되는데, 스크립트의 경우 파싱 단계에서 버려지므로 문제가 되지 않는다. 그러나 하이퍼링크를 사용할 경우 제거되면 안 되는 데이터임에도 불구하고 제거되게 된다. 이와 같은 문제를 해결하기 위하여 추출규칙의 마지막에 휴리스틱을 적용하여 문제를 해결하였다.

추출된 데이터는 문서 요약 시스템의 입력으로 전달되어 요약되며, 요약 시스템은 입력 데이터가 가진 클러터가 적을수록 요약의 성능이 높아지므로 태그정보 기반의 뉴스 텍스트 추출 시스템에서 최대한 많은 클러터를 제거해야 한다.

```
void ExtractFunc (Node node)
  WHILE(if node is NULL end roof)
    if node.nodeType == HTMLDefs.PCDATA THEN
      if node.attribute != HTMLAttr.HREF THEN
        if Heuristic(node) == TRUE THEN
          Extract TEXT
        ENDIF
      ENDIF
    ENDIF
    FOR cnt=0 TO node.childCnt DO
      ExtractFunc(node.childNode[cnt])
    ENDFOR
  ENDWHILE
ENDFUNCTION
```

그림 4. 추출 알고리즘의 의사코드
Fig 4. Pseudo code of extraction algorithm

3.2 공기관계 그래프에 의한 요약

3.2.1 공기관계 그래프

본 논문에서는 문단 위주의 핵심 문장 추출 방식이 아닌 대상 문서에서의 문장 및 문서를 구성하는 단어 간의 공기 관계를 나타내는 그래프(co-occurrence relation graph)를 생성하고, 이를 이용하여 핵심어들을 추출한 후, 이를 기반으로 하여 핵심 문장을 추출하는 방식을 사용하였다[5, 9-10].

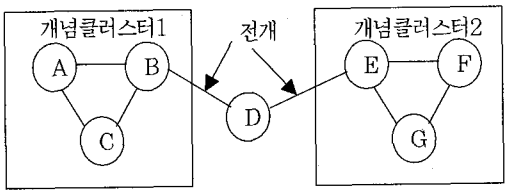

그림 5. 공기관계 그래프의 구성
Fig 5. Construction of the co-occurrence relation graph

이 방식은 문서가 기사 작성자의 독자적인 생각을 주장하기 위해 쓰여 졌다는 가정 하에 문서상의 저자의 주장을 대표하는 키워드 추출에 효과적인 방식이다. 공기 관계 그래프의 구조는 문서를 크게 개념클러스터(Mean-Cluster), 주장(Insistence), 전개(Deployment)로 구분한다.

공기 관계 그래프는 문서에서 출현 빈도가 높은 단어들의 공기 관계를 이용하여 구성한다. 문서에서 출현 빈도가 높은 단어들은 개념클러스터 및 키워드와 주장의 후보 단어가 된다. 출현 빈도가 높은 단어들은 문서의 요점과는 상관관계가 적을 수는 있으나 기사 작성자가 주장을 펼치거나 혹은 설명을 하는데 있어서 당연시되는 전제들이다. 공기 관계 그래프에서 각각의 구성요소는 다음과 같이 정의된다.

① 개념클러스터

어떤 문장에서 함께 나타나는 단어들이 다른 문장에서도 같이 나타난다면, 그 단어들은 그만큼 문서전체에서 중요성을 나타내는 개념들로 생각된다. 따라서 개념클러스터는 문서의 여러 문장에서 공통적으로 나타나는 단어들의 집합이라고 할 수 있다. 이 단어들의 집합들은 공기 그래프 상에서 루프를 형성하게 된다. 그래프 상에서 루프를 형성한다는 것은 문서상에서 같은 문장 혹은 문단에서 동시에 나타나는 것을 의미하며, 어떠한 문장 혹은 문단이 자주 출현한다는 것은 기사 작성자가 주장하거나 혹은 설명하는데 있어서 주된 내용과 밀접한 연관을 가지고 있음을 의미한다.

② 주 장

기사 작성자가 의도하는 문서의 핵심이 될 수 있으며, 개념 클러스터 사이에 강하게 연결되어 문장을 통합하는 역할을 한다.

③ 전 개

개념클러스터간의 연결을 나타내는 것으로서 문서에서 중요한 내용의 흐름을 표현한다. 그림 5에서 B와 E는 문서

의 키워드가 되는데 이는 문서에서 주장과 강하게 연결되어 기사 작성자의 의도를 뒷받침하거나 혹은 강조하는 역할을 하는 단어를 의미한다.

3.2.2 전처리 단계

요약할 문서는 키워드를 추출하기 위해서 우선 형태소 분석 혹은 영어의 경우에는 스태밍 작업을 우선 수행하며, 불용어(Stop List)등 불필요한 단어를 제거하고, 단어의 중복을 제거한다. 본 논문에서는 숫어의 요소는 고려하지 않았으며, 단어의 선정도 명사와 형용사, 동사로 제한하였다. 후보 단어의 집합은 식(1)과 같이 표현한다.

$$\text{문서} = W_1, W_2, W_3, W_4, \cdots W_n \quad \cdots\cdots (1)$$

3.2.3 키워드 추출

① 개념클러스터의 형성

개념클러스터를 형성하기 위해서 우선 각 단어들로 이루어진 공기 관계 그래프를 생성한다. 그래프를 생성하기 위해 대상 문서에서 출현빈도의 순위별로 후보 단어 집합으로부터 상위 30위까지의 단어를 추출하여 그래프의 Vertex로 간주한다. 만약 30위에 해당하는 단어가 복수 개일 경우 해당 단어를 모두 선택한다. 그래프를 형성하는데 있어서 출현빈도가 높은 단어들을 추출하는 이유는, 기사 작성자가 문서에서 어떠한 주요 내용을 서술할 때는 중요한 내용과 밀접한 관련을 가지는 단어들이 반드시 그 문장 내에 존재하기 마련이며, 이러한 단어들이 반드시 문서 내에서 자주 나타난다고 볼 수는 없으나, 중요한 내용에 해당하는 단어 자체만으로 문장을 구성할 수는 없으므로 반드시 그와 함께 자주 나타나는 단어들은 존재하고, 따라서 문서 내에서 자주 사용되는 단어일수록 그러한 중요한 단어와 같이 쓰일 확률이 높은 것으로 간주할 수 있다. 이는 자주 출현하는 단어들의 집합에선 추출된 개념클러스터내의 단어가 문서 내에서 기사 작성자가 표현하려는 내용과 밀접한 관계를 가지는 단어와 함께 출현할 확률이 높은 단어들로 간주될 수도 있음을 나타낸다.

본 논문에서는 추출된 Vertex들의 공기관계를 나타내기 위해서 Vertex들의 쌍을 만들어 공기도를 계산하고, 계산된 공기도가 유효한 경우 즉 문서에서 한번이상 동일한 문장에서 나타난 경우에만 Vertex간의 관계를 나타내는 Edge를 생성해준다. 공기도는 두 개의 단어가 대상 문서내의 동일한 문장에 대해 동시에 출현한 횟수의 합을 의미하고 식(2)와 같이 나타낸다.

$$\mathrm{Co}(W_i, W_j) = \sum_{S \subset D} |W_i, \| W_j | S \quad \cdots\cdots (2)$$

S : 대상 문서에 존재하는 문장
D : 대상 문서
$|W|s$는 문장 S에서 단어 W의 출현 빈도

일단 그래프를 생성하면 그래프 상의 단어들의 연결 관계를 분석하여 개념클러스터를 생성한다. 개념클러스터의 단어들은 문서에서 주장, 키워드의 후보단어가 된다. 각각의 개념클러스터는 문서에서 중요한 내용의 흐름을 나타내는 문장 혹은 문단에 속하는 단어들의 집합을 의미한다

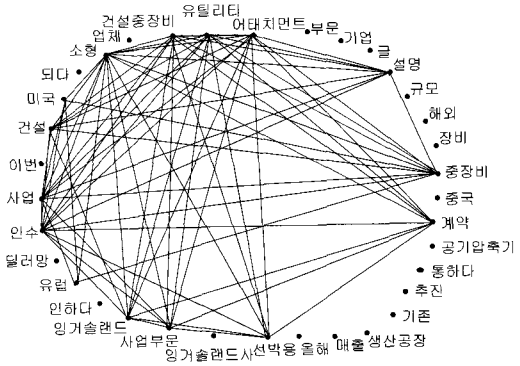

그림 6. 공기관계 그래프 예
Fig 6. Example of co-occurrence relation graph

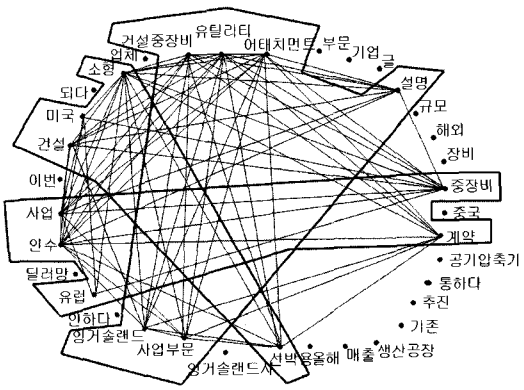

그림 7. 개념클러스터의 형성
Fig 7. Formation of mean cluster

개념클러스터는 그래프 상의 그림6에서 루프를 형성하는 단어들의 집합으로 표현된다. 개념클러스터를 찾기 위해서

우선 각각의 노드가 가지는 연결 노드 정보를 가지고 우선 노드 3개로 구성된 가장 간단한 루프(loop)의 형태들을 찾는다. 이러한 각각의 루프들은 서로 동일한 Edge를 가지는 루프들과 다시 합쳐 새로운 형태의 루프가 형성될 때까지 반복한다. 그림 7의 굵은 선으로 표시된 부분 중 돌출된 부분에 해당하는 단어들의 집합이 하나의 개념클러스터를 형성한다. 그림 7에는 총3개의 개념클러스터가 존재하며 각각의 개념 클러스터 별로 최적화한 결과는 그림 8과 같다.

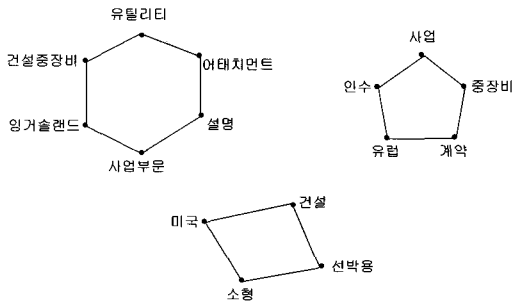

그림 8. 최적화된 클러스터의 집합
Fig 8. Set of optimized cluster

② 주장의 형성

주장은 문서 내에서 기사 작성자가 의도하는 내용의 흐름을 나타내는 단어로서 개념클러스터와 개념클러스터를 강하게 연결시켜주는 역할을 한다. 주장은 각각의 개념클러스터 상에서의 존재 여부에 관계없이 각각의 개념클러스터들과 강하게 연결된 단어를 나타낸다. 개념클러스터는 기사 작성자가 의도하는 중요한 내용의 일부로서 각각의 개념클러스터를 서로 연결 시켜준다는 것은 문서에서 중요한 내용들을 기사 작성자가 의도하는 내용상의 일괄된 흐름을 가지고 연결 시켜주는 것을 의미한다. 주장은 개념클러스터의 단어들과 강한 연결성 즉 높은 공기도를 가지는데 그림 7에서 링크의 밀집도가 높은 단어들이 이에 해당하며, 또한 두 개 이상의 개념클러스터내의 단어들에 대해서 동시 출현 확률이 높다. 주장을 계산하기 위한 함수 $Key(W)$는 임의의 단어 W가 각각의 개념클러스터에 속하는 단어들에 대해서 동시에 출현할 수 있는 확률을 식(3)과 같이 계산하며, W가 특정 개념클러스터에 해당할 경우에도 자신이 속한 개념클러스터에 대한 동시 출현 확률을 계산한다. 본 논문에서는 $Key(W)$값이 높은 순위부터 상위 12위까지의 단어를 주장으로서 선택하였다.

$$Key(W) = \{1 - \prod_{c}^{Cluster} (1 - f(w, c)/F(c))\} \quad \cdots\cdots\cdots\cdots (3)$$

$$F(c) = \sum_{s \subset D} \sum_{w \subset D} |C\text{-}W|s$$

$$f(w,c) = \sum_{s \subset D} |W|s\,|C\text{-}W|s$$

$$|C\text{-}W|s = |C|s - |W|s \quad if \quad W \subset C$$

$$= |C|s \quad if \quad W \not\subset C$$

$$|C|s = \sum_{X \subset C} |X|s$$

$Clusters$: 개념클러스터의 개수
$f(W,C)$: 단어 W와 개념클러스터 C와의 공기도
$F(C)$: 모든 단어와 모든 개념클러스터의 공기도의 합.
S : 문장 W : 단어 C : 개념클러스터
X : 개념클러스터 C에 포함되는 단어.
$|X|s$: 문장 S에 대한 개념클러스터 C에 포함되는 단어 X의 출현빈도

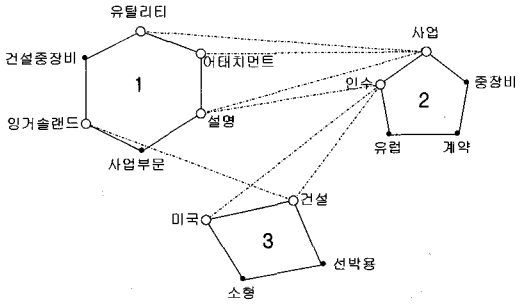

그림 9. 주장과 전개의 추출
Fig 9. Extraction of Insistence and deployment

그림 9는 $Key(W)$에 의하여 계산된 각각의 그래프상의 노드로부터 주장과 전개를 추출한 결과를 보여주고 있다. 그림 9에서 투명한 점으로 표시된 노드들이 주장에 해당하며, 각각의 점선 화살표들은 주장과 개념클러스터사이의 전개를 나타낸다.

③ 키워드 추출

키워드는 문서에서 기사 작성자의 주장을 강하게 뒷받침하거나 혹은 자세한 설명을 위해 쓰여진 단어로서 그래프 상에서 주장과 강하게 연결되어지는 개념클러스터내의 단어를 의미한다. 즉 주장과 함께 자주 출현하는 단어는 그만큼 기사 작성자의 의도를 설명하기 위해 자주 사용되었다는 것을 의미하며 이는 키워드로 간주될 수 있다. 키워드는 그래프

상에서 주장과의 공기도가 높은 개념클러스터내의 단어를 의미한다. 단, 어떤 주장1이 다른 주장2와 높은 공기도를 가지며, 주장1과 주장2가 모두 특정 개념클러스터내의 단어일 경우 주장1과 주장2는 우선적으로 키워드로서 간주된다.

키워드의 추출은 그래프를 이용해서 얻어진 주장중의 단어 W_i와 개념클러스터에 포함되는 단어 W_j사이의 강도를 계산하여 강도가 높은 단어를 키워드로 선택한다. 본 논문에서는 강도가 높은 순위부터 상위 12위까지의 단어를 키워드로 선택하였다. 주장과 개념클러스터의 강도 *INTENSITY(K)*는 식(4)와 같이 개념클러스터에 있는 임의의 단어 K가 각각의 주장과 가지는 강도의 합을 나타내며, 단어 K의 가중치로 간주된다.

$$Intensity(K, INS) = \sum_{s \subset D} |K \| INS| S$$

$$INTENSITY(K) = \sum_{i}^{Insistence} Intensity(K, INS_i) \cdots\cdots (4)$$

$$KeyWord = min(12, |Document|)$$

K : 각각의 개념클러스터에 존재하는 키워드 후보 단어
INS : 주장.
Insistence : 주장의 개수

앞의 식에서 나타난 바와 같이 키워드의 강도 *INTENSITY(k)*를 계산하는 과정에서 하나의 키워드 후보가 모든 주장과 가지는 공기도의 합을 키워드의 강도로서 취하는 이유는 하나의 개념클러스터에 대해 한 개 이상의 주장이 강하게 연결될 수 있기 때문이다. 그림 9에서 주장과 각각의 개념클러스터 사이의 강도를 계산한다. 추출한 키워드의 주장과 *Key(W)* 값을 표 2에 나타내었다.

표 2. 키워드 및 주장과의 강도
Table 2. Intensity of keyword and insistence

주장	Key(W)
유틸리티	0.012854
어태치먼트	0.012854
잉거솔랜드	0.010218
설명	0.005274
미국	0.020105
건설	0.008899
인수	0.024720
사업	0.016150

3.2.4 핵심 문장의 추출

추출된 키워드들은 핵심 문장 추출기의 후보 문장을 선정하는 기준이 된다. 핵심 문장은 우선 문장의 흐름을 대표할 수 있는 주장에 해당하는 단어를 포함하며, 주장과 강하게 연결되어지는 다른 키워드를 포함 할수록 해당 문서에서의 중요도가 높다. 즉 본 논문에서는 키워드를 포함하는 모든 문장을 우선 핵심 문장의 후보로서 추출한다. 이것을 식(5)와 같이 표현한다.

$$SK = sk_1 + sk_2 + sk_3 + sk_4 + \cdots + sk_n \cdots\cdots (5)$$
$$(SK \subset D)$$

SK : 키워드를 포함하는 문장의 집합.
sk_n : 키워드를 포함하는 문장.
D : 대상 문서

일단 키워드를 포함하는 문장들을 추출하고 나면 추출된 문장들 중에서 주장을 포함하는 문장을 식(6)과 같이 찾는다.

$$SI = si_1 + si_2 + si_3 + si_4 + \cdots + si_n \cdots\cdots (6)$$
$$(SI \subset SK)$$

SI : 키워드를 포함하면서 주장도 같이 포함하는 문장의 집합
si_n : 키워드를 포함하면서 주장도 같이 포함하는 문장

추출된 키워드와 주장을 모두 갖는 문장들을 추출한 후에는 추출된 문서 각각에 포함된 키워드들이 갖는 가중치의 합이 높은 문장에 중요도를 높게 부여하는 방식으로 핵심 문장을 추출한다. 문장의 중요도를 구하는 함수 *Importance(SI)*는 식(7)과 같다.

$$\text{Importance(SI)} = \sum_{k \in SI} INTENSITY(k)$$
K : Si에 포함된 키워드의 집합 ..... (7)

IV. 시스템 구현

본 논문에서 제시한 시스템은 크게 웹 뉴스 콘텐츠의 추출과 요약을 수행하는 두개의 시스템으로 구성되어 있다. 웹 뉴스 콘텐츠의 추출 시스템은 JAVA 1.4 플랫폼 기반이며, 요약 시스템은 Microsoft Win32 SDK 기반으로 제작되었다.

그림 10은 조선일보의 인터넷 신문인 www.chosun.com의 콘텐츠를 브라우저를 통해 본 화면이며, 그림 11은 웹 뉴스 콘텐츠 추출 시스템을 통해 추출된 기사이다.

그림 10. chosun.com의 브라우저
Fig 10. Browser of chosun.com

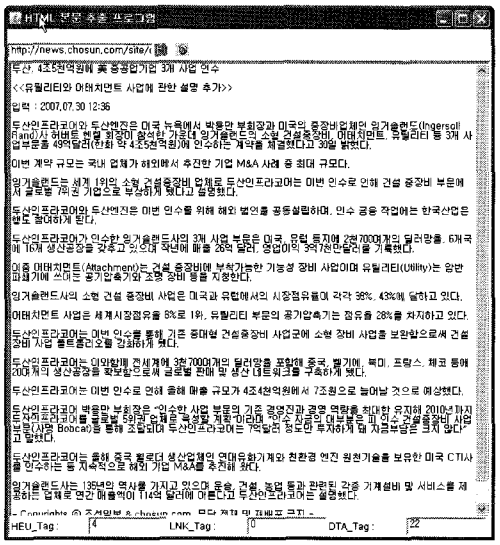

그림 11. 추출 시스템을 통해 추출된 뉴스
Fig 11. Extracted news through the extraction system

그림 10과 그림 11의 예에서 볼 수 있듯이 거의 대부분의 뉴스 사이트의 콘텐츠는 중앙의 텍스트를 제외하곤 모두 클러터로 간주할 수 있다.

뉴스 콘텐츠의 추출시스템을 통해 출력된 데이터는 요약 시스템으로 전달되어 핵심문장이 추출된다.

그림 12는 그림11에서 추출된 웹 뉴스를 입력받아 본 시스템을 사용하여 요약한 결과이다. 화면 하단좌측은 요약할 문서의 원문을 보여주며, 우측은 요약 시스템에 의해 추출된 핵심 문장들을 보여준다.

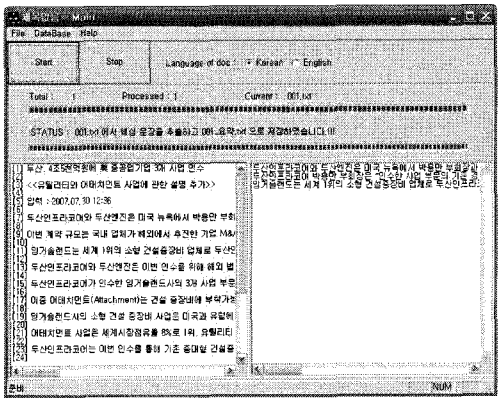

그림 12. 추출된 뉴스의 요약 결과
Fig 12. Summarized result of the extracted news

본 논문에서는 시스템이 추출한 핵심 문장을 수작업에 의한 결과와 비교하여 평가를 수행하였다. 평가 기준은 다음의 4개를 기준으로 하였다. 수작업에 의한 결과는 동일한 문서에 대해 두 명의 사람이 추출한 핵심 문장을 근거로 하였으며, 두 개의 수작업에 의한 추출 문장의 일치율은 48.6%였다.

- Optimistic: 각각의 수작업에 의한 결과와 두 개의 시스템의 결과를 비교하여 일치도가 높은 것을 선택한다.
- Pessimistic: 각각의 수작업에 의한 결과와 두 개의 시스템의 결과를 비교하여 일치도가 낮은 것을 선택한다.
- Union: 각각의 수작업에 의해 추출된 핵심 문장들 모두를 시스템의 결과와 비교한다.
- Intersection: 각각의 수작업에 의해 추출된 핵심 문장들 중 일치하는 것만을 추출하고 이를 시스템의 결과와 비교한다.

표 3. 수작업 결과와 시스템 결과의 비교
Table 3. Comparison manual and system result

Machting rate between two manual extracts : 48.6%				
	Opti. (%)	Pessi. (%)	Inter. (%)	Union (%)
공기그래프	49	31	50	54

표 3의 결과에서 Intersection 부분을 살펴볼 경우 두 개의 시스템 모두 수작업에 의해 추출한 경우와 비슷한 수준을 보이고 있음을 알 수 있다.

논문의 작성 시점에서 자바 플랫폼과 윈도우즈 플랫폼 간의 데이터 교환을 위한 시스템을 구성하지 못하였다. 따라서 현재 추출 시스템과 요약 시스템은 파일처리를 통해 이루어지고 있으며, 향후 과제로서 시스템을 구성할 계획이다.

V. 결론

뉴스 콘텐츠 등 웹을 통해 제공되는 많은 정보들은 불필요한 클러터를 많이 포함하고 있다. 이러한 클러터들은 문서의 요약, 추출, 검색과 같은 자동화된 정보처리 시스템의 구축을 어렵게 한다. 이러한 이유로 구조와 콘텐츠가 혼합된 HTML 문서의 처리를 위해 많은 연구가 이루어지고 있다.

본 논문에서는 이러한 클러터들을 효과적으로 제거하기위해 대부분의 클러터가 가진 HTML 태그의 하이퍼링크 속성에 관심을 두었다. HTML 파서가 구축한 요소 트리에서 링크속성을 가진 클러터를 제외하고 휴리스틱 규칙을 적용하여 하이퍼링크 속성을 가지지 않은 클러터를 제외 하였다. 이를 통해 뉴스 콘텐츠의 본문을 추출 한 후 이를 문서의 요약 시스템의 입력으로 사용하였다. 요약 시스템은 공기관계 그래프를 통해 구성되었으며, 요약된 핵심 문장은 SMS와 같은 메시지 서비스를 통하여 PDA와 모바일 폰 등에 전송된다.

현재의 문서 추출 시스템은 두개의 휴리스틱이 적용되어 동작한다. 그러나 많은 뉴스 사이트의 뉴스 콘텐츠를 처리하기 위해서는 더 많은 휴리스틱 규칙이 필요하며, 각종 스크립트와 문법에 맞지 않는 HTML 문서 등은 HTML 파서의 성능에도 큰 영향을 미치게 된다.

현재 구현된 시스템은 자바 기반의 태그정보 기반 텍스트 추출 시스템과 마이크로소프트 윈도 기반의 문서요약 시스템으로 이루어져 있다. 현재 두개의 어플리케이션으로 동작하고 있으나, 향후 하나의 시스템으로 통합하여 시스템의 복잡도를 줄여야 할 것이다. 또한 현재 이슈가 되고 있는 시멘틱웹이나 웹2.0과 관련해서 RDF와 RSS 등의 다양한 기술을 논문의 시스템에 접목하여 보다 지능적인 정보검색 시스템을 구축하고자 한다.

참고문헌

[1] Suhit Gupta, Gail Kaiser, David Neistadt and Peter Grimm, "DOM-based Content Extraction of HTML Documents" Proc. Int'l Conf. World-Wide Web Conf., 2003.

[2] Bok Keun Sun, Je Ryu and Kwang Rok han, "Structure Detection System from Web Documents through Backpropagation Network Learning", Proc. Int'l Conf. of Advances in Artificial Intelligence, pp.1281-1287, 2006.

[3] N. Ashish and C.A. Knoblock, "Semi-Automatic Wrapper Generation for Internet Information Sources," Proc. Int'l Conf. Cooperative Information System, pp.160-169, 1997.

[4] K. Papadakis, D Skoutas, K Raftopoulos and A. Varvarigou, "STAVIES: A System for Information Extraction from Unknown Web Data Sources through Automatic Web Wrapper Generation Using Clustering Techniques," IEEE Transactions on Knowledge and Data Engineering, Vol 17, No. 12, pp1638-1652, 2005.12

[5] Ryu Je, Han Kwang Rok, "A Study on Automatic Document Summarization using Graph Combining Method", Proc. Int'l Conf. of ISMIS, 2004.

[6] K.R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, M.Y. Kan, B. Schiffman and S. Teufel. "Columbia Multi-document Summarization: Approach and Evaluation", In Document Understanding Conf., 2001.

[7] A. F. R. Rahman, H. Alam and R. Hartono. "Content Extraction from HTML Documents", In 1st Int. Workshop on Web Document Analysis (WDA2001), 2001.

[8] http://www.openxml.org/

[9] 류제, 한광록, 손석원, 임기욱, "단어의 공기 관계 그래프를 이용한 문서의 핵심문장 추출에 관한 연구", 한국정보처리학회논문지, Vol.7, No.11, pp.3427-3437, 2000.

[10] Yukio Ohsawa, Nels E. Benson, Masahiko Yachida, "Automatic Indexing by Segmentation and Unifing Co-occurrence Graphs", IEICE, D-1 Vol. J82-D-I, No.2, pp391-400, 1992.

저 자 소 개

한광록

1989년 8월 : 인하대대학교 정보전공 공학박사
1991년 ~현재 : 호서대학교 교수
관심분야 : 정보검색, HCI, 멀티미디어, 시멘틱웹

선복근

2006년 2월 : 호서대학교 컴퓨터공학과 공학박사
2005년 ~ 현재 : 호서대학교 컴퓨터공학과 프로그램전문강사
관심분야 : HCI, 시멘틱웹

유형선

1996년 2월 : 고려대학교 문학박사
1998년~현재 : 순천향대학교 부교수
관심분야 : 형태소분석, 문헌분류, 온토로지