ONASys: 온라인 뉴스기사의 재구성

최주원, 나종열, 최동현, 여명숙, 신지애*, 최기선
한국과학기술원 전자전산학과 전산학전공
*한국정보통신대학교 공학부

ONASys: Online News Adaptation System

Joo-Won Choi, Dong-Hyun Choi, Jong-Ryul Nah, Myeung-Sook Yoh, Ji-Ae Shin, Key-Sun Choi
Division of Computer Science, KAIST

요 약

웹에는 방대한 양의 정보가 있지만 현재 그 활용은 제한적이다. 본 연구는 기존에 PC를 통해서만 접근 가능했던 웹 정보를 TV나 모바일 기기를 통해서 볼 수 있게 함으로써, 사용자로 하여금 시공간의 제약으로부터 자유로운 정보 환경을 제공하려는 취지에서 수행되었다. 현재의 웹 정보를 TV나 모바일 기기와 같은 일상적인 매체를 통해서도 활용할 수 있기 위해서는 각 디바이스의 화면에 적합하게 필터링과 요약을 통해 재구성 하는 기술이 요구된다. 본 논문에서는 자연어처리 기술에 기반하여 온라인 뉴스 페이지를 개인 디바이스에 적합한 형태로 내용과 레이아웃을 재구성하는 시스템인 ONASys(Online News Adaptation System) 에 대해 소개한다. ONASys는 공간적, 내용적 속성을 이용하여 웹페이지 상의 중요 정보를 선별하는 '컨텐츠 필터링 모듈', 요약 패턴의 분석과 공기 정보를 통한 문법적 주 요소 인식을 통해 문장을 요약하는 '문장 요약 모듈', 제목과 문장의 연관성을 통하여 문장의 중요도를 판단하는 '중요도 결정 모듈'로 구성된다.

1. 서 론

주지하다시피 웹은 정보의 바다이지만 그 많은 정보를 유의미하게 활용하는 데는 많은 제약이 따른다. 그 이유는 현재의 웹구조 자체에서 찾을 수도 있지만, 또 다른 중요 원인은 현재 웹 정보에 접근하는데 필요한 디바이스가 극히 제한되어 있다는 사실에서 찾아볼 수 있다. 본 논문의 공학적 관심은 바로 후자의 문제를 완화시키는데 초점 맞추고 있다. 특정 웹 정보를 PC로부터 가져오지 않고서도 직접 TV나 휴대폰을 통해서 볼 수 있다면, 사람들은 정보검색에 소요되는 시간과 비용을 대폭 줄일 수 있게 될 것이다.

예컨대, 가수 '비'가 나오는 TV의 가요프로를 보다가 갑작스레 '비'의 미국 콘서트 정보를 알고 싶어진 경우를 생각해보자. 이때 다른 채널에서 관련된 영상을 방영해 주고 있지 않다면 사람들은 PC를 켜고 인터넷으로 뉴스 검색을 시작할 것이다. 하지만 이는 매우 번거로운 일이다. 만약 인터넷에 떠도는 '비'에 대한 정보를 직접 TV로 가져 올 수 있다면 굳이 컴퓨터로 검색하지 않고도 큰 화면으로 원하는 영상을 감상할 수 있을 것이다. 마찬가지로 모바일 디바이스로도 사용자가 보고 싶은 인터넷상의 정보를 원하는 즉시 찾아볼 수 있다면 특정 시공간에서 사용자가 누릴 수 있는 행동의 자유도는 더 높아질 것이다.

이를 가능케 하는 시스템이 필수적으로 갖추어야 하는 기능은 검색한 웹 컨텐츠의 내용을 TV나 모바일 디바이스에 적합한 형태로 재구성하는 기능이다. 그림 1의 왼쪽 사진에서 보듯, 대부분의 웹페이지는 특정 기사와 관련 없는 메뉴 및 다른 기사로의 링크 등을 많이 가지고 있다. 또한 기사의 본문이 너무 길어서 이를 그대로 출력하면 시청자가 그것을 한 눈에 알아 보기 어렵다. 이러한 상황을 개선하여 오른쪽 사진과 같이 사용자가 원하는 컨텐츠만을 뽑아내어 한 눈에 알아 볼 수 있도록 TV와 모바일 디바이스 등에 보여주는 것이 바람직하다.

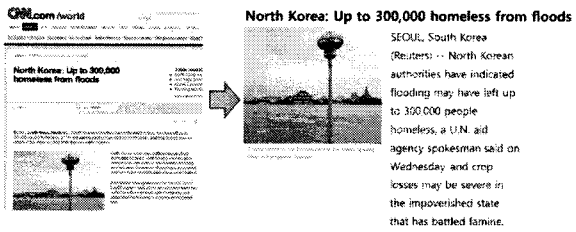

그림 1. 온라인 뉴스기사와 재구성 된 뉴스기사

본 논문에서는 이처럼 PC이외의 디스플레이 가능한 각종 디바이스에 적용될 수 있도록 온라인 뉴스를 재구성하는 새로운 시스템을 제안하며, 이를 ONASys(Online News Adaptation System)라고 부른다. ONASys는 크게 세가지 모듈로 이루어져 있다. 첫째, 컨텐츠 필터링 모듈은 각 뉴스 페이지의 기사와 관련 없는 메뉴나 광고 등을 제거하는 기능을 한다. 이 모듈을 통해 기사의 주제와 직접적 관련 있는 페이지의 일부가 남게 된다. 다음의 두 단계인, 컨텐츠 요약 모듈과 중요도 결정 모듈에서는 기사 본문의 문장을 요약하고 각 문장의 중요도를 결정함으로써 TV 등의 각종 디바이스에서 보여주기 적합한 최소화된 기사를 만들어낸다. ONASys의 모든 과정을 거치게 되면 그림1의 오른쪽처럼 재구성된 페이지를 얻을 수 있다.

뉴스기사의 재구성을 위하여 ONASys의 각 모듈에서는 자연어 처리를 이용하였다. 컨텐츠 필터링 모듈에서는 기사의 주제와 관련된 정보를 추출하는 과정에서 단어의 유사도를 이용하였고, 컨텐츠 요약 모듈은 기존의 문장 요약(gist)의 한 방법을 바탕으로 그것을 발전시켰다. 끝으로 중요도 결정 모듈에서는 새로운 기사 제목, 4W[1] 등을 참조하여 문장의 중요도를 결정하는 새로운 방식을 제안하였다.

처음의 연구목표 달성과는 별도로 본 연구를 진행하는 과정에 얻은 예상외의 수확은 ONASys를 일반 웹검색에 활용할 때의 가치이다. ONASys를 웹크롤러에 적용할 경우, 각 뉴스 페이지마다 재구성된 페이지를 얻을 수 있는데, 이는 각 기사의 주제와 관련된 정보만 담고 있어서 효율적인 검색환경을 제공할 수 있을 것이다. 만약 재구성된 페이지를 이용한 검색을 할 경우 기사와 관련된 정보만을 참조하여 매칭하기 때문에 사용자는 일반 페이지를 텍스트 매칭하는 것보다 좀 더 정밀한 결과를 예상해 볼 수 있다.

2. 관련 연구

2.1. 웹페이지 컨텐츠 필터링

웹페이지 내용을 IPTV, PDA와 같은 다른 디바이스로 적합하게 재가공 하기 위해서는 우선 의미블락[2]을 찾아내어 그 의미블락들간의 상대적인 중요도를 판단하는 것이 필요하다[1]. 의미블락의 중요도는 각 의미블락이 해당 웹페이지의 주제와 연관성이 얼마나 있느냐에 따라 정한다. 예를 들어, 어떤 기사에 대한 웹페이지가 있다고 하자. 해당 기사의 제목과 본문 및 기사 관련 멀티미디어에 해당하는 의미블락은 가장 중요도가 높다. 또한 기사 작성일, 기자 이름 등 간접적으로 관련된 부분은 그 다음으로 중요도가 높다. 광고, 로그인 박스 등의 기사 주제와 관련이 전혀 없는 부분은 가장 중요도가 낮다.

의미블락을 나누는 잘 알려진 방법 중에 하나는 HTML 태그, 배경색, 텍스트 길이와 같은 페이지의 표면적 구조 정보에 근거하여 의미블락을 정하는 것이다[2].

정해진 각 의미블락들 간의 중요도를 정하는 방법들 중에서 이 논문에서는 같은 의미블락이 반복되는 횟수가 높을수록 중요도가 낮다고 가정을 근거에 기초한다[4]. 한편, 의미블락의 공간적 속성과 링크의 개수와 같은 내용적 속성들이 의미블락의 중요도를 구하는 데 사용되었다[5].

본 시스템에서는 이런 공간적 속성과 내용적 속성을 사용하여 SVM을 학습시켰다.

2.2. 문장 요약

문장 요약과 관련된 연구로는 비주요 문법 요소[3] 에 대해서만 단어의 빈도수에 근거하여 요약하는 방법[6]과 구문 트리의 서브 트리가 변화할 수 있는 방법을 이용한 요약 방법[7]이 있다. 또한 [7]의 방법에 개체명 인식(named entity) 정보와 하위범주화(subcategorization) 정보를 추가한 시도가 있다[8]. 본 시스템에서는 [7]의 방법을 확장하여 비주요 문법 요소를 요약하는 데 사용하였다.

2.3. 주제 관련 문장 선택

Goldstein[9]은 코사인 유사도, TF-IDF 가중치 및 문장의 위치 등의 통계적 특성(statistical features)와 인용문, 경어적 표현, 품사 등의 언어학적 특성(linguistic features)에 근거한 방법을, Hovy[10]는 의미의 다양성이 제한되는 단어들의 쌍(전치사, 동사 또는 명사)을 이용한 방법을 제시 하였다. Metzler[12]는 의존관계에 있는 명사들이나 형용사로 수식이 되는 명사 등의 독립적인 명사구 단위들을 사용하여 정보량을 판단하는 방법을 제시하였다.

본 연구에서는 의존문법으로 분석된 제목의 구문트리를 인접행렬(adjacency matrix)을 이용하여 단어 집합의 주제와의 연관성을 판단한다. 또한 행렬의 정보로부터 4W와 관련된 의존성을 띄는 단어가 연관성 판단에 반영한다.

3. ONASys: 시스템 구조

시스템의 전체 구조는 그림 2 의 모듈로 구성된다.

컨텐츠 필터링 모듈

각 페이지의 주제와 관련된 의미블락을 추출한다.

컨텐츠 요약 모듈

문장 단위로 핵심내용을 찾아낸다.

중요도 결정 모듈

기사 제목과 4W 에 근거하여 문장의 중요도를 결정한다.

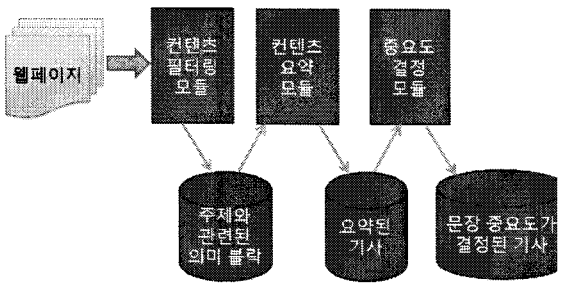

그림 2. ONASys 시스템 구조

3.1. 웹페이지 컨텐츠 필터링 모듈

이 모듈은 세가지 기능을 수행한다.

첫째, RSS 를 통해 실시간으로 저장되는 뉴스 페이지를 의미블락 단위로 나누는 기능이다. 이것은 VIPS[2]를 사용하여 각 페이지의 의미블락을 정한다.

둘째, 의미블락의 중요도는 공간적 속성과 내용적 속성에 의해 정해진다. 공간적 속성에는 의미블락의 위치와 크기가 있다. 내용적 속성에는 순수 텍스트의 길이, 폰트, 링크의 개수, 멀티미디어의 개수, Table 의 개수 등이 있다. 이러한

[1] 4W: Who, When, Where, What

[2] 의미블락: 웹페이지의 일부분으로 비슷한 기능을 하는 원소(element)의 집합이다.

[3] 주요 문법 요소: 문장 이해를 위하여 필수적인 문법적 요소로 문법적 주 요소와 같은 의미. 비주요 문법 요소는 이와 반대되는 의미이다.

속성에 근거하여, SVM 을 학습시킨다. 웹페이지 컨텐츠 필터링 모듈은 한 페이지 내에서 굳이 보여줄 필요가 없는 부분을 제거하여 중요한 정보만 남도록 하는 과정이다. 자세한 과정은 그림 2 와 같다. 실시간으로 모이는 뉴스 페이지를 VIPS 를 통해 작은 단위의 의미블락으로 나눈다[2]. 그 다음 각각의 의미블락에서 속성을 추출하게 되며, 그 속성을 추출한 결과를 SVM 에 돌려 각 의미블락의 중요도를 얻게 된다[5]. 중요도가 정해진 각각의 의미블락에 대해 중요도가 낮은 것은 버리고 높은 것은 비슷한 의미를 가지는 의미블락끼리 묶게 되면 전체 과정이 마무리 된다.

셋째, VIPS 는 표면적 구조 정보에 근거하여 의미블락을 나누기 때문에 사이트에 따라서는 의미블락의 크기가 적절하게 나누어 지지 않을 수도 있다. 이 문제를 본 시스템에서는 각 의미블락의 주제와의 연관성과 의미블락의 분류를 이용하여 해결하였다. 이 연관성은 각 의미블락에 나타나는 명사들과 기사 제목의 의미적 유사도를 워드넷을 참조하여 결정하였다. 또한 각 의미블락을 제목, 본문, 멀티미디어 블락으로 분류 하였다. 연관성이 높은 의미블락 중에 같은 분류에 속하는 것들은 하나의 의미블락으로 재구성 된다.

3.2. 컨텐츠 요약 모듈

이 모듈은 두 가지 단계로 구성된다.

첫째, 코퍼스를 사용하여 요약에 사용될 구문 트리 변화 규칙을 찾는다. 이 코퍼스는 각 문장과 그것의 요약된 문장으로 구성되어 있다. 구문분석된 문장들을 이용하여 원래 문장에서 요약된 구조를 만들기 위한 규칙들을 찾는다[7].

규칙들은 다음 세 가지 변환을 나타낸다.

REDUCE: 요약된 내용에 포함된 단어들 또는 부분 구문 분석 트리를 묶어 새로운 구문 분석 트리를 만든다.

SHIFT: 해당 단어가 요약된 내용에 포함된다.

DROP: 해당 단어가 요약된 내용에 포함되지 않는다.

이 규칙은 다음과 같은 한계점을 가지고 있다.

- 동사의 공기 정보를 포함하고 있지 않다. 예를 들어 'expect'는 주로 to-부정사와 사용되지만, 일반적으로 동사는 이를 필요치 않는다. 따라서 문법적으로 맞지 않는 변환들이 만들어지게 된다.
- 코퍼스를 이용한 방법의 한계로서, 이 규칙들이 모든 가능한 요약 방식을 찾을 수 없다.

둘째, 이 규칙들을 사용하여 문장을 요약한다. 규칙을 적용할 때 공기 정보를 포함하여 위의 한계점들을 극복한다.

- 주요 문법 요소를 표시한다. 만약 이러한 부분들에 대해서 DROP 명령어가 실행되어야 한다면, 해당 단어를 강제로 SHIFT시킨다.
- 어떤 동사가 요약된 문장에 포함되면(SHIFT), 워드넷에서 추출한 그 동사의 공기 정보를 이용하여 필요한 부분을 표시한다.
- What, Which와 같이 육하원칙에 해당하는 단어가 DROP되는 경우와 관사의 품사를 가지는 단어가 DROP된다면, 해당 구는 요약에서 포함되지 않는다.

3.3. 중요도 결정 모듈

각 문장의 주제와의 관련성의 정도의 결정은 주제와 관련되는 단어들의 집합을 구하는 단계와 문장들의 정보 함유량을 판단하는 두 단계를 거친다.

첫째, 뉴스제목에서 VNA[4] 집합을 구하고 각 VNA원소들의 패턴이 나타나는 본문 문장에 가중치 점수를 부여한다. VNA 집합은 동사, 명사, 형용사 및 부사의 묶음이 원소로 이루어진 집합이며 중요도에 따라 원소의 레벨이 나뉘어 진다. VNA집합을 구하기 위해 뉴스 제목을 의존문법 파서로 분석하여 얻은 의존함수들을 근거로 인접행렬을 생성한다. 그림 3(a)는 제목의 파싱 결과이며 이를 통해 의존함수들이 인접행렬을 표현한다. 그림3(b)의 인접행렬에 표현된 내용은 다음과 같다.

- $A := (a_{i,j})_{n \times n}$, n 은 제목의 토큰 갯수
- $a_{i,j} = k$, 토큰 i가 토큰 j에게 k의 의존성을 갖음
- 토큰 m이 root이면 $a_{m,m} = *$
- 토큰 m이 문장부호 이면 $a_{m,m}$ = 문장부호

본 연구에서는 기존의 의존문법 분석이 트리의 너비우선탐색 (BFS)으로 되어 온 것을 시간적으로 개선하는 방법을 제시한다.

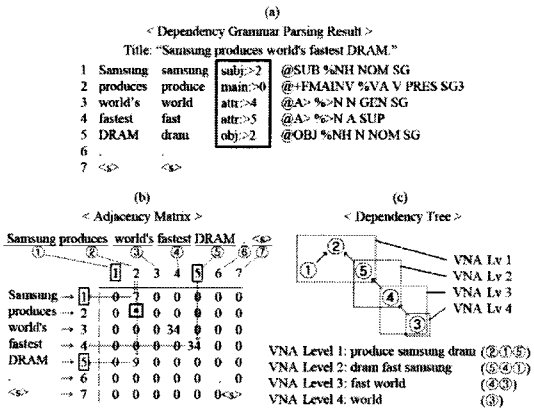

그림 3. 중요도 결정 모듈

(a): 각 단어의 토큰번호, 원형, 의존함수, 형태소 정보

(b): 의존함수에 의해 생성된 인접행렬(adjacency matrix)

(c): VNA원소들에 대한 위상적 표현

- 루트로 지정된 단어(그림 3(b)의 토큰번호②)는 제목문장의 중심 동사이므로 루트에 의존하는 단어와 함께 핵심주제를 표현한다
- 핵심 주제를 찾기 위해 그림 3(b)와 같이 루트의 토큰번호 열에 있는 단어(토큰번호 ①②⑤)들을 찾으면

[4] VNA: 동사, 명사, 형용사

이 원소들이 VNA Level 1이 되며 제목의 주요 내용을 나타낸다. ("produces Samsung DRAM")

- 핵심주제와 관련된 내용에 대한 문장들을 찾기 위해 ①번 열과 ⑤번 열에 존재하는 단어 토큰번호들(④)을 찾음으로써 VNA Level 2를 구한다. ("Samsung DRAM fastest")
- 의존관계가 끝날 때까지 재귀적으로 주제와 관련된 단어들을 찾으면 VNA집합의 모든 원소들을 구하게 된다. (그림 3(c))

의존관계가 끝날 때까지 재귀적으로 주제와 관련된 단어들을 찾으면 그림 3(c)와 같이 VNA집합의 모든 원소들을 구하게 된다. VNA집합을 모두 구한 다음 각 문장마다 주제 관련성 가중치를 부여한다. VNA Level 1의 단어들이 모두 들어가 있는 문장에 가장 높은 점수를 부여하고 레벨이 높아지면서 점수를 낮추어 가중치를 부여하면 모든 문장의 중요도를 결정 할 수 있다.

둘째, 뉴스본문의 문장에서 when, where, why, who와 관련된 의존형태를 나타내는 단어들을 찾아서 해당 점수를 문장의 주제 연관성의 판단에 반영한다. 뉴스는 시간, 장소, 인물, 사건을 중심으로 어떻게 또는 왜 일어났는지에 대한 설명을 하고 있다. 그러므로 시간, 장소, 인물, 사건에 대한 단어가 많은 문장이 정보 응집력이 크다. 한 단어가 전치사구의 의존성을 보이면 주체에 대한 보완이 대부분이므로 when, where의 정보가 따를 확률이 높아진다. 주어의 의존성을 보이면 동사에 연관된 주어이므로 who의 정보를, 목적어의 의존성을 보이는 단어는 타동사와 연관이 있으므로 what의 내용을 담고 있다고 판단될 수 있다.

4. 실험결과 및 분석

4.1. 웹페이지 컨텐츠 필터링 모듈

SVM은 다음과 같은 학습용 집합을 이용하여 학습되었다.

속성(Features): 3.1 에서 언급한 속성들.

클래스(Class)

Class 1: 메뉴바, 헤더, 하단의 저작권 등 모든 페이지에 반복되는 정보

Class 2: 최신뉴스, 인기 있는 뉴스 등 대부분의 페이지에 반복되나 가치 있을 수 있는 정보

Class 3: 저자, 날짜, 태그 등 기사와 간접적으로 관련이 있는 자료

Class 4: 기사의 제목, 본문, 관련된 이미지나 영상

학습용 집합은 ZDNet[5]뉴스의 300 개의 기사가 사용되었고, 테스트 집합으로는 ZDNet 뉴스와 CNN[6] Tech 뉴스의 각 100 페이지가 SVM 이 구한 의미블락의 중요도를 평가하기 위해 사용되었다. 학습, 테스트 집합은 같은 의미블락이 반복되는 횟수가 높을수록 중요도가 낮다는 가정에 근거하여 만든 후 사람의 확인 절차를 거쳤다. SVM 프로그램은 libsvm[7]을 사용하였고, SVM Type 은 C_SVC, Kernel Type 은 RBF 가 사용되었다.

표 1. SVM으로 구한 의미블락 중요도의 평가

	ZDNet News	CNN TECH
SVM 으로 구한 중요도와 테스트 집합과의 일치도	94.5194%	82.1341%

표 1 에 따르면 ZDNet 에 대해서만 학습하였지만, CNN 페이지에 대해서도 높은 확률로 중요도를 예측 하는 것을 알 수 있다. 이는 뉴스 페이지의 구성이 대체로 비슷하기 때문이다.

4.2. 컨텐츠 요약 모듈

3.2에서 서술한 방법은 결정 트리(decision tree)로 학습되었다. Ziff-Davis 코퍼스[8]가 학습을 위하여 사용되었고, Korean Herald에서 무작위로 가져온 996개의 문서가 테스트를 위해 사용되었다[9]. 또한 스탠포드 구문 분석기[10]가 사용되었다.

표 2. 실험결과 비교: 압출율, 중요도, 문법성

	압축율	중요도	문법성
방법1	55.03±23.58%	5.16±1.99	6.35±1.92
방법2	73.13±27.65%	6.86±2.69	7.41±2.24
사람	62.91±19.65%	6.94±1.22	8.21±0.94

방법1은 [8]의 결과이고, 방법2는 본 모듈의 결과이다. 본 모듈의 요약 결과물은 압출율은 비교적 낮으나, 기존 알고리즘에 비해 문법적으로 올바르며 원문의 내용을 사람이 요약한 것과 비슷한 수준으로 담고 있는 것을 알 수 있다.[13]

표 3. 문장 요약 결과

Original: However , **SK Telecom** , the nation ` s largest mobile carrier with a 52 percent market share , and thus subject to the government ` s plan , **expressed skepticism .**
Reduced: SK Telecom expressed skepticism
Original: **The two mobile carriers are expected to have the necessary infrastructure ready by October** this year .
Decision-Tree Only : The two mobile carriers are expected
Reduced : The two mobile carriers are expected to have the necessary infrastructure ready by October

[5] http://www.zdnet.com/

[6] http://www.cnn.com/TECH/

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[8] [7] 및 [8]에서 사용되었다.

[9] 현재 Marcu[2002] 및 Nguyen[2004]의 결과와 비교하기 위해 저자와 연락 중이며, 최종본에는 비교 결과를 넣을 수 있을 것이다.

[10] http://www-nlp.stanford.edu/downloads/lex-parser.shtml

Original : As the Science Ministry is involved in R&D when it comes to policies on climate change , **the ministry will** also **establish a climate change R&D consultation group** , possibly **headed by Science Minister Kim Woo-sik** , according to ministry official Kim Sang-jun .

Reduced : the ministry will establish a climate change R&D consultation group headed by Science Minister Kim Woo-sik

첫 번째는 요약이 잘 된 예이다. 두 번째 문장은 동사에 대한 공기 정보를 사용했기 때문에 주요 문법 요소가 삭제되지 않은 경우이다. 세 번째는 문법상으로는 요약이 잘 되었으나, 'the ministry'가 지시하는 바가 'the Science Ministry'라는 것을 알지 못하기 때문에 요약된 문장의 뜻이 달라진다

4.3. 문장의 주제와의 연관성 판단

총 31명에게 뉴스기사를 2개씩 제공하고 제목과 관련이 높은 문장을 선택하도록 설문조사를 실시했었다. 문장의 중요성 판단 모듈의 성능을 평가 하기 위해 설문조사에서 선택된 중요문장들과 문장 중요도 결정 모듈의 선택문장들을 비교한 결과 평균 68%의 일치도를 보였다. VNA집합 단어들의 동의어나 머리글자에 의한 단어를 인식 못하기 때문에 성능 저하를 보였다.

5. 결론 및 향후 연구

본 논문에서는 각종 디바이스로 웹 컨텐츠를 보여주기 위해 웹페이지를 재구성하는 시스템을 새로이 구축했다. 컨텐츠 필터링 모듈은 웹페이지의 공간적 속성, 내용적 속성을 바탕으로 SVM 을 통해 학습하는 방식을 취하였다. 문장 요약 모듈은 요약 패턴의 학습과 공기 정보를 통한 주요 문법 요소의 인식을 통해 요약된 문장을 만든다. 중요도 결정 모듈은 제목의 단어들과 4W 와 관련된 단어에 대하여 각 문장의 의존성을 분석함으로써 중요도를 결정한다. 세 모듈은 서로 독립적으로 작동하며, 사이트의 특성에 따라서는 모든 모듈이 작동될 필요가 없는 경우와 각 모듈의 순서가 바뀌면 좋은 경우가 있다. 이는 적용 프로그램에 따라 조정 가능하다.

또한 뉴스가 아닌 일반 웹사이트로의 확장 가능성을 생각해 볼 수 있다. 컨텐츠 필터링 모듈의 경우 아직은 학습량이 충분하지 못하기 때문에 학습용 집합의 도메인을 벗어나는 부분은 10% 정도 떨어지는 결과가 나오나 여러 도메인의 웹페이지 학습과 SVM 매개변수 조정을 통해 각종 도메인에 적용 가능한 결과를 얻을 수 있을 것이다. 다만 포탈의 메인 페이지와 같이 한 페이지에 여러 가지 주제가 존재하는 경우에는 컨텐츠 필터링 모듈이 잘 작동하지 않는다. 그러나 굳이 포탈의 메인 페이지와 같은 복잡한 페이지를 재구성 할 필요 없이, 하나의 주제를 가지는 페이지들만 ONASys 를 적용 한 후 이들의 연관성을 만들고 이를 각 디바이스에 적합한 형태로 재구성하여 사용자들이 한 눈에 볼 수 있는 결과물을 만들 수 있다. 문장 요약 모듈과 중요도 추출 모듈은 일반적인 텍스트에 적용 시킬 수 있는 모듈이므로 확장가능성은 충분하다. 다만 문장 요약 모듈의 경우 지시사가 무엇을 가리키는 지 찾아내고, 주요 문법 구조를 정밀하게 판별함으로써 요약 결과를 향상시킬 수 있을 것이다. 중요도 추출 모듈은 워드넷을 이용하여 VNA 원소들의 동의어와 및 약어를 찾아냄으로써 VNA 집합을 확장시키면 더욱 나은 결과를 얻을 수 있을 것이다.

이렇게 세 가지 모듈을 향상시키면 하나의 주제를 가지는 웹페이지의 경우에는 각 페이지의 주제와 연관된 중요도가 높은 부분만 추출할 수 있게 되어 재구성 여부에 따라서 어떤 디바이스에서도 적합한 형태로 보여줄 수 있을 것이다.

향후 연구: 온톨로지를 이용한 기사간의 연관성 생성

ONASys 를 각종 디바이스에 적용하기 위해선 재구성 된 페이지들 중에 무엇을 사용자에게 보여줄 지 결정해야 한다. 서론의 예와 같이 '마이클 잭슨'의 영상을 원하는 경우에는 '마이클 잭슨'의 특정 영상을 찾아서 보여줘야 한다. 하지만 단순 텍스트 매칭을 사용할 경우에는 실제로는 '마이클 잭슨'의 영상이 아닌데, 내용 중에 '마이클 잭슨'이라는 단어가 포함 되어 검색 결과로 선택 될 수 있다. 개인 컴퓨터의 경우에는 많은 결과를 보여주고 사용자가 선택할 수 있기 때문에 이런 경우가 크게 문제되지 않으나, 단지 몇 개의 결과만 출력할 수 있는 TV 나 모바일 디바이스의 경우에는 문제가 될 수 있다. 이를 해결하기 위해선, 각 기사의 중요한 단어를 찾아 그 단어만 검색에 이용할 경우 잘못된 결과를 줄일 수 있다. 하지만 이 경우에는 너무 검색되는 정보가 적을 수 있다. 그래서 향후 연구에서는 두 가지 단점을 모두 보완하기 위해, 하나의 기사에서 여러 가지 속성을 추출하여 이를 온톨로지와 매핑함으로써 다각적인 검색을 가능하게 하고, 온톨로지 추론을 통한 검색을 이용하여 질문에서 원하는 답을 정확히 가져오는 것을 목표로 한다. 여기서 속성이라 함은 다음 예제의 것들을 말한다.

주체: 삼성, LG, SK 등
사건: 신제품 개발, 정전, 주가 급등 등
시간: 2007/8/16 등
장소: 서울, 뉴욕 등

주체, 사건, 장소 등을 온톨로지와 그림 5 와 같이 매핑할 수 있다. 그리고 "삼성에서 일어난 정전사고"에 관한 기사를 찾고자 하면 주체 온톨로지의 '삼성'과 사건 온톨로지의 '정전'과 매핑되어 있는 기사를 모두 검색하면 된다. 또한 "2007 년 8 월 6 일 주가가 상승한 기업"을 찾으려면 시간, 사건, 주체가 각각 2007/08/06, 주가상승, 기업과 매핑되어 있는 기사를 검색하면 된다.

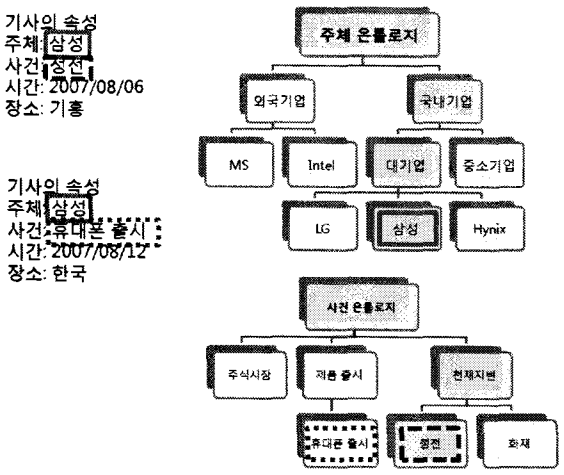

그림 4. 기사의 속성과 온톨로지의 매핑

이 방식을 따르면 원하는 것과 관련 없는 결과는 모두 제외하고 검색할 수 있다. 또한 이를 인터넷 검색에 적용하여도 위에서 말한 정확성과 다각적 각도의 검색이라는 두 가지 장점을 그대로 가질 수 있다. 단지 이를 구현함에 있어서 어려운 것은 웹페이지에서 사건 등의 속성을 추출하는 것이다. 하지만 꼭 사건이 아니라 기사가 쓰인 날짜와 행위 주체 등의 쉽게 얻을 수 있는 속성만 이용해도 행위 주체에 따라 시간순서로 사건의 흐름을 보여 줄 수 있기 때문에 사용자에게 양질의 검색결과를 제공할 수 있다.

그리고 모든 웹페이지에 대해 속성을 매칭시키려면 정치, 경제, 스포츠 등 각종 분야의 온톨로지가 필요하다는 문제점이 있다. 그러나 온톨로지는 전세계적으로 구축되고 있는 상황이고 각 응용분야에 따라 적합한 온톨로지를 찾아 매핑 시키면 원하는 검색결과를 찾아낼 수 있을 것이다.

감사의 글

본 논문은 정통부 및 정보통신연구진흥원의 정보통신선도기반기술개발사업의 연구결과로 수행되었습니다.

참 고 문 헌

[1] Xing Xie, Chong Wang, Li-Qun Chen, Wei-Ying Ma, "*An adaptive web page layout structure for small devices*", Multimedia Systems, Volume 11, Number 1, pp. 34-44, November 2005.

[2] Deng Cai, Shipeng Yu, Ji-Rong Wen, Wei-Ying Ma, "*VIPS: a Vision-based Page Segmentation Algorithm*", Microsoft Technical Report, MSR-TR-2003-79, pp. 28, November 2003.

[3] Nanno Tomoyuki, Saito Suguru, Okumura Manabu, "*Structuring Web Pages based on Repetition of Elements*", Transactions of Information Processing Society of Japan, Volume 45, Number 9, pp. 2157-2167, 2004.

[4] Sandip Debnath, Prasenjit Mitra, Nirmal Pal, C. Lee Giles, "*Automatic Identification of Informative Sections of Web Pages*", IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 9, pp. 1233-1246, 2005.

[5] Ruihua Song, Haifeng Liu, Ji-Rong Wen, Wei-Ying Ma, "*Learning Block Importance Models for Web Pages*", International World Wide Web Conference Proceedings of the 13th international conference on World Wide Web, pp. 203-211, 2004.

[6] Hongyan Jing, "*Sentence Reduction for automatic text summarization*", Proceedings of the sixth conference on Applied natural language processing, pp. 310-315, 2000

[7] Kevin Knight, Daniel Marcu, "*Summarization beyond sentence extraction: A probabilistic approach to sentence compression*", Artificial Intelligence, Volume 139, Issue 1, pp. 91-107, 2002

[8] Minh Le Nguyen, Akira Shimazu, Susumu Horiguchi, Bao Tu Ho, Masaru Fukushi, "*Probabilistic Sentence Reduction Using Support Vector Machines*", The 20th International Conference on Computational Linguistics COLING 2004, Geneva, pp.23-27, 2004

[9] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, Jaime Carbonell, "*Summarizing Text Documents: Sentence Selection and Evaluation Metrics*", Annual ACM Conference on Research and Development in Information Retrieval Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pp. 121-128, 1999.

[10] Eduard Hovy, Chin-Yew Lin, Liang Zhou, "*Evaluating DUC 2005 using Basic Elements*", In Proceedings of the Fifth Document Understanding Conference (DUC '05) , 2005.

[11] Ralph Debusmann, "*An Introduction to Dependency Grammar*", Hausarbeit fur das Hauptseminar SoSe 99. Univeristat des Saarlandes, 2000.

[12] D. P. Metzler, T. Noreault, L. Richey, B. Heidorn, "*Dependency Parsing for Information Retrieval*", Proc. of the third joint BCS and ACM symposium on Research and development in information retrieval, pp. 313-324, 1984.

[13] 최동현, 신지애, 최기선, "뉴스 기사의 문장 요약", 한글 및 한국어 정보처리 학술대회 2007

[14] 나종열, 신지애, 최기선, "VNA 집합을 이용한 뉴스기사의 중요문장 추출", 한글 및 한국어 정보처리 학술대회 2007