웹 문서로부터 잡음 영역의 클리닝을 위한 규칙기반 방법

이민형 김연석° 이경호
연세대학교 컴퓨터과학과
{mhlee, yskim°}@icl.yonsei.ac.kr, khlee@cs.yonsei.ac.kr

A Rule-based Method for Cleaning Noisy Areas from Web Documents

Min-Hyeong Lee, Yeon-Seok Kim° Kyong-Ho Lee
Dept. of Computer Science, Yonsei University

요 약

본 논문에서는 웹 문서의 논리적 구조분석을 위한 전처리 과정으로서 웹 문서에 포함된 잡음 영역을 제거하는 클리닝 방법을 제안한다. 제안된 방법은 잡음 영역을 내비게이션 영역, 광고 영역, 상호작용 영역, 특수정보 영역, 그리고 스크립트/스타일 영역의 5종류로 구분한 후, 이를 제거할 수 있는 규칙을 제안한다. 제안된 방법의 성능을 평가하기 위하여 웹으로부터 추출된 726개의 다양한 문서를 대상으로 실험한 결과, 91.16%의 정확률을 보였다.

1. 서론

웹 문서는 전달하고자 하는 주 내용은 물론이고 시각적인 편의를 위한 내비게이션 도구나 광고와 같은 부가적인 정보들을 포함한다. 그러나 이러한 부가 정보들은 유용한 정보의 추출 및 재사용 측면에서는 불필요하다. 따라서 정보의 효과적인 추출을 위해서는 먼저 이러한 불필요한 영역을 제거할 필요가 있다.

기존에 웹 문서 클리닝을 위한 연구결과가 다수 발표되었다. (표 1 참조) 그러나 이들 대부분은 특정한 구조를 갖는 문서만을 대상으로 하거나 혹은 사용자의 개입이 필요하다는 단점을 가지고 있다.

본 논문에서는 이러한 기존 연구의 문제점을 보완하며, 논리적 구조분석[16]의 전처리 과정으로 불필요한 잡음 영역을 제거하는 규칙 기반의 방법을 제안한다. 제안된 방법은 웹 문서에 포함될 수 있는 잡음 영역을 내비게이션 영역, 광고 영역, 상호작용 영역, 특수정보 영역, 그리고 스크립트/스타일 영역으로 구분한다. 제안된 방법의 성능을 평가하기 위해서 웹으로부터 추출된 726개의 다양한 문서를 대상으로 실험한 결과, 91.16%의 정확률을 보여 기존 연구보다 우수하였다.

2. 웹 문서 클리닝

제안된 방법은 그림 1과 같이 전처리와 잡음 영역 제거의 두 단계로 구성된다. 전처리 단계는 웹 문서를 가공하기 위한 전단계로서 HTML 문서를 XML 적격성(well-formedness) 요건을 만족하도록 변경하고 DOM(Document Object Model)을 사용하여 트리 모델로 구조한다. 잡음 영역 제거 단계에서는 제안된 웹 문서 클리닝 규칙을 사용하여 잡음 영역을 제거함으로서 논리적 구조분석에 적합한 문서를 생성한다. 각 단계에 대한 자세한 설명은 다음과 같다

2.1 전처리

제안된 방법은 HTML 문서의 구조화를 위하여 DOM 트리를 사용한다. DOM은 HTML 문서를 구성하는 태그, 속성(attribute), 그리고 텍스트 정보를 노드로 트리를 구성한다.

한편, HTML 문서로부터 DOM 트리를 구성하기 위해서는 HTML 문서가 XML 적격성 요건을 만족하여야 한다. 이를 위해서 HTML Tidy를 적용한다.

표 1. 웹 문서 클리닝

관련논문	특 징
[1]	HTML 태그와 텍스트를 비트로 표현하고 이를 이용한 누적분포 함수 그래프를 그려 주 내용 추출
[2]	웹 문서를 분할하고 각각의 영역에 특징을 부여, 결정트리를 사용하여 역할 구분
[3]	<TABLE>태그만을 기준으로 내용 블록을 분할
[4]	빈도 기반의 데이터 마이닝 기법을 사용
[5]	레이아웃이 비슷한 이웃 문서를 찾아내어 두 문서의 트리매칭을 통해 공통된 부분을 제거
[6]	웹 문서를 분류하기 위해 웹 문서의 시각적 정보를 사용
[7]	문서를 시각적인 위치정보를 통해 5개의 고차원 내용 블록으로 나눔
[8]	유사한 템플릿을 갖는 웹문서에 대해 스타일 트리를 정의한 다음 중요도를 계산하여 불필요한 부분 제거
[9]	유사한 템플릿을 갖는 웹문서에 대해 HTML문서의 table 태그의 반복빈도를 체크하여 내비게이션 바를 제거
[10]	태그 자체와 태그의 특징을 사용한 필터링 기법을 사용하여 불필요한 영역 제거
[11]	웹 문서의 구조를 구성할 수 있는 태그들을 선택하고 그 서브트리의 유사도를 구하여 영역구분
[12]	노드의 특징과 편집거리 사용
[13]	웹 문서를 기본 요소들로 구분하고 각각의 요소에 대해 랭킹 알고리즘을 적용하여 중요부분 추출
[14]	공간적인 특징과 내용적인 특징을 추출하여 기계학습법을 사용하여 영역의 중요도를 구함
[15]	유사한 구조를 갖는 웹 문서들의 패턴정보에 와일드 카드를 이용하여 표현하고, 표현된 템플릿과 대상 문서의 트리편집 거리를 이용하여 주 내용 추출

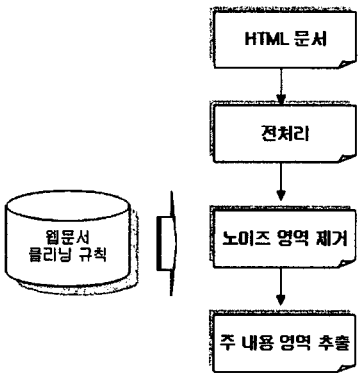

그림 1. 제안된 웹 문서 클리닝 과정

2.2 잡음 영역 제거

제안된 방법은 잡음 영역을 기하적 특성을 이용하여 제거할 수 있는 영역과 태그 정보만을 이용하여 제거할 수 있는 영역으로 나누어 처리한다. 기하적 특성을 이용하여 제거할 수 있는 영역이란 웹 문서에 제거되어야 할 영역 중에 시각적으로 특정한 위치에 존재하는 영역을 말한다. 이러한 영역에는 내비게이션 영역, 광고 영역 그리고 특수정보 영역이 있다. 제안된 방법은 이러한 영역을 식별하기 위해 우선 DOM 트리의 노드에 좌표 정보를 할당한다. 좌표 정보에 따라 각각의 노드를 상하좌우, 중앙 등 다섯 개의 영역으로 구분한다. 잡음 영역에 있어서 이러한 위치정보는 가장 일반적인 특징이다. 제안된 방법에서는 각 영역의 일반적인 특징을 통해 후보 노드를 선정하고 각각의 후보 노드에 대해 세부적인 규칙을 적용하여 잡음 영역을 제거한다.

영역을 식별하는데 기하적 정보가 필요하지 않은 영역에는 상호작용 영역과 스크립트/스타일 영역이 있다. 이러한 영역은 DOM 트리를 하향식 너비우선 탐색을 통해 해당 태그나 영역을 제거하는 방식으로 규칙을 적용한다.

2.2.1 내비게이션 영역 제거

웹 문서에서 사용자의 편의를 위해 다른 문서로 연결을 하거나 같은 문서 내에서 다른 영역으로 이동할 수 있도록 하는 영역을 내비게이션 영역(navigation area)이라고 한다. 이러한 내비게이션 영역은 웹 문서의 내용에 관련이 있는 부분이 아니라 단순한 연결을 위한 부분이다. 따라서 내비게이션 영역은 웹 문서의 논리적 구조분석의 성능을 저하시키는 요인이 될 수 있기 때문에 제거되어야할 대상이다. 이를 위하여 제안된 방법은 내비게이션 후보노드 선택규칙 1,2,3과 텍스트 메뉴 링크규칙을 적용하여 내비게이션 후보노드를 선택한 다음 내비게이션 영역 인식규칙 1~6을 적용하여 내비게이션 영역을 설정한 후 이를 제거한다.

2.2.2 광고 영역 제거

일반적으로 웹 문서는 이미지나 플래시 등을 사용한 광고 또는 로고를 포함한다. 본 논문에서는 주 내용과는 다른 광고 또는 로고를 표현한 영역을 광고 영역(advertisement area)으로 정의한다. 광고 영역은 웹 문서를 통해 수익을 얻거나 부가정보를 표현하기 위한 영역이기 때문에 사람들에게 상품을 광고한다는 측면에서는 의미가 있지만 웹 문서로부터 논리적 구조를 추출한다는 측면에서는 불필요한 영역이다. 이를 위하여 제안된 방법은 광고 후보노드 선택규칙 1,2,3을 적용하여 광고 후보노드를 선택한 다음 광고 후보영역 인식규칙 1~5를 적용하여 광고 영역을 설정한 후 이를 제거한다.

2.2.3 특수정보 영역 제거

특수정보 영역(special information area)에 포함되는 저작권, 버전, 업데이트, 연락처 등의 정보는 웹 문서를 방문한 사용자가 부가적인 정보를 얻기 위해 필요한 정보이기 때문에 주 내용과는 관련이 없다.

한편, 저작권 정보나 업데이트 정보와 같은 특수 정보들은 웹 문서의 하단에 위치하기 때문에, 이러한 특징을 기반으로 한 특수정보 후보노드 선택규칙을 적용하여 후보노드를 선택하고 다시 특수정보 노드 규칙 1,2를 적용하여 특수정보 노드를 설정한 다음 특수정보 영역 인식 규칙을 적용하여 불필요한 텍스트들을 모두 제거한다.

2.2.4 상호작용 영역 제거

웹 문서에는 사용자 로그인이나 검색과 같은 사용자와 상호작용을 통해 동적으로 상태가 변하는 영역이 있다. 제안된 방법에서는 이러한 영역을 상호작용 영역(interactive area)이라고 정의한다. 상호작용 영역은 태그 <FORM>의 안에서 여러 가지 태그들과 속성들의 구성으로 이루어진다. 따라서 태그 <FORM>과 태그 <FORM>이 둘러싸고 있는 영역을 제거하여 상호작용 영역을 제거한다. 그러나 HTML 문서는 문법 검사가 유연하기 때문에 태그 <FORM> 없이도 GUI 컴포넌트에 해당하는 태그를 포함할 수 있다. 상호작용 제거 규칙(interactive area remove rule)에서는 태그 <FORM>과 더불어 다른 상호작용 관련 태그들이 존재할 경우 그 영역을 제거한다.

2.2.5 스크립트/스타일 영역 제거

웹 문서에서 태그 <SCRIPT>는 HTML에서 제공하는 동적인 기능보다 복잡한 기능을 제공하기 위하여 사용된다. 스크립트 태그는 브라우저상에서 시각적으로 표현되지도 않고, 논리적 구조분석에 필요 없는 정보를 제공하기 때문에 제거되어야 한다. 태그 <STYLE> 역시 텍스트의 스타일을 정의할 때 쓰이는 태그로 직접적으로 브라우저상에 표현되지 않고, 웹 문서가 제공하고자 하는 내용과는 관련이 없기 때문에 제거되어야 한다. 스크립트/스타일 제거규칙(script/style remove rule)은 이러한 정보를 제공하는 태그의 이름을 갖는 노드를 제거한다.

3. 실험결과

제안된 방법은 웹 문서 클리닝의 성능을 평가하기 위하여 웹으로부터 표 2와 같이 726개의 다양한 주제의 문서를 수집하여 실험하였다. 성능을 평가한 결과는 표 3과 같다.

제안된 방법은 동일한 유형의 템플릿을 갖는 문서집합을 대상으로 하는 잡음 제거에 대한 기존연구[4][5][8][9][15]와는 다르게 단일 문서를 대상으로 한다는 장점이 있다. 즉, 잡음 제거를 위해 동일한 템플릿을 갖는 문서의 수집이 필요가 없다.

또한 Gupta[10]등의 방법은 대부분이 링크로 구성된 문서가 입력으로 주어질 경우 모든 내용을 제거하는 문제점이 있었다. 제안된 방법은 이러한 문제점을 해결하기 위하여 위치 정보를 사용하여 후보영역을 제한하고 링크를 구성하는 단어의 개수에 제한을 두어 메뉴가 될 수 있는 링크를 한정하였다.

제안된 방법은 기존 연구들에서 다루었던 모든 잡음 영역을 포괄함으로써(표 4 참조) 웹 문서 내에 존재 할 수 있는 모든 잡음 영역을 제거한다. 또한 Lin과 Ho[3]의 방법과 Ma[9]등의 방법과는 달리 잡음 영역이 될 수 있는 대상 노드를 특정 노드로 한

표 2. 주제별 웹 문서의 수

주제	웹 문서 수
University	121
News	106
Organization	113
Syllabus	143
Design	161
Etc.	82
합계	726

표 3. 웹 문서 클리닝 방법 성능평가

기준	제거해야 할 영역의 수	제안된 방법을 사용하여 제거된 영역의 수	정확률 (%)
내비게이션 영역	1158	1066	92.06
광고 영역	1284	1134	88.32
특수정보 영역	310	294	94.84
상호작용 영역	157	157	100
스크립트/스타일 영역	347	347	100
전체 잡음 영역	3271	2982	91.16

표 4. 식별하는 영역에 따른 분류

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
주 영역	O	O	O	O	O	O	O	O	
내비게이션 영역		O							O
광고영역									
특수정보 영역		O							
상호작용 영역		O							
스크립트 영역									

	[10]	[11]	[12]	[13]	[14]	[15]	제안된 방법
주 영역	O	O	O	O	O	O	O
내비게이션 영역	O	O					O
광고영역	O						O
특수정보 영역		O					O
상호작용 영역	O	O					O
스크립트 영역	O						O

정하지 않고 위치, 크기와 같은 기하정보를 통해 후보노드를 선택하여 좀 더 다양한 형태의 잡음 영역을 인식할 수 있다.

4. 결론

본 논문에서는 논리적 구조분석의 전처리 과정으로서 불필요한 잡음 영역을 제거하는 규칙 기반의 방법을 제안하였다. 또한 Tidy를 이용하여 적격성을 갖춘 HTML 문서에 대하여 DOM 트리를 구성하고, 파싱된 DOM 트리에를 하향식 너비우선 탐색하면서 제안된 웹 문서 클리닝 규칙을 적용하여 잡음 영역을 제거하였다.

제안된 방법은 기존 연구들과 비교하여 이질적인 구조를 갖는 문서에 대해 보다 정교하게 잡음 영역을 인식하였으며 실험결과, 91.16%의 정확률을 보여 우수하였다.

참고문헌

[1] Aidan Finn, Nicholas Kushmerick, and Barry Smyth, "Fact or Fiction: Content Classification for Digital Libraries," Proc. Joint DELOS-NSF Workshop on Personalisation and Recommender Systems in Digital Libraries (Dublin), 2001.

[2] Jinlin Chen, Baoyao Zhou, Jin Shi, HongJiang Zhang, and Qiu Fengwu, "Function-based Object Model Towards Website Adaptation," Proc. World Wide Web Conf., pp. 587-596, 2001.

[3] Shian-Hua Lin and Jan-Ming Ho, "Discovering Informative Content Blocks from Web Documents," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 588-593, 2002

[4] Ziv Bar-Yossef and Sridhar Rajagopalan, "Template Detection via Data Mining and its Applications," Proc. World Wide Web Conf., pp. 580-591 2002.

[5] Jiying Wang and Fred H. Lochovsky, "Data-rich Section Extraction from HTML Pages," Proc. Int'l Conf. Web Information Systems Engineering, pp. 313-322, 2002.

[6] Milos Kovacevic, Michelangelo Diligenti, Marco Gori, and Veljko M. Milutinovic, "Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification," IEEE Int'l Conf. Data Mining, pp. 250-257, 2002

[7] Yu Chen, Wei-Ying Ma, and HongJiang Zhang, "Detecting Web Page Structure for Adaptive Viewing on Small Form Factor Devices," Proc. World Wide Web Conf., pp. 225-233, 2003

[8] Lan Yi and Bing Liu, "Eliminating Noisy Information in Web Pages for Data Mining," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 296-305, 2003.

[9] Ling Ma, Nazli Goharian, and Abdur Chowdhury, "Automatic Data Extraction from Template Generated Web Pages," Int'l Conf. Parallel and Distributed Processing Techniques and Applications, pp. 642-648, 2003

[10] Suhit Gupta, Gail Kaiser, David Neistadt., and Peter Grimm, "DOM-based Content Extraction of HTML Documents," Proc. World Wide Web Conf., pp. 207-214, 2003.

[11] Vincenza Carchiolo, Alessandro Longheu, and Michele Malgeri, "Extracting Logical Schema from the Web," Appl. Intell., pp. 341-355, 2003.

[12] Bing Liu, Robert Grossman, and Yanhong Zhai, "Mining Data Records in Web Pages," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 601-606, 2003

[13] Xinyi Yin and Wee Sun Lee, "Using Link Analysis to Improve Layout on Mobile Devices," Proc. World Wide Web Conf., pp. 338-344,2004.

[14] Ruihua Song, Haifeng Liu, Ji-Rong Wen, and Wei-Ying Ma, "Learning Block Importance Models for Web Pages," Proc. World Wide Web Conf., pp. 203-211, 2004.

[15] Davi de Castro Reis, Paulo B. Golgher, Altigran S. da Silva, and Alberto H. F. Laender, "Automatic Web News Extraction Using Tree Edit Distance," Proc. World Wide Web Conf., pp. 502-511, 2004.

[16] 이민형, 이경호, "특정 주제 웹문서의 논리적 구조 분석," 한국정보과학회 춘계학술대회, Apr. 2004.