정보 추출을 위한 트리거에서 HTML 이미지 태그 정보의 이용

김연정° 박제현 최중민
한양대학교 컴퓨터공학과
{yeonjung°, jhpark, jmchoi}@cse.hanyang.ac.kr

Application of the HTML Image Tag on Triggers for Describing Rules for Information Extraction

Yeonjung Kim° Jeahyun Park Joongmin Choi
Dept. of Computer Science and Engineering Hanyang University, Korea

요 약

웹 문서를 대상으로 하는 정보 추출이나 웹 마이닝에 관한 연구가 활발히 진행되면서 특히, 웹에서 나타나는 구조적 패턴을 이용해 정보를 추출하는 방법에 대한 연구가 이루어 지고 있다. 하지만 구조적 패턴을 이용할 수 없는 경우 텍스트에 의존한 추출기를 생성할 수 밖에 없는데 웹 문서에서 시각적 요소가 강해지면서 트리거가 단순 텍스트가 아닌 이미지로 처리되는 경우가 있다. 기존의 연구들은 트리거를 단순 텍스트로 보는 관점에서의 연구가 많았고 이러한 접근 방법은 트리거가 이미지인 경우에 정확한 정보를 추출해 낼 수 없었다. 이 논문에서는 트리거가 텍스트가 아닌 이미지인 경우에도 필요한 정보를 잘 추출해 내기 위한 방법에 관해 제안하고자 한다.

1. 서 론

인터넷이 월드 와이드 웹이라는 기술에 의해 폭발적인 성장을 하면서 우리의 일상이나 업무에서 차지하는 비중은 날로 커져만 간다. 그리고 매일 새롭게 생성되는 많은 웹 문서에서 사용자가 실제로 필요로 하는 정보는 일부 문서, 또 일부 문서 중 극히 일부 내용에 국한된 경우가 많다. 이렇게 많은 문서 중에서 사용자가 필요로 하는 정보만을 찾아내기 위해서 웹 마이닝이라는 분야에서 많은 연구가 진행되고 있다. 규칙에 기반한 랩퍼를 만들어 원하는 정보만을 추출하는 정보 추출 기법[1,2] 및 패턴이나 구조적 정보를 이용해 정보를 추출하는 웹 컨텐트 마이닝[4,7]에 이르기까지 방대한 영역에서 연구되고 있다. 찾고자 하는 정보가 특정 구조에 의존하지 않고 단순한 트리거에 의한 정보라면 후자의 방법은 적절치 않다. 전자의 방법을 사용하더라도 정보를 찾아내기 쉽지 않다. 가장 큰 이유는 웹 문서가 더 이상 텍스트에 기반하지 않는다는 사실에 있다. 좀 더 시각적으로 사람들의 시선을 끌기 위해 화려해지고 있다. 그래서 단순히 트리거를 텍스트 정보에만 의존해 왔던 기존의 연구만으로는 실제 타겟 정보를 추출하기란 쉽지 않다. 따라서 이 논문에서는 웹 문서의 HTML 태그정보의 일부를 이용하여 타겟 정보 추출의 트리거가 이미지라도 타겟 정보 추출이 가능한 정보 추출기법을 제안하고자 한다.

2. 제안 시스템의 소개

정보 추출을 위한 단계는 그림 1과 같이 진행된다. 전처리 단계를 거쳐 학습 데이터를 수집한 후, 학습 데이터의 타겟정보 앞에 위치하는 트리거 후보들이 단순 텍스트

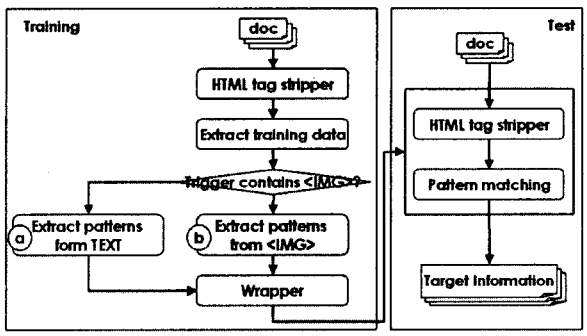

그림 1. 전체 정보 추출기 개관도

이면 a단계를, 이미지가 포함되어 있으면 b단계를 지나게 되고 a와b를 통해 생성된 패턴 모두 적용해서 실제 타겟 정보를 추출해 내게 된다. 본 논문에서 다루고자 하는 바는 트리거가 이미지로 처리된 경우이므로 그림 1의 b 단계에 대해서만 구체적인 방법을 제시하고자 한다.

2.1 HTML 태그 제거

본 논문에서 대상으로 하는 문서는 웹 문서이므로 전처리 단계가 필요하다. 정보 추출을 위해 구조적 정보는 필요치 않으므로 HTML 태그 제거단계가 필요하다. 하지만 서론에서도 이야기 했듯이 트리거가 되는 대상 중 일부는 이미지로 처리된 경우가 있으므로 태그 제거 단계에서 IMG 태그는 남겨 두어야 한다. 특히 IMG 태그 중 다른 속성은 필요치 않으므로 src 속성만 남기게 된다.

2.2 학습 데이터 추출

태그 제거 단계를 거친 문서에서 우리가 추출하고자 하는 타겟 정보와 이의 트리거에 관한 패턴을 추출하기 위한 학습 데이터를 결정하기 위해 타겟 정보와 타겟 정보의 앞에 k개의 토큰을 추출하여 학습 데이터로 삼는다. 이 논문에서는 k를 2로 설정하였다.

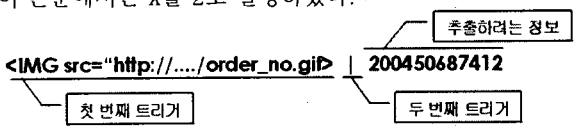

그림 2. 규칙 학습 예제

2.3 패턴 추출

전 단계에서 결정한 학습 데이터에서 트리거에 해당하는 부분이 단순 텍스트인지 이미지인지에 따라 다른 패턴 추출 단계를 거치게 된다. 이 논문에서는 트리거가 이미지인 경우의 패턴 추출 방법(그림 1의 b)만 다루며, 세부 과정에 대해서는 다음 장에서 좀 더 자세히 알아본다.

3. 이미지를 트리거로 하는 규칙 추출

이미지를 트리거로 하는 규칙을 추출하기 위해서는 실제 파일이름을 분리해내는 전처리단계가 필요하다. 이 전처리단계를 거쳐 분리된 파일이름을 통해 규칙을 추출하게 된다. 이 경우 이미지 이름이 어느 정도 실제 이미지의 의미를 담고 있다는 가정 하에 이러한 작업이 가능하다. 주요 예는 다음과 같다.

예) mailOrderNo, new_product, prodStatus_0426, …

만약 파일 이름이 ' 1111' 과 같은 의미 없는 이름으로 작성된 경우라면 이러한 과정을 적용하여도 기존의 시스템과 성능상의 큰 차이를 보이지 못할 것이다. 하지만 정보 추출의 대상이 상업용 페이지인 경우가 많고, 개발상의 용이함을 고려한다면 위와 같은 가정은 상당히 설득력이 있다고 볼 수 있다.

3.1 파일이름을 분리해내기 위한 전 처리단계

트리거가 이미지인 경우 IMG 태그에서 src에 해당하는 내용 중 실제 파일이름만을 남겨 놓는다. 예를 들어
<IMG src="http://.../.../mail_order_no.gif">
라는 태그가 트리거로 존재 한다면 접두어와 경로를 제외한 mail_order_no만 남긴다. 그런 다음 mail_order_no를 다시 mail, order, no로 분리해낸다. 만약 파일이름이 mailOrderNo이었다면 mail, Order, No으로 분리될 것이다. 파일 이름이 분리되는 전체 흐름은 그림 3과 같다. 자세한 알고리즘은 그림 4를 참고한다.

3.2 분리된 파일 이름을 이용한 규칙 생성

3.1에서 파일 이름이 분리되면 타겟 정보 앞에는 k개의 단어 집합이 존재 할 것이다. 여기서 정보 추출 규칙을 구성할 타겟 정보와 그 앞에 위치하는 k개의 트리거를 도식

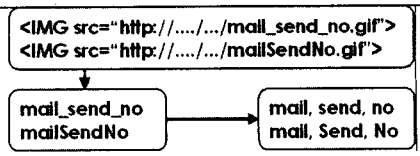

그림 3. 파일 이름 분리하기

Algorithm : FileNameDevider
Input : <IMG> 태그정보
Output : 분리된 파일 워드들

A ← { }
a ← { }

1. **src의 실제 파일 이름만 남기고 모두 제거**
2. **파일 이름 중 파일 확장자에 해당하는 정보 (예 : .gif)도 모두 제거**
3. **'_'를 딜리미터로 적용하여 파일이름을 분리하여 a 집합에 추가**
4. **while a != { }**
   1. **a에서 원소 하나를 꺼낸다.**
   2. **대문자가 포함되어 있으면 대문자 이전까지를 하나의 파일 워드로 분리, 분리되는 모든 워드는 A 집합에 추가**
5. **Return A**

그림 4. 파일 이름 분리 알고리즘

그림 5. 패턴 생성을 위한 학습 데이터 모습

화하면 그림5와 같다. k=2일 때, 그림 5와 같이 타겟 정보로부터 앞 쪽으로 첫 번째에 존재하는 단어 집합을 " 1" 로, 두 번째에 존재하는 파일 단어 집합을 " 2" 로 나타내었을 때, 특별히 " 1" 의 위치에 집합의 원소가 하나이고 길이가 1인 :,| 와 같은 특수 기호가 존재할 경우 이는 규칙 상의 부가적인 부분으로 처리하고, 이 트리거를 제거한다. 그렇게 되면 학습 데이터는 " 1" 자리에 " 2" 가 위치한 트리거 후보가 하나밖에 없는 모습을 띄게 될 것이다.

규칙 생성을 위한 학습 예제가 m개 존재한다면 이제 이 m개의 예제를 통해서 규칙을 생성할 것이다. 이 논문에서는 m개의 데이터 중 $m\theta$개 ($0 \le \theta \le 1$) 이상의 학습 예제에서 나타나는 단어가 규칙을 구성하도록 하였다. θ가 작으면 작을수록 좀 더 많은 규칙이 생성될 것이고 이를 이용한 정보 추출기는 θ가 비교적 큰 경우보다 좀 더 많은 결과물을 추출할 것이다. 그림 6에서 m개의 학습 예제에서 추출한 후보들을 비교할 때 완전히 동일한 예제만 추출한 것이 아님을 알 수 있다. 여기서 각 예제의 유사도는 편집 거리[3]를 사용하였으며, 편집 거리가 2 이내인 단어들은 서로 같다고 판단하였고 대소문자의 구분은 없다. 이러한 기준에서 $m\theta$번 나온 예제에서만 규칙을 생성하였다. 그림 6에서 추출된 규칙을 보면 <,>로 묶인 부분은 반드시 있어야 하는 필수 부분이고 [,]로 싸여진 묶인 부분은 부가적인 부분으로서, 이후에 대상 문서에서 이 부분에 해당하는 트리거가 존재하지 않아도 이 규칙에 해당하는 정보로서 인정하게 된다. 그리고 추출된 트리거는 논리곱연산으로 결합하게 되는데 이는 정확하게 규칙에

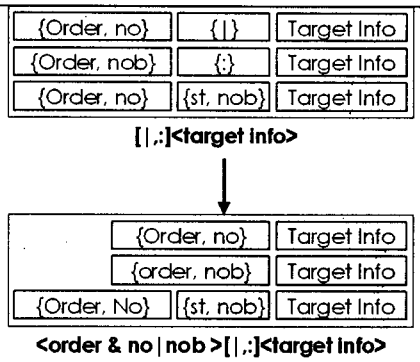

그림 6. 패턴 생성 단계

적용되는 타겟 정보만을 추출하기 위한 방법이다.

3.3 개선된 규칙 생성

이미지 태그의 파일 이름에서 추출한 단어를 조합하는 과정에서 논리곱연산을 이용한 결합을 사용함으로써 지나치게 엄밀한 규칙이 생성되는 경우가 나타나게 된다. 이러한 단점을 보완하기 위해 다음과 같은 경우 규칙이 성립했다고 인정한다.

생성된 규칙에서 한 트리거의 단어 집합에 존재하는 단어가 n개일 때, 이 단어 중 $\lfloor n*\theta+0.5 \rfloor$개 이상 일치되면 타겟 정보로 추출.

θ는 3.2에서와 같은 값을 사용하며, 이 경우 규칙은 다음과 같은 형태를 가진다.

<order , no | nob >[|,:]<target info>

그림 7. 개선된 패턴

4. 결론 및 향후 과제

웹 문서들이 점점 시각적인 요소가 강해짐으로써 웹을 단순히 꾸며주는 부가적인 기능으로만 존재하던 이미지들은 웹 문서의 상태 특성이나, 표현하고자 하는 정보와 밀접한 관계를 갖게 되었다. 기존의 텍스트에 기반한 정보 추출기법으로는 이러한 경우에 타겟 정보를 추출하는 것이 매우 어렵다. 이 논문에서는 이러한 상황에서 타겟 정보의 손실 없는 추출을 위해 이미지 파일 이름에서 단어를 추출하여 규칙을 생성하는 기법을 제안하였다.

향후 과제로서, 단순한 트리거로서 뿐만 아니라 어떤 문서의 상태 정보를 알고 싶을 때에도 이 논문에서 제시한 이미지 태그를 이용한 방법이 유용하게 사용되리라고 본다. 상태 정보를 실제 텍스트가 아니라 이미지로서 처리하는 페이지에 적용될 수 있는데 그림 8과 같은 경우, 각 페이지는 A, B, C, D, E, Z라는 이미지와 실제 내용으로 구성된다. 여기서 각 상태를 분석하기 위해 텍스트만 사용해서 부족한 경우 각 상태에 의존적인 C, D, E를 부가적으로 이용해 상태를 알 수 있다. 이 때, 이 논문에서 제시한 IMG 태그를 이용한 분류와 텍스트를 이용한 분류를 혼합하여 사용하면 좀 더 정확한 상태 분류가 가능할 것이다.

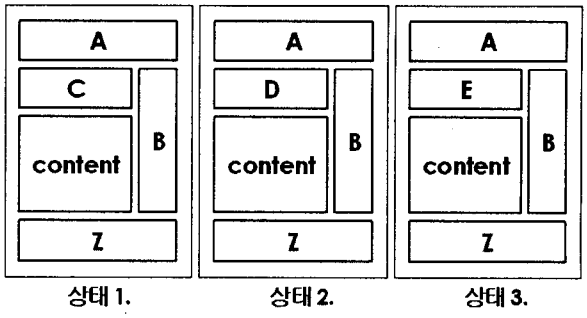

그림 8. 이미지를 이용해 상태를 나타낸 경우

5. 참고문헌

[1] Xiaoying Gao, Mengjie Zhang, "Learning knowledge bases for information extraction from multiple text based Web sites", IEEE/WIC, pp.119-125, 2003

[2] Jaeyoung Y., Joongmin C., "Agents for Intelligent Information Extraction By Using Domain Knowledge and Token-based Morphological Patterns", 6th Pacific Rim International Workshop on Multi-Agents, 2003

[3] W. J. Masek and M. S. Paterson. " A faster algorithm computing string edit distances" . Journal of Computer and System Sciences, Vol.20: pp.18-31, 1980

[4] H. Davulcu, S. Koduri, S. Nagarajan, "Datarover: a taxonomy based crawler for automated data extraction from data-intensive websites ", the 5th ACM International Workshop on Web Information and Data Management, pp. 9 - 14, 2003

[5] Lan Yi., Bing Liu, Xiaoli Li., "Eliminating Noisy Information in Web Pages for Data Mining", the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003

[6] Bing Liu, Grossman, R., Yanhong Zhai, "Mining Web Pages for Data Records", Intelligent Systems and Their Applications, IEEE, Vol. 19, Issue 6, pp.49-55, 2004

[7] Chia-Hui Chang, Shih-Chien Kuo, "OLERA: Semisupervised Web-Data Extraction with Visual Support", Intelligent Systems and Their Applications, IEEE, Vol. 9, Issue 6, Nov.-Dec., pp.56-64, 2004