정보 추출을 위한 의미 있는 블록 검출 방법

강진범[O1] 양재영[2] 최중민[1]
[1]한양대학교 컴퓨터공학과 지능시스템 연구실
{midgetfx, jmchoi}@hanyang.ac.kr
[2](주) 코리아 와이즈넛
isconan@gmail.com

Detecting Informative Blocks for Information Extraction

Jinbeom Kang[O1] Jaeyoung Yang[2] Joongmin Choi[1]
[1]IS Lab, Department of Computer Sci. & Eng., Hanyang University
[2]Korea Wisenut Co.

웹 정보 추출은 대상 정보를 정확하게 추출하는 것이 주요 문제점이다. 대부분의 지도 학습(supervised learning)인 경우 사용자에 의해 수집된 학습 자료가 필요하고, 비지도 학습(unsupervised learning)인 경우 학습 자료가 필요하지 않더라도 도메인 지식이 필요하거나 리스트 혹은 테이블과 같이 특수한 형태에 대해서만 추출이 가능하다는 문제가 있었다. Wrapper Induction은 지도학습 접근방법으로써 반자동(semi-automatic)하게 수행된다. 이것은 수동적으로 꼬리표달린 문서(labeled page)나 자료 레코드의 집합으로부터 추출 규칙을 학습하고 이후 구조적으로 유사한 새로운 웹 문서에서 관련 정보를 추출하는 방법이다. 최근 웹 문서는 다양한 형태로 표현됨으로써 대상 정도 추출하는 것은 보다 복잡하며 어려워졌다. 인터넷 보급이 확대되고 인터넷을 이용하는 사용자도 증가하면서, 최근 웹 문서에서는 표현할 수 있는 콘텐츠의 종류(동영상, 이미지, 플래시 등) 또한 다양해 졌다. 그의 예로, 최근 웹 문서 디자이너는 사용자에게 보여지는 시각적인 측면을 강조하고 있다. 뿐만 아니라 보다 다양한 정보를 사용자에게 제공할 수 있도록 디자인하고 있다. 웹 문서가 다양한 콘텐츠를 포함함에 따라 HTML 구조는 복잡해 졌고 그렇기 때문에 이러한 웹 문서에서 사용자가 요구하는 콘텐츠가 어떠한 것인지를 선별하는 것 또한 어려워 졌다. 만약 우리가 콘텐츠라는 정보를 미리 파악해 줄 수 있다면, Wrapper Induction은 실제 대상 정보를 추출하기 위해 보다 쉽게 규칙을 생성할 수 있고, 소요시간을 감소시킬 수 있다.

본 논문에서 시각적 문서 분할 방법을 이용하여 웹 문서를 분석하고 구조적으로 유사한 블록을을 군집한 뒤, 의미있는 군집을 식별할 수있는 RIPB(Recognizing Informative Page Blocks) 알고리즘을 제안한다. RIPB 알고리즘은 시각적으로 의미 있는 정보와 구조적 성질들을 고려한 편집 거리(levenshtein distance) 알고리즘을 이용해 유사한 블록들을 군집한다. 하지만 방대한 정보를 담고 있는 콘텐츠에 대해서 동일한 구조적 성질을 가지고 있음에도 불구하고 편집 거리 값이 켜져 동일한 정보를 담고 있는 군집인지 판단하기 어렵다. 그래서 편집 거리를 트리 편집 거리와 같이 구조적 성질 비교를 위해 토큰 단위로 비교하고 콘텐츠 정보 및 구조적 성향에 영향을 미치지 않는 태그, 속성 정보는 제거한다. 결과적으로 구조적으로 복잡하게 만들면서 의미가 없는 태그와 속성, 콘텐츠를 제거한다. 구조적 성질과 연관 없는 정보를 제거한 문서에서 토큰 기반 편집 거리를 통해 군집이 이루어진다.

구조적으로 유사한 블록들을 군집한 후, 구조적으로 의미 있는 군집(informative cluster)을 식별한다. 쇼핑몰에서 상품 목록 문서의 경우 의미 있는 군집은 상품 목록이 되며, 상세 정보 문서의 경우, 가격, 상품평과 같은 상품 상세 정보이다. 이와 같은 도메인 경우, 상품 정보에 대해 반복적인 패턴이 존재하며 상세 문서에서도 상품평과 같은 반복적인 패턴이 존재한다. 하지만 뉴스 도메인의 기사 문서에서는 기사 내용만 존재하며 반복적인 패턴을 찾을 수 없다. 그래서 단일 웹 문서를 분석하여 의미있는 블록을 식별하거나 패턴을 얻는 것은 매우 어렵다. 하나의 웹 문서를 분석하여 의미 있는 군집을 식별하기 위해 단순히 블록의 수가 가장 많은 군집을 의미 있는 군집이라 할 수 없다. 이와 같은 문제점을 해결

하기 위해 얼마나 많은 콘텐츠를 포함하고 있는지를 평가하였다. 군집에 속하는 모든 블록들의 단어 수와 이미지의 면적이 가장 큰 군집을 의미 있는 군집으로 평가한다. 쇼핑몰의 상품 목록이 가장 큰 군집이 될 것이며 또한 많은 정보를 포함하고 있게 된다. 기사문서의 경우 유사 패턴이 존재하는 블록은 존재하지 않기 때문에 군집에 속하는 블록은 하나만 형성되지만, 많은 정보를 포함하게 된다.

정보 추출 과정의 규칙 생성 단계는 사용자가 추출하고자 하는 대상 영역을 선택하고 wrapper는 사용자가 선택한 아이템이 속하는 블록을 시각적 문서 분할을 한다. 분할된 블록의 HTML 코드에서 구조적으로 의미없는 불필요한 태그와 모든 속성들을 제거하여 태그 시퀀스(Tag Sequence)로 인코딩한다. 태그 시퀀스에서 사용자가 선택한 아이템을 중심으로 유일한 패턴 규칙을 생성한다. 생성된 규칙은 추출 단계에서 RIPB를 통해 식별된 의미있는 블록들을 대상으로 정보를 추출한다.

RIPB를 이용한 정보 추출의 성능을 입증하기 위해 총 8개의 사이트에서 성능평가를 하였다. 각 사이트마다 목록 문서와 상세 문서로 구성이 되며 각 사이트 마다 무작위로 100개의 문서를 수집하여 총 800개의 문서를 평가하였다. 실험은 RIPB를 통해 의미있는 블록 식별 성능과 정보 추출에 적용했을 때의 성능을 비교 분석하였다. (1) 의미있는 블록을 식별하는 실험에서 대체로 쇼핑몰이 높은 성능을 보였으며 뉴스 사이트는 낮은 성능을 보였다. 뉴스 사이트 경우 복잡한 구조로 다양한 정보를 담고 있었다. 예를 들어 기사의 중간에 광고가 있는 경우와 기사와 연관 있는 링크 집합이 나타나는 경우에도 발생하였다. 그래서 블록 형성 시 관련 없는 정보가 많이 포함하고 있어서 정확률이 낮게 나왔다. 이것은 VIPS의 문서 분할 결과가 좋지 못한 경우로 DoC(Degree of Coherence)의 값에 따라 영향을 받는다. DoC는 블록들 간 어느 정도 응집도가 존재하는 지를 나타내며, DoC 값이 커지면 블록들을 세밀하게 분할하고 작으면 큰 블록으로 형성된다. 문서 마다 블록들의 간격 또는 사용 태그들이 다르기 때문에 동일한 DoC 값을 사용하면 좋은 성능을 얻을 수 없다. 하지만 DoC의 값을 기계적으로 판단하여 설정하는 것은 매우 어렵기 때문에 전체적인 성능을 높이는 값 6으로 설정하였다. 실험을 통해 성능을 낮추는 요인을 분석하였다. 첫째, VIPS가 올바르게 분할하지 못한 경우, 둘째, 의미 있는 블록이 하나만 존재하여 패턴을 얻을 수 없는 경우, 셋째, 구조적으로 비슷하지만 연관 없는 블록이 군집 되는 경우로 분석되었다. 평균적으로 60%이상의 F-measure 성능을 보였다. (2) RIPB를 정보 추출 성능 실험은 순수 wrapper와 RIPB를 이용한 wrapper간의 정보 추출 성능을 비교하였다. 순수 wrapper가 쇼핑몰에서 RIPB기반 wrapper에 비해 20% 가까이 성능이 좋지 않게 나왔다. 그 이유는 2가지로 분석할 수 있다. 첫째, 규칙을 생성하기 위한 자료가 시각적 문서 분할을 통해 간단한 HTML 구조를 가지기 때문에 규칙이 간단, 명료하다. 더욱이 쇼핑몰은 복잡한 구조로 디자인 되어 있기 때문에 RIPB에서 생성한 규칙이 순수 wrapper에서 관련 없는 자료까지 추출하게 되었다. 둘째, 만약 시각적 문서 분할을 하였더라도 모든 블록을 대상으로 평가를 하면 비슷한 결과를 얻게 된다. RIPB는 사용자가 정보 추출을 할 가능성이 있는 의미있는 블록들을 식별하기 때문에 명확한 정보를 추출할 수 있다. 그래서 의미있는 블록 식별을 잘못하게 되면 불순한 정보까지 추출되어 정확도를 떨어트릴수 있다.

RIPB를 이용한 Wrapper의 이점은 (1) 생성된 규칙이 간단하고 이해하기 쉬우며 피드백이 손쉽다. (2) 더불어 규칙 생성 과정이 매우 짧으며, 하나의 규칙만으로도 많은 정보를 명확하게 추출할 수 있다. (3) 그리고 RIPB가 의미있는 블록을 식별해 줌으로써 잘못된 정보를 추출할 가능성을 방지하고 있다.

하지만 문서 분할이 잘못되어 개별 블록으로 형성할 수 없거나, 의미있는 블록임에도 불구하고 구조적으로 유사하지 않아 의미있는 군집에 속하지 않는 경우 성능을 저하시키는 요인이 된다. 그래서 시스템의 전체 성능에 영향을 미치는 VIPS 알고리즘의 DoC 값을 결정은 매우 중요하다. 향 후 이와 같은 문제점을 해결하기 위해 문서 분석(page analysis)을 통해 VIPS의 DoC를 자동적으로 결정하는 방법과 시각적 기반 유사도 평가 및 군집하는 방법을 연구할 것이다.

후 기

본 논문은 "국가IT온톨로지 인프라 기술개발"정보통신부 선도과제 성과의 일부입니다.