정보추출을 위한 학습 가능한 인터페이스 에이전트

김용기[0] 양재영 최중민
한양대학교 컴퓨터공학과
{ykkim, jyyang, jmchoi}@cse.hanyang.ac.kr

Trainable Interface Agents for Information Extraction

Yongkee Kim[0] Jaeyoung Yang Joongmin Choi
Dept. of Computer Science and Engineering, Hanyang University

요 약

본 논문의 목적은 기계 학습 방법을 이용하여 정보 추출 규칙의 패턴을 학습할 수 있는 인터페이스 에이전트의 개발에 있다. 인터페이스 에이전트는 사용자와 상호작용이 가능한 지능형 에이전트이다. 사용자는 인터페이스 에이전트와 상호작용을 하게 되며 에이전트는 이 상호 작용에서 사용자가 원하는 정보 추출 규칙을 학습하게 된다. 사용자는 웹 문서에서 원하는 정보의 위치를 지정하여 데이터를 인터페이스 에이전트에게 학습시킨다. 인터페이스 에이전트는 학습된 추출 규칙으로부터 사용자가 원하는 정보를 추출한다.

1. 서론

문서의 의미를 분석하고자 하는 초창기의 연구들은 자연어 처리분야에서 시작되었다. 현재는 웹의 발달로 웹상의 문서를 분석하고자 하는 연구가 증가하고 있다. 이것은 분석의 대상인 문서의 종류가 변하고 있음을 의미한다. 초기의 문서는 비 구조화된 문서로 자연어 처리를 통해서 필요한 정보를 추출할 수 있었다. 하지만 근래에는 비 구조화된 문서보다는 준 구조화(semi-structured)된 문서의 비중이 높아지고 있다. 준 구조화된 웹 문서에서 사용자가 원하는 정보를 찾는 문제는 wrapper induction 방법을 이용하여 해결할 수 있다. 현재의 기술로는 웹페이지의 50%정도에 대해서만 자동으로 추출 규칙을 생성할 수 있다[1,2]. 이것은 자동 wrapper 시스템의 한계로 이 한계를 극복하기 위해서는 수동으로 생성된 많은 온톨로지와 미리 고정된 문자열 분석도구가 필요하다. 또한 사용자가 시스템에게 원하는 정보를 표현하는 방법에도 많은 문제점이 존재한다. 본 논문에서는 이러한 문제점을 해결하기 위하여 사용자와 상호작용이 가능한 인터페이스 에이전트를 이용하고자 한다. 인터페이스 에이전트는 문서에서 정보추출이 가능한 상황이나 상태가 발생하면 사용자대신 적절한 행동을 취하게 된다.

2. 관련연구

2.1 정보추출규칙의 반 자동생성

준구조화된 문서에서 정보를 추출하기 위해서는 정보 추출 규칙을 사용한다. 정보 추출 규칙은 생성 방식에 따라 다음과 같이 구분 되어 진다.

1)정보추출규칙의 수동생성: 정보 소스마다 사용자가 직접 어떤 정보를 얼마만큼 추출 해야 하는지에 대한 규칙을 직접 작성해야 한다.
2)정보추출규칙의 반 자동생성: 도메인 지식이나 미리 설정된 몇가지 규칙을 통해서 문서를 분석하고 분석하지 못한 부분은 사용자가 수동으로 추출 규칙을 작성한다.
3)정보추출규칙의 자동생성: 정보 추출 시스템이 자동으로 문서를 분석하여 정보 추출 규칙을 생성한다.

준구조화된 문서에서 정보 추출을 위한 규칙을 수동으로 생성한다는 것은 어려운 일이다. XWRAP[3]시스템은 웹상의 정보 소스의 정보 추출 규칙을 반 자동으로 생성하기 위해 사용자와 상호작용하는 방안을 제시하고 있다. 이 연구의 목적은 변환하기 어려운 HTML문서를 프로그램 친화적인 XML출력으로 변환하는 것이다. 여기서 웹 문서에서 의미있는 부분을 인식하는 것이 어렵기 때문에 이 부분에 대해서 수동으로 입력을 받아서 처리하고 있다. 그러나 사용자의 입력과 HTML태그에 의존적이며 계층구조로 나타나지 않는 웹페이지에서는 정보를 추출하지 못하는 문제점이 있다.

2.2 자연어 처리를 통한 Wrapper 자동 생성

일반적인 정보 추출 시스템들은 주로 키워드 매칭(matching)과 토큰 구분자(token delimiter)에 의존해 정보 추출 규칙을 학습한다. 이러한 시스템들은 대부분 다음 단계들을 수행한다.

① Token and Tagging: 입력 문서에서 각 단어들로 나눌 수있고, 각 단어들은 품사와 의미로 분류된다.
② 추출: 하나의 추출 규칙을 문자에 일치시켜 본다. 만약 문장과 추출 규칙이 일치하면, 패턴의 구성요소를 기반으로 관련된 정보만을 식별한다.
③ 결과생성: 정보 추출 시스템은 결과를 출력하기 위해 미리 정의된 형태에 추출된 정보를 채우는 방식을 이용한다.

이러한 시스템으로 AutoSlog[5]와 CRYSTAL[6,7]이 있다.

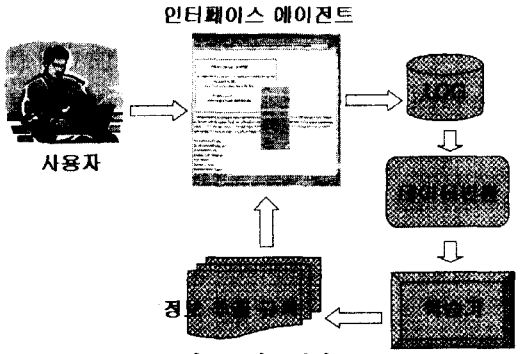
2.3 구조화 및 준구조화 문서에서 정보추출규칙의 자동생성

자연어만으로 이루어진 문장이 아닌 경우에는 자연어 처리를 위한 알고리즘을 적용할 수는 없다. 따라서 자연어가 아닌 문서의 정보 추출 규칙의 자동생성을 위해서는 각 문서의 형태에 따른 문서 분석 연구가 필요하다. WIEN[1,2]은 인터넷상의 많은 정보들이 상관관계가 있는 데이터로 존재한다고 보고, wrapper학습에 이러한 상관관계를 이용한다. 이 시스템은 wrapper의 귀납적 자동 생성방법을(induction)을 제시하였다. 이러한 시스템은 확장성이나 관리자의 노력을 많이 필요로 하지 않는 시스템이지만 정확성(precision)은 50% 내외이다. 그러나 특정 도메인에서는 높은 정확성을 나타낸다.

3. 시스템구조

본 논문에서 제안하는 학습가능한 인터페이스 에이전트의 구조는 그림 1과 같으며 인터페이스에이전트의 기능은 다음과 같이 요약할 수 있다

1)추출할 정보의 종류 획득
2)학습한 추출 정보와 같은 상태나 상황에 대한 판단
3)추출한 결과에 대한 사용자의 Feedback수용
4)도메인 지식의 갱신



<그림 1> 시스템의 구조

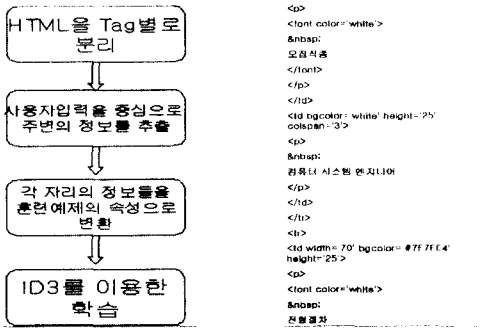
3.1 인터페이스 에이전트

본 논문에서 가장 중요한 인터페이스 에이전트의 기능은 "추출할 정보의 종류 획득"에 있다. 사용자는 팝업메뉴의 아이템을 선택함으로써 자신이 원하는 정보에 대한 종류를 입력한다. 본 논문에서 사용자로부터 입력받는 정보는 두 가지이다. 첫째는 사용자가 선택한 정보이며, 둘째는 그 주변에 나타나는 정보이다. 이것은 어떤 정보는 주변에 나타나는 다른 정보와 관련성을 가지고 나타난다는 가정하에 주변 정보도 함께 입력 받는 것이다.

인터페이스 에이전트는 사용자로부터의 입력을 positive example로 인식하고 이를 정보 추출 학습을 위해 기록한다. 이때 인터페이스 에이전트는 사용자가 입력한 정보만을 기록하지 않고 입력한 정보의 주변 상황을 기록하게 된다. 사용자의 입력에 의해 획득된 정보는 정보 추출을 학습하기 위해 사용된다.

3.2 데이터 변환기

본 연구에서 사용하고자 하는 학습 알고리즘은 decision tree[12]이다. Decision tree를 사용하기 위해서는 먼저 사용자가 입력한 정보들을 훈련예제(training example)로 변환해야 한다.



<그림 2> 훈련예제의 속성추출 <그림 3>훈련예제로 변환

훈련예제로 변환은 그림 2와 같은 과정을 통하여 이루어진다. 각 자리의 정보들을 훈련 예제의 속성으로 변환하는 과정은 그림 3과 같다. 그림 3에서 사용자가 추출하고자 하는 정보가 "모집직종"인 경우 사용자는 그림 3과 같이 인터페이스 에이전트를 이용하여 "컴퓨터 시스템 엔지니어" 부분을 선택했다고 가정하자. 준 구조화 및 구조화된 문서의 특성상 추출하고자 하는 값이 있으면 그 값을 기술해 주는 label이 존재한다. 즉 label과 value로 구성된 하나의 단위로 볼 수있다. 따라서 n과 m은 사용자가 입력한 단어의 앞과 뒤에 나타나는 단어 중 선택된 단어의 수이다. 이렇게 결정된 n과 m사이의 정보에서 특정 도메인이 아닌 일반적인 정보 추출 시스템에서 사용할 수 있는 속성을 추출할 수 있는 정보에 대해 연구하였다.

3.3 학습기

이렇게 만들어진 훈련예제를 중심으로 식 1과 같은 엔트로피를 이용하여 학습에 용이한 분류자(classifier)를 추출하고 식 2와 같은 information gain 함수를 이용하여 최적화된 정보 추출 규칙을 만들어 내고자 한다.

$$Entropy\ (s) \equiv \sum_{i=1}^{c} - p_i \log_2 p_i \qquad (1)$$

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \qquad (2)$$

시스템의 구조에서는 실제 정보를 추출하게 될 정보 추출기가 존재하지 않는다. 왜냐하면 인터페이스 에이전트 내부에 정보추출을 위한 규칙 해석기와 정보 추출기가 존재하기 때문에 좋은 정보 추출 규칙을 위한 규칙을 생성하기 위해서는 좋은 훈련 예제들이 많이 필요하다. 이것은 overfitting된 훈련 예제로 인한 적용범위가 적은 규칙을 생성을 방지하기 위함이다.

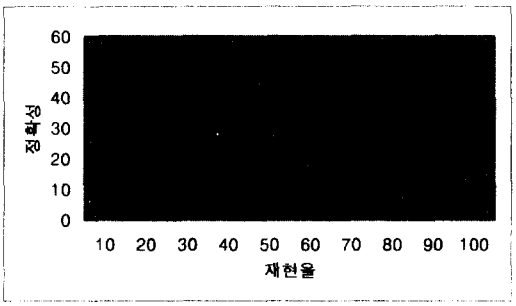
사용자 피드백을 이용하여 정확도가 높은 정보 추출 규칙을 생성해 내고자 한다. 이를 위해 인터페이스 에이전트는 사용자가 인터페이스 에이전트를 훈련시키기 위해 사용한 문서와 유사한 페이지로부터 학습한 정보 추출 규칙을 사용하여 사용자에게 추출 결과를 제공한다. 만약 사용자가 정보 추출 결과에 만족한다면 정보추출 규칙에는 변화가 생기지 않는다. 그러나 사용자의 명시적인 명령없이 인터페이스 에이전트가 자율적으로 행한 결과에 만족하지 않는다면 그 상황을 인터페이스 에이전트는 negative agent로 받아들이도록 하여 정보 추출 규칙의 정확도를 높이게 된다. 이것은 인터페이스 에이전트가 훈련예제로 사용자로부터 positive example만 받아들여 overfitting되는 결과를 막기위해서 사용된다.

4. 시스템 성능 평가

(1)정확성(Precision): 인터페이스 에이전트가 추출한 정보의 정확성을 평가한다. 정확성은 $\frac{\text{정확한정보의수}}{\text{추출한정보의수}}$ X 100으로 계산한다.

(2)재현률(Recall): 인터페이스 에이전트가 추출한 정보가 얼마나 많은 정보를 추출할 수 있는지를 평가한다. 재현률은 $\frac{\text{추출한정보의수}}{\text{추출되어야할정보의수}}$ X 100 으로 계산한다.

(3)안정성을 갖춘 규칙 생성을 위한 예제 수: 정확성과 재현률의 임계값(threshold)을 만족시키기 위해 사용되는 예제수를 바탕으로 보다 빠르고 안정화된 학습 방법 평가한다.



학습 데이터는 각 학회에서 제공하는 call for paper를 사용하였다. 사용자는 논문을 제출할 목적으로 웹페이지를 열어 보았다는 가정하에 submission에 관한 정보를 추출하는 규칙을 생성하는 wrapper를 생성해 보았다. 특히 사용자는 논문제출 시한과 양식을 알고자 할 것이다. 각 페이지에서 고르게 하나 이상의 정보를 추출할 수 있었으나 주변에 나타나나는 정보의 순서와 어휘의 종류에 따라 정확성이 영향을 받는 것을 알 수 있었다. 테스트를 위해 사용한 웹페이지 수는 80개이며 재현률에 대하여 40~50%정도의 정확도를 보여준다.

5. 결론 및 향후연구

준구조화된 웹 문서에서 사용자가 원하는 정보를 찾는 문제에 대한 해결책으로 wrapper 기술을 이용할 수 있다. 본 논문은 기계 학습 방법을 이용하여 정보 추출 규칙(wrapper)의 패턴을 학습할 수 있는 인터페이스 에이전트를 제시하였다. 사용자는 웹 문서에서 원하는 정보의 위치를 지정하여 데이터를 인터페이스 에이전트에게 학습시키고, 인터페이스 에이전트는 학습된 추출 규칙으로부터 새로운 웹 문서에 대하여 사용자가 원하는 정보를 찾아 준다.

향후 과제로 사용자가 알려주는 정보를 좀더 세분화하여 레이블(Label)과 타입(type)을 사용할 필요가 있다. 대개의 경우 사용자는 하나의 단어보다 블럭단위로 자신이 원하는 정보를 지정하기를 원할 것이다. 따라서 인터페이스 에이전트에서 데이터를 블록단위로 입력할 수 있도록 확장할 필요가 있다. 또 사용자가 선택한 부분과 선택한 부분의 앞과 뒤의 단어를 따로 나누어 분석 한다면 좀 더 정확한 결과를 얻을 수 있을 것으로 예상된다. 또 학습기에 사용되는 알고리즘도 변경할 필요가 있다. ID3 알고리즘이 가지는 문제점은 특정 단어를 선택했을 경우 그 앞과 뒤에 나타나는 단어가 한정되어 있고 또 반복되어야 좋은 결과를 얻을 수 있는데 반해 어떤 문서에서는 단어가 반복해서 나타나지 않았다. 이러한 문제점을 해결할 수 있는 알고리즘으로 대체해야 할 것이다.

6. 참고문헌

[1] Nicholas Kushmerick, "Wrapper Induction for Information Extaction", Proceedings of 15th International Conference on Artificial Intelligence (IJCAI-95), pp.729-735.

[2] Nicholas Kushmerick, "Wrapper Induction for Information Extaction", University of Washington Department of Computer Science Ph.D thesis, 1997

[3] Ling Liu, Calton Pu, Wei Han. "XWrap: An XML-enabled Wrapper Construction System for Web Information Sources", In *Proceedinghs of the 16th International Conference on Data Engineering (ICDE 2000)*, March, 2000, San Diego, CA (IEEE CS Press). pp611-621

[4]Ion Muslea, "Extraction Patterns for Information Extraction Tasks: A Survey,"Proceedings of the 16th Conference on Artificial Intelligence (AAAI-99), pp1-6, 1999

[5] S. Huffman, "Learning Information Extraction Patterns from Examples," Workshop on New Approaches to Learning for Natural Language Proceessing, pp. 127-142, 1996

[6] S. Soderland, "CRYSTAL: Inducing a Conceptual Dictionary," Proceeding of 15th International Conference on Artificial Intelligence, pp.1314-1319, 1995

[7] S. Soderland, D. Fisher, and W.Lehnert., "Automatically Learned vs. Hand-crefted Text Analysis Rules," Technical Report TE-44 at Center for Intelligent Information Retrieval, University of Massachusetts, 1997.

[8] S. Soderland, "Learning Text Analysis Rules for Domain-Specific Natural Language Processing," University Massachusettes Amherst, Department of Computer Science Ph.D thesis, 1997.

[9] S. Soderland, "Learning Information Extraction Rules for Semi-Structured and Free Text," http://www.cs. washington. edu/homes/soderland/WHISK.ps.

[10]. 최중민, "인터넷 정보 가공을 위한 에이전트," 정보처리학회지, Vol.4, No5, pp101-109, 1997.

[11]. 양재영, "전자상거래에서 상점 Wrapper 생성을 위한 지능형 에이전트의 학습 방안 연구," 한양대학교 전자계산학과 석사학위 논문, 2000.

[12]. Tom M. Mitchell, "Machine Learning," McGraw-Hill. pp52-80, 1997