논문 2009-46CI-3-2

은닉 마코브 모델을 이용한 인터넷 정보 추출

(Hidden Markov Model-based Extraction of Internet Information)

박 동 철*

(Dong-Chul Park)

요 약

본 논문은 은닉 마코브 모델을 이용한 인터넷 정보 추출 방법을 제안하고, 인터넷상의 웹 사이트에서 상품가격을 효율적으로 추출하는 문제에 적용되었다. 제안된 방법에서 시스템으로 입력되는 데이터는 검색엔진의 인터페이스 URL 인데, 상품의 이름을 포함하며, 시스템의 출력은 추출된 각 상품의 상품명, 가격, 사진, 그리고 URL을 목록형태로 보여준다. 주어진 관찰 데이터를 이용해, 은닉 마코브 모델의 학습단계에서는 Maximum Likelihood 알고리듬과 Baum-Welch 알고리듬이 학습에 사용되었으며, 학습된 은닉 마코브 모델을 이용하여 시스템의 출력을 찾는 방법으로는 Viterbi 알고리듬이 사용되었다. 제안된 HMM 기반의 정보 검출기는 실제상황에서 수집된 관찰데이터에 대해 실험이 수행되었는데, 기존의 PEWEB 알고리듬에 비해 검출도와 정확도에서 매우 향상된 결과를 보이고 있으며, 특히 정확도에서는 99%이상의 높은 결과를 보여주고 있다. 한편, 보다 충실한 학습을 위해 학습 데이터의 수를 800개 이상으로 증가시켰을 때 검출도 역시 약 93%로 향상된 성능을 보여주었다.

Abstract

A Hidden Markov Model(HMM)-based information extraction method is proposed in this paper. The proposed extraction method is applied to extraction of products' prices. The input of the proposed IESHMM is the URLs of a search engine's interface, which contains the names of the product types. The output of the system is the list of extracted slots of each product: name, price, image, and URL. With the observation data set, Maximum Likelihood algorithm and Baum-Welch algorithm are used for the training of HMM and The Viterbi algorithm is then applied to find the state sequence of the maximal probability that matches the observation block sequence. When applied to practical problems, the proposed HMM-based system shows improved results over a conventional method, PEWEB, in terms of recall ration and accuracy.

Keywords : HMM, internet, data extraction, Baum-Welch algorithm, Viterbi algorithm

I. 서 론

일반적으로 정보의 추출 (Inforamtio Extraction: IE)은 구조적으로 완전하지 않은 자료에서 필요한 정보 또는 지식을 추출하여, 정형화된 형태의 데이터베이스로 만드는 과정으로 규정지을 수 있다. 웹 페이지와 같이 구조적으로 불완전한 자료에서 정보를 추출하는 모델에 대한 연구는 웹의 폭발적인 사용으로 지난 십여 년 동안 집중적으로 진행되어 왔는데, 수많은 인터넷 상거래 웹 사이트에서 자동적으로 상품의 정보를 추출하는 것은 웹 사용자를 도울 뿐 아니라, 인터넷 상거래를 촉진시키는 요소가 되고 있다[1~9]. 웹에서 상품의 정보를 추출하는 과정은 웹문서에서 어떻게 상품구역(product region)을 인식해 내는 가, 어떻게 상품구역을 몇 개의 레코드로 정확하게 분리해 내는 가, 그리고, 어떻게 광고 등의 상품정보와는 관계없는 것들 제거하는 가에 주안점이 있다고 할 수 있다.

목표상품을 검색하기 위해, 탐색 엔진 (Search engine)으로 부터 검색된 웹 페이지에서의 정보추출 과정은 특별히 웹 정보 추출 (Web information extraction)로 규정될 수 있는데, 초기의 연구에서는 주로 IE를 위한 규칙의 생성과 이용에 의존하고 있었다.

* 정회원, 명지대학교 정보공학과
(Dept. of Information Eng., Myong Ji University)
※ 본 연구는 한국과학재단 특정기초연구 (R01-2007-000-20330-0) 지원 에 의한 것임.
접수일자: 2009년3월27일, 수정완료일: 2009년5월4일

즉, 초기의 연구에서는 도메인 온톨로지와 휴리스틱을 이용하는 방법[1], 이를 발전시켜 몇가지 휴리스틱을 추가하는 대신 도메인 온톨로지를 사용하지 않는 방법[2], 클러스터링에 의한 문법유도 방법[3] 등이 있었는데, 이들 각각의 특징적인 장점에도 불구하고, 모두 추출률과 정확도의 성능이 만족스럽지 않았다[5]. 이들 중 특별히 HTML에 기반한 기술의 연구는 2000년도 초반에 몇 가지의 자동화 또는 그에 가까운 IE 방법들로, IE based on Pattern Discovery (IEPAD)[6], Object Mining and Extraction System (OMINI)[2], Mining Data Records (MDR)[5], 그리고 Product Extraction from the Web (PEWEB)[7]이 제안되었다.

IEPAD는 웹 페이지의 HTML 내용을 분석(parsing)하여 스트링의 형태로 만들고, 목적 패턴에 해당할 수 있는 후보 패턴을 찾기 위해 Patricia 트리[10]를 이용한다. 한편, OMNI는 주어진 입력 웹 페이지에서 태그 트리 (tag tree)를 만들고, 몇 가지의 관련 정보(heuristics)를 이용하여, 필요한 데이터 레코드를 포함할 것 같은 sub-tree를 추출한다. 추출된 sub-tree는 다시 추가의 관련 정보 (heuristics)에 의해 생성된 분리기(separator)를 이용해 데이터 레코드로 변환된다. 그러나 OMNI에서 사용된 분리기는 단 한 개의 HTML tag만을 가지는 것이기 때문에, 그 성능에 제한을 받게 된다. 한편, PEWEB에서는 목표상품을 포함할 가능성이 있는 부분을 찾아내기 위해 엔트로피의 개념을 도입한다. PEWEB에서 결과는 엔트로피 비율이 큰 sub-tree node들로 주어진다. 한편, MDR은 edit distance를 사용하는 스트링 매칭 방법을 이용하여 일반화된 노드를 가지는 데이터 구역을 찾아내고, 각각의 일반화된 노드에서 휴리스틱을 사용해 원하는 데이터 레코드를 찾아낸다. 그러나, MDR은 edit distance를 사용하기 때문에 속도가 느리고, 휴리스틱의 정교함에 의해 정확도가 매우 좌우되는 경향이 있다[7].

본 논문에서는 사용자의 편의성과 검색 속도를 최대로 보장하면서, 결과의 정확도에서 선행연구의 방법들과 차별되는 은닉 마코브 모델에 기반한 인터넷 정보추출 방법을 제안한다. 제안된 방법은 인터넷 상에 있는 어떤 상품의 가격과 관련된 수많은 웹 페이지에서 목적하는 상품의 규격에 최대한의 정확도를 지니는 상품의 목록을 추출하기위한 방법이다.

본 논문의 Ⅱ장에서는 제안된 방법의 기반이 되는 은닉 마코브 모델에 대해 간단히 살펴보고, Ⅲ장에서는 제안된 은닉 마코브 모델 기반의 인터넷 정보추출 시스템을 요약하여 설명한다. Ⅳ장에서는 제안된 방법의 성능을 평가하기 위하여, 웹 상에서 실제 상품의 가격을 추출하고, 결과를 가장 최신의 알고리듬의 하나인 PEWEB의 결과와 비교 분석하며, Ⅴ장에서는 결론을 내린다.

Ⅱ. 은닉 마코브 모델

은닉 마코브 모델(Hidden Markov Model: HMM) HMM에서 확률적으로 어떤 분포를 갖는 각각의 상태(state)는 직접적으로 파악되지 않지 않지만, 그 상태에 의해 영향 받는 매개변수들의 파악은 가능하다. HMM은 음성 데이터 등 시간적으로 변화를 가지는 데이터의 인식에 오랫동안 성공적으로 사용되어 왔다[11~14].

1차 HMM에서 이산 출력은 유한 상태 오토마타이며, 5개의 변수인 $\{S, V, \Pi, A, B\}$로 표현할 수 있는데, 여기서, S는 N 개의 은닉상태 값의 집합인 $\{s_1, s_2, ..., s_N\}$이고, V는 M 개의 관측가능관측 값의 집합인 $\{v_1, v_2, ..., v_M\}$이며, Π는 모든 상태의 초기 확률의 집합인 $\{\pi_1, \pi_2, ..., \pi_N\}$이고, A는 각 상태변이의 확률을 정의하는 맵핑, $\{P(q \rightarrow q')\}$,이며, B는 각 상태에 대한 각각의 관측의 출구 확률(emission probability)을 정의하는 맵핑인 $\{P(q \uparrow \sigma)\}$를 나타낸다[14].

주어진 HMM을 사용하기 위해서는 목적에 따라 다음의 세 가지 문제를 풀어야한다.

1. 문제 1:
   주어진 T개의 관측 sequence, $O = O_1 O_2 ... O_T$, 과 모델, $\lambda = (A, B, \Pi)$,에 대해서, T개의 관측확률인 $P(O|\lambda)$를 구한다.
2. 문제 2:
   주어진 T개의 관측 sequence, $O = O_1 O_2 ... O_T$, 과 모델, $\lambda = (A, B, \Pi)$,에 대해서, 관측 sequence를 발생시키는 최적의 상태 sequence인 $Q = q_1 q_2 ... q_T$를 구한다. 여기서 최적도를 측정하는 척도는 Maximum Likelihood 이다.
3. 문제 3:
   모든 관측 sequence에 대해 likelihood를 최대화 시키는 $\lambda = (A, B, \Pi)$의 파라메터를 구한다.

본 논문에서는 주로 모델 파라메터를 학습하는 문제 3과 데이터 레코드를 추출하는 문제 2에 집중하게 된다.

Ⅲ. 은닉 마코브 모델 기반의 정보추출 시스템

본 논문에서 제안되는 시스템의 대략적인 구성도는 그림 1과 같은데, Page retrieval, Segmentation and Parser, Segment 필터, 관측 생성기, 그리고 추출기(extractor)의 다섯 가지 부분으로 구성되어 있다. 시스템으로 입력되는 데이터는 검색엔진의 인터페이스 URL 인데, 상품의 이름을 포함하며, 시스템의 출력은 추출된 각 상품의 상품명, 가격, 사진, 그리고 URL을

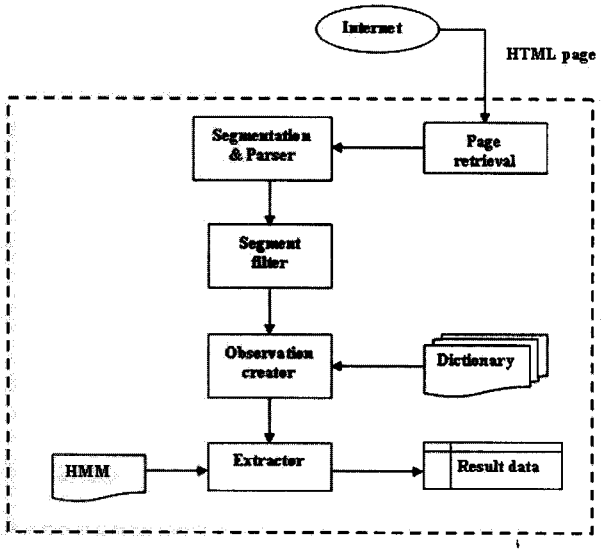

그림 1. HMM 기반의 정보추출기 개략도
Fig. 1. Overview of HMM-based information extraction system.

목록형태로 보여준다.

일반적으로 웹 페이지는 tag로 구성되는 sequence의 HTML 문서인데, HTML segment 트리는 웹 페이지의 HTML tag들로 만들어진다. 한 개의 segment는 기본 요소의 그룹을 나타내며, 본 논문에서는 이를 기본 단위로 사용한다. 본 논문에서 사용하는 네 가지의 주요 segment는 paragraph, table, list, 그리고 heading 이며, HTML 문서는 가장 큰 segment이다. 그림 2는 HTML 페이지와 대응되는 segment 트리의 한 예를 보여주는데, 상품 영역은 다음의 성질들을 가지는 tag 노드와 비슷한 노드가 여러 개 존재하는가를 찾아서 결정한다:

1) 모든 노드는 같은 부모 노드를 갖을 것
2) 모든 노드들은 서로 인접할 것
3) 각 노드는 상품명 또는 가격을 가지고 있을 것

Segment 필터는 어떤 segment 가 상품 segment인가를 결정하며, 상품의 이름, 설명, 사진, 가격을 포함하는 인접한 segment는 보다 큰 segment로 확장되어 관측생성기에 사용 된다. 본 논문에서 각 segment는 학습된 HMM에 대한 입력 관측데이터로 사용된다. 각각의 상품에 대해 HMM이 각각 존재하는데, 각 HMM은 같은 구조를 갖는다. 추출기는 target 상태들에 의하여 모델링되는데, 가장 가능성 높은 상태의 token은 Viterbi 알고리듬을 사용하여 찾게 된다[14].

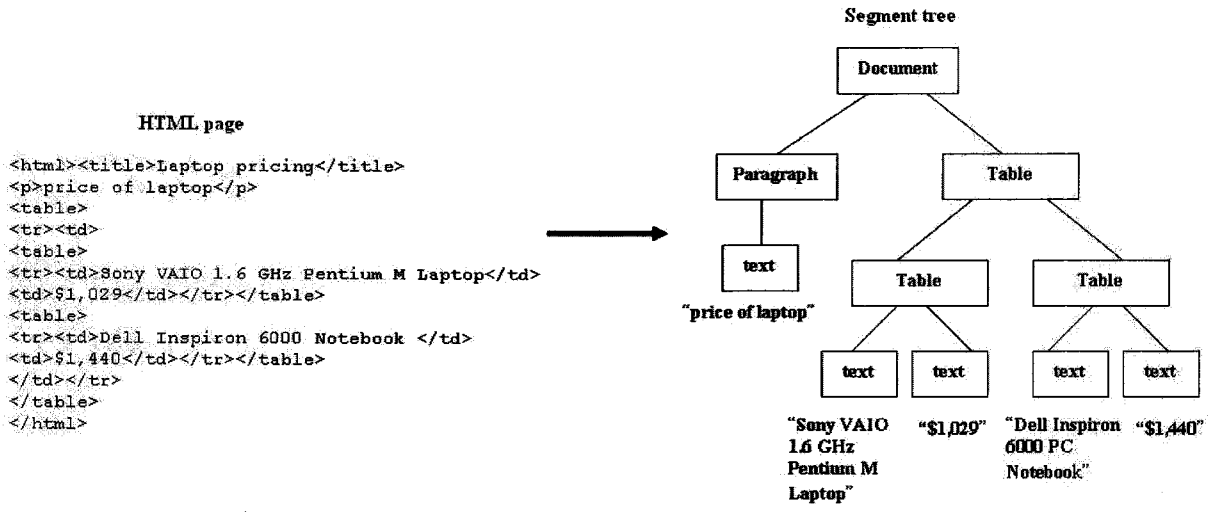

그림 2. HTML 문서와 대응되는 segment 트리 의 예
Fig. 2. A HTML document and corresponding segment tree.

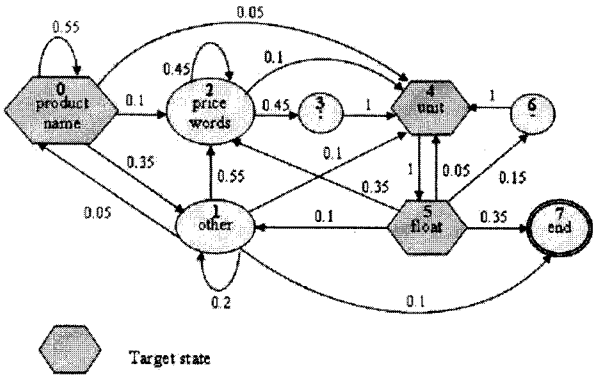

그림 3. USB 메모리 추출을 위해 학습된 HMM
Fig. 3. A trained HMM for extracting USB memory problem.

본 논문에서는 노트북 컴퓨터, USB 메모리, 웹 카메라 등의 각각의 상품 에 대한 HMM 을 학습시키기 위하여, 일반적인 상업용 검색엔진을 이용해 각 상품에 대해 추출된 HTML 페이지를 관측자료로 사용하였으며, Maximum Likelihood (ML) 알고리듬과 Baum-Welch 알고리듬이 학습에 사용되었다. HMM의 학습과정에는 다음의 세 가지 파라메터가 결정된다. 먼저 초기 확률 π 은 다음의 식으로 구해진다.

$$\pi_i = \frac{I(i)}{\sum_{j=1}^{N} I(j)}, \quad 1 \le i \le N \tag{1}$$

여기서, $I(\bullet)$ 는 어떤 상태에서 시작하는 초기 확률이며, N은 HMM 모델의 총 상태의 수이다.

한편, 상태 전이 확률 매트릭스 A의 요소 a_{ij}는 다음과 같이 정의 된다:

$$a_{ij} = \frac{C_{ij}}{\sum_{k=1}^{N} C_{ik}}, \quad 1 \le i,j \le N \tag{2}$$

여기서, C_{ij} 는 상태 s_i 에서 상태 s_j로 전이하는 경우의 수를 말한다.

또한, emission 확률 매트릭스 B를 구하기 위해, emission 확률 요소 $b_j(v_k)$ 는 다음으로 정의된다:

$$b_j(v_k) = \frac{E_j(v_k)}{\sum_{i=1}^{M} E_j(v_i)}, \quad 1 \le j \le N, \quad 1 \le k \le M \tag{3}$$

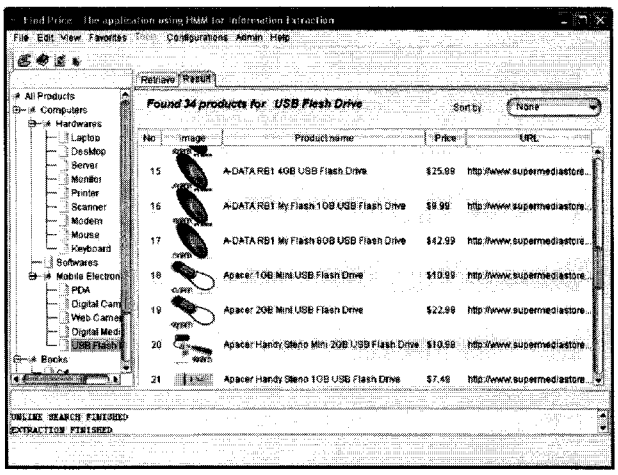

그림 4. USB 관련 정보추출 화면의 예
Fig. 4. A screen of USB product extraction.

여기서, $E_j(v_k)$ 는 v_k 가 s_j에서 나온 경우의 수이다.

그림 3에서는 " USB 메모리"를 추출하는 문제에 대한 학습된 HMM을 보여주고 있는데, 한 가지 관찰에 대한 emission 확률은 K가 그 관찰에 대한 경우의 수일 때, 다음의 식으로 표현된다:

$$b_j(o_t) = \sum_{k=1}^{K} b_j(o_{tk}) \tag{4}$$

이렇게 모든 파라메터들을 구하고 나면, 한 덩어리의 관찰 sequence와 확률적으로 가장 비슷한 상태 sequence를 Viterbi 알고리듬을 이용해 구해낸다[14]. 추출된 데이터 레코드는 표의 형태로 저장되는데, 그림 4는 "USB 메모리"를 추출하는 문제에 대한 결과화면의 예이다.

IV. 실험 및 결과

4.1 성능 평가 도구

인터넷 정보추출기의 성능평가를 위해 정보추출기의 성능평가에 보편적으로 사용되는 다음 정의의 검출도와 정확도가 사용되었다[4].

검출도 = CE/CO, 정확도=CE/EO

여기서, CE 는 정확하게 추출된 관찰의 총 수, EO는 그 페이지에서 추출된 관찰의 총수, 그리고 CO 는 정확한 관찰(목표 관찰)의 총 수를 나타낸다.

4.2 실험 및 성능 평가의 결과

제안되는 HMM 을 이용한 인터넷 정보추출 방법에 대해, 그 유용성을 평가하기 위하여, 실제 데이터를 이용해 실험을 수행하였다. 실험의 결과는 PEWEB의 결과와 비교하였다. 성능분석을 위해 OMINI[2~3], IEPAD[6], MDR[5], PEWEB[7]등의 기법들과의 성능비교가 중요하지만, 이미 기존의 연구 결과에서 IEPAD 과 OMNI 보다 MDR이 월등한 성능 향상이 보고되었고[5], PEWEB는 MDR에 비해 검출도와 정확도에서 더욱 향상된 결과를 보고하였으므로[7], 본 논문에서는 HMM 기반의 방법과 PEWEB의 성능 비교 분석에 주안점을 두고자한다. 본 연구에서 PEWEB에 관련된 실험은 일반에게 개방되어 있는 실행코드를 이용하여 HMM기반의 모델과 동일한 학습/검증 환경에서 실행하였다[8].

PEWEB에서의 데이터 block은 상품의 이름, 제조사, 가격, 사진 등의 유용한 정보와 함께, 광고, 관련 사이트 링크 등의 불필요한 정보들을 포함하고 있으며, 어떤 경우에는 추출된 레코드가 추가적인 상품의 내용을 포함하는 경우도 있다. 따라서 유용한 데이터 필드만을 가지는 데이터 레코드를 추출하지 못할 가능성이 있다. PEWEB 와는 다르게, 제안되는 HMM 기반의 정보추출 시스템은 노이즈에 해당되는 정보들을 제외한 필요한 내용만을 가지는 데이터 레코드를 추출하므로, 웹페이지를 좀 더 충실하게 다룰 수 있다.

PEWEB와의 성능비교 실험을 위해, 각각의 상품에 대해 상업용 일반 검색엔진인 Google로부터 먼저, 총 200개의 URL을 관측자료로 수집하였다. 표 1은 이렇게 수집한 웹사이트를 보여주는데, 이 리스트상의 각 웹사이트는 상품의 특징자료와 가격을 충분히 포함하고 있었는데, 여기에 없는 두 개의 사이트는 부정확한 검색으로 관련정보를 포함하고 있지 않아서, 리스트에서 제거 되었다. HMM의 학습에서는 8개의 상태를 가지는 모델을 사용하였다. 실험에서 무작위로 추출된 100개의 관측자료를 HMM의 학습 데이터로 사용하고, 나머지 100개의 관측자료는 학습된 HMM을 검증하는 테스트 데이터로 사용하였다.

표 1은 제안된 HMM 기반의 인터넷 정보추출기와 PEWEB의 실험에 대한 성능의 비교 분석 자료이다. 실험의 대상은 다른 포맷과 상품정보를 갖고 있는 총 18개의 웹 사이트이다. 표 1에서 2번째 열은 각 웹 사이트의 URL 주소인데, 일부는 URL이 너무 길어서 생략한 부분도 있다. 이들 웹 사이트는 Google로 검색한 결과로 얻어진 것들로, 각 웹 페이지는 usb 메모리, 노트북

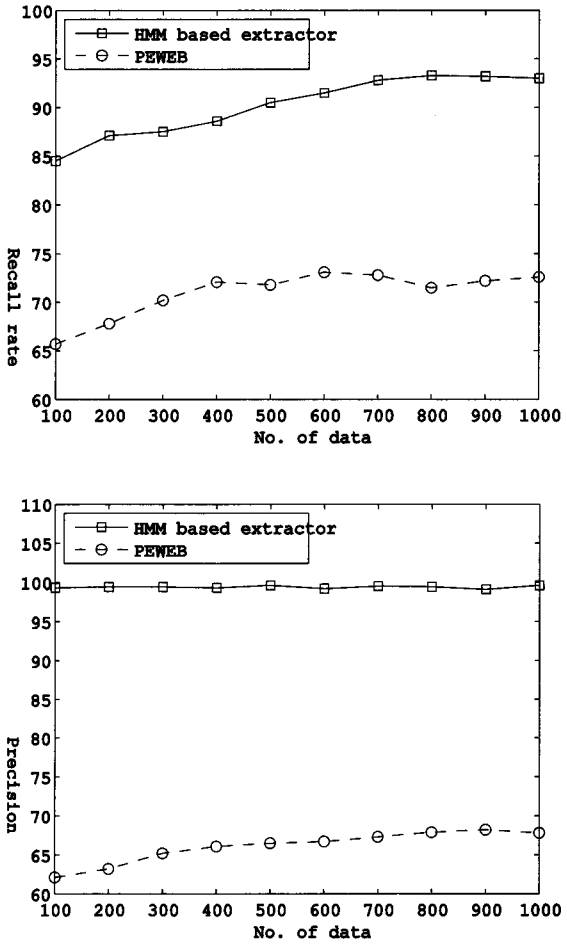

그림 5. Observation data 수의 변화에 따른 성능의 변화: (a) 검출도 (b) 정확도

Fig. 5. Performance vs. Number of observation data : (a) Recall (b) Accuracy

컴퓨터, 웹 카메라 등의 상품에 대한 가격, 사진, 설명 등의 충분한 정보를 가지고 있다. 실험에서는 컴퓨터와 관련된 몇 개의 상품을 query로 사용하여 정보추출 시스템의 성능을 평가하는데 사용하였다. 3번 째 열은 해당되는 URL에 있는 목표 상품의 수를 나타낸다. 4번째와 5번 째 열은 제안된 HMM 기반의 정보추출기에 의해 추출된 상품의 수와 정확하게 추출된 상품의 수를 각각 나타낸다. 제 6열과 제 7열은 PEWEB에 의해 추출된 상품의 수와 정확하게 추출된 상품의 수를 각각 나타낸다.

표 1의 결과에서 보듯이, PEWEB의 경우 평균 검출도가 65.7%인 것에 비해, HMM 기반의 검출기는 84.5%를 나타내었다[15]. 즉, HMM 기반의 검출기가 PEWEB에 비해 평균적으로 약 18.8%의 검출도 향상을 보여주고 있다. 한편, 정확도의 측면에서는 PEWEB가

표 1. 실험에 사용된 자료와 실험결과
Table 1. Experimental data and results of experiments

번호	웹 사이트	상품수	HMM 기반		PEWEB	
			검출수	참검출수	검출수	참검출수
1	www.flash-memory-store.com	21	16	16	26	20
2	www.tigerdirect.com	19	16	16	11	2
3	www.supermediastore.com	32	30	28	38	29
4	www.usbflashdrivestore.com	25	24	24	27	24
5	www.buy.com	15	14	14	9	0
6	www.ecost.com	25	22	22	25	25
7	www.overstock.com	11	10	10	14	10
8	usbflashstore.com	25	24	24	27	25
9	shopping.aol.com	16	16	16	9	7
10	www.pricespider.com	10	10	10	15	0
11	www.usanotebook.com	27	21	21	28	27
12	www.dealtime.com	21	17	17	16	5
13	www.geeks.com	28	12	12	55	28
14	www.nextag.com	11	10	10	15	0
15	www.kenzo.com	16	13	13	20	16
16	www.mysimon.com	25	21	21	0	0
17	www.pricewatch.com	15	15	15	33	15
18	computing.kelkoo.co.uk	20	17	17	5	5
총계		362	308	306	378	238
평균 검출도/정확도			검출도 : 84.5% 정확도 : 99.3%		검출도: 65.7% 정확도: 62.1%	

평균 62,1%를 보이는데 비해, HMM 기반의 검출기는 99.3%의 정확도를 보이고 있어, 제안된 HMM 기반의 검출기가 매우 정확한 검출 결과를 보이고 있음을 알 수 있다. 이는 HMM의 기본적인 특성인 학습의 성능에 따른 기대되는 정확한 인식능력에 기인한다고 할 수 있다. 한편, PEWEB가 검출에서 실패하는 경우는 데이터 범위가 한 개의 데이터 레코드로 구성되었을 때, 데이터 범위를 인식하지 못 하는 경우가 많았다. 또한, 한 개 이상의 광고가 연속되는 데이터 레코드를 두 개 또는 그 이상의 데이터 레코드로 분리할 때 PEWEB는 검출에 실패할 경우가 많았다.

일반적으로, HMM 기반의 인식기는 학습에 사용되는 관측자료(Observation)의 질과 양에 의해 그 성능이 영향을 받기 때문에, 관측자료의 증가에 따른 검출도와 정확도의 변화를 측정하기 위해, 관측자료를 추가로 발생시켜 학습 자료의 수를 100개에서 1000개로 변화 시킨 후, 같은 실험을 반복하였는데, 그 결과가 그림 5에 주어진다. 그림 5에서 보듯이 PEWEB는 검출도와 정확도 모두 관측자료 수의 증가에 따라 어느 정도 향상을 가져오고 있다. 그러나 약 600개 이 후에서는 결과의 향상이 눈에 띠게 줄어서 포화된 결과를 보이고 있다. HMM 기반의 검출기에서는 검출도는 관측자료 수의 증가에 따라 검출도가 눈에 띠게 증가하지만, 정확도는 100 개의 경우와 대동소이하다. 검출도에서도 약 800 개 이후에서는 포화된 결과를 보여주고 있다. HMM이 학습데이터에 의해 영향을 받는 모델이기 때문에, 학습 데이터의 수가 매우 중요한 역할을 하고 있음이 이 경우에서도 관찰되고 있다.

V. 결 론

본 논문에서는 은닉 마코브 모델 기반의 인터넷 정보 추출 방법이 제안되었는데, 인터넷 웹 사이트에서 상품 가격을 효율적으로 추출하는 문제에 적용되었다. 제안된 방법은 상품의 레코드를 포함하는 데이터의 범위를 정확하게 인식해낼 수 있는 장점이 있다. 대부분의 전통적인 방법에서는 데이터 범위가 한 개의 데이터 레코드로 구성되었을 때, 데이터 범위를 인식하지 못 하는

경우가 많다. 또한, 한 개 이상의 광고가 연속되는 데이터 레코드를 두 개 또는 그 이상의 데이터 레코드로 분리할 때 문제가 발생된다. 그러나 위의 두 문제에 대해, 제안된 방법은 문제가 되지 않음이 밝혀졌다. 제안된 HMM 기반의 추출기에서의 독특한 장점으로는 추출된 데이터 레코드가 relational DB에 쉽게 저장될 수 있는 형태로 저장될 수 있다는 것이다. 따라서 이렇게 저장되는 추출 데이터는 필요에 따라 여러 가지 형태의 지식검색에 쉽게 응용될 수 있다는 장점이 있다. 실제 상황의 실험 결과에서 보듯이 제안된 HMM 기반의 정보검출기는 PEWEB에 비해 검출도와 정확도에서 매우 향상된 결과를 보이고 있으며, 특히 정확도에서는 99% 이상의 높은 결과를 보여주고 있다. 한편, 보다 충실한 학습을 위해 학습 데이터의 수를 800개 이상으로 증가시켰을 때 검출도 역시 약 93%의 성능을 보여주었다. HMM 의 구조를 보다 다양한 형태로 변화시켜, 정확한 학습에 의해 검출도를 높일 수 있는 방안에 대한 연구가 계속되어야 할 것이다.

참 고 문 헌

[1] D. Embley, Y. Jiang, Y and Y. Ng, "Record-boundary discovery in Web documents," *Proc. of SIGMOD-99*, 1999.

[2] D. Buttler, S. Liu, and C. Pu, "A Fully Automated Extraction System for the World Wide Web", *Proc. of IEEE ICDCS*, pp. 361-370, 2001.

[3] http://omini.sourceforge.net/

[4] K. Lerman, S. Minton, and C. Knoblock, "Wrapper Maintenanc: A machine learning approach," *J. of Artificial Intelligence Research*, V. 18, pp. 149-181, 2003.

[5] B. Liu, R. Grossman, and Y. Zhai, "Mining Data Records in Web Pages," *IEEE Intelligent Systems*, V. 19, No.6, pp. 49-5, 2004.

[6] C. Chang, C. Hsu, and S. Lui, "Automatic information extraction from semi-structured Web pages by pattern discovery", *Decision Support Systems*, Vol. 35, No.1, pp. 129-147, 2004.

[7] X. H. Phan, S. Horiguchi, and T. Ho, "PEWEB: Product Extraction from the Web Based on Entropy Estimation", *Proc. of the 2004 IEEE/WIC/ACM International Conference on the Web Intelligence*, pp. 590-593, 2004.

[8] http://www.jaist.ac.jp/~hieuxuan/softwares/peweb/

[9] 노수호, 박병준, "Stochastic 프로세스 모델을 이용한 웹 페이지 추천 기법," *전자공학회논문지*, 제42권 CI편 제6호, pp. 37-46, 2005.

[10] D. Gusfield, *Algorithms on strings, tree, and sequence*. 1997.

[11] 석현택, 곽경섭, "인체에 투사된 스트라이프 파형의 HMM을 이용한 인식방안," *전자공학회논문지*, 제42권 CI편 제1호, pp. 51-58, 2005.

[12] 양욱일, 손광훈, "방사 기저 함수 신경망을 이용한 3차원 얼굴인식," *전자공학회논문지*, 제 44권 SP편, 제2호, pp. 82-92, 2007.

[13] 박창현, 송명선, "인지 무선 시스템을 위한 채널 집합 관리기의 개발 및 성능 분석," *전자공학회논문지*, 제45권 CI편 제5호, pp. 8-14, 2008.

[14] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. of IEEE*, Vol.7, No. 2, 57-286, 1989.

[15] D.-C. Park, et al., "Information Extraction System Based on Hidden Markov Model", *Proc. of ISNN 2009*, (accepted for presentation).

저 자 소 개

박 동 철(정회원)

1980년 서강대학교 전자공학과 학사 졸업.

1982년 한국과학기술원 전기 및 전자공학과 석사 졸업.

1990년 Univ. of Washington, Seattle, Dept. of Electrical Eng. 박사 졸업.

2009년 현재 명지대학교 정보공학과 교수.

<주관심분야 : 지능컴퓨팅, 신호처리>