음식메뉴 개체명 인식을 위한 음식메뉴 사전 자동 구축 방법

구영현[0], 유성준[1)]
세종대학교, 컴퓨터공학과
yhgu@sejong.ac.kr, sjyoo@sejong.ac.kr

Automatic Construction of Restaurant Menu Dictionary

Yeong-Hyeon Gu[0], Seong-Joon Yoo
Sejong University, School of Computer Science

요 약

레스토랑 리뷰 분석을 위해서는 음식메뉴 개체명 인식이 매우 중요하다. 그러나 현재의 개체명 사전을 이용하여 리뷰 분석을 할 경우 구체적이고 복잡한 음식메뉴명을 표현하는데 충분하지 않으며 지속적인 업데이트가 힘들어 새로운 트렌드의 음식 메뉴명 등이 반영되지 않는 문제가 있다. 본 논문[2)]에서는 레스토랑 전문 사이트와 레시피 제공 사이트에서 각 레스토랑의 메뉴 정보와 음식명 등을 래퍼기반 웹 크롤러로 수집하였다. 그런 다음 빈도수가 낮은 음식메뉴와 레스토랑 온라인 리뷰에서 쓰이지 않는 음식메뉴를 제거하여 레스토랑 음식 메뉴 사전을 자동으로 구축하였다. 그리고 레스토랑 온라인 리뷰 문서를 이용해 음식메뉴 사전의 엔티티들이 어느 유형의 레스토랑 리뷰에서 발견되는지를 찾아 빈도수를 구하고 분류 정보에 따른 비율을 사전에 추가하였다. 이 정보를 이용해 여러 분류 유형에 해당되는 음식메뉴를 구분할 수 있다. 실험 결과 한국관광공사 외국어 용례사전의 음식 메뉴명은 1,104개의 메뉴가 실제 레스토랑 리뷰에서 쓰인데 비해 본 논문에서 구축한 사전은 1,602개의 메뉴가 실제 레스토랑 리뷰에서 쓰여 498개의 어휘가 더 구성되어 있는 것을 확인 할 수 있었다. 이와 아울러, 자동으로 수집한 메뉴의 정확도와 재현율을 분석한다. 실험 결과 정확률은 96.2였고 재현율은 78.4, F-Score는 86.4였다.

주제어: 음식메뉴 사전, 개체명 인식

1. 서론

레스토랑 음식 메뉴는 레스토랑이라는 의견 대상의 속성들 중 가장 높은 비율을 차지하며 많은 정보를 포함하고 있다. 대부분의 레스토랑의 경우 여러 가지 음식메뉴를 갖고 있는데 레스토랑에 대해서만 의견 분석을 할 경우에는 레스토랑 자체에 대한 의견 분석만 가능하다. 그러나 음식메뉴라는 레스토랑의 속성을 가지고 의견 분석을 하면 음식메뉴들에 대해 개별적인 긍정/부정 정보를 알 수 있다. 예를 들어 음식 메뉴 속성을 이용하여 분석을 하면, 메뉴1은 부정 의견이고 메뉴2는 긍정 의견인 경우에 각각의 평가 정보를 얻을 수 있지만 레스토랑을 대상으로 분석을 하면 어느 메뉴가 더 평가가 좋은지 알 수 없다.

현재 한국어 음식메뉴사전은 전자통신연구원에서 구축한 한국어 개체명 사전 DB[1]와 한국관광공사에서 구축한 관광용어 외국어 용례사전[2]이 있다. 전자통신연구원에서 개발한 사전은 200개의 엔트리를 가지고 있으며 추상물, 구체물, 표상물, 조직 등으로 분류하였다. 그 중 음식과 관련된 엔트리들은 구체물의 음식으로 분류하여 관리한다. 한국관광공사에서 구축한 사전은 외국 관광객들을 위한 표준화된 레스토랑 메뉴판 작성을 목적으로 개발되었으며 음식 분류정보도 함께 제공된다.

전자통신연구원에서 구축한 사전은 일반적인 목적의 개체명 사전이다. 일반적인 목적의 개체명 사전과는 다르게 레스토랑 도메인의 음식메뉴명은 재료와 조리방법에 따라 많은 파생어들이 있기 때문에 구체적이고 복잡한 음식메뉴명을 표현하는데 충분하지 않다. 그리고 한국관광공사에서 구축한 사전은 음식메뉴명과 함께 음식메뉴의 분류를 제공하지만 실제 분류와 다른 경우가 발생한다. 또한 음식메뉴명은 주기적으로 새로운 음식메뉴가 업데이트 된다는 특징이 있는데 이를 반영하기 위해 지속적인 업데이트를 하려면 시간과 비용이 많이 소요된다.

이러한 문제점들을 해결하기 위해 본 논문에서는 레스토랑 전문 사이트와 레시피 제공 사이트를 이용해 자동으로 레스토랑 음식메뉴 사전을 구성하는 방법을 제안한다.

이를 위해 복수의 레스토랑 전문 사이트들과 레시피 제공 사이트에서 래퍼 기반 웹 크롤러를 이용해 메뉴판 정보, 인기메뉴, 추천메뉴 등과 같은 레스토랑 음식메뉴와 관련된 정보를 주기적으로 추출한다. 그런 다음 빈도수를 측정하여 특정 빈도수 이하의 정보를 제거한다. 그리고 추가적으로 온라인 레스토랑 리뷰와 비교하여 실제로 리뷰어들이 쓰는 어휘인지 분석한 다음 사용 빈도수가 낮은 어휘들은 제거하여 음식메뉴 사전을 자동으로 구축한다.

---

1) 교신 저자

2) 본 연구는 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10044454, 기기 정보뿐 아니라 사용자의 환경/감성/인지 정보에 적응적으로 반응하는 정보기기용 원격 UI 기술 개발]

음식메뉴는 한식, 양식, 일식, 주점 등과 같은 분류 유형 정보를 가지고 있다. 자동으로 구축된 음식메뉴 사전의 엔티티들이 어느 유형의 레스토랑 리뷰에서 사용되는지를 빈도수 기반으로 측정하여 음식 유형 분류 정보를 제공한다.

하나의 음식메뉴는 여러 개의 분류 정보를 가질 수 있는데 이러한 문제를 해결하기 위해 유형별로 음식메뉴가 발견된 비율을 같이 표현하여 확률적으로 어느 유형에 해당하는지를 알 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 알아보고 3장에서는 본 논문에서 제안하고 있는 음식메뉴 사전 자동 구축을 설명한다. 4장에서는 실험을 통해 기존 음식메뉴 사전과 비교하고, 마지막으로 5장에서는 결론 및 향후 과제를 제시한다.

2. 관련 연구

일반적인 웹 크롤러는 자동적으로 웹서버를 순회하며 웹페이지의 내용을 분석하고, 그 안에 포함되어 있는 URL들을 추출한다. 그런 다음 그 URL들로 하나씩 이동하면서 웹 문서를 수집한다. 이렇게 수집된 다양하고 많은 양의 웹 문서는 검색 엔진에서 이용된다.

반면에 특정한 주제의 문서만을 수집할 수 있는 형태의 크롤러도 있다. Focused Crawler와 Topical Crawler는 크롤러 자체에 document classifier나 수집해야 할 문서의 규칙들을 탑재하여 수집을 한다.[3, 4] 그러나 분류기의 성능이 좋지 못하고 규칙의 양이 불충분하다면 사용자가 원하는 데이터를 수집할 수 없다.

이러한 이유로 인해 수집하고자 하는 특정 사이트의 구조를 분석하여 직접적으로 데이터를 수집해올 수 있는 wrapper에 대한 연구가 진행되었다. Wrapper를 이용한 크롤러는 필요한 정보를 추출하기 위해 규칙을 만들고 자동으로 사이트의 컨텐츠를 수집한다.[5,6,7,8,9,10]

전자통신연구원에서 구축한 한국어 개체명 사전은 음식메뉴명 정보만 제공할 뿐 음식 분류를 제공하지 않는다. 그리고 한국관광공사에서 구성한 음식메뉴명 사전은 음식 분류를 제공하지만 실제 분류와 다른 경우가 발생한다. 예를 들어 '알밥'은 일식으로 분류되어 있지만 리뷰들을 분석해보면 분식, 한식 등에서 더 많이 발견된다. 뿐만 아니라 온라인 리뷰에는 '갈비살'과 '소갈비살'처럼 같은 음식메뉴인데 다르게 표현되는 경우도 있다.

또한 레스토랑 리뷰 문서의 음식메뉴 어휘들을 살펴보면 기존의 음식메뉴와 관련한 개체명 사전에 없는 경우가 있다. 이러한 이유는 개체명 사전은 일반명사들로 이루어져 있는데 반해 레스토랑에서는 신메뉴가 지속적으로 개발이 된다. 그럼에도 불구하고 개체명 사전을 수작업으로 구축하는 것은 많은 시간과 비용이 발생하기 때문에 업데이트가 지속적으로 이루어지지 않기 때문이다.

3. 레스토랑 메뉴 사전 자동 구축 알고리즘

표 1은 레스토랑 메뉴 사전 구축 알고리즘에 대한 설명이다.

<표 1> 레스토랑 메뉴 사전 구축을 위한 알고리즘

1. 래퍼 기반 크롤러를 이용해 레스토랑 전문 사이트에서 메뉴 정보 추출
2. 래퍼 기반 크롤러를 이용해 레시피 정보 제공 사이트에서 메뉴 정보 추출
3. 추출된 메뉴명의 빈도수 측정
4. Threshold 이상의 빈도수를 갖는 메뉴만 DB에 저장
5. 수집된 레스토랑 리뷰에서 음식메뉴 포함한 리뷰의 빈도수 측정
6. 음식메뉴 분류 정보에 따른 빈도수와 비율 DB에 업데이트
7. 레스토랑 리뷰에서 발견되지 않은 음식메뉴 제거

3.1 래퍼기반 메뉴 정보 추출

기존의 일반적인 웹 크롤러는 모든 문서를 대상으로 하기 때문에 성능이 좋지 못하다. 이러한 문제를 해결하기 위해 수집하고자하는 특정 사이트들의 구조를 분석해 직접적으로 원하는 데이터를 수집하는 래퍼(Wrapper) 기반 웹 크롤러 연구가 있다.

원하는 데이터를 정확하게 수집하는 방법은 해당 사이트의 html code를 이용하여 사이트의 구조를 분석하는 것이다. 그런 다음 원하는 데이터가 있는 부분만 접근할 수 있도록 해야 한다. 사이트의 구조 정보를 토대로 데이터를 수집할 수 있는 프로그램이 wrapper이다. 본 논문에서는 이러한 wrapper 모델을 기반으로 음식점 검색 사이트에서 음식 메뉴를 수집하였다.

레스토랑 전문 사이트는 일반적으로 공통된 구성으로 되어 있다. 레스토랑 전문 사이트는 다수의 레스토랑 s로 이루어져 있다.

$$S = \{s_1, s_2, \dots, s_n\}$$

레스토랑 s는 레스토랑명, 전화번호, 주소, 추천메뉴 등과 같은 속성 정보 a로 구성되어 있다.

$$A_i = \{a_1, a_2, \dots, a_o\}$$

레스토랑 S는 p개의 리뷰 r을 가지고 있다.

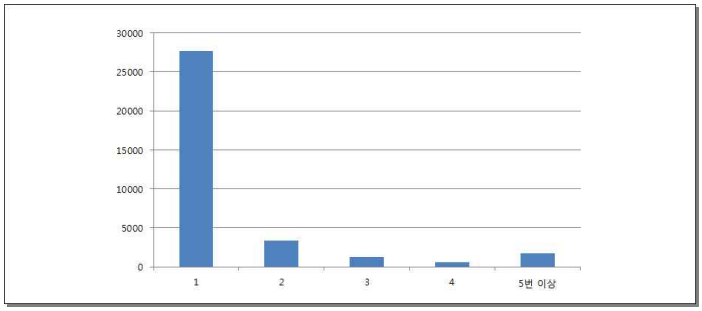
$$R_j = \{r_1, r_2, \dots, r_p\}$$

메뉴는 주로 레스토랑 사이트 내의 추천메뉴, 인기메뉴, 메뉴판 등 정보의 위치와 구조 정보를 분석해보면 수집이 가능하다. 본 논문에서는 국내의 5개의 레스토랑 전문 사이트 중에서 메뉴 정보를 제공하는 2개의 사이트를 대상으로 수집하였다.

3.2 빈도수 기반 메뉴 선정

수집된 음식메뉴는 모두 34,506개였다. 이렇게 수집된

음식메뉴의 빈도수를 측정하고 빈도수 순으로 정렬하였다. 빈도수를 측정한 이유는 여러 레스토랑의 메뉴에서 빈번하게 나타난 음식 메뉴는 좀 더 일반적인 음식 메뉴일 확률이 높기 때문이다. 그림 1은 추출된 메뉴명의 빈도수를 나타내는 그림인데 빈도수가 1인 메뉴들이 전체의 대다수를 차지한다.



[그림 1] 추출된 메뉴명의 빈도수

수집된 음식 메뉴 중에서 빈도수가 3 이상의 음식 메뉴를 선정하였다. 표 2는 빈도수 순으로 정렬된 음식메뉴의 예이다.

<표 2> 빈도수 순으로 정렬한 상위 20개의 음식메뉴

순위	메뉴명	빈도수	순위	메뉴명	빈도수
1	돼지갈비	522	11	우럭	190
2	삼겹살	499	12	항정살	187
3	갈비탕	318	13	김치찌개	173
4	생삼겹살	300	14	차돌박이	167
5	냉면	285	15	생등심	185
6	갈비살	240	16	삼계탕	165
7	광어	230	17	된장찌개	164
8	회덮밥	221	18	등심	152
9	육회	213	19	생갈비	151
10	닭도리탕	198	20	갈매기살	149

이렇게 빈도수가 높은 음식 메뉴만을 선택한 이유는 빈도수가 낮은 음식 메뉴들은 대다수의 음식점에서 사용하는 일반적인 음식 메뉴가 아니었기 때문이다. 예를 들어 '메뉴1'을 A음식점에서 'A 음식점 메뉴1'이라 부른다고 하자. 두 음식 메뉴는 실제로는 같음에도 불구하고 '메뉴1'은 일반적인 음식 메뉴이기 때문에 빈도수가 높고 'A 음식점 메뉴1'은 A 음식점에서만 사용하기 때문에 빈도수가 낮다. 실제로 레스토랑에 따라 같은 메뉴에 레스토랑 상호를 붙이거나 새로운 이름으로 부르기도 한다. 그러나 온라인 리뷰를 분석한 결과 실제로 사용자들은 리뷰를 작성할 때 이러한 특이하거나 독특한 메뉴명은 잘 사용하지 않는다. 그렇기 때문에 빈도수가 낮은 음식메뉴명은 사전 목록에서 제외하도록 하였다.

3.3 음식메뉴의 분류 정보 추출

수집된 음식 메뉴 중에는 음료수와 술과 같이 음식 메뉴로 부적합한 것들이 있다. 이렇게 부적합한 음식 메뉴들을 제거하고 나머지를 음식 메뉴 사전에 등록한다. 이렇게 구축된 음식 메뉴 사전에도 확인결과 빠진 메뉴들이 존재했다. 이러한 문제를 해결하기 위해 레시피 제공 사이트를 크롤링해서 메뉴를 추가하였다.

음식메뉴는 한식, 양식, 중식과 같은 고유의 분류 정보를 가지고 있다. 그러나 일반적인 분류 체계와 실제 메뉴의 분류 정보가 다른 경우가 있다. 이러한 문제를 해결하기 위해 음식메뉴 사전의 엔티티들이 어느 유형의 레스토랑 리뷰에서 발견되는지를 찾아 빈도수를 구하고 분류 정보에 따른 비율을 사전에 추가하였다.

음식메뉴 M은 음식메뉴명, 기타/양식/일식/주점/중식/카페/한식 리뷰에서의 음식메뉴 빈도수, 전체 리뷰에서의 음식메뉴 빈도수, 기타/양식/일식/주점/중식/카페/한식 리뷰에서의 음식메뉴 비율로 이루어진다.

$$M = \{ food_name, freq_{etc}, freq_{wst}, freq_{jpn}, freq_{bar}, freq_{chi}, freq_{caf}, freq_{kor}, freq_{tot}, rate_{etc}, rate_{wst}, rate_{jpn}, rate_{bar}, rate_{chi}, rate_{caf}, rate_{kor}, rate_{tot}, \}$$

아래의 표 4는 분류 정보 분석 결과를 포함한 음식 메뉴 사전 중 리뷰에서의 분류 정보에 따른 비율의 예를 보여주고 있다. 이 분류 정보를 이용하면 정보 추출을 통해 음식 메뉴를 추출할 경우 부가적으로 분류 정보도 얻을 수 있다. 표 4의 '알밥'의 경우 일식과 한식에서 비슷하게 분포되어 있는 것을 볼 수 있는데 이렇게 여러 분류에 해당되는 음식메뉴도 이 방법을 이용해 구분할 수 있다.

<표 3> 카테고리 분석 정보가 포함된 음식메뉴 사전의 예

구분	기타 리뷰 비율	양식 리뷰 비율	일식 리뷰 비율	주점 리뷰 비율	중식 리뷰 비율	카페 리뷰 비율	한식 리뷰 비율
갈비탕	0.0	0.7	0.7	0.7	0.3	0.7	97.0
미소라멘	0.0	1.1	85.1	11.7	0.0	0.0	2.1
모듬회	0.0	0.0	20.0	28.0	0.0	0.0	52.0
안창살	0.0	2.1	0.0	2.1	0.0	0.0	95.7
매운탕	0.0	0.0	4.6	37.1	1.7	2.2	54.5
낙지볶음	0.0	2.4	0.8	1.6	0.0	3.3	91.9
해물탕	1.3	8.5	1.3	19.6	19.0	2.6	47.7
알밥	1.0	0.5	43.4	4.0	0.0	2.0	49.0
감자탕	0.3	0.9	1.5	0.3	0.6	0.3	96.0

4. 실험

실험을 위해 레스토랑 온라인 리뷰를 래퍼 기반 웹 크롤러를 이용해 수집했다. 크롤링 된 리뷰 중 1,000개의 리뷰를 뽑아 다시 문장 단위로 나누었다. 수작업으로 분류를 한 결과 음식 메뉴를 포함한 리뷰 문장이 1,129개였고 음식 메뉴를 포함하지 않은 리뷰 문장은 2,393개였다.

음식 메뉴는 품사 중에서 명사에 해당되기 때문에 분

류한 리뷰 문장을 형태소 분석한다. 그런 다음 명사만 추출하고 음식 메뉴 사전과 비교하여 일치하는 음식메뉴를 추출하여 리뷰 문장의 음식메뉴를 얼마나 정확히 추출하는지를 실험하였다. 그 결과 True Positive는 778, False Positive는 0, False Negative는 351, True Negative는 2,393이었다. <표 5>는 리뷰를 형태소 분석한 다음 명사를 음식메뉴 사전과 비교해 성능을 평가한 결과이다.

<표 4> 온라인 레스토랑 리뷰의 형태소 분석결과를 이용한 사전 성능 평가

추출 결과	실제 Class		Total
	Positive	Negative	
Positive	778	0	778
Negative	351	2,393	2,744
Total	1,129	2,393	3,522

온라인 리뷰의 특성상 띄어쓰기가 제대로 되어 있지 않은 문장이 많기 때문에 추출이 되지 않는 문장들이 있었다. 이러한 문제점을 해결하기 위해 형태소 분석을 하지 않고 문장의 공백을 제거한 다음 Java의 indexOf() 메소드를 이용해 문자열 비교를 해 음식메뉴 사전에 있는 엔티티를 포함하는 문자열을 뽑도록 하였다. 이 방법은 포함 관계인 음식 메뉴가 모두 추출되는 문제가 있다. 예를 들어 '바지락칼국수'는 '바지락칼국수', '칼국수', '국수'가 모두 추출된다. 이러한 문제점을 막기 위해 여러 개의 음식 메뉴가 추출되는 경우 가장 문자열이 긴 음식 메뉴를 추출하도록 한다.

이 방법으로 온라인 레스토랑 리뷰 문장의 음식메뉴를 얼마나 정확히 추출하는지를 실험하였다. 그 결과 True Positive는 884, False Positive는 35, False Negative는 245, True Negative는 2,358이었다. <표 6>은 리뷰를 음식메뉴 사전과 문자열 비교를 하여 성능을 평가한 결과이다.

<표 5> 온라인 레스토랑 리뷰의 문자열 비교를 통한 사전 성능 평가

추출 결과	실제 Class		Total
	Positive	Negative	
Positive	884	35	919
Negative	245	2,358	2,603
Total	1,129	2,393	3,522

좀 더 쉽게 두 가지 방법을 비교하기 위해 F-Score를 구하였는데 형태소 분석기를 이용한 방법은 81.6이었고 문자열 비교를 한 방법은 86.4로 상대적으로 나은 성능을 보였다.

그러나 문자열 비교를 이용한 음식 메뉴 추출 방법은 '죽', '난', '전', 떡'등과 같이 음식 메뉴명의 문자열 길이가 짧은 경우에는 잘못된 음식메뉴명을 추출하기도 한다. 또한 '라면', '파이','스프' 등도 잘못된 음식 메뉴명을 추출하는 주원인이었다. <표 7>은 문자열 비교를 통해 음식 메뉴명을 잘못 추출하는 경우의 예이다.

<표 6> 잘못 추출된 음식 메뉴명의 예

'고향의 맛이라 생각하고 먹을수도 있겠지만 그게아니<u>라면</u> 먹기가 좀 부담스럽습니다.'
'와이<u>파이</u>가 잘 터지네요.'
'이 집 에<u>스프</u>레소가 맛있네요.'

그 밖에도 외국 음식명은 외래어 표기 방법에 따라 다양하게 표현되어 추출의 성능을 나쁘게 하였다. 예를 들어 'cake'는 '케잌', '케이크', '케익' 등 레스토랑이나 리뷰어에 따라 다르게 표현된다.

5. 결론

본 논문에서는 레스토랑 전문 사이트와 레시피 제공 사이트에서 각 레스토랑의 메뉴 정보와 음식명 등을 웹 크롤러로 수집하였다. 그런 다음 빈도수가 낮은 음식메뉴와 레스토랑 온라인 리뷰에서 쓰이지 않는 음식메뉴를 제거하여 레스토랑 음식 메뉴 사전을 자동으로 구축하였다. 그리고 레스토랑 온라인 리뷰 문서를 이용해 음식메뉴 사전의 엔티티들이 어느 유형의 레스토랑 리뷰에서 발견되는지를 찾아 빈도수를 구하고 분류 정보에 따른 비율을 사전에 추가하였다. 이 정보를 이용해 여러 분류 유형에 해당되는 음식메뉴를 구분할 수 있다.

성능 측정을 위해 구축된 음식 메뉴 사전을 이용해 레스토랑 리뷰로부터 음식 메뉴를 추출하였다. 추출 방법은 형태소 분석 방법과 문자열 비교 방법 두 가지를 비교하였는데 실험 결과 형태소 분석 방법의 정확률은 100, 재현율은 68.9, F-Score는 81.6이었고 문자열 비교 방법의 정확률은 96.2, 재현율은 78.4, F-Score는 86.4로 상대적으로 문자열 비교 방법의 성능이 더 좋았다.

이렇게 구축된 음식 메뉴 사전으로 레스토랑 리뷰 분석을 통한 인기메뉴 추출이나 의견대상에 대한 의견 검출 등을 할 수 있다.

참고문헌

[1] 전자통신연구원 한국어 어휘 사전 DB, http://www.itec.re.kr/itec/sub02/sub02_01_1.do?t_id=5674#

[2] 한국관광공사 관광용어 외국어 용례사전, http://kto.visitkorea.or.kr/kor/translation/translation/main.kto

[3] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused Crawling : A New Approach to Topic-Specific Web Resource Discovery", Computer Networks, Vol.31, No. 11-16, pp.1623-1640, 1999

[4] Ziyu Guan, Can Wang, Chun Chen, Jiajun Bu, Junfeng Wang, "Guide Focused Crawler Efficiently and Effectively Using On-line

Topical Importance Estimation", In Proc. of ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 757-758, 2008
[5] S. Chakrabarti, Mining the Web. Discovering Knowledge from Hypertext Data, Morgan Kaufmamm, pp. 257-287, 2003
[6] J. Cho, H. Garcia-Molina, and L. Page, “Efficient Crawling through URL Ordering”, Computer Networks, Vol.30, No.1-7, pp. 161-172, 1998
[7] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled Shaalan, “A Survey of Web Information Extraction Systems”, IEEE Transaction on Knowledge and Data Engineering, Vol.18, No. 10, pp.1411-1428, 2006
[8] Jaeyoung Yang, Tae-Hyung Kim, and Joongmin Choi, “An Interface Agent for Wrapper-Based Information Extraction”, In Proc. of the International Confenrence on Principles of Practice in Multi-Agent Systems(PRIMA'04), pp.291-302, 2004
[9] Claudio Bertoli, Valter Crescenzi, and Paolo Merialdo, “Crawling programs for wrapper-based applications”, In Proc. of IEEE International Conference on Information Reuse and Integration(IRI'08), pp.160-165, 2008
[10] Stephen Soderland, Claire Cardie, and Raymond Mooney, "Learning information extraction rules for semi-structured and Free Text", Machine Learning, Vol. 34, No.1-3, pp.233-272, 1999