특정 도메인 문서 내 관계 트리플 추출

이효갑*, 김판구**
*조선대학교 대학원 컴퓨터공학과
**조선대학교 컴퓨터공학부
e-mail : leoscientist@gmail.com

The Triple Relationship Extraction from Domain Article

Hyokab Lee*, Pan-Koo Kim**
*Graduate School of Computer Engineering, Chosun University
**Division of Computer Science and Engineering, Chosun University

요 약

최근 정보의 의미적 검색을 위해 국내외 포털을 중심으로 시멘틱 웹 검색과 관련된 연구가 진행되고 일정부분 사용되고 있다. 이러한 시멘틱 웹 검색은 기존 작성된 정보를 인간이 가장 쉽게 이해할 수 있는 subject, predicate, object로 구성된 RDF Triple형태로 바꾸어 이를 쉽게 접근하고자 하는 연구가 필요하다. 본 논문에서는 정보의 재가공 후 문서 내 트리플 관계를 추출하는 과정을 실험을 통해 제시한다. 이를 통해 웹 도메인과 컨텐츠 정확한 검색을 가능하게 하고 검색 시간을 단축시켜 효율을 높여주는 계기가 될 것이다.

1. 서론

최근 정보의 의미적 검색을 위해 국내외 포털을 중심으로 시멘틱 웹 검색과 관련된 연구가 진행되고 일정부분 사용되고 있다. 이러한 시멘틱 웹 검색은 기존 작성된 정보를 인간이 가장 쉽게 이해할 수 있는 subject, predicate, object로 구성된 RDF[1] Triple형태로 바꾸어 이를 쉽게 접근하고자 하는 연구가 필요하다. subject는 표현하고자 하는 대상이나 객체를 나타내며predicate는 subject와 object의 관계를 기술한다. object는 predicate에 의해 기술되는 내용이나 값, 혹은 predicate에 의해 관계를 가지게 되는 대상을 말한다.

본 논문에서는 HTML로 구성된 정보를 분석하여 이를 RDF Triple 형태로 생성 할 수 있는 기본 관계 구조를 추출하는 것이 목적이다. 이를 위해 도메인 문서의 핵심용어의 빈도수 분석과 각 어휘의 품사 분류 그리고 문서의 핵심어 선정 및 추출과정이 필요하다. 또한 문장의 토큰화와 태깅을 통해 중요문장 추출과 분류가 필수적이라 할 수 있겠다.

2. 관련 연구

2.1 Wikipedia 문서를 이용한 온톨로지 자동 구축

위키피디아는 웹기반의 다국적 언어의 자유로운 콘텐츠를 지향하는 백과사전 프로젝트로부터 시작되어 오늘날 가장 많이 접속하는 웹사이트중의 하나이다. 또한 전 세계 모든 사용자들이 정보의 생산자 혹은 가공자로 참여하여 웹 2.0을 대표하는 대중의 지혜(The wisdom of crowds), 혹은 집단지성(The collective intelligence)이 가장 잘 반영된 곳으로, 특정 도메인의 개념과 관계를 추출하는데 있어서 훌륭한 대상이다.

이러한 장점으로 인해 이를 이용한 많은 연구들이 수행되었으며 대표적으로 Harvesting Wiki Consensus[7]는 위키피디아 자체가 온톨로지 체계를 구성한다는 점에 착안하여 포함된 개체를 재사용하여 의미명확화에 관해 연구하였다. Robust Minimal Recursion Semantics(RMRS) [2] 시스템은 위키피디아의 biological내용으로부터 12,000개의 동물과 관련된 위키피디아 페이지로부터 개념들의 관계를 추출하고 이를 통해 온톨로지를 구축 하였으며. F Wu[3]의 연구에서는 위키피디아의 infobox의 class들을 SVM과 HMM등을 통해 IS-A관계 등으로 추출하고 wordnet[4]과의 매핑을 통하여 온톨로지를 구축하였다.

3. Wikipedia 문서에서 관계 Triple 추출

3.1 Wikipedia 문서 전처리

위키피디아 문서의 구성은 주요 기사의 제목과 이를 설명하는 Text body와 그림, 표, 목차, reference, category 들로 이루어져 있으며 중요한 특징으로 기사내용에 대한 분류항목을 지니고 있다. 목차에서는 기사내용을 순차적으로 기술하고 있고 문서 내부에 분류항목으로 infobox나 navbox와 같은 box형식의 표, 그림을 통해 세부 분류와

특징을 표현하고 있다. infobox와 navbox는 특정양식에 따라 사용자들에 의해 작성되는 것으로 전문적인 사전이나 서적, 연구문헌, 뉴스 등을 참고 하여 기재된다. (그림 1)은 위키피디아에서 "England"에 관한 정보로써 텍스트, 이미지, 표를 비롯한 각종 정보로 구성되어 있다.

(그림 1) 위키피디아에서 'England'에 관한 정보

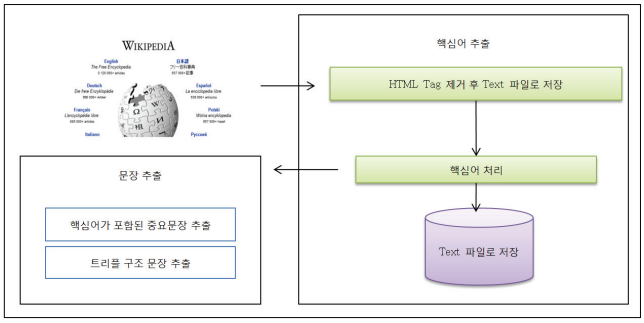
HTML 문서의 구조는 Tag 정보가 포함되어 있어 이를 제거하는 과정이 필요하다. 이를 위해 <표 1>와 같은 과정으로 통해 Tag를 제거할 수 있으며 이를 통하여 특정 도메인에 대한 순수 Text 정보를 추출 할 수 있다.

<표 1> 웹 문서 전처리 과정

```
out_file = open("d:/docu/hyogap/Engand_ori.txt","w")
  noo=0
  for aj;in;kk
     if (aj=="")
            continue
     else:
            Tsen = re.sub("(\[+\d+\])","",aj)
            Tsen = re.sub("[^a-z',.\"\- A-Z',.W-0-9]+", " ",Tsen)
            Tsen = re.sub("[\n]+" , "", Tsen)
            noo=noo+1
            dict[noo]=Tsen
            #print 'Sentence %d: %r' %(noo, Tsen)
            out_file.write(Tsen+"\n")
     out_file.close()
```

3.2 중요문장 추출 및 Triple 추출

중요문장 추출을 위해 다음 (그림 2)와 같은 과정을 통해 처리된다. 이는 기존 HTML 문서의 전처리 구정을 거쳐 중요문장 및 핵심어를 추출하는 과정으로 진행된다.



(그림 2) 핵심어 추출 과정

중요문장 추출은 추출된 문장 단위의 구조에서 품사태거로 'penn treebank project'에서 품사 태거를 수정하여 적용하였다. 복합명사를 추출하기위해서 태그기호 JJ(형용사)와 NN(명사)을 이용하였다.

<표 2> 전문용어 추출 알고리즘

```
while taggedTerms:
    term, tag, norm = taggedTerms.pop(0)
    if state == SEARCH and tag.startswith('N'):
    // 검색 중이고, 명사로 시작되는 경우
        state = NOUN //  복합 어절 탐색 상태
        _add(term, norm, multiterm, terms)
    elif state == SEARCH and tag == 'JJ' and term[0].isupper():
    // 검색 중이고, 형용사로 시작되는 경우
        state = NOUN
        _add(term, norm, multiterm, terms)
    elif state == NOUN and tag.startswith('N'):
    // 복합 어절 탐색 중이고, 명사로 시작되는 경우
        _add(term, norm, multiterm, terms)
        // 검색된 용어와 복합 어절 용어를 각각 추가
    elif state == NOUN and not tag.startswith('N'):
    // 복합 어절이 아닌 경우
        state = SEARCH
        if len(multiterm) > 1:
            word = ' '.join([word for word, norm in multiterm])
            terms.setdefault(word, 0)
```

<표 3> Penn treebank project에서 품사 tag 정의

태그	품사	태그	품사
DT	Detrminer	NN	Sigular Noun
IN	Preposition	NNP	Sigular Proper Noun
JJ	Adjective	NNSS	Plural Proper Noun
JJR	Comparitive Adjective	NNS	Plural Noun
JJS	Superlative Adjective	POS	Possessive Ending
MD	modal verb	RB	Adverb

중요 문장은 주로 위의 방법을 사용하여 태깅된 어휘를 사용하였고 반복되는 핵심어를 출현빈도를 이용하여 이를 선정하였다. 이래 <표 4>과 같이 출현빈도(freq)와 구성단어 수(Multi number)를 분석하고 이를 활용하여 중요문장을 추출하였다.

<표 4> 어휘 발생빈도 측정

Number	Term	Frequency	Mumti number
101	world Wide Web	1	3
102	Retrieved	180	1
103	precision engineering	1	2
104	English World Heritage Sites	1	4
105	royal house	1	2
106	prison	3	1
107	east	3	1
108	New Zealand	1	2
109	people	27	1
110	Parliament George Frideric Handel	1	4
111	Bahamas Barbados Belize Bermuda	1	4
112	manufacturing	3	1
113	expensive stadium	1	2

<표 5> “England"의 Triple 추출 예

English | established | by de facto usage.: 1
Population Estimates. | Assigned | on a UK basis, : :3
England l nd help info | is | a country : 4
a country | is | part of the United Kingdom. : 4
the United Kingdom. | shares | land borders with Scotland : 5
The area | called | England : 7
England | settled | by people of various cultures : 7
the Germanic | settled | during the 5th : 7
England | became | a : 8
a | unified | state in AD : 8
Discovery, | began | during the 15th century, : 8
the 15th century, | had | a significant cultural : 8
the world | developed | in England, : 9
government | adopted | by other nations. : 9
the High Middle Ages | originating | from Brythonic traditions:383
Brythonic traditions | entered | English folklore the Arthurian myth. : 383
These | derived | from Anglo-Norman, French:384
Welsh sources, | featuring | King Arthur, Camelot, Excalibur, Merlin : 384
British tradition, King Cole, | based | on a real figure from Sub-Roman Britain.:386
Britain, a collection | shared | British folklore. : 387
Morris dance, an English folk dance Some folk figures | based | on semi:388

이러한 과정을 통해 HTML 문서를 통한 핵심어 및 핵심 중요문장을 추출하였고, 이를 활용하여 RDF Triple로 문장을 재구성하여 특정 도메인에 대한 온톨로지 구축에 활용할 수 있는 정보를 제공할 수 있다.

중요문장을 통해 다음과 같이 Triple을 추출하였는데, <표 5>의 “England” Triple 추출 예로써 subject, predicate, object와 그 문서에서 추출한 결과를 보이고 있다.

4. 실험 및 평가

실험을 위해 웹 도메인 문서 중 wikipedia의 국가별 정보를 사용하영 실험하였다. 실험 결과 국가당 평균 2,640여개의 명사(복합명사)가 추출되었으며, 추출된 어휘를 바탕으로 어휘의 출현 빈도수를 조사한 결과 주요명사가 포함된 주요 문장은 1,550여개의 문장 추출 되었고 Triple 구조의 문장 400여개를 수동으로 검사하였다.

위와 같이 추출된 Triple은 어느 정도 만족할 만한 성능을 보였으나 품사 tag에 정의되지 않는 문장의 경우 추출할 수 없는 문제를 보였다. 또한 가장 큰 오차 발생은 대명사와 접속사들로 인한 주어, 목적어 파악 문제와 긴 문장으로 인한 잘못된 Link설정이었다. 주어 인식은 It, That과 같은 대명사를 주어로 인식하여 의미파악이 불분명한 경우가 약 50여개 Triple에서 관측되었다. 긴 문장으로 인한 오류는 다양한 Triple을 추출하기 위해 설정한 패턴이 잘못된 Triple을 생성하는 경우에 나타났다.

5. 결론 및 향후 과제

본 논문에서는 HTML문서 중 대표적인 위키피디아를 이용하여 HTML Tag를 제거하고 중요문장과 핵심어를 추출하여 Triple를 생성하였다. 이는 현재 웹이 향후 시멘틱 웹으로 나가는데 있어서 기존 웹 페이지를 이용한 의미기반 검색이 가능하게 하는 기초가 되는 연구로 시멘틱 웹의 지향하는 기존 문서들의 공유와 재사용이라는 목적에도 부합하는 연구라고 판단된다.

향후 실험 및 평가에서 나온 문제점을 해결하기 위해 품사 tag의 종류를 다양화 할 필요성이 있으며 대상사와 접속사들로 인한 주어와 목적어 판단에 관한 패턴설정이 필요하다.

Acknowledgements

“이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (No.2009-0064749).”

참고문헌

[01] http://www.w3.org./RDF/
[02] Robust Minimal Recursion Semantics WORKING PAPER Ann Copestake January 2004
[03] F Wu, D Weld. "Automatically refining the Wikipedia infobox Ontology". portal.acm.org2008.
[04] http://meta.wikimedia.org/wiki/List_of_Wikipedias
[05] http://www.cis.upenn.edu/~treebank/
[06] 정보통신정책동향 정보통신정책 제19권 13호통권 420호 위키피디아 활용 현황 및 활성화 요인
[7] M Hepp, K Siorpaes, D Bachlechner. "Harvesting WikiConsensus:Using Wikipedia Entries as Vocabulary