구 기반 색인 시스템의 구현

이충희∪ 김현진 장명길
한국전자통신연구원 언어공학연구부
{forever, jini, mgjang}@etri.re.kr

Implementation of Phrase-based Indexing

Chung-Hee Lee∪ Hyun-Jin Kim Myung-Gil Jang
Linguistic Engineering Department, ETRI

요 약

정보 검색 결과의 정확성을 높이기 위해서는 상위수준의 색인 정보를 이용한 검색 기법이 요구된다. 상위수준의 색인을 하기 위해서는 구문 분석을 이용할 필요가 있지만 웹 페이지를 이용하는 웹 검색에서는 웹 페이지 자체의 오류 때문에 구문 분석을 할 때 실패할 확률이 높으므로 견고한 구문 분석이 요구된다. 본 논문은 구, 문장에 기반한 색인 기법 및 기존 색인 방법을 병행해서 사용하는 시스템에 대하여 소개한다. 본 논문에서 소개하는 시스템은 5가지 방법의 색인 기법을 사용한다. 각 색인 기법은 적용될 분야 또는 범위에 따라 선택적으로 사용될 수 있다. 색인 기법은 1)명사 색인 2)명사+용언 색인 3)명사+용언+문장정보 색인 4)명사구 색인 5)중심어-종속어(Head-Modifier) 색인으로 나누어진다. 색인 기법 중 4와 5의 경우, 구문 분석된 결과를 사용하여 특정 명사구 및 중심어-종속어 관계를 고려함으로써 문서의 특성을 잘 나타내는 색인어를 추출할 수 있고 그러므로 정보검색의 성능을 향상시키는 기반 기술로 사용될 수 있다.

1. 서론

인터넷을 사용하는 인구가 급증하면서, 사용자가 원하는 정보를 빠르고 정확하게 찾아주는 검색엔진은 없어서는 안될 존재가 되었다. 초기 검색엔진은 원하는 정보가 포함되기만 하면 되도록 가능한 많은 웹 페이지를 찾아주면 되었고, 인터넷에 등록된 정보의 양이 그렇게 많지 않았기 때문에 검색 결과의 양도 많지가 않았다. 하지만 전세계적으로 인터넷의 이용이 폭발적으로 증가하면서 인터넷으로부터 얻을 수 있는 정보의 양 또한 무한히 증가하게 되었다. 그러므로 검색결과에 대한 사용자의 요구사항이 정확한 몇 개의 웹 페이지만을 검색하는 방향으로 바뀌고 있다. 이러한 사용자 요구사항을 만족시키는 검색엔진을 만들기 위해서는 상위수준의 색인 정보를 사용한 검색 기법이 필요하다.

정보검색 시스템에서 색인어를 선정하는 방법은 검색 성능과 직접적인 관련이 있다. 지금까지 상용화된 시스템들은 대부분 색인어를 추출한 후 문헌 내의 색인어 출현 빈도에 따른 가중치를 주어 색인 시스템을 구현하였다. 이러한 시스템들에서 사용된 언어학적 방법은 단순히 형태소 분석을 통해 명사들을 색인하는 정도이다. 문장의 주요 의미를 명사들이 나타내기 때문에 명사만으로 색인을 하더라도 검색에서 일정 수준까지의 성능은 기대할 수 있다. 하지만 명사만으로 색인을 해서는 검색 성능에 한계가 있으므로 명사 이외의 정보를 색인할 필요가 있다. 이러한 요구 사항을 만족시켜 줄 정보로는 구문 정보가 훌륭한 대안이 될 것이다.

본 논문에서 소개하는 시스템에서는 기존의 검색엔진에서 사용하던 색인 방법에 구문 분석을 이용한 색인 방법을 추가한 시스템을 구현하여 검색 성능을 향상시키고자 한다.

2. 구기반 색인 방법

본 시스템에서 사용하는 색인 단위[1]는 자연어처리 수준에 따라 5가지이며 각각에 대해 알아보면 다음과 같다.

2.1 명사 색인(SIU 1)

형태소 분석 결과 중 명사만을 추출하며 수사는 제외한다. 복합명사는 한 어절에 2개 이상의 명사들이 붙어있는 경우만을 고려하며, 형태소 분석 및 태깅에서 분리시키더라도 결합하여 하나의 색인어로 추출한다. 미등록어는 모두 명사로 추정하여 추출한다. 하다_명사의 경우 문장에서 용언으로 쓰였더라도 명사로 추정하여 추출한다.

2.2 명사+용언 색인(SIU 2)

명사와 용언을 색인어로 추출한다. "~을 읽어 준다"에서 '주다'와 같은 보조용언은 추출하지 않는다.

2.3 명사+용언+문장정보 색인(SIU 3)

명사와 용언을 추출하며 색인을 문서 단위가 아닌 문장 단위로 한다. 즉, 명사 또는 용언 외에 문장 번호를 같이 색인하여 같은 명사 또는 용언이라도 다른 문장에서 발생하는 경우 다른 색인어로 추출된다. 이러한 문장 정보는 색인어간의 거리 또는 관련도 정보로 이용될 수 있다. 예를 들어, 자연어 질의를 처리할 경우, 질의어에 같이 들어있는 명사나 용언들이 특정 문서의 한 문장에 동시에 나타날 경우 해당 색인어들에게 더욱 높은 가중치를 줄 수 있다. 즉, 질의어가 "최근에 출시된 워크맨은?"이고 특정 문서의 같은 문장에서 '최근', '출시', '워크맨'들이 색인된다면 결국 그 문서에 더욱 높은 가중치가 주어지게 된다.

2.4 명사구 색인(SIU 4)

2.4.1 명사구 정의

명사구를 인식하기 위해서는 구문 분석을 수행하거나 별도의 구 묶음(chunking) 모델을 만들어야 한다. 구문 분석을 사용할 경우에도 명사구만을 추출할 때는 전체 문장을 완전 분석할 필요 없이 부분 분석만으로 명사구 인식이 가능하다.

별도로 명사구만을 인식하는 방법으로는 최대 엔트로피 모델을 이용한 방법[1], 규칙과 어휘 정보를 이용한 방법[2] 등이 있다.

본 시스템에서는 부분 구문 분석된 결과로부터 명사구를 추출하여, 명사구를 구성하는 명사들로 색인어를 생성한다.

본 시스템에서 정의되어 사용된 명사구는 다음과 같다.

1) 명사 수식어(-의, -인)를 포함한 구
2) 접미사 적/화/상 포함 어구
3) 명사 병렬 어구
4) 명사(복합명사 포함)

2.4.2 서브 트리 타입

구문 분석 결과는 이진 트리 구조를 가지는데 각 트리의 표층 구조로 각 노드의 서브 트리 타입을 결정하여 명사구 또는 중심어-종속어 추출 시에 이용한다. 서브 트리 타입은 다음과 같이 5가지를 정의하여 사용한다.

```
STT_CNN: 복합명사(종이 호랑이)
STT_DET: 관형사(새 책)
STT_PAR: 명사병렬어구(철수, 영희, 민수와 만수)
STT_NDN: 속격 어구(철수의 옷)
STT_DCL: 관형 어구(아름다운 영희)
```

위에서 정의된 타입이 아닌 경우에는 서브 트리 타입을 할당하지 않는다.

2.4.3 명사구 색인어 추출

구문 분석 결과인 구문 트리를 탐색하며 각 트리 노드의 구문 태그 및 서브 트리 타입 정보를 이용하여 추출하고자 하는 명사구를 결정한다. 사용된 구문 태그는 다음과 같다.

```
NO: 명사구
CO: 격구조(명사에 격조사가 붙은 구)
VO: 동사구
DO: 관형사구
AO: 부사구
SO: 문장
끝부분이 '-'인 경우: 형식형태소 (예: N-, C-..)
```

본 시스템에서는 현재노드의 구문 태그가 'NO'이고,

1) SIU(Semantic Indexing Unit): 이후부터는 SIU로 명기한다.

서브 트리 타입이 STT_DET를 제외한 4가지 중 하나인 경우에 하위 노드들로부터 명사들을 추출하여 색인어로 생성한다.

2.4.1절에서 정의된 명사구 정의 중 1), 2)번의 명사들은 순서에 의미가 있는 것으로 색인어 생성 시 각 명사는 '+'로 연결되고 3)번과 같은 명사들은 순서에 의미가 없는 것으로 '_'로 연결된다.

3개 이상의 연속된 복합명사의 경우, 전체를 하나로 추출한 뒤 인접한 순서대로 2개씩 색인어로 추출한다. 예들 들어 "정보검색시스템"은 '정보','검색','시스템'의 세 가지 단일 명사로 이루어진 복합명사로서 "정보검색시스템"이 하나의 색인어를 형성하고 "정보검색", "검색시스템" 또한 색인어로 추출된다. 명사구 색인어 외에 각각의 명사들도 색인어로 추출한다.

2.5 중심어-종속어 색인(SIU 5)

2.5.1 중심어와 종속어 관계

SIU5는 문장으로부터 중심어와 종속어를 추출한다. 추출되는 중심어-종속어 관계는 다음과 같은 경우로 제한한다.

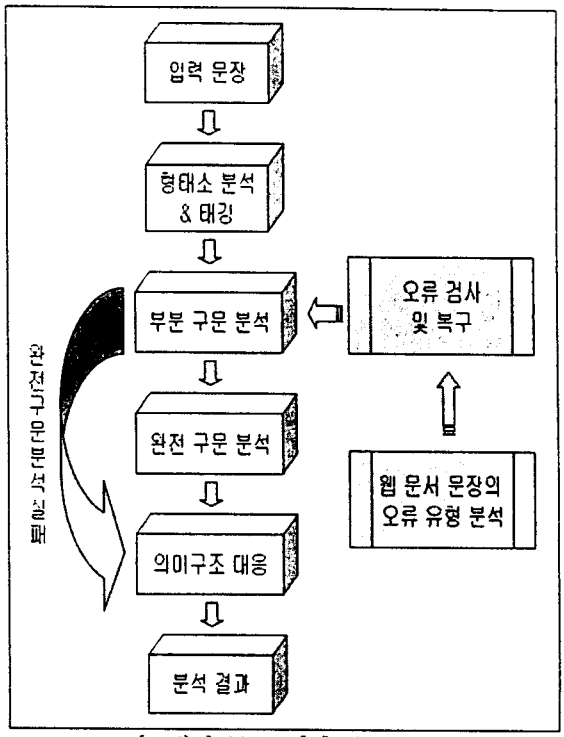
1) 서술어-주어
2) 서술어-목적어
3) 서술어-보어
4) 피수식어-수식어

추출되는 색인어의 품사는 명사, 동사, 형용사로 한정하며 SIU 1과 SIU 2에서 이용한 규칙을 그대로 적용한다.

2.5.2 완전 구문 분석 이용

SIU 5를 추출하기 위해서는 완전구문분석을 할 필요가 있다. 완전구문분석을 이용하는 방법 중에는 구문분석결과의 중의성을 해소하는 경우와 해소하지 않는 경우로 나눌 수 있다. 중의성 해소를 하지 않는 경우에는 불필요한 색인어들이 너무 많이 발생하여 검색 성능을 저하시키므로 중의성 해소를 통하여 정확률을 높일 수 있다. 하지만 구문분석기의 중의성해소 정확률이 너무 낮을 경우에는 도리어 검색 성능을 저하시킨다. 그러므로 이용할 분야에 따라 구문 분석기의 종류 및 중의성 해소를 할 것인지를 선택하여야 한다.

본 시스템에서는 구문 분석 결과의 중의성을 해소한 후 중심어와 종속어를 추출한다.



[그림1] SIU 5 추출 흐름도

2.5.3 정규화 과정

색인어 추출 시에 구문 분석 결과로부터 일정한 정규화 과정을 거칠 수 있다.

정규화 과정으로는 서술어 정규화와 관형절 정규화 과정이 있다.

서술어는 대부분 중심어의 역할을 하는데 한국어 서술어 표현은 매우 복잡하므로 서술어 정규화 과정에서 보조용언 및 기능어를 처리하여 단순화시킨다. 예를 들어, "반발을 하고 나서다"는 "반발하다"로, "이삿짐을 싸 주다"는 "이삿짐을 싸다"로 정규화를 시켜서 서술어를 단순화시킬 수 있다.

관형절 정규화의 경우를 보면, "예쁜 영희"의 중심어는 '영희'이고 종속어는 '예쁘다'이다. "영희가 예쁘다"는 반대로 중심어가 '예쁘다'이고 '영희'가 종속어이다. 하지만 의미상 두 문장은 같은 중심어-종속어 구조를 가진다고 할 수 있다. 그러므로 관형절을 구성하는 술어가, 수식하는 단어의 서술어로 판단될 수 있을 경우, 관형절의 서술어를 중심어로 선택한다. 즉, "예쁜 영희"와 "영희가 예쁘다"는 모두 '예쁘다'를 중심어로 가진다.

2.5.4 중심어-종속어 색인어 추출

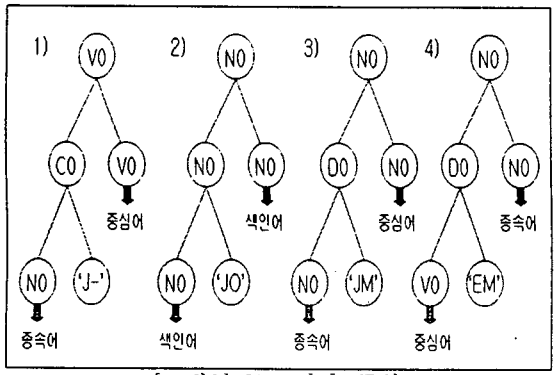
SIU 5는 그림1과 같은 과정을 통해 추출된다. 이 때 전처리 과정을 거쳐서 웹 문서의 오류를 줄이고, 중의성 해소를 통해 구문 분석 결과의 정확률을 높인다.

구문 분석된 결과에 2.4.2절의 서브 트리 타입을 적용한 후 현재노드 및 하위노드의 서브 트리 타입 및 구문 태그를 고려하여 색인어를 추출한다. 이때, 정규화 과정은 관형절 정규화, 서술어 정규화 등이 적용된다.

SIU5는 다음과 같은 네가지 경우에 추출된다.

1) 술어+명사(복합명사 포함)(V-N)
 - 예: 사람이 과자를 먹었다.
  → 먹다-사람, 먹다-과자
2) 복합명사 또는 명사병렬어구(N-NIL)
 - 예: 사과, 딸기, 배를 먹다.
  → 사과+딸기+배-NIL
  먹-사과, 먹-딸기, 먹-배
3) 속격 조사(N-N)
 - 예: 사람의 관절을 연구한다. → 관절-사람
4) 관형절(V-N)
 - 예: 아름다운 사람을 보았다.
  → 사람-아름답다 → 아름답다-사람

다음 그림2는 각각의 추출 규칙들을 보여준다.



[그림2] SIU5 추출 규칙

3. 색인 시스템의 구현

3.1 형태소 분석

구문 분석 시스템의 입력으로 사용되는 형태소 분석 및 태깅 결과는 ETRI에서 제작한 SCAN 시스템을 사용하여 생성한다.

3.2 형태소 태그 및 구문 태그 매핑

형태소 분석된 결과를 사용하는 구문 분석기는 GLR 파싱 알고리즘을 사용한다. 구문 분석기에서 사용하는 형태소 태그 집합과 SCAN에서 사용하는 태그 집합이 다르다. 그러므로 중간에서 태그를 변환시킬 필요가 있다. 또한, 형태소 분석 결과를 구문 분석을 위해 제정된 문법의 구문 태그와 매핑 시켜 실제 구문 분석에서 이용한다. 표1은 구문 분석기에서 사용되는 구문태그를 나타낸다.

[표1] 구문 태그

IC^{-1}	호격 구문(기능어)
IC^{0}	호격 구문(내용어)
VC^{-1}	동사구문(기능어)
VC^{0}	동사구문(내용어)
NC^{-1}	명사구문(기능어)
NC^{0}	명사구문(내용어)
DC^{-1}	관형구문(기능어)
DC^{0}	관형구문(내용어)
AC^{-1}	부사구문(기능어)
AC^{0}	부사구문(내용어)
SC^{-1}	문장(기호)
SC^{0}	문장

3.3 구문 분석

SIU 4와 5는 문장의 의미를 잘 나타내기 위해 구문 분석 결과를 사용한다.

인터넷 정보 검색에서 처리하는 문서들은 웹 문서이기 때문에 자체적으로도 많은 오류들을 가지고 있고 HTML 등의 markup 언어로 된 내용을 일반 텍스트로 변환하는 과정에서도 오류가 많이 발생한다. 그러므로 구문 분석이 성공적으로 수행되기 힘든 문제점을 가지고 있다. 이러한 문제점을 해결하기 위해 본 시스템에서는 완전 구문 분석이 실패할 경우, 분석이 성공한 부분까지만 사용하는 기법을 이용하여 견고성을 높인다.

색인어 SIU4와 SIU5를 추출하기 위해 이용되는 구문분석의 전체적인 흐름은 다음과 같다.

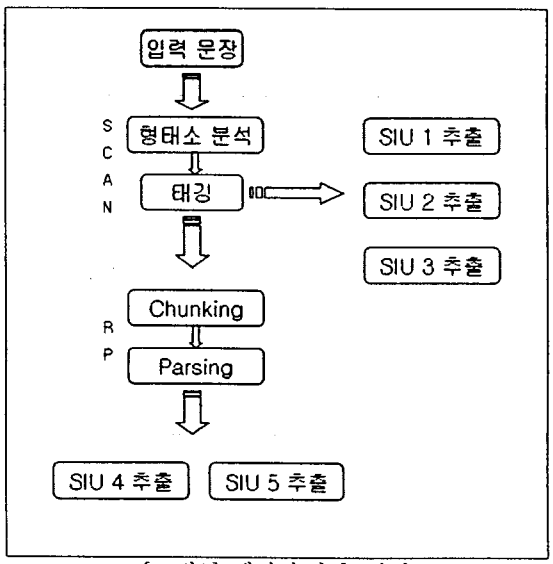
1) 입력 문장의 형태소 분석

2) 형태소 분석결과를 단위 구문(chunk)으로 분할
3) 단위구문단위를 최소 단위로 한 구문분석을 실시
4) 각각의 단위 구문 내부에 대한 구문 분석
5) 완전구문분석이 성공한 경우 전체 분석 결과를 이용
6) 완전구문분석이 실패한 경우 분석이 성공한 각 단위 구문 내부에 대한 결과만을 이용

부분 구문 분석 단계에서는 입력된 문장을 단위 구문으로 분할하여 완전구문분석단계에서 사용하는 알고리즘의 부하를 줄이고 분석 수를 줄여준다. 완전구문분석단계에서는 단위 구문을 최소단위로 한 구문 분석을 실시한 후, 각각의 단위 구문의 내부에 대해 또다시 분석을 실시하여 전체 분석 결과를 만들어 낸다. 완전구문분석에 실패한 경우, 부분 분석된 결과만을 사용한다.

단위 구문 인식은 서로 겹치지 않은 범위에 있는 구문 단위들을 인식하는 것으로 대상은 명사구(NP), 동사구(VP), 부사구(AP), 관형구(DP)이다.

SIU 4는 명사구만을 이용하므로 부분 구문 분석 결과만을 사용하고 SIU 5는 완전 구문 분석 결과를 이용하여 중심어-종속어 관계를 찾아낸다.



[그림3] 색인어 추출 과정

3.4 전체 과정

그림3은 문장을 입력 받아서 SIU1에서 SIU5까지 색인어를 추출하는 전체과정을 보여준다.

4. 실험 및 분석

현재 구단위 색인 및 검색에 대하여 객관적인 평가를 할 수 있는 테스트 집합 및 정답 집합이 없으므로 색인 정확도 및 검색 성능 향상에 대한 객관적인 평가를 할 수 없다. 그러므로 색인 및 검색 성능에 대해서는 객관적인 실험을 하지 않고 SIU 4, 5를 중심으로 색인 속도 및 견고성을 측정한다. 테스트 집합은 HANTEC 2.0의 12만 문서 모두 및 일부를 대상으로 한다.

4.1 실험

4.1.1 전처리 작업

HANTEC 2.0 문서들에 대하여 형태소 분석 및 구문 분석을 하기 위하여 문서에 들어있는 구조정보 태그를 제거한다. 이 때 웹 문서의 특성상 문장 구분이 제대로 되어 있지 않은 점을 고려하여 특정 태그를 삭제할 때는 문장을 구분하기 위하여 마침표를 삽입한다[3].

4.1.2 색인 실험

전체 12만 문서들에 대하여 색인을 하여 전체 시스템(형태소 분석기, 구문 분석기, 색인 추출기)의 견고성 및 속도를 실험하였다.

SIU 4,5는 시간 관계상 12만 문서 중 일부분에 대해서만 실험하였다.

4.2 실험 결과

전체 문서들에 대한 색인속도 측정결과는 표2와 같다.

[표2] 색인어 추출 시간

	추출 시간	색인 시간(DB 기록까지)
SIU 1	120,000문서 45분	엔트리수: 1,073,382 180분
SIU 2	120,000문서 68분	엔트리수: 1,066,403 205분
SIU 3	120,000문서 56분	엔트리수: 213,975 100분
SIU 4	21,000문서 960분	엔트리수: 1,132,390 1200분
SIU 5	5,000문서 330분	엔트리수: 117,898 20분

4.3 평가(실험 결과 분석)

4.3.1 구 색인으로 인한 기대효과

명사단위 색인 및 검색의 문제점은 문장 전체의 구조를 파악할 수 없으므로 관련이 없는 많은 문서들을 추출하는 것이다. 구단위 색인 및 검색은 이러한 문제점을 보완할 수 있다.

구단위 색인을 통한 검색 성능 향상에 대하여 SIU 4,5에 대하여 각각 알아보면 다음과 같다.

4.3.1.1 SIU 4에 의한 성능 향상

질의어로 "정보의 검색에 관한 사이트"를 넣었을 때 사용자는 '정보의 검색'이라는 단어들에 대하여 검색하기를 원한다. 즉, '정보'와 '검색'이 같은 문장에서 나올 경우 그 의미가 있다. 하지만 명사 색인만을 할 경우, 문서에 '정보', '검색', '사이트' 중 한가지의 단일명사만 많이 나오는 경우나, 같은 문장이 아닌 다른 문장이나 단락에 각 단어들이 많이 나오는 경우에도 해당 문서를 질의어와 관련된 문서로 검색한다. 그러므로 관련이 없는 문서들이 많이 검색되거나, 관련된 문서의 중요도가 낮아져 하위에 랭크될 확률이 높아질 수 있다. 구단위 색인은 이런 문제를 보완한다.

SIU 4의 색인 결과에 대한 예를 보면 다음과 같다.

질의어	색인 결과
모니터의 성능을 점검할 수 있는 업체의 주소는?	'모니터+성능', '업체+주소'
사용한 책을 사고, 집까지 배달된 물건을 확인할 수 있는 사이트는?	'사용+책', '배달+물건'
홍길동이 사용한 물건을 파는 장소?	'사용+물건'

4.3.1.2 SIU 5에 의한 성능 향상

SIU 4의 경우는 명사구만을 고려하기 때문에 문장 전체에 대한 의미를 반영할 수 있는 색인어를 추출하여 검색에 이용할 수는 없다. 즉, "홍길동이 사용한 물건을 파는 장소?"라는 질의어에서 '홍길동'+'사용'이라는 단어쌍이 이 문장의 의미를 가장 잘 나타내지만 SIU 4에서는 추출되지 않는다. 이런 문제는 SIU 5에서 보완한다.

4.3.1.1절의 질의어들에 대한 SIU 5 추출 결과는 다음과 같다.

질의어	색인 결과(Head|Modifier)
모니터의 성능을 점검할 수 있는 업체의 주소는?	'성능|모니터', '점검|성능', '업체|점검', '주소|업체'
사용한 책을 사고, 집까지 배달된 물건을 확인할 수 있는 사이트는?	'사용|책', '사|책', '배달|집', '배달|물건', '확인|물건', '사이트|확인'
이순신이 사용한 물건을 파는 장소?	'사용|이순신', '사용|물건', '팔|물건', '장소|팔'

4.4 색인 정확도

구단위 색인은 구문분석기에 전적으로 의존한다. 현재 구문 분석기의 성능이 보통 80% 이하이므로 색인 정확도가 80%를 넘기는 힘들 것으로 보인다.

구단위 색인에 대해서 객관적으로 테스트할 자료가 없으므로 자체적으로 몇 개의 문서에 대해서 정답 집합을 만들어 색인 정확도를 측정한 결과 65%정도의 성능을 보였다. 테스트한 문서의 수가 너무 적기 때문에 성능을 정확하게 나타낸다고는 할 수 없지만 대략적인 성능은 판단할 수 있다.

정확한 색인 정확도 및 검색 성능을 알 수 있도록 테스트 및 정답 집합을 만들어 테스트할 예정이다.

4.5 문제점

SIU 4,5 추출 중 발생한 문제점은 대부분 구문 분석 속도와 관련이 있다. 문장의 어절이 특정 개수 이상일 경우, 구문 분석 속도가 현저히 저하된다. 문장이 긴 경우는 다음과 같다.

1) 웹 문서에 문장 종결 기호가 찍히지 않아 여러 개의 문장을 하나의 문장처럼 처리하는 경우
2) 리스트나 표에 있는 명사들이 긴 명사열을 구성하는 경우
3) 한 문장 자체가 매우 긴 경우

1)번과 2)번의 경우, 구문 분석의 속도를 저하시킬 뿐만 아니라 분석 성능에도 커다란 영향을 미치므로 반드시 처리되어야 한다. 현재는 웹 문서를 전처리하는 과정 중, 특정 태그를 삭제하면서 마침표를 삽입하여 처리하고 있다.

3)번은 문장 분리기를 사용하여 복문을 단문으로 분리시켜 구문 분석 속도를 향상시키는데, 아직은 문장 분리기 자체의 성능이 저조하여 속도 향상보다는 구문 분석

성능을 도리어 저하시킬 수 있으므로 많은 연구를 필요로 한다.

5. 결론

본 시스템에서 이용한 접근 방법의 특징은 2단계 구문 분석을 사용함으로써 완전 분석 단계에서의 부하를 줄임과 동시에 전체 구문 분석이 실패하더라도 실패시점까지의 분석 결과를 이용할 수 있다는 것과, 구 및 문장에 기반한 색인을 하여 문장의 의미를 정확히 나타내는 색인어를 추출한다는 것이다.

본 연구를 통해 개발된 색인 시스템은 웹 문서의 오류를 다소 극복함과 동시에 고품질의 웹 문서 검색을 수행할 수 있는 기반 시스템을 개발하는데 기여를 할 것이라 생각된다.

앞으로는 5.4절에서 나타난 문제점을 해결할 수 있도록 문장 분리기를 더욱 보완하고, 색인 및 검색 성능을 측정할 수 있는 테스트 및 정답 집합을 만들어 실험할 예정이다.

6. 참고문헌

[1] 강인호, 전수영, 김길창, "최대 엔트로피 모델을 이용한 한국어 명사구 추출", *제12회 한글 및 한국어 정보처리 학술대회 논문집*, pp.127-132, 2000.

[2] 김미영, 강신재, 이종혁, "규칙과 어휘정보를 이용한 한국어 문장의 구묶음(Chunking)", *제12회 한글 및 한국어 정보처리 학술대회 논문집*, pp.103-109, 2000.

[3] 심준혁, 차정원, 이근배, "웹 인덱싱을 위한 통합 전처리 시스템의 개발", *제12회 한글 및 한국어 정보처리 학술대회 논문집*, pp. 216-223, 2000

[4] 원형석 외 2, 복합명사 분할과 명사구 합성을 이용한 통합 색인 기법, *정보과학회논문지:소프트웨어 및 응용* 제27권 1호, 2000.1.

[5] 장명길, 김현진, 오효정, "HANTEC 3.0에서의 키팩트 기반 텍스트 검색 방법에 관한 연구", *제5회 한국 과학기술 정보인프라 워크샵 학술발표 논문집*, pp. 203-221, 2000.12.

[6] 장명길, 김현진, 장문수, 최재훈, 오효정, 이충희, 허정, "의미기반 정보검색", *정보과학회지 10월호 한글정보처리 특집*, 2001.

[7] A.T. Arampatzis, T. Tsoris, C.H.A. Koster and Th.P. van der Weide, "Phrase-based Information Retrieval", *Journal of Information Processing & Management*, Volume 34, Issue 6, Pages 693-707, November 1998.

[8] Jose Perez-Carballo and Tomek Strzalkowski, "Natural language information retrieval: progress report", *Information Processing & Management*, Vol. 36, Issue 1, Pages 155-178, January 2000.