특집

정보검색을 위한 자연어처리

정 경 택
한국전자통신연구원 자연어처리연구실

I. 서 론

정보검색은 통계적 방법을 사용하는 전통적인 방법으로부터 시작하여 현재는 자연어처리 기술을 적용하는 단계에까지 와 있다. 단순히 단어를 키워드로 사용하여 빈도수와 가중치에 의존하던 방법이 갖고있는 검색 효율을 개선하기 위한 노력이 자연어처리 기술을 정보검색에 접목하기에 이른 것이다. 이미 오래 전부터 많은 관련 연구가 진행되어 왔지만 대용량의 다양한 문서와 여러 계층의 사용자를 위해 그리고 통계적인 정보검색 방법이 다다른 한계상황의 해결책으로 자연어처리 기술이 부각되고 있다.

본 논문의 2장에서는 정보검색의 정의와 역사 그리고 단계별로 대표적인 기술을 살펴보고, 3장에서는 자연어처리 기술이 정보검색에 활용되어야 하는 요구사항과 현재까지의 연구 사례를 간단히 살펴본다. 4장에서는 자연어처리 기술을 활용한 키팩트기반의 새로운 정보검색 기술을 설명하고, 5장에서 결론을 맺도록 한다.

II. 정보검색

1. 정보검색이란

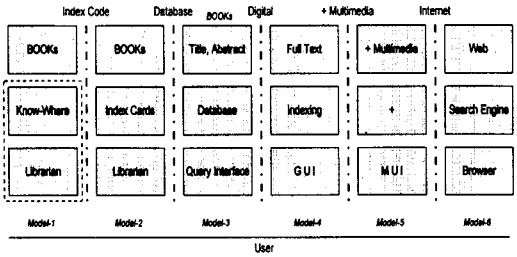
정보검색은 거시적으로는 문서 검색(Document Retrieval)으로부터 지식기반 검색(Knowledge Retrieval)까지의 모든 것을 말한다.[1] 문서 검색이란 용어는 특별한 경우를 제외하고는 텍스트 검색(Text Retrieval)이란 말과 같은 의미로 사용된다. 정보검색은 일반적으로는 문서 검색을 뜻하며 본 논문에서도 문서 검색의 의미로 사용할 것이다.

우리가 현재 인터넷을 사용하고 있는 환경에서 실질적인 정보검색의 의미는 사용자가 원하는 어떤 주제에 대해 관련 있는 문서를 찾아서 자기가 원하는 것을 문서에 포함된 가시적인 정보로부터 혹은 문서가 제공하는 의미를 유추하여 습득하는 과정과 그 결과를 말한다. 즉 사용자는 정보검색이

라는 기능으로부터 직접적인 답(Answering)을 원하는 것이 아니라 어쩌면 막연한 주제로부터 시작하여 자기가 원하는 목표를 구체화시켜 나가면서 그 결과를 만들어 가는 것이다.

2. 정보검색의 역사

도서관의 변화 관점에서 검색 기술의 발전을 살펴보자. 우리는 이미 도서관이라는 전통적인 개념이 바뀐 현재의 환경에서 살고 있지만 이 역사를 살펴보는 것은 정보검색의 역사와 기술의 변화를 쉽게 알 수 있다. 각각의 모델은 사용자 인터페이스, 인덱싱, 그리고 저장 미디어의 3가지의 측면에서 기술되어 있다. 그림 1의 가로축은 시간을 나타내고, 세로축의 점선은 기술적인 변화의 단계를 표시하고 있다.



〈그림 1〉 도서관에 비춰 본 정보검색의 역사

초창기의 도서관은 서고에 장서를 구비하고 사서를 통해 사용자가 원하는 정보(책)를 제공하였다(Model-1). 특히 분류 코드가 제대로 마련되기 전에는 사서의 지식(Know-Where)이 많은 역할을 하였다. 즉 검색 기술은 사서의 머리 속에 존재하였으며 사용자는 그 정보와 기술을 갖고 있는 사서에 의존하여야만 하였다.

이후 분류 코드가 정의됨에 따라 사서가 갖고 있던 검색 정보는 목록 카드로 만들어지고 사용자는 사서의 도움 없이 직접 목록 카드(Index Card)를 이용하여 원하는 정보를 찾을 수 있었다(Model-2). 사서의 머리 속에 있던 검색 기술은 분류 코드에 따른 목록 카드로 정리된 것이다. 사용자에게 사서뿐만 아니라 새로운 검색 도구를 사용할 수 있는 방법이 제공된 것이다.

70년대에는 도서관 전산화에 힘입어 데이터베이스 등이 구축되면서 도서의 타이틀 검색 기능과 일부 초록이 단말기 등으로 사용자에게 제공이 될 수 있었다(Model-3). 컴퓨터가 활용됨으로써 수동으로 입력된 키워드 등을 활용한 검색 기술과 일부 초록의 내용을 미리 확인하는 기능으로서 정확한 정보를 제공받을 수 있었다. 그러나 여전히 사용자가 원하는 정보는 책으로 제공이 되었다.

80년대에 들어서면서 컴퓨터 기술의 발전과 정보화 기기의 가격 하락으로 많은 정보가 디지털화 될 수 있었고, 따라서 장서의 전문이 데이터로 구축이 되어 초보적인 그래픽 사용자 인터페이스를 통해 제공이 될 수 있었다(Model-4). 많은 정보와 전문 정보가 구축되어야 함으로써 이를 위한 자동 인덱싱 처리 기술들이 적용되었다. 이로써 사용자는 데이터베이스를 통해 모든 데이터를 디지털 정보로 직접 제공받을 수 있는 환경이 되었다.

90년대 초반에 이르러 멀티미디어 기술이 정보처리 기술과 통합이 되어 사용자는 소리와 영상까지 포함하는 멀티미디어 정보(Multimedia Information)를 접하게 되었으며(Model-5), 최근 들어 무한히 발전하고 있는 인터넷에 힘입어 도서관은 통신망 속으로 사라져 버렸다(Model-6). 즉 웹이라는 거대한 망으로 연결된 컴퓨터들이 정보의 바다라는 보이지 않는 서고를 제공하고 있다. 이러한 환경 속에서 여러 검색 기술과 검색 엔진들이 개발되었으며 현재도 계속 발전하고 있다.

3. 정보검색 기술

전통적인 방식의 검색 기술은 4가지로 요약된다. 첫번째는 전문 주사방식(Full text scanning)으로서 찾고자 하는 모든 대상 문서에 대해 원하는 단어가 포함되어 있는 지를 스트링 검색으로 찾아보는 것이다. 이 방식 내에서의 성능은 스트링 검색 알고리즘에 좌우된다[3,4]. 이 방식이 다른 방식과 비교하여 유리한 점은 인덱스 정보를 필요로 하지 않기 때문에 별도의 저장 공간을 필요로 하지 않고, 대상 문서의 추가와 삭제에 부담이 없다는 것이다. 그러나 대상 문서가 많을 경우 성능면에서 치명적일 수 밖에 없다. 따라서 대부분 다른 검색

방식의 제한된 영역 내에서 활용된다.

두 번째는 시그니춰 비교 방식(Signature matching)으로 각 문서는 그에 포함된 단어들을 대상으로 해싱 등의 방법을 이용하여 비트열(Signature)를 만든 다음 별도의 화일(Signature file)을 생성하여 이들을 비교하는 것이다[4]. 시그니춰 화일은 원문보다 훨씬 작은 공간을 차지하며, 검색 성능은 비교적 빠르다. 이 방식은 문서의 크기가 큰 경우 성능이 저하되는 단점이 있으나, 구현이 용이하고 대상 문서의 추가와 불완전한 질의어 처리에 용이한 장점이 있다.

세 번째는 가장 널리 사용되고 있는 방식인 인덱스 검색(Inverted file search) 방식으로서, 각 문서는 그 문서를 대표하는 키워드로 표현되고, 키워드를 중심으로 그 키워드가 사용되고 있는 문서번호를 저장하는 인덱스 화일을 생성한다[5]. 이 방식의 단점은 인덱스 화일의 과도한 크기로 인한 저장 장소의 과중과 검색 대상 문서가 수시로 변하는 동적인 환경에서 인덱스(키워드)의 추가 및 삭제에 따른 관리 부담 등을 들 수 있다. 반면 비교적 구현이 쉽고, 속도가 빠른 장점이 있다.

네 번째는 벡터 공간 모델(Vector space model) 방식으로서, 각 문서와 질의어는 선정된 단어들의 용어 벡터(Term vector)로 표현되어 유사성을 검사하는 방식이다. 단어들 간의 독립성이나 벡터 방식의 이론적인 증명이 부족한 관계로 응용 도메인에 따라 사용자의 경험에 바탕한 판단에 의존해야 하기는 하지만, 반면 모델이 간단하고 가중치에 따른 검색 순위 매김이 용이하고 또한 관련 반응(Relevance feedback)과 같은 동적 환경에 잘 적응하는 장점이 있다.

III. 자연어처리

전통적인 정보검색 기술들이 비교적 만족할 만한 결과를 보여주고는 있지만, 문서가 포함하고 있는 제한된 단어 자체만을 기반으로 한 검색의 한계가 있다. 이를 극복하기 위한 여러 가지 방법들이 제안되고 있고 여기서는 자연어처리 기술이 요구되고 있는 이유와 자연어처리 기술을 적용한 정보검색 연구 사례를 살펴본다.

1. 자연어

정보검색과 자연어처리 기술은 불가분의 관계에 놓여 있는 것으로 봐야 한다. 초창기에 사용자는 그가 원하는 내용을 사서에게 자연스런 말(Natural language)로 표현하면서 정보를 요구하였다. 여기서 사용된 자연어는 바로 질의어가 단어의 단순 나열과 몇몇의 불리언 연산자로 구성된 것이 아니라 사용자가 원하는 의미를 담고 있는 하나의 문장인 것이다.

자연어 질의는 사용자가 원하는 정보의 내용에 대해 가장 많은 의미 정보를 보유하고 있다. 성능과 효율을 추구하면서 단순화되고 정형화되었던 질의어는 이제 다시 자연어 질의라는 원점으로 돌아가고 있는 것이다.

2. 사용자

정보검색은 얼마 전까지만 해도 학생이나 전문가들이 독점하는 것으로 인식되었던 것이 사실이다. 그러나 인터넷의 발전으로 심지어는 전화선을 통해 연결된 컴퓨터는 일반인들을 포함한 모든 사람에게 많고도 다양한 정보를 아주 가까이 가져다 주었다. 이제 정보는 주부나 초등학생부터 과학자나 전문 검색사에 이르기까지 다양한 사용자를 위해 존재하고 있다. 이들 다양한 계층의 사용자를 위한 검색 언어는 어떤 하나로 고정시킬 수는 없다. 왜냐하면 그들 각 계층의 요구를 만족하기란 쉽지 않은 일이기 때문이다. 따라서 이러한 요구를 충족시키기에 충분한 언어는 바로 자연어이다.

3. 비정형 문서

타이틀이나 초록 정보와 같이 어느 정도 정형화되어 있고 제한된 문서 정보에 대해 검색 기술을 적용하는 것은 비교적 쉬운 일이었다. 그러나 전문으로 제공되는 문서의 자유스러운 표현 방식과 다양한 분야의 문서 내용은 어느 일정한 패턴으로 검색 정보를 구축할 수는 없다. 이들을 위해서 가

장 그 문서들을 대표할 수 있는 단어를 선택하고 검색 정보를 구축할 수 있는 것은 자연어처리 기술을 이용한 문서 분석을 통하는 것이다. 이를 통해 인덱싱이 추구하는 목표인 재현률과 정확도를 높일 수 있다.

4. 사전 확장

기존의 검색 기법은 제한된 단어(키워드)를 중심으로 인덱싱을 구축하고 단어 벡터를 계산한다. 신문과 News 등과 대상 문서가 한 없이 늘어나고 새로운 단어가 만들어지는 경우 고정적인 단어 집합(Dictionary)으로는 이들을 처리하는데 문제가 있다. 따라서 이러한 정보들의 사전 정보를 자동으로 원활히 구축할 수 있는 방법은 대상 문서를 자연어처리 기술로 분석하여 의미 있는 새로운 단어와 정보를 추출하여 사전에 추가하는 것이다.

5. 표현의 다양성

하나의 의미는 다양한 표현을 통해 구사될 수 있다. 다음과 같이 여러 문장에서 사용된 여러 가지 표현은 자연어처리를 통해 하나의 의미군으로 취급할 수 있어야 한다.

- 정보 검색
- 정보검색
- 정보를 검색하고
- 정보가 검색되고
- 검색된 정보

〈그림 2〉 표현의 다양성

각각을 세밀하게 분석하면 차이가 있기는 하지만 현재의 정보검색을 위한 자연어처리 기술에서 이들을 동일한 의미로 취급할 수 있다면 검색 효율을 높일 수 있다.

6. 적용 방법

자연어처리 기술은 별도로 정보검색에 적용하기보다는 기존의 검색 방법의 효율을 올릴 수 있는 부가적인 방법으로 사용하는 것도 바람직하다. 예를 들어 질의어만을 자연어 처리 기법을 적용하여 사용자에게 자연어 질의 기능을 제공하거나, 혹은 기존의 검색 방법으로 검색된 1차 결과에 대해 자연어처리 기법을 적용하여 사용자가 가장 원하는 문서를 제공하는 것을 말한다.

7. 적용 수준

자연어처리는 몇 개의 수준으로 분류해 볼 수 있다. 이들은 음운론(phonological), 형태론(morphological), 어휘론(lexical), 문장론(syntactic), 의미론(semantic), 그리고 실용론(pragmatic) 수준들이다[6].

음운론 수준은 음성 이해나 음성 생성 시스템에 필요한 음성(소리)을 다룬다. 이것은 문서 검색과는 직접적인 관계는 없다. 형태론 수준은 단어의 형태와 의미 있는 단어의 일부를 처리한다. 접두어, 접미어의 처리와 단어의 변형 처리는 형태론에 기반 한다. 어휘론 수준은 문장의 모든 단어를 대상으로 문장의 기본적인 구조를 파악하고 각 단어의 품사를 나눌 수 있다. 정보 검색에서 각 단어의 사전 정보 처리, 전거어 처리, 그리고 시소러스 활용 등이 이에 해당한다. 문장론 수준은 문장을 문법 구조 정보로 표현하고 전치사구, 주부-목적부-술부 등으로 처리하는 것을 말한다. 이는 단어가 문장 내에서 구문적인 역할에 의해 성격이 결정되는 것을 이용한 것으로 많은 발전이 있는 분야이다. 의미론 수준은 문장론 처리에 문맥상의 지식을 부가하여 같은 의미를 갖는 새로운 문장으로 재생성하는 것이다. 예를 들어 "사과는 빨갛다". 라고 할 때, 이 문장을 의미론적 처리를 한다면 "빨간 것은 사과다". 라는 동등한 의미를 갖는 문장을 만들 수 있다. 실용론 수준에서는 실세계의 지식과 경험에 의한 지식이 문장 해석에 적용된다. 예를 들어 "가지가 나무에서 잘려나갔다". 라고 할 때, "가지"의 의미는 "나무"와의 현실적인 관계, 그리고 "잘리다"라는 행위의 관계와 함께 다른 의미의 "가지"와 구별될 수 있는 것이다.

기존의 정보 검색이 형태론, 어휘론 수준으로 연구 개발되어 왔지만, 현재의 방향은 문장론, 의미론, 실용론 수준으로 진행되고 있다.

8. 연구 사례

질의어의 의미적 내용과 문서의 의미적 내용이 부합하는 것을 찾으므로서 성능을 높이기 위한 자연어처리 기술을 적용한 연구가 있었고[7], Text Retrieval Conference(TREC) 대용량 corpus에도 적용이 되어 왔다. 단어를 기준으로 인덱싱하는 단계를 넘어 구(Phrase)를 인덱싱 단위로 사용하는 방법으로서 Croft 등이 문장(Sentence)을 인덱싱 단위로 사용하기도 하였다[8]. 이러한 방법은 좀 더 많은 정확한 정보를 담기는 하지만 우선순위 매김 및 부합하는 문서 찾기 성능에는 저하된 결과를 초래하기도 하였다. Rau 와 Jacobs 는 정보검색에서 쓰이는 중요한 단어가 문장에 따라 다른 의미를 갖는 다의성을 갖는 경우에, 단어 어근, 단어 의미, 구문론적 및 의미론적 정보를 갖고 있는 사전을 활용하여 이들을 구별할 수 있도록 자연어처리 기법을 적용하였다[9]. CMU에서는 HELIOS를 개발하면서 정보검색을 위해 자연어처리 소프트웨어인 CLARIT을 사용하여 내용에 기반한 인덱싱을 제공하고 있다[10]. 국내에서도 연구소와 많은 학교에서 자연어처리 기술을 적용한 정보검색 연구에 노력하고 있다.

간단한 자연어처리 기술은 이미 정보검색 시스템에 활용이 되어 왔다. 키워드를 추출하기 위해서는 형태소 분석기 등은 기본이기 때문이다. 근래에 와서는 고급의 자연어처리 기술과 같은 의미 분석 기술을 이용하여 정보검색에 활용하고자 하는 연구가 진행되고 있다.

IV. 키팩트기반 정보검색

본 장에서는 자연어처리 기술을 적용한 새로운 방식의 키팩트(Keyfact)기반 정보검색에 대하여 그 기술 내용을 간략히 설명하도록 한다.

1. 검색 수준

이미 2장에서 정보검색의 정의를 내린 바 있지만, 다시 한번 정보검색의 수준에 대해서 설명하고자 한다. 이를 통해 키팩트의 정의를 쉽게 이해할 수 있다. 예를 들어 다음과 같은 질의어가 있다고 하자.

정보를 검색하는데 사용되는 편리한 도구는 무엇입니까?

● "정보를 검색하는데 사용되는 편리한 도구는 무엇입니까"

〈그림 3〉 자연어 질의 예

이 자연어 질의 문장에 대하여

● 키워드기반 정보검색은

'정보', '검색', 그리고 '도구' 등의 단어가 들어있는 문서를 검색해 주는 수준,

● 키팩트기반 정보검색은

'정보를 검색하고', '정보의 검색', 그리고 '편리한 도구'라는 문장이 들어있는 문서를 검색해 주는 수준,

● 지식기반 정보검색은

구체적인 답으로 정보검색 도구인 'Netscape', 'Explorer' 등을 대답해 주는 수준으로 정의할 수 있다.

2. 키워드기반의 문제점

기존의 검색 시스템이 안고 있는 문제점을 다음의 측면에서 정리해 보았다.

● '의'로 묶어진 복합명사

'파리의 성당'이라는 문장에서 '파리'와 '성당'은 별개의 독립 단어(Keyword)로 취급됨으로써 단순히 빈도수에 의한 통계치로 관련 문서를 찾을 경우 오류를 발생시킬 확률이 커지게 된다.

● 관형사를 가지는 명사

사용자가 보통 일반적인 '산'이 아니라 '아름다운 산'이라는 특별한 산을 찾고 싶을 때 현재의 키워드기반 정보검색 시스템에서는 형용사나 동사를 키워드로 사용하지 않기 때문에 쉽게 찾을 수 없다.

● 명사를 목적으로 '하는' 하다동사가 결합되어 복합명사가 되는 경우

'정보를 검색'하고라는 문장이 들어있는 문서가 있을 때 현재의 키워드기반 정보검색 시스템에서는 '정보검색'이라는 복합명사로 된 질의어로는 이 문장을 찾을 수 없다.

● 같은 의미를 가지는 문장이 서로 다른 통사적 형태로 존재하고 있는 경우

'인터넷은 발전하고', '인터넷의 발전'등의 같은 내용이 '인터넷 발전' 혹은 '발전된 인터넷'이라는 질의를 사용해서 찾을 수 없다. 이러한 4가지 표현이 하나의 사실(fact)를 나타내기 때문에 어떠한 표현이라도 같은 의미로 색인되어 찾아져야 할 것이다.

● 같은 의미를 가지는 문장이 서로 다른 표현방법으로 존재하고 있는 경우

'새로운 기술'이라는 내용이 들어있는 문서를 '첨단 기술'이라는 질의어로 찾으려고 할 때 현재의 키워드기반 정보검색 시스템에서는 불가능하다. 이를 위해서는 의미적으로 같은 표현들은 의미 표준화를 통하여 같은 의미로 취급되어져야 할 것이다.

● 같은 의미는 아닐지라도 의미적으로 매우 가까운 키워드를 가지는 경우

'의사'라는 내용이 들어있는 문서를 찾을 때 정확하게 그 '의사'라는 단어가 없거나 의미적으로 주위에 있는 단어들이 필요한 경우 '간호사'나 '병원' 등의 의미적으로 관련된 질의어를 사용해서도 그 문서를 찾을 수 있어야 한다.

3. 키팩트 정의

위의 예에서 알 수 있듯이 키워드기반 정보검색 시스템에서는 해결할 수 없는 문제점들이 많이 있다. 이러한 것을 해결하기 위하여 문서의 내용을 대표하는 것이 'word(단어)'가 아니라 'fact(사실)'가 되어야 하는데 관심을 둔 것이 바로 키팩트(Keyfact) 기반 정보검색 기술이다[11].

문장에서 표현하는 방법은 여러 가지가 있을 수 있지만 그것이 나타내는 내용(사실)이 의미적으로 동일하다면 같은 키팩트라고 할 수 있다. 그러니까 하나의 키팩트는 어휘적으로나 통사적으로 같은 형태를 가진다고 할 수 없다. 왜냐하면 하나의 키팩트를 표현하는데는 여러 가지 형태의 표현 방법이 존재할 수 있기 때문이다.

어떤 문서에서 기존의 방법으로 키워드를 추출하여 놓고 그 키워드만으로 원래의 문서를 추론하는 것과 명사구를 추출한 후 그 명사구만으로 원래의 문서를 추론하는 것을 시험해본 결과 후자의 방법이 원래의 문서를 더 잘 대표하고 있다는 것이 증명되었다. 그러나 색인어의 조건은 첫째 문서를 대표하여야 하고 둘째 다시 나타날 확률이 있어야 한다. 이 조건을 만족하기 위해서는 통사론적인 명사구는 어느 정도 문서를 대표하기는 하나 다시 나타날 확률은 거의 없다. 예를 들어 "물에 매우 잘 녹고 또 공기보다 가벼운 기체"를 모을 때 쓰이는 방법 이라는 명사구는 문서의 내용을 잘 나타내고 있지만, 다른 문서나 질의어에서 다시 나타날 확률은 거의 없다. 키팩트기반 정보검색에서는 이러한 명사구는 여러 개의 키팩트로 추출되어야 할 것이다.

이런 의미에서 키팩트는 짧고 단순한 표현 양식을 가지고 있으면서 의미적인 내용을 잘 표현할 수 있어야 한다. 그리고 키팩트는 키워드의 수퍼세트이므로 키팩트로 구성되지 않는 키워드들은 기존의 키워드기반 색인으로 처리되어야 한다.

4. 키팩트 색인

2절에서 열거한 문제들을 해결하기 위한 하나의 방법이 문서 혹은 문장으로부터 의미 있는 사실을 추출할 수 있는 키팩트 색인이다. 키워드들의 단순한 집합만으로도 하나의 문서를 대표한다고 하지만 그것보다는 문서로부터 추출된 것(키팩트)이 완전한 명사구의 형태가 아니더라도 단순한 키워드의 형태보다는 문서를 대표하는데 유리하다면 그것을 추출하여 색인하는 것이 필요하다는 것이다.

색인과정을 살펴보면 먼저 하나의 문장에서 명사, 복합명사(명사와 명사가 띄어쓰기와 '의'로 연결된 경우만), 형용사, 동사에 해당하는 것을 원형과 같이 추출한다. 그리고 명사 뒤에 있는 조사도 같이 추출한다. 이것들을 문장 단위로 나열한 다음, 2-gram Keyfact Grammar로 재구성하여 키

팩트를 추출한다. 이 중 원래의 문장에서는 서로 문법적으로 혹은 의미적으로 연결이 되었으나 재구성 후 제대로 연결되지 않은 것들도 존재할 수 있다. 이러한 것들을 감안하더라도 명사들로만 이루어진 키워드들보다는 우수한 성능을 발휘할 수 있게 된다. 키팩트기반 정보검색에서는 검색의 기본단위가 키워드를 나타내는 스트링의 표현과는 달리 키팩트를 나타낼 수 있는 프레임과 같은 형태의 표현방법을 사용하여야 한다.

이러한 방법으로 만든 정보검색 시스템이 앞에서 설명한 키워드기반 정보검색의 문제점들을 모두 해결할 수 있을 지는 알 수 없다. 그러나 수동으로 해 본 좋은 결과에 비추어 보아 좋은 결과가 나올 수 있을 것으로 기대하고 있다.

V. 결 론

정보검색의 효율(재현률과 정확률)을 높이기 위한 신경망과 잠재의미 인덱싱(Latent Semantic indexing)과 같은 여러 가지 방법 및 기술들도 연구되고 있다[12]. 이들은 자연어처리 기술과는 달리 아직 큰 붐은 일으키지 않고 있으나 많은 가능성을 갖고 있다. 자연어처리 기술은 많은 응용 분야에 활용되고 있으며, 특히 정보검색과 관련된 다른 분야로서는 정보 추출(Information Extraction)에 그 기술이 적용되고 있다[13].

자연어처리 기술은 많은 기본적인 기술을 바탕으로 가져야만 그 진가를 발휘할 수 있다. 특히 의미를 다루는 것을 목표로 하고 있는 분야에 있어서 사전의 역할은 매우 중요한 것이다. 시소러스와 같은 의미망도 구축이 되어 왔지만, 정보검색은 그 나름대로 특별한 구조의 사전을 필요로 한다. 본 논문에서는 사전의 내용에 대해서는 깊이 다루지 않았지만 기본적으로는 그들을 바탕으로 진가를 발휘한다고 인식하여야 한다.

참 고 문 헌

[1] Lewis D. D., Jones K. S, Natural Language Processing for Information Retrieval, Commun. ACM, Vol. 39, pp. 92-101, January 1996.

[2] D.E. Knuth, J. H. Morris, and V. R. Pratt, Fast pattern matching in strings, SIAM J. Comput, Vol. 6(2), pp. 323-350, June 1977.

[3] R. S. Boyer and J. S. Moore, A fast string searching algorithm, CACM, Vol. 20(10), pp. 762-772, October 1977.

[4] C. T. Meadow, Text Information Retrieval Systems, Academic Press, Inc., pp. 201-211, 1992.

[5] G. Salton, Automatic Text Processing, Addison-Wesley Publishing Company, pp. 229-236, pp. 313-345, 1989.

[6] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill Book Company, pp. 257-302, 1983.

[7] E. B. Wendlandt and J. R. Driscoll, Incorporating a semantic analysis into a document retrieval strategy, Proc. of ACM SIGIR, pp. 270-279, October 1991.

[8] W. B. Croft, H. R. Turtle, and D. D. Lewis, The use of phrases and structured queries in information retrieval, Proc. of ACM SIGIR, pp. 32-45, October 1991.

[9] L. F. Rau and Paul S. Jacobs, Creating segmented databases from free text for text retrieval, Proc. of ACM SIGIR, pp. 337-346, October 1991.

[10] Galloway, Edward A., and G. V. Michalek, The Heinz Electronic Libray Interactive Online System(HELIOS) : Building a Digital Archive Using Imaging, OCR, and Natural Language Processing Technology, The Public-Access Computer Systems

Review 6, no. 4, 1995.
[11] H. W. Jang and S. Y. Park, Keyfact Concept for an Information Retrieval System, NLPRS 95, pp. 510-513, 1995.
[12] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, Indexing by latent semantic analysis, Journal of the American Society for Information Science, Vol. 41(6), pp. 391-407, 1990.
[13] J. Cowie and W. Lehnert, Information Extraction, Commun. ACM, 39, pp. 80-91. January 1996.

저 자 소 개

鄭 慶 澤

1959年 9月 24日生
1982年 2月 경북대학교 전자공학과 학사
1984年 2月 KAIST 전산학과 석사

1984年 2月～1987年 2月 현대전자산업주식회사 정보기술연구소 연구원
1987年 3月～현재 한국전자통신연구원 자연어처리연구실 선임연구원

주관심 분야 : 정보검색, 자연어처리, 멀티미디어 기술