자연언어를 이용한 연구보고서 검색효율성 측정 사례연구

Case Study on Retrieval Effectiveness of Technical Reprots by Natural Language

김 재 수, 국 방 과 학 연 구 소

KIM, JAE SOO, Agency for Defense Development

우리나라 연구소의 소규모 검색체계에서는 시소러스를 이용하지 않고 검색체계를 유지해도 별 문제가 없다는 생각을 가져온 것이 사실이다. 그러나 현실적으로는 검색효율이 극히 저조하고 잡음율이 높을 뿐만 아니라 필요한 정보의 접근이 불가능한 경우까지도 있다. 그래서 과연 현 체계대로 검색했을때 검색효율 즉 적합율과 재현율은 어느 정도 인가를 실험을 통해서 측정해 보았더니 극히 저조하다는 결론을 얻었고 그 원인을 분석해 보았다.

1. 실 험 환 경

1. 1. 실 험 대 상

자연언어를 사용하였을 경우의 검색 효율성을 측정하기 위한 실험 대상 자료를 연구 보고서로 정하였다. 그 이유는 보고서를 작성 제출한 저자(연구과제 책임자나 연구자들)로 하여금 실험에 절대적으로 필요한 해당 보고서의 키워드와 초록을 작성하도록 요청할 수 있기 때문이었다. 실험 대상 보고서는 연구 보고서의 주제 분야를 "미사일 공학"분야로 정하고 등록 대장에서 추출하였다. 추출된 160개의 미사일 공학 관련 보고서에 대한 키워드와 초록을 저자로 하여금 작성하도록 하였다. 작성 요령은 키워드를 5개 내외로 제한하고, 초록은 통보적 초록 형식에 따라 200-400자의 길이로 제한하였다.

1. 2. 기본파일 구축

본 실험을 위한 파일 구성 요소로서 160건의 미사일 공학 관련 보고서의 보고서번호, 저자명, 제목, 초록 및 저자 자신이 제시한 키워드를 포함시켰다. 표제를 파일 구성 요소로 포함시킨 이유는 표제는 저자가

본문의 내용을 최대한 압축하여 다른 정보와의 구별을 용이하게 한것이며 또한 이용자로 하여금 본문 내용을 어느 정도 파악할 수 있도록 하는 1차 식별자 역할을 수행하기 때문이다.

보고서 저자들로 부터 회수된 160건의 입력 양식에 기재된 내용으로 부터 색인어를 추출한 결과는 다음과 같다.

(1) 표제 : 표제에서 띄어쓰기를 기준으로 단어를 분리한 결과 829개 단어가 추출되었다.

(2) 초록 : 초록에서 띄어쓰기를 기준으로 단어를 분리한 결과 12,315개의 단어가 추출되었다.

이상 표제로 부터의 829개 단어, 초록으로 부터의 12,315개 단어와 저자가 제시한 823개 키워드를 합쳐 13,967개 단어가 1차적으로 파일 구축을 위한 기본 단어가 되었다.

2. 색인어 추출

2.1 특수 문자 제거 단계

단어의 앞뒤에 있는 구두점이나 기호 등의 특수문자를 제거 하였다.

2.2. 조사(助詞) 분리 단계

(1) 조사 역순 배열 테이블

비교 회수를 줄이기 위하여 첫째 조사테이블에 있는 조사들을 역순으로 배열한 조사 역순 배열 테이블을 작성하고, 둘째 최장 일치의 원칙(the principle of the longest match)을 적용시키기 위하여 음절 길이가 긴것부터 먼저 놓는다.

(2) 다음은 추출된 단어들을 역순으로 배열한다.

(3) 역순의 추출 단어 첫 글자가 조사 역순 배열 테이블에 있는가를 비교하고 역순으로 된 조사와 완전하게 일치하는가를 순차적으로 비교하여 일치하는 것이 있으면 원래 추출 단어의 뒤를 조사의 길이 만큼 절단하여 조사를 분리한다.

2.3. 불용어 제거 단계

(1) 좌절단 불용어 제거

좌절단 불용어 테이블을 작성하고 조사 절단 방법과 동일한 방법으로 비교하여, 좌절단 불용어 테이블에 있는 키워드를 제거하였는데 총 2,846개가 처리되었다.

(2) 우절단 불용어 제거

다시 우절단 불용어 테이블을 작성하여 우절단 불용어를 제거하였는데 총 1,170개가 처리되었다.

(3) 조사 절단 불용어 제거

다음으로 조사 절단 불용어 테이블을 작성한 후, 키워드가 조사 절단 불용어인가를 먼저 확인한다. 만약 불용어일 경우에, 조사 절단 불용어 다음에 조사가 오면 불용어로 처리하고 명사가 오면 키워드로 선정하였다. 처리 결과 총 661개가 처리되었다.

(4) 일반 불용어 제거단계

일반 불용어는 불용어 테이블에

등록하고, 키워드가 불용어 테이블에 있는 단어이면 제거하였는데 총 818개가 처리 되었다. 일반 불용어의 제거 기준이 되는 불용어테이블은 산업기술정보원에서 정한 100여개의 불용어와 국내 문헌에 나타나는 불용어를 중심으로 추가 또는 삭제하였다.

지금까지 기계적인 처리 방법에 의해서 추출된 키워드는 6,706개 였다. 이 가운데서 단순 중복 등 여러가지 부적합한 단어들을 수작업을 통하여 최종 점검한 결과 최초의 13,967 단어에서 6,119 단어가 남았다. 결국 43.8%가 남은 셈이다. 최종 6,119개의 키워드 가운데 저자가 제시한 색인어, 표제에서 추출한 색인어, 초록에서 추출한 색인어별로 문헌당 추출된 평균 색인어수를 조사해 본 결과는 (표1)와 같다. 여기에서 중복 단어들 제거할 때의 우선 순위는 저자의 색인어, 표제와 초록에서 추출한 색인어 순으로 하였다.

(표 1). 출처별 문헌당 키워드 수

구 분	키워드수	평균 (문헌당)
저자가 제시한 색인어	823	5.14
표제에서 추출한 색인어	653	4.08
초록에서 추출한 색인어	4,643	29.02
합 계	6,119	38.24

3. 측정 기준 설정

검색 효율 측정의 대표적인 몇가지 방법으로 재현율 및 정확율 이외에 대체 효율 척도, 복합 척도, 재현율과 정확율의 평균치 산출 기법등이 있는데 본 실험에서는 「재현율과 정확율의 평균치 산출기법」을 검색 효율의 측정 방법으로 하고자 한다. 다른 방법에 비해 본 실험에서와 같이 복수의 질문인 경우에는 이 방법이 가장 적합하다고 판단되어 이 방법을 택하였다.

본 실험에서는 질문 i에 대한 재현율과 정확율을 각각 더하여 평균을 낸 다음 공식을 이용하여 검색 효율을 측정하고자 한다.

$$\text{평균 재현율} = \frac{1}{m}\sum_{i=1}^{m}\frac{\text{검색된 적합 문헌 수}_i}{\text{적합 문헌 총수}_i}$$

$$\text{평균 정확율} = \frac{1}{m}\sum_{i=1}^{m}\frac{\text{검색된 적합 문헌 수}_i}{\text{검색 문헌 총수}_i}$$

이때 n은 전체 질문 수를 나타낸다.

4. 실험/평가

160건의 "미사일 공학" 관련 보고서의 보고서 번호, 저자, 6,119개의 키워드, 초록 등을 입력하여 구축한 마스타 파일을 대상으로 한 검색 과정은 다음과 같다.

먼저 "미사일공학" 전공자 3명으로 하여금 아래의 질문으로 검색을 수행하도록 하였다. 검색과정은 재현율을 높이기 위해 통상 이용하는 우측 절단 기법을 사용하였으며 다양한 동의어들을 합 집합으로 결합시켰다.

질문 1. 미사일 공력 해석에 관련된 보고서는 ?
질문 2. 미사일 유도조종에 관련된 보고서는 ?

질문 3. 미사일 발사 장치에 관련된 보고서는?

질문 1. 미사일 공력 해석에 관련된 보고서는 ?
- o 파일내의 적합 문헌 수 : 14건
- o 검색 문헌 수 : 11건
- o 검색된 적합 문헌 수 : 9건

질문 2. 미사일 유도 조종에 관련된 보고서는 ?
- o 파일내의 적합 문헌 수 : 35건
- o 검색 문헌 수 : 31건
- o 검색된 적합 문헌수 : 18건

질문 3. 미사일 발사 장치에 관련된 보고서는 ?
- o 파일내의 적합 문헌 수 : 35건
- o 검색 문헌 수 : 18건
- o 검색된 적합 문헌 수 : 9건

그 결과 각 질문별 재현율과 정확율은 (표2)와 같다.

이상 3개의 질문에 의한 검색 결과의 평균 재현율과 평균 정확율을 산출하면 다음과 같다.

$$\text{평균 재현율} = \frac{64.2 + 51.4 + 25.7}{3} = 47.13$$

$$\text{평균 정확율} = \frac{81.8 + 58.0 + 50.0}{3} = 63.23$$

지금까지 실험결과에서 보는 바와 같이 보고서를 쓴 저자가 준 초록, 키워드 제목을 전혀 가공하지 않은 순수 자연언어 상태로 색인하고 탐색하였을때는 검색효율이 극히 저조하다는 결론이 나왔다. 그 원인을 분석해 보면 다음과 같다.

첫째 초록의 작성이 전문가에 의하지 않고 저자들 자신이 각자 작성했기 때문에 그 수준이 다양하다.

둘째 키워드의 선정 또한 같은 방법으로 처리했기 때문에 키워드의 인식도가 부족한 탓으로 그대로 이용하는 데는 많은 문제가 있다.

셋째 탐색측면에서도 탐색자의 전공, 업무 배경에 따라 용어의 인식 차이가 많이 있다.

넷째 이번 실험에서의 특수사항이긴 하지만 많은 부분에서 위장 명칭이 사용되었기 때문에 그것을 모르는 탐색자는 탐색이 불가능한 경우가 있다.

위에서 나타난 문제점을 해결하는 방법은 한글 시소러스의 구축을 시도하는 것이라고 판단된다.

(표 2) 질문별 검색 결과치

질 문	적합문헌총수	검 색 기 준 치		결 과	
		검색문헌총수	검색된 적합문헌수	재현율	정확율
1	14	11	9	64.2	81.8
2	35	31	18	51.4	58.0
3	35	18	9	25.7	50.0