Posting File을 이용한 구절 검색 방법

박대원°, 박민식, 박진희, 권혁철
부산대학교 전자계산학과
dwpark@solge.cs.pusan.ac.kr

Phrase search using posting file in Korean Information Retrieval System

Dae-Won Park°, Min-Sik Park, Jin-Hee Park, Hyuk-Chul Kwon
Dept. of Computer Science, Pusan National University

요 약

Posting file은 문서 내의 색인어와 색인어의 위치 정보-문장번호, 어절 번호 등-로 구성한 문서별 색인어 역파일(inverted file)이다. 본 논문에서는 posting file을 구성하고 이를 정보검색시스템에 적용하여 색인어의 어절 거리 계산에 의한 구절 검색이 가능한 정보검색시스템을 소개한다. 또한 사용자 질의문과 가장 유사한 문장을 검색결과 대표문장으로 제시하여 사용자가 검색결과를 쉽게 확인할 수 있는 방법을 제시한다.

1. 서론

오늘날 인터넷의 발달로 웹 문서는 기하급수적으로 증가하고 있다. 수많은 문서들 속에서 원하는 정보를 효과적으로 검색할 수 있는 정보검색시스템의 중요성이 더욱 증가하고 있다.

그러나 질의어 나열에 의한 단어 위주의 정보 검색만으로는 사용자가 원하는 정보를 정확히 검색하기 어렵고 검색 결과 또한 상당한 양으로 사용자가 원하는 정보인지를 쉽게 판단하기가 어렵다.

본 논문에서는 기존의 단어 위주의 정보검색의 한계를 극복하기 위한 방법으로 문서별 posting file을 구성하고 이를 정보검색시스템에 적용하여 구절 검색이 가능한 정보검색시스템과 효과적인 검색결과 디스플레이를 위한 방법을 제시한다.

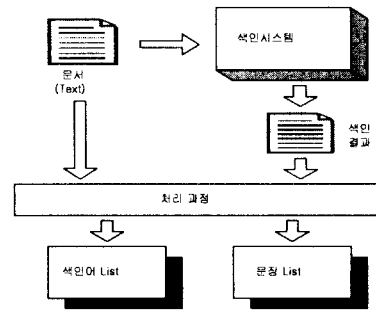
2. Posting File

기존 정보검색시스템은 검색대상 문서집합(Document Set)을 대상으로 색인어 역파일을 생성하여 검색에 이용하였다. 색인어 역파일은 문서집합(Document Set)에서의 색인어의 tf(term frequency), idf(inverse document frequency) 정보로 구성되어 문서 내의 색인어 위치와 같은 개별 문서 정보는 얻을 수 없다. 따라서 개별 문서를 대상으로 posting file을 구성하여 색인어 역파일과 함께 정보검색시스템에 이용한다.

Posting file은 하나의 문서(One Document)를 대상으로 문서 내 색인어와 색인어의 위치 정보로 생성한 문서별 색인어 역파일이다. 본 시스템에서는 posting file에 문서를 구조적으로 저장하기 위한 문장 정보 리스트를 추가로 구성하였다.

2.1 posting file 구성

Posting file은 원 문서(original text document)와 원 문서의 색인 결과를 이용하여 구성한다. Posting file은 문서에 나타난 색인어와 색인어의 위치 정보를 저장하는 색인어 정보 리스트와 문서의 구조적 저장을 위한 문장 정보 리스트로 구성된다. [그림-1]은 posting file을 구성하는 과정이다.



[그림-1] Posting file 구성

색인어 정보 리스트는 색인어(indexword), 문서 내 색인어 출현 빈도(frequency)와 색인어의 문서 내 위치 정보로 구성된다. 색인어 위치 정보는 색인어를 포함하는 문장의 문장번호와 문장 내 어절번호 tuple로 이루어진다. [그림-2]는 posting file의 색인어 정보 리스트이다.

[그림-2] 색인어 정보 리스트

문서는 색인어 정보 리스트를 이용하여 문장을 색인어와 비색인어 영역으로 구분하여 전체 문서를 문장별로 구조화하여 저장한다. [그림-3]은 posting file의 문장 정보 리스트이다.

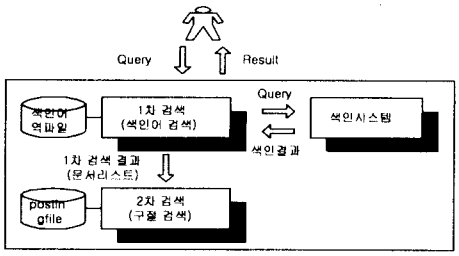
[그림-3] 문장 정보 리스트

2.2 문장 복원

구조화되어 저장된 문장의 복원은 문장 정보 리스트에서 해당 문장 스트림의 색인어 대체 과정을 통해 이루어진다. 색인어 부분은 색인어 ID로 색인어 정보 리스트에서 해당 색인어로 바꾸고 비색인어 영역은 그대로 문장으로 복원한다.

3. 정보검색시스템

정보검색은 색인어 역파일을 이용하는 1차 검색 과정과 1차 검색 결과에 대한 posting file의 색인어 어절 거리 연산의 2차 검색 과정을 통해 사용자가 원하는 정보를 정확히 검색한다. [그림-4]는 정보검색시스템의 검색과정이다.



[그림-4] 정보검색시스템의 검색과정

3.1 1차 검색 - 색인어 검색

정보검색시스템에서 검색을 위해서는 사용자 질의문의 색인이 필요하다. 이는 정보검색시스템이 색인어 역파일을 이용하여 검색하기 때문이다. 1차 검색은 사용자 질의문을 색인하고, 색인어 역파일에서 사용자 질의문의 색인어를 포함하는 문서를 검색하여 문서 리스트를 생성하고 문서 내 색인어 출현빈도와 문서 빈도를 이용하여 사용자 질의문과의 유사도를 계산하여 문서 순위를 결정한다.

예를 들어, 사용자 질의문이 "밤으로의 긴 여행"일 때 사용자 질의문의 색인 결과는 {"밤", "여행"}이다. 1차 검색은 "밤"과 "여행"을 포함하는 문서를 색인어 역파일을 이용하여 추출하고 각 색인어의 문서 빈도와 문서 내 출현 빈도를 이용하여 유사도를 계산, 문서의 순위를 결정한다. 문서 유사도 계산은 벡터스페이스 모델이나 P-norm 모델을 적용한다.

3.2 2차 검색 - 구절 검색(Phrase Search)

1차 검색은 일반적인 정보검색 과정으로 색인어의 문서 내 출현 여부와 빈도로 검색을 한다. 따라서, 검색 결과는 색인어 역파일에서 사용자 질의문의 색인어를 포함하는 문서를 추출한 문서리스트가 된다.

2차 검색은 1차 검색과 달리 사용자 질의문을 단순한 색인어의 나열로 보지 않고 하나의 구절로 인식하여 검색한다. 즉 사용자 질의문의 색인어 뿐 아니라 색인어 간의 상대 거리를 검색에 이용하여 검색의 정확도를 높인다. 본 논문에서는 사용자 질의문을 구절로 인식하고 색인어의 상대 거리를 이용하는 검색을 어절 거리에 의한 구절 검색이라 정의한다.

1차 검색의 결과 사용자 질의문의 색인어를 포함하는 문서 리스트를 얻었다. 구절 검색은 1차 검색의 결과를 대상으로 posting file을 이용하여 사용자 질의문의 색인어 간 어절 거리로 각 문서와 사용자 질의문과의 유사도를 구하고 문서의 순위를 조정한다.

예를 들어, 사용자 질의문 "밤으로의 긴 여행"의 색인 결과는 { "밤", "여행" }이고 두 색인어간 어절 거리는 2이다. 1차 검색 결과 색인어 역파일로부터 "밤"과 "여행"을 포함하고 있는 문서를 얻었다. 2차 검색에서는 posting file을 이용하여 1차 검색 결과 문서에서 "밤"과 "여행"의 어절 거리가 2인 문장이 있는지를 검사하고 문서의 가중치를 조정한다.

즉, "밤 OOOO 여행"이 "밤 여행", "밤여행", 또는 "여행 밤"보다는 사용자 질의문 "밤으로의 긴 여행"과 유사하므로 이를 포함하는 문서는 사용자가 원하는 문서일 가능성이 높다. 따라서 이들 문서에 대한 가중치를 높여주어 검색의 정확도를 높인다.

다음은 구절 검색을 위한 posting file의 색인어간 어절 거리 연산 함수들로 Adjacent, Near_N, Next_N 등이 있다.

3.2.1 Adjacent

Adjacent는 두 색인어의 어절 간 거리가 1 이하로, 같은 어절에 위치하거나 앞 뒤 어절에 연속하여 위치한 문장을 찾는다. 즉, Adjacent(색인어 1, 색인어 2)는 색인어 1과 색인어 2의 어절번호가 같거나 어절 간 거리가 1인 문장을 검색한다.

예) Adjacent(학교, 생활)
검색 가능한 구절
-> { 학교생활, 생활학교, 학교 생활, 생활 학교 }

3.2.2 Near_N

두 색인어의 어절 간 거리가 N이하인 문장을 찾는 함수이다. 색인어의 위치와는 상관없이 두 색인어의 어절 간 거리가 N 이하인 문장이 이에 해당한다. 색인어의 어절 간 거리를 중요시된다.

예) Near_N(학교, 생활, 3)
검색 가능한 구절
-> { 학교생활, 생활학교, 학교 생활, 생활 학교,
학교 OO 생활, 생활 OO 학교, 학교 OO OO 생활,
생활 OO OO 학교 }

한편 Adjacent는 Near_N의 특별한 형태로 어절 간 거리가 1 이하인 경우이다. Adjacent(색인어1, 색인어2) = Near_N(색인어1, 색인어2, 2)

3.3.3 Next_N

Adjacent와 Near_N은 색인어의 순서와는 관계없이 색인어의 어절 간 거리만을 계산하지만, Next_N은 어절 간 거리 뿐 아니라 색인어의 순서도 고려한다. Next_N의 첫 번째 색인어는 두 번째 색인어의 앞에 위치하고, 두 번째 색인어는 첫 번째 색인어의 위치에서 N어절만큼 떨어져 있어야 한다.

예) Next_N(학교, 생활, 3)
검색 가능한 구절 -> { 학교 OO OO 생활 }

3.3 검색 결과 Display

사용자의 질의문과는 상관없이 동일한 대표문장을 제시하는 기존 정보검색시스템과는 달리 사용자 질의문에 따라 다른 대표문장을 제시한다. 즉, posting file을 이용하여 사용자 질의문과 가장 가까운 문장을 대표문장으로 제시하여 사용자가 쉽게 검색 결과를 확인할 수 있도록 한다.

Posting file 연산을 통해 사용자 질의문과 유사한 문장의 문장 번호를 추출하고 문장리스트에서 해당 문장의 복원과정을 거쳐 검색결과의 대표문장으로 제시한다. 대표문장은 결과화면에서 사용자 질의문의 색인어를 다른 색으로 표시하여 사용자가 쉽게 검색 결과를 확인할 수 있다.

4. 실험

Posting file의 구성은 텍스트 문서 10만 건을 대상으로 실험하였다. Posting file은 원 문서(original text)의 평균 1.33배의 디스크 공간을 차지하였다. 문서는 압축 저장하지 않고 구조화하여 저장하였다. Posting file 구성 시간은 문서 1만 건당 평균 33.96초를 필요로 하였다.

(단위: KB)

문서 수	원 문서 (텍스트)	posting file		비율
		색인어	문장	
10000	16138	10616	11136	1.34
20000	32476	21256	22384	1.34
30000	47904	31264	32808	1.33
40000	63309	41248	43112	1.33
50000	78829	51608	53704	1.33
60000	94380	62192	64440	1.34
70000	110120	72968	75240	1.34
80000	126513	84080	86624	1.34
90000	139789	92576	94800	1.34
100000	153308	100840	102808	1.32

[표 1] Posting file 등록

5. 결론 및 향후 연구

본 논문에서는 정보검색시스템에 posting file을 적용하여 사용자 질의문을 구절 검색하는 방법과 사용자 질의문과 유사한 문장을 검색 결과 대표 문장으로 제시하여 사용자가 쉽게 검색 결과를 판단할 수 있는 방법을 제시하였다.

Posting file을 이용한 구절검색이 가능한 정보검색시스템은 색인어 중심의 검색 결과를 대상으로 2차 검색을 시도한 것으로 검색속도 면에서는 기존 시스템과 큰 차이를 보이지 않았다. 본 연구에서는 명사 색인어만으로 posting file을 구성하여 완전한 구절 검색이 이루어지지 않았다. 향후에는 명사 뿐 아니라 용언 색인 결과의 posting file 적용과 대용량 posting file의 분산 저장을 통한 성능 개선이 필요하다.

참고문헌

[1] Robert R. Korfhage, "Information Storage and Retrieval", Willey Publishing, 1997

[2] Ian H. Witten, Alitair Moffat, Timothy C. Bell, "Managing Gigabytes", Van Nostrand Reinhold, 1994

[3] Dornna Harman, "An Experiment Study of Factors Important in Document Ranking", Information Retrieval, 186-193, 1986, 8.

[4] Gerard Salton, Michael J. McGill, "Introduction to Modern Information Retrieval", McGrawHill, 1983.

[5] 정영미, 정보검색론, 구미무역 출판부, 1993.