한글 문서 검색에서 n-Gram 색인방법의 성능 분석

이 준 규*, 심 수 정**, 박 혁 로***
전남대학교 자연과학대학 전산학과
전화 : (062) 530-0311 / 팩스 : (062) 530-3439
H.P 번호 : 016-638-1423

※ 저자분은 www.ieek.or.kr 에 가서서 회원정보를 정확하게 입력하시기 바랍니다.

Performance Analysis of n-Gram Indexing Methods for Korean text Retrieval

Jun-Gyu Lee*, Su-Jung Sim**, Hyuk-Ro Park***
Computer Science Department, Chonnam National University

E-mail : *jglee266@netian.com, **sjsim@chonnam.ac.kr, ***hyukro@chonnam.ac.kr

Abstract

The agglutinative nature of Korean language makes the problem of automatic indexing of Korean much different from that of Indo-Eroupean languages. Especially, indexing with compound nouns in Korean is very problematic because of the exponential number of possible analysis and the existence of unknown words. To deal with this compound noun indexing problem, we propose a new indexing methods which combines the merits of the morpheme-based indexing methods and the n-gram based indexing methods. Through the experiments, we also find that the best performance of n-gram indexing methods can be achieved with 1.75-gram which is never considered in the previous researches.

I. 서론

문서 작성에 있어서 컴퓨터의 사용이 일반화되고 개별 컴퓨터들이 인터넷으로 연결됨에 따라서 온라인에서 사용 가능한 정보량은 기하급수적으로 증가하고 있다. 이러한 정보의 홍수 속에서 사용자가 필요한 정보를 신속하고 정확하게 접근하는 능력은 현대 사회의 중요한 생존 경쟁력이 될 정도로 점점 중요성이 커지고 있다. 정보검색 시스템은 이러한 정보 접근 문제를 해결하기 위한 기술로서 사용자가 대용량의 데이터로부터 필요한 정보를 효율적으로 발견할 수 있도록 도와준다 [1].

정보검색 시스템에서는 사용자 질의와 문서와의 빠른 비교를 위하여 문서와 질의의 내용을 색인어 집합으로 표시한다. 따라서 어떤 색인어를 추출하는가는 검색 성능에 결정적인 영향을 끼치게 된다.

컴퓨터가 문서의 내용을 분석하여 색인어를 추출하는 자동색인의 경우 주로 구미어를 대상으로 이론이 개발되어 왔다. 구미어 자동색인의 경우 먼저 색인어로써 전혀 가치가 없다고 판단되는 관사, 접속사 등과 같은 단어(불용어)를 제거한다. 이후 남은 단어에 대해 Stemming이라는 접사 제거 과정을 통하여 어근으로 변환한 후 이 어근을 색인어로 사용한다.

한글 자동색인 문제에 이러한 구미어 방법을 적용하는 것은 몇 가지 문제점을 가지고 있다. 먼저 불용어의 일종인 조사, 어미 등이 어근에 연결되어 어절을 형성하기 때문에 불용어를 제거하기 위해서는 어절 분석이 필요하다. 게다가 한글에서는 접사의 종류가 매우 다양하고, 이들 접사들끼리 매우 복잡한 형태로 결합될 수 있으며, 어근과 접사와의 결합과정에서 음운변화가 생기기도 하여 접사처리 과정이 영어에 비해 훨씬 복잡하다.

한글 문서 자동색인 방법으로는 자연언어 처리 기술의 일종인 형태소 분석을 이용한 방법 [2,3,4,5]과 어절을 일정한 크기로 분할한 단어 조각을 색인어로 사용하는 n-Gram 방법 [6,7,8]이 있다.

일반적으로 한글 문서의 경우 명사만을 색인어로 사용한다. 형태소 분석을 수행하면 비교적 정확하게 명사를 추출할 수 있지만 추출된 명사들 중 복합명사를 판별하고 분석해야 하는 새로운 문제가 발생한다. 이러한 문제를 해결하기 위해 본 논문에서는 형태소 분석을 통하여 명사를 추출한 후 복합명사 분석 단계에서 n-Gram 기법을 적용하는 색인 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 벡터 공간 모델과 기존의 한글 자동 색인 방법을 살펴본다. 3장에서는 본 논문에서 제안하는 형태소 분석 후 n-Gram을 적용하는 색인 방법에 대해 설명한

다. 4장에서는 한글 테스트 컬렉션인 Hantec 2.0을 사용한 실험 결과를 기술하고 5장에서는 결론을 맺는다.

II. 관련연구

2.1 벡터 공간 모델

벡터 공간 모델은 가장 널리 사용되는 정보 검색 모델 중 하나로서 사용자 질의와 시스템에 저장되어 있는 각 문서들을 색인어를 축으로 하는 n차원 공간상에 한 벡터로 표시한다.

벡터 모델에서 용어, 문헌 쌍(k_i, d_j)의 가중치 $w_{i,j}$는 양의 비이진 값이며, 질의 색인어도 가중치를 가진다. $[k_i, q]$의 가중치를 $w_{i,q}$라 하면, 질의 벡터 $\bar{q}$는 $\bar{q}=((w_{1,q}, w_{2,q}, \cdots, w_{t,q})$로 정의 되며, 여기서 t는 시스템 내의 전체 색인어 수이다. 문헌 d_j 벡터는 $\bar{d}=(w_{1,j}, w_{2,j}, \cdots, w_{t,j})$로 표현된다.

벡터 모델에서 문헌 d_j와 질의 q의 유사도 측정은 두 벡터 $\vec{d_j}$와 $\vec{q}$ 사이 각의 코사인 값으로 정량화할 수 있다.

$$sim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} = \frac{\sum_{t=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{t=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{t=1}^{t} w_{i,q}^2}}$$

벡터 공간 모델에서 각 색인어의 가중치 계산에는 tf-idf 방법이 일반적으로 널리 사용된다. tf-idf 가중치 계산 방법에서는 한 색인어 k_i의 가중치 w_{ij}는 이 색인어가 문서 d_j에서 출현한 빈도 tf_{ij}에 비례하고, 이 색인어가 출현한 문서 수 df_i에 반비례한다. 이를 식으로 표현하면 다음과 같다. 이 식에서 N은 데이터베이스 내 총 문서 수이며 $maxtf_j$는 문서 d_j에서 최대 tf 값을 표시한다.

$$w_{ij} = \frac{tf_{ij}}{\max tf_j} \log \frac{N}{df_j}$$

벡터 공간 모델의 주요 장점은 질의와 문서 간 부분 매칭이 가능하며, 이 둘 사이의 유사도에 따라서 검색 결과를 정렬할 수 있다는 점과 실수 가중치를 도입함으로써 색인어들의 상대적인 중요도를 고려하여 검색 성능이 개선된다는 점을 들 수 있다. 본 논문에서는 이러한 벡터 공간 모델을 기반으로 하여 검색 성능 평가를 실시한다 [9].

2.2 기존의 한글 자동 색인 방법

한글 문서에서 문서의 내용을 나타내는 것은 주로 명사이므로, 기존의 한글 자동 색인 방법들은 명사나 명사구 추출에 중점을 두어왔다. 따라서 문장 내에서 조사들을 제거하여 명사를 추출함으로 정보검색에 이용하였고, 이러한 한글 자동 색인 방법들은 추출되는 색인어의 단위에 따라 크게 어절 단위 색인법과 형태소 단위 색인법이 있다. 어절 단위 색인 방법은 문장 내에서 어절을 추출한 다음 비색인 분절을 제거하여 색인을 하는 방법으로서 이때 비색인 분절이란 체언의 뒤에 붙여 쓰이지만 색인어에 포함되지는 않은 조사, 어미, 접미사 등의 음절을 말한다. 형태소 단위 색인 방법은 문장의 모든 어절들에 대해 형태소 해석을 수행하여 최소 의미의 명사들을 문서와 질의의 표현을 위한 색인어로 선정한다. 이 방법은 형태소 해석단계, 애매성 제거단계, 명사추출 단계, 불용어 제거 단계로 이루어진다.

n-Gram 기반 색인 방법은 어절 단위 색인 방법을 개선한 방법으로 비색인 분절을 제거한 후 남은 색인어를 일정 크기의 단어 조각으로 분해한 후 이 분해된 단어 조각을 색인어로 사용하는 방법이다.

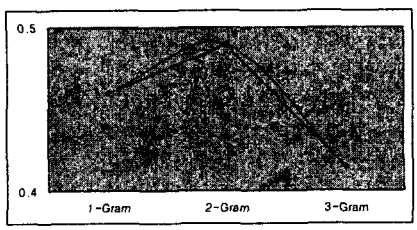
III. 제안하는 방법

3.1 동기

색인어를 추출함에 있어 n-Gram방법은 Stopword와 비색인 분절 문제를 처리하는데 단점을 가지고 있고, 형태소 분석은 복합명사, 신조어, 외래어 처리의 문제를 가지고 있다. 따라서, 형태소 분석을 통하여 Stopword와 비색인 분절 문제를 해결하고, n-Gram 방법을 통하여 복합명사, 신조어, 외래어 처리 문제를 해결하기 위하여 형태소 분석을 통한 n-Gram 방법을 제안한다.

3.2 제안하는 n-Gram

아래 (그림 1)과 같이 1-Gram, 2-Gram, 3-Gram 에 대한 정확률을 표시했을 때 2-Gram이 가장 좋은 성능을 보이지만 곡선 그래프로 표시한 경우에는 2-Gram 주위에서 높은 정확률을 보일 수 있다는 가정 하에 1.5-Gram 방법과 1.75-Gram 방법을 제안한다.



(그림 1) 제안하는 방법

제안 하는 방법은 형태소분석을 통하여 복합명사와 그에 해당하는 색인분절을 추출한다. 이때 3음절이상의 복합명사일 경우, 1.5-Gram은 1과 2를 Random하게 선택하여 1인 경우에는 1음절씩, 2인 경우는 2음절씩, 복합명사를 분절하는 방법이다. 그리고, 1.75-Gram

은 1과 2와 3을 Random하게 선택하여 1인 경우에는 1음절씩, 2와 3인 경우에는 2음절씩, 복합명사를 분절하는 방법이다.

3.3 1.5-Gram 색인방법의 예

<표 1> 1.5-Gram List Document

가계약상태 3 가 계약 상태
가계약서 3 가 계 약서
가계약은 2 가계 약은
가계약체결 3 가계 약 체결
가계약체결분 3 가계 약체 결분
가계약현황 4 가 계약 현 황
가계여유자금 5 가계 여 유 자 금
가계영업 3 가계 영 업
가계예금 2 가계 예금
가계용 2 가계 용
가계용품 2 가계 용품

위 <표 1>에서와 같이 '가계약상태'라는 복합명사는 '가', '계약', '상태' 3개의 음절로 분리되는데, 이것은 1, 2, 2라는 숫자가 Random하게 선택된 결과이다. 또, '가계여유자금'이라는 경우에는 2, 1, 1, 1, 1과 같이 Random하게 선택되었기 때문에 2음절, 1음절, 1음절, 1음절, 1음절과 같이 분리된다.

3.4 1.75-Gram 색인방법의 예

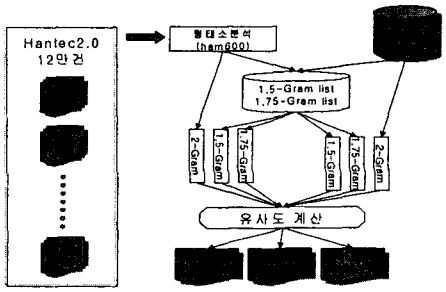
<표 2> 1.75-Gram List Document

가격점검 2 가격 점검
가격점검결과 4 가 격점 검 결과
가격정 2 가격 정
가격정보 3 가 격정 보
가격정보등 3 가격 정보 등
가격정보망 3 가격 정보 망
가격정보망등 4 가격 정 보 망등
가격정상화 3 가격 정 상화
가격정책 2 가격 정책
가격정책보 3 가격 정책 보
가격정책비 3 가격 정 책비

위 <표 2>에서와 같이 '가격점검'이라는 복합명사는 '가격', '점검' 2개의 음절로 분리되는데, 이것은 3, 2라는 숫자가 Random하게 선택된 결과이다. 또, '가격정보망'의 경우에는 3, 2, 1와 같이 Random하게 선택되었기 때문에 2음절, 2음절, 1음절과 같이 분리된다.

Ⅳ. 실험 및 평가

4.1 실험방법



(그림 2) 전체구조도

시스템의 전체 구성은 (그림 2)와 같다. 그리고, 실험에 사용한 문서는 한국과학기술정보연구원(KISTI)에서 제작한 Hantec 2.0 테스트 컬렉션을 사용하였다. Hantec 2.0 컬렉션은 일반, 사회과학, 과학기술 분야에 속하는 문서 12만 건 및 50개의 질의로 구성되어 있으며, 문서들이 각 분야별로 40,000건씩 균등하게 선정되어 특정 분야에 편중되지 않고 고른 분포를 가지고 있다. 질의문서는 전체 50개의 질의에 대한 적합문서집합을 사용하였다. 평가문서는 평점 5점을 받은 매우 적합문서 즉 G5문서를 사용하였다.

본 논문에서는 문서의 순위 결정 방법을 제공하는 검색 시스템은 보간 기법을 사용하여 고정된 재현율에 대한 정확률을 계산하는 11-Point 평균 정확률을 사용하여 검색 효과를 측정하였다.

용어가중치(TW)를 구성하는 요소는 단어빈도(TF), 역문헌빈도(IDF), 문헌길이 정규화(DL)의 3가지이다.

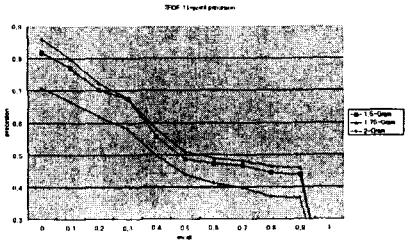
본 논문에서는 출현빈도(Term Frequency) 장서빈도(Collection Frequency) 정규화 방법을 아래 <표 3>의 3가지 방법을 사용하여 측정하였다.

<표 3> 색인어 가중치 부여 기법의 구성 요소

출현 빈도(Term Frequency)	
n tf	문서내에서 색인어의 출현 빈도
a $0.5 + 0.5\frac{tf}{\max tf}$	보강된 정규화 출현빈도(tf를 $\max tf$로 나누고, 그 결과가 0.5 ~ 1.0의 값을 갖도록 정규화)
o $\frac{tf}{2+tf}$	2-포아송 모형을 적용하는 OKAPI 시스템에서 사용하는 공식
장서 빈도(Collection Frequency)	
t $\ln\frac{N}{n}$	색인어 출현 빈도와 역문헌 빈도를 곱한다.(N은 전체 문서들의 수이며, n은 그 색인어를 포함하고 있는 문서들의 수이다.)
정규화(Normalization)	
c $\frac{1}{\sqrt{\sum_{vector} w_i^2}}$	SMART 시스템에서 사용되는 방법으로서 TF와 IDF의 조합인 TF*IDF값인 w를 해당 문헌 내 모든 단어의 w값의 제곱의 합의 제곱근으로 나눈값

4.2 실험 결과 및 평가

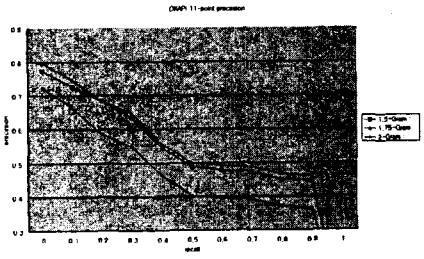
4.2.1 TF-IDF



(그림 3) 11-Point 표준 정확률(TF-IDF)

TF-IDF의 가중치를 사용하였을 경우 1.5-Gram방법은 2-Gram방법보다 15% 향상된 결과를 가져왔고, 1.75-Gram방법은 2-Gram방법보다 19% 향상되었다. 1.5-Gram방법보다는 1.75-Gram방법이 더 높은 정확률을 갖는 것으로 나타났다.

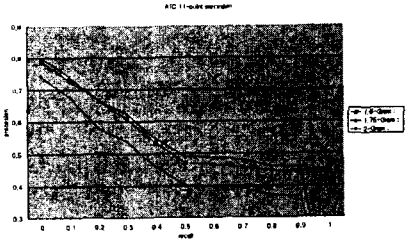
4.2.2 OKAPI



(그림 4) 11-Point 표준 정확률(OKAPI)

가중치를 $\frac{tf}{2+tf}$사용하였을 경우 1.5-Gram방법은 2-Gram방법보다 16% 향상된 결과를 가져왔고, 1.75-Gram방법은 2-Gram방법보다 19% 향상되었다. 1.5-Gram방법보다는 1.75-Gram방법이 더 높은 정확률을 갖는 것으로 나타났다.

4.2.3 ATC



(그림 5) 11-Point 표준 정확률(ATC)

가중치를 $0.5+0.5\frac{tf}{\max tf}$으로 주었을 경우 1.5-Gram방법은 2-Gram방법보다 14% 향상된 결과를 가져왔고, 1.75-Gram방법은 2-Gram방법보다 17% 향상되었다. 1.5-Gram방법보다는 1.75-Gram방법이 더 높은 정확률을 갖는 것으로 나타났다.

<표 4> 성능비교 결과

정확률 / 재현율	TFIDF-G5			OKAPI-G5			ATC-G5		
	1.5-Gram	1.75-Gram	2-Gram	1.5-Gram	1.75-Gram	2-Gram	1.5-Gram	1.75-Gram	2-Gram
0	0.8191	0.8604	0.7070	0.7735	0.7972	0.7237	0.7870	0.7983	0.7421
0.1	0.7710	0.7986	0.6652	0.7277	0.7472	0.6849	0.7277	0.7404	0.6737
0.2	0.7027	0.7230	0.6184	0.6647	0.6845	0.5886	0.6656	0.6851	0.5902
0.3	0.6748	0.6751	0.5816	0.6361	0.6473	0.5474	0.6091	0.6240	0.5370
0.4	0.5572	0.5787	0.4973	0.5444	0.5550	0.4646	0.5323	0.5452	0.4573
0.5	0.4861	0.5108	0.4393	0.4891	0.4993	0.4018	0.4703	0.4913	0.4014
0.6	0.4723	0.4881	0.4133	0.4791	0.4924	0.3951	0.4655	0.4872	0.3955
0.7	0.4662	0.4797	0.3967	0.4718	0.4862	0.3925	0.4574	0.4812	0.3936
0.8	0.4434	0.4647	0.3707	0.4493	0.4653	0.3709	0.4454	0.4606	0.3692
0.9	0.4392	0.4588	0.3668	0.4456	0.4611	0.3656	0.4413	0.4586	0.3670
1	0.0120	0.0120	0.0185	0.0134	0.0104	0.0137	0.0112	0.0081	0.0125
평균	0.5313	0.5500	0.4613	0.5195	0.5316	0.4479	0.5103	0.5237	0.4481

V. 결론

한글 문서 자동색인에 있어서 복합명사 분석 문제는 사전에 등록되지 않은 미등록어나 외래어 출현, 지수적인 분석 모호성 발생 등으로 인하여 매우 어려운 문제 중의 하나이다. 본 논문에서는 이러한 복합명사 색인의 어려움을 해결하기 위하여 형태소 분석을 통한 명사 추출 후, n-Gram 색인 방법을 적용할 것을 제안하였다.

여러 가지 가능한 n-Gram 방법들 중 가장 좋은 성능을 보이는 n-Gram을 찾기 위해 실험한 결과 기존의 연구에서 제시되지 않은 1.5-Gram, 1.75-Gram 등이 기존의 방법보다 좋은 성능을 보임을 알 수 있었다.

참고문헌(또는 Reference)

[1] Salton, G. "Historical Note: The Past Thirty Years in Information Retrieval, Journal of the American Society for Information Science." Vol.38, No.5, 1987

[2] 강승식, 권혁일, 김동혁, "한국어 자동 색인을 위한 형태소 분석 기능", 한국정보과학회 봄 학술발표논문집. 제22권 1호. 930-932

[3] 강승식, 한국어 형태소 분석과 정보검색, 홍릉과학출판사, 2002

[4] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회논문지(B), 25권 1호, pp.172-182, 1998.

[5] 안현수, "한글 문헌의 자동색인에 관한 실험적 연구", 정보관리학회지, 제3권 2호 108-306

[6] Cavnar, W.B. N-Gram-Based Text Filtering for TREC-2. In Proceedings of the Second Text Retrieval Conf.(TREC-2), NIST Special Publication 500-215. 171-179, 1994

[7] Damashek, M. "Gaushing Similarity with n-Grams: Language-Independent Categorization of Text," Science. Vol. 267. 843-848, 1995

[8] 이준호, 안정수, 박현주, 김명호 "한글 문서의 효과적인 검색을 위한 n-Gram 기반의 색인 방법", 정보관리학회지 제13권 제1호, p47-63, 1996

[9] 김명철, 김덕봉, 김유성, 김재훈, 박혁로, 이하규 공역 최신정보검색론, 홍릉과학출판사, 2001