키팩트의 가중치 부여에 대한 연구

김수희* : shkim@dogsuri.hoseo.ac.kr ,
남효돈* : hera@hscs.hoseo.ac.kr , 정경택** : ktchong@mail.etri.re.kr
*호서대학교 전자계산학과 데이터베이스 연구실, **ETRI 지식정보 연구팀

A Study for Weight Assignments on Keyfacts

Su-Hee Kim*, Hyo-Don Nam*, Kyung-Taek Chong**
* Department of Computer Science, Hoseo University.
**ETRI Knowledge Information Research Department.

요 약

정보검색에서 궁극적으로 지향하는 바는 질의에 대한 정확률과 재현률을 동시에 높이는 것이다. 본 논문에서는 [중심어, 종속어]로 이루어지는 키팩트를 그 유형에 따라 9가지 형태로 분류하였으며, 이 유형들의 주요도를 반영하여 키팩트의 가중치를 계산하는 방법을 개발하였다. 키팩트 유형들에 주요도 값들을 할당한 방법을 검정하기 위한 실험은 질의문들을 이용하여 평균 정확률와 평균 재현률을 계산함으로써 수행되었다. 9개의 키팩트 타입에 6가지의 주요도 값을 할당하는 방법을 실험하였고 그 결과를 분석하였다. 본 논문의 결과는 기존의 키워드 기반 정보검색에서 문제시되고 있는 정확률을 키팩트 기반 정보 검색에서 향상할 수 있는 가능성을 시사하고 있다.

1. 서론

기존의 키워드를 기반으로 하는 정보검색 시스템은 그 정확성에 있어 여러 문제들을 가지고 있다. 특히 정보의 양이 많아짐에 따라 정보검색의 전통적인 성능분석 척도인 재현률과 정확률의 잣대 중에서 정확률에 더 많은 노력을 기울이게 되었다. 한 예로 지금 인터넷상에 '전자 도서관'이라는 검색어를 입력했을 경우, 그 검색결과가 무려 2만여 건에 이르고 있다. 이렇게 검색되는 문서의 수가 많아짐에 따라, 재현률은 이에 비례하여 높아지는 경향이 있지만, 반면에 정확률은 매우 낮아진다. 정확률을 높이기 위해서 복합 명사들을 인덱스로 추출하는 연구가 활발히 이루어지고 있다[1, 2, 3, 4]

전자통신 연구원에서는 문서의 주된 내용을 대표하는 키팩트들을 [중심어, 종속어]의 형태로 추출하는 키팩트 추출기를 개발하였다[5]. 중심어는 주로 명사로 구성되고 종속어는 주로 명사, 관형사, 동사로 구성된다. 한 문서가 키팩트 제작기에 입력되면 [중심어, 종속어]의 리스트로 이 문서의 대표군이 생성된다. 본 논문에서는 전자통신 연구소에서 제작한 키팩트 추출기에 의해 생성되는 키팩트들을 몇 가지의 유형으로 분류하고 가중치를 계산하는 모델을 개발하고, 이를 검정하기 위해 질의문들을 이용하여 정확률과 재현률을 계산하여 비교 분석하고자 한다.

2. 키팩트

문장에서 표현방법은 여러 가지이지만 그것이 나타내는 내용(사실)이 의미적으로 동일하다면 같은 키팩트라고 할 수 있다. 키워드가 아닌 새로운 색인 대상 단위 즉 사실 기반 색인 단위를 키팩트라 부르고 있다[6]

1997년 전자통신 연구원에서는 문서의 주된 내용을 대표하는 키팩트들을 [중심어, 종속어]의 형태로 추출하는 키팩트 추출기를 개발하였다. 중심어는 주로 명사로 구성되고 종속어는 주로 명사, 관형사, 동사로 구성된다. 한 문서가 키팩트 추출기에 입력되면 [중심어, 종속어]들의 리스트로 이 문서의 대표군이 생성된다[7].

3. 키팩트의 유형분류 및 키팩트의 가중치

이 논문에서 사용한 키팩트 추출기는 다양한 형태의 키팩트들을 생성한다. 이 절에서는 키팩트를 몇가지의 유형으로 분류하고, 유형별 주요도를 부여한 가중치 계산법을 개발한다.

3.1 키팩트 유형 분류

- 유형 1) **[N, Nil]**, N은 명사
- 유형 2) **[N, Y]**, N1, N2는 명사, Y는 서술격의 종속어
- 유형 3) **[N1, N2]**, N1, N2는 명사
- 유형 4) **[N1 N2, Nil]**, N1, N2는 명사
- 유형 5) **[N1 N2, Y]**, N1, N2는 명사, Y는 서술격 종속어
- 유형 6) **[N1 N2, N3]**, N1과 N2는 명사, N3는 명사형 종속어
- 유형 7) **[N1 N2 N3, Nil]**, N1,N2 그리고 N3는 명사
- 유형 8) **[N1 N2 N3, Y]**, N1,N2 그리고 N3는 명사, Y는 서술격 종속어
- 유형 9) **[N1 N2 N3, N4]**, N1,N2 그리고 N3는 명사, N4는 명사형 종속어

4개 이상의 명사로 중심어가 되는 키팩트는 유형 7, 유형 8 그리고 유형 9중의 하나로 분류할 수 있다. 그러므로 모든 키팩트는 지금까지 분류한 9가지 유형중 하나의 유형에 속하게 된다.

3.2 키팩트의 가중치

가중치란 문서를 대표하는 인덱스의 중요한 정도를 나타내는 값으로 가중치를 구하는 방법은 여러 가지가 있지만 그 중에서도 빈도수에 근거를 둔 $w(tf*idf)$를 본 논문에서는 사용하여 키팩트 가중치를 개발한다[8]. 인덱스의 빈도수를 계산하여 가중치를 구하는 $w(tf*idf)$를 간단히 소개하면 다음과 같다[8].

$$w_{ij} = tf_{ij} \times \log \frac{N}{df_j} \quad (1)$$

식 (1)에서

N : 문서의 총갯수

w_{ij} : i번째 문서에서 j번째 인덱스의 가중치

tf_{ij} : i번째 문서에서 j번째 인덱스 t가 나타나는 빈도수

df_j : 총문서에서 j번째 인덱스 t가 나타나는 문서의 수

$\log \frac{N}{df_j}$: j번째 인덱스 t의 문서들의 식별자로서의 값

이다.

앞 절에서 키팩트들을 모두 9개의 유형으로 분류하였다. 각 유형의 주요도를 정량적으로 나타내기 위한 기호를 q라 하자. 즉 $q_l, 1 \le l \le 9$, 는 l 유형의 키팩트의 주요도이다. 빈도수로 계산할 수 있는 가중치 w에 키팩트의 유형별 주요도 q를 부여한 키팩트 가중치를 s라 하자. s는 다음과 같이 나타낼 수 있다.

$$s = q \times w = q \times tf \times idf \quad (2)$$

즉,

$$s_{ij} = q_{cj} \times w_{ij} = q_{cj} \times tf_{ij} \times \log \frac{N}{df_j} \quad (3)$$

라고 표현할 수 있다.

식 (3)에서

s_{ij} : i번째 문서에서 j번째 키팩트 t의 가중치

q_{cj} : j번째 키팩트 t의 유형별 주요도

이다.

4. 실험 및 결과 분석

4.1 모듈 개발

키팩트의 각 유형에 다양한 방법으로 유형별 주요도를 부여하고, 정확률과 재현률을 계산하는 실험을 수행하기 위해 C++로 유형별 주요도 부여 모듈, 특정 키팩트가 문서당 나타나는 빈도수 계산 모듈, 특정 키팩트가 나타나는 문서의 빈도수 계산 모듈, 키팩트 가중치 계산 모듈, 질의와 코퍼스내에 있는 문서들간의 내적 유사도 계산모듈, 유사도의 순위대로 검색하는 모듈들을 개발하였다[9, 10]. 두 문서간의 유사한 정도를 나타내는 값으로, 유사도를 계산하기 위해서 일반적으로 사용되고 있는 내적(Inner Product) 방법을 사용하였다.

4.2 사용한 코퍼스와 질의문

실험을 위해 사용한 코퍼스는 계몽사 대백과 사전의 일부인 250개의 문서이다. 전자통신연구소에서 이들을 바탕으로 46종류의 질의문을 개발하였으며, 각 질의문과 밀접한 관련이 있는 문서들의 목록을 관련 순위별로 작성하여 제공하였다[11].

4.3 키팩트 유형별 주요도 부여

제 3.1절에서 분류한 9 가지의 키팩트의 유형에 따라 다음과 같이 유형별 주요도를 부여하는 방법을 고안하였다.

● 방법 1

코퍼스 내에서 나타나는 빈도수가 가장 높은 유형 즉, [단순 단어, Nil]에 가장 큰 유형별 주요도를 부여하고, 그 다음으로 서술적 종속어가 있는 유형에 빈도수에 관계없이 큰 값을 부여하는데, 중심어가 단일 명사, 2중 복합 명사 그리고 3중 복합 명사인 경우에 따라 차이를 둔다.

● 방법 2

방법 1)과는 반대로 코퍼스 내에서 나타나는 빈도수가 가장 높은 유형 즉, [단순 단어, Nil]에 제일 적은 유형별 주요도를 부여하였다. 그 외의 유형에는 중심어에서는 명사의 갯수와 따라, 종속어가 있는 경우에는 종속어가 없는 경우보다는 큰 값을 부여한다.

● 방법 3

코퍼스 내에서 나타나는 빈도수가 가장 높은 유형 즉, [단순 단어, Nil]에 제일 적은 유형별 주요도를 부여하고, 방법 2)의 경우를 바탕으로 각 유형별의 값의 차이를 더 크게 한다.

● 방법 4

[N, Nil] 에만 주요도 1을 부여한다. 이 방법은 키워드기반 정보 검색과 유사하다.

● 방법 5

종속어가 없는 유형에만 균등한 주요도 1을 부여한다.

● 방법 6

이 방법은 앞서 소개한 5가지의 방법의 실험을 수행하여, 정확률과 재현률이 좋은 방법을 적절하게 반영한 방법이다([표 1] 참조). 9개의 키팩트 유형중 종속어가 Nil이 아닌 유형에 매우 높은 주요도를 부여하고, 그렇지 않은 유형에 매우 낮은 주요도를 부여한다.

유사도가 매우 낮은 문서인 경우에는 검색에서 제외하는 것이 바람직하다. 본 실험에서는 이러한 경계(*threshold*)를 가장 높은 유사도의 15% 이상으로 제한한다.

키팩트 유형 \ 유형별 주요도	빈도수 비율(%)	방법 1	방법 2	방법 3	방법 4	방법 5	방법 6
[단일 명사, Nil]	65.4%	1	0.01	0.01	1	1	0.01
[단일 명사, 서술격 종속어]	0.9%	0.7	0.87	0.4	0	0	1
[단일 명사, 명사형 종속어]	5.4%	0.4	0.9	0.9	0	0	0.9
[단1 단2, Nil]	16.6%	0.5	0.7	0.3	0	1	0.2
[단1 단2, 서술격 종속어]	0.4%	0.8	0.96	0.7	0	0	1
[단1 단2, 명사형 종속어]	2.3%	0.5	0.96	0.7	0	0	0.9
[단1 단2 단3 이상, Nil]	8%	0.7	0.9	0.5	0	1	0.5
[단1 단2 단3이상,서술격종속어]	0.1%	0.9	1	0.9	0	0	1
[단1 단2 단3이상,명사형종속어]	1%	0.6	1	1	0	0	0.9

[표 1] 키팩트 유형별 주요도 부여

[표 1]는 지금까지 논의한 6가지의 방법을 토대로 구체적인 주요도 값을 부여한 예이며, 이 값들을 실제 실험에 적용하였다.

4.4 실험

앞 절에서 고안한 6가지 방법별로 실제 주요도 값을 부여한 [표 1]를 적용하여, IBM PC 펜티엄III-450 MHz 상에서 실험을 수행하였다

46개의 질의문을 각 방법에 적용하였으며, 각 질의문마다 검색되는 문서들을 대상으로 정확률과 재현률을 계산하였다. 정확률과 재현률의 계산은 질의와 이와 관련한 문서들을 순위별로 나타내고 있는 자료를 활용하였다. [표 2]은 몇 개의 질의문에 대해 계산한 정확률(P)와 재현률(R)을 나타낸다.

질의 No.	방법 1		방법 2		방법 3		방법 4		방법 5		방법 6	
	P	R	P	R	P	R	P	R	P	R	P	R
1	0.36	1	1	0.6	1	0.6	0.36	1	0.38	1	1	0.6
21	0.3	0.75	0.75	0.5	0.75	0.5	0.3	0.4	0.27	0.4	0.75	0.5
37	0.7	0.6	1	0.6	1	0.6	0.46	0.86	0.63	1	1	0.6
42	0.35	1	0.75	0.85	0.75	0.85	0.35	1	0.35	1	0.75	0.85

[표 2] 유형별 주요도에 대한 정확률과 재현률

그 질의문들은 다음과 같다.

'질의문 1. 광개토 대왕 집권시기의 고구려의 영토는?',
'질의문 21. 지구 자전과 공전에 대해 설명해 주시오.',
'질의문 37. 우리나라 구석기 시대 유물이 발견된 곳은?'
'질의문 42. 신석기 시대의 토기 모양은?'

4.5 결과 분석

다음의 [표 3]는 [표 1]의 주요도 값들을 적용하여, 유형별 주요도 부여에 대한 6가지 방법별로, 46개의 질의문들의 정확률의 평균과 재현률의 평균을 나타낸다.

방법 \ 평균	평균 정확률	평균 재현률
방법 1	0.45	0.82
방법 2	0.54	0.75
방법 3	0.56	0.74
방법 4	0.45	0.82
방법 5	0.47	0.84
방법 6	0.58	0.76

[표 3] 각 방법에 따른 평균 정확률과 평균 재현률의 분석

방법 1은 평균 정확률이 45%, 평균 재현률은 82%로 재현률에서 우수한 결과를 보였다. 방법 2는 평균 정확률 54%, 평균 재현률 75%로 방법 1에 비하여 정확률은 9%가 증가한 반면, 재현률은 7% 감소하였다. 방법 3은 평균 정확률 56%, 평균 재현률 74%로 방법 2와 근사한 값들을 보였다. 방법 4는 키워드 기반 정보검색과 유사한 형태로 평균 정확률 45%, 평균 재현률 82%로서 평균 재현률은 높은 반면 낮은 정확률 값을 보였다. 방법 5는 평균 정확률 47%, 평균 재현률 84%로 모든 방법들 중 재현률에서 가장 우수한 결과를 보였으나 정확률에서는 키워드 기반 정보검색과 다를 바 없이 나왔다. 방법 5까지 보면 평균 정확률 값이 높으면 반대로 평균 재현률 값이 낮아지고, 평균 재현률 값이 높으면 평균 정확률 값이 낮아지는 경향을 보였다.

이런 경향을 극복하기 위하여 나온 방법으로 방법 6은 위의 5가지 방법들을 실험하며 나온 결과들을 재분석하여 주요도를 부여하였으며 평균 정확률 58%, 평균 재현률 76%로 정확률에서 가장 높은 값을 보였으며, 재현률에서는 다른 방법들의 평균치 값을 보였다. 실험한 방법들 중 재현률과 정확률이 동시에 높게 산출된 방법이다. 비교적 만족스러운 정확률과 재현률의 범위가 약 50% ~ 60%이라고 볼 때, 이 결과는 매우 고무적이다[8]. 본 연구 결과는 또한 기존의 키워드 기반 정보검색에서 문제시되고 있는 정확률을 키팩트 기반 정보검색에서 향상할 수 있는 가능성을 시사하고 있다.

[단일 명사, Nil]의 유형에만 가중치 값을 부여한 방법 4는 키워드 기반 정보검색과 유사한 방법으로서 키팩트 기반 정보검색인 방법 6과 비교하면, 평균 정확률에서 13%의 차이를 보이며 후자의 방법인 키팩트 기반 정보검색이 우수하게 나왔고, 평균 재현률에서는 6%의 차이를 보이며 키워드 기반 정보검색과 유사한 전자의 방법이 우수하게 나왔다. 이 결과는 기존의 키워드 기반 정보검색에서 문제시되는 정확률을 키팩트 기반 정보검색에서 개선할 수 있는 가능성을 시사하고 있다.

마지막으로, 종속어가 없으며 단일 명사로 이루어진 유형에만 주요도 값을 부여하는 방법 4와 종속어가 없으며 단일 명사나 복합 명사로 이루어진 유형에 주요도 값을 부여하는 방법 5를 비교하면, 후자의 방법에서 평균 정확률과 평균 재현률이 각각 2%의 차이로 더 높음을 알 수 있다. 이 결과는 단일 명사로만 색인하는 것보다 복합 명사로 색인하는 것이 더 우수하다는 것을 보여주는 또 하나의 예이다.

5. 결론

본 논문에서는 키팩트를 이루는 [중심어, 종속어]의 유형을 9가지로 분류하고, 6가지의 유형별 주요도 값을 부여하는 방법을 고안하였다. 그리고 빈도수에 근거를 둔 가중치 $w(tf\times idf)$에 유형별 주요도 q_i을 반영하여 키팩트 가중치를 계산하는 모델

$$s = q_i \times w = q_i \times tf \times idf \qquad (4)$$

을 개발하였다.

본 연구를 바탕으로 대규모의 코퍼스를 대상으로 정확률과 재현률을 계산하여 보다 이상적인 유형별 주요도의 부여와 키팩트 가중치 모델의 개발이 필요하다.

참 고 문 헌

[1] 박영찬, 최기선, "통계적 명사패턴 분류를 이용한 복합 명사 검색 모델", 제 8회 한글 및 한국어 정보처리 학술발표 논문집, 1996.

[2] 이현아, 이종혁, 이근배, "구분분석과 공기정보를 이용한 개념기반 명사구 색인방법", 제 7회 한글 및 한국어 정보처리 학술대회 논문집, 1996.

[3] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 공학박사 학위논문, 1993.

[4] Yasushi OGAWA, Avako BESSHO, Masako HIROSE. "Simple Word Strings as Compound Keywords' An Indexing and Ranking Method for Japanese Texts.", Proceedings of the sixteenth annual international ACM SIGIR conference on Research and Development in Information Retrieval, 1993.

[5] 한국전자 통신 연구원, 내용기반 멀티미디어 정보검색 기술 개발의 "의미정보 기반 검색 시스템 개발" (15 - 125), 정보통신부, 12월 1997.

[6] 오길록, 최기선, 박세영, 한글공학, 대영사, 1994.

[7] 한국전자 통신 연구원, 내용기반 멀티미디어 정보검색 기술 개발의 "내용기반 멀티미디어 정보검색 기술 개발" (3 - 7), 정보통신부, 12월 1997.

[8] Salton, G., Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer, Addison - Wesley Publishing Company, 1989.

[9] Stephen Prata, C++ Primer Plus second edition, Waite Group Press, 1995.

[10] 이경호, 파일처리론, 정익사, 1997.

[11] 계몽사 편집부, 계몽사 학생백과사전 CD, 계몽사. 1991.