정보검색에서 정확률의 향상을 위한 키팩트의 가중치 부여

(Weight Assignments on Keyfacts for Enhancing Precision in Information Retrieval)

김 수 희 † 남 효 돈 ††
(Su-Hee Kim)(Hyo-Don Nam)

요 약 정보검색에서 궁극적으로 지향하는 바는 질의에 대한 정확률과 재현률을 동시에 높이는 것이다. 이 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 검증하기 위한 실험은 질의문들을 이용하여 평균 정확률와 평균 재현률을 계산함으로써 수행되었다. 9개의 키팩트 타입에 9가지의 주요도 값을 할당하는 방법을 실험하였고 그 결과를 분석하였다. 이 논문의 결과는 기존의 키워드 기반 정보검색에서 문제시되고 있는 정확률을 키팩트 기반 정보 검색에서 향상할 수 있는 가능성을 시사하고 있다.

Abstract The main consideration in information retrieval is the precision and the recall of queries. In this paper, key-facts consisting of [main word, sub-word] are classified into nine types and a way to compute key-fact weights is developed based on the significance values of these types. An experiment to test one way for the assignment of significance values to key-fact types was performed by computing the precision and recall rates for the queries. Nine different ways for assigning significance values to nine key-fact types were experimented with and the results were analyzed. The result of this paper suggests that the precision rate can be improved in the method of information retrieval based on key-facts.

1. 서 론

인터넷의 발달로 일상생활에서부터 전문적인 연구에 이르기까지 필요한 정보를 수집하기 위해, 인터넷상에서 서비스되고 있는 각종 정보검색 시스템을 사용하는 추세가 일반화되어 가고 있다. 사용자가 원하는 정보를 보다 효율적으로 검색하도록 하기 위해서는 먼저 사용자가 질의를 쉽게 표현할 수 있도록 하여야 하며, 그 다음으로 정보 검색시스템이 이 질의를 분석하여 가장 적절한 정보를 제공하는 것이 필요하다.

† 정 회 원 : 호서대학교 컴퓨터공학부 교수
shkim@office.hosco.ac.kr
†† 비 회 원 : 한국보쉬기전전산실 연구원
HyoDon.Nam@KR.BOSCH.COM
논문접수 : 1999년 11월 29일
심사완료 : 2000년 10월 24일

전자통신연구원에서는 문서의 주된 내용을 대표하는 키팩트들을 [중심어, 종속어]의 형태로 추출하는 키팩트 추출기를 개발하였다[1]. 중심어는 주로 명사로 구성되고 종속어는 주로 명사, 관형사 그리고 동사로 구성된다. 한 문서가 키팩트 제작기에 입력되면 [중심어, 종속어]의 리스트로 이 문서의 대표군이 생성된다. 마찬가지로 문장 형태의 질의문도 [중심어, 종속어]의 리스트로 그 대표군이 생성될 수 있다.

이 논문에서는 전자통신연구원에서 제작한 키팩트 추출기에 의해 생성되는 키팩트들을 몇 가지의 유형으로 분류하여 가중치를 계산하는 모델을 개발하고, 이를 검증하기 위해 질의문들을 이용하여 정확률과 재현률을 계산하여 비교 분석하고자 한다.

이 논문의 구성은 다음과 같다. 제 2절에서는 정보검색에서 사용되는 기본용어에 대하여, 제 3절에서는 정보검색의 추세와 문제점들을 살펴본다. 제 4절에서는 키팩

트의 정의를 비롯하여 전자통신연구원에서 구현한 키팩트 추출기에 대하여 간단히 소개한다. 제 5절에서는 키팩트들을 유형별로 분류하고, 유형별 가중치 계산법을 개발한다. 제 6절에서는 실험을 수행하기 위해 개발한 모듈들에 대해 설명한다. 제 7절에서는 키팩트 유형별 주요도를 부여하여 실험과 분석을 하며, 마지막으로 결론을 맺는다.

2. 정보검색에서의 기본용어

내용 분석이나 인덱스 시스템의 효율성은 주로 인덱스의 완전성(exhaustivity) 혹은 철저성과 인덱스에 사용하는 용어의 특정성(specificity)에 의해 좌우된다고 볼 수 있다. 여기서 말하는 인덱스의 완전성은 원문의 각 요소가 실제로 인덱스 시스템에서 어느 정도 인식되는가 하는 정도를 반영하며, 용어의 특정성은 인덱스에 사용되는 용어들이 포괄하는 범위의 정도를 의미한다. 매우 좁은 범위의 용어들이 인덱스로 선택되었을 때, 오직 관련있는 몇 개의 문서들만이 검색될 것이고 실제로 관련있는 많은 문서들이 제외될 가능성이 높다. 일반 사용자들은 높은 정확률과 높은 재현률을 동시에 성취하기를 원하지만 사실상 이것은 불가능하므로 적당한 절충이 필요하다.

● 정확률 (Precision)과 재현률 (Recall)

정확률과 재현률은 정보검색 시스템의 성능을 평가하는 전통적인 주요 척도이다. 정확률은 검색된 문서 내에서 질의와 관련이 있는 문서들의 수를 검색된 총 문서들의 수로 나누어 표현할 수 있으며, 재현률은 검색된 문서들 중에서 질의와 관련이 있는 문서들을 전체 문서들 중에서 질의와 직접적 관련이 있는 문서들의 수로 나누어 표현할 수 있다[2]. 이들을 간단하게 수식으로 다음과 같이 나타낼 수 있다.

정확률 = c / b

재현률 = c / a

여기서 a는 전체 문서들 중에서 찾고자하는 정보와 관련이 있는 문서들의 수이고, b는 검색되어진 문서들의 수이며, c는 검색되어진 문서들 중 찾고자하는 정보와 관련이 있는 문서들의 수이다.

● 가중치

문서를 대표하는 인덱스의 중요한 정도를 나타내는 값으로, 사용자에게 질의와 관련이 높은 문서들을 제공하기 위해서는 각 문서를 대표하는 인덱스들을 잘 선정하여야 하며, 또한 선정된 인덱스들의 주요도를 변별력이 있도록 할당하는 것은 매우 중요하다.

가중치를 구하는 방법은 여러 가지가 있지만 그 중에서도 빈도수에 근거를 둔 $w(tf*idf)$를 이 논문에서는 사용하여 키팩트 가중치를 개발한다[2]. 인덱스의 빈도수를 계산하여 가중치를 구하는 $w(tf*idf)$를 간단히 소개하면 다음과 같다[2].

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

식 (1)에서

N : 코퍼스에 있는 문서들의 총 수

w_{ij} : i번째 문서에서 j번째 인덱스의 가중치

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

df_j : 코퍼스에서 j번째 인덱스 t가 나타나는 문서들의 수

$\log \frac{N}{df_j}$: j번째 인덱스 t의 문서들에 대한 식별자 값

이다.

● 유사도

두 문서간의 유사한 정도를 나타내는 값으로, 유사도를 계산하기 위해 일반적으로 사용되고 있는 방법들은 내적(Inner Product), Dice 계수, Cosine 계수, Jaccard 계수 등의 방법이 있다[2].

이 논문에서는 질의문과 검색의 대상이 되는 문서들과의 유사도를 계산하기 위해 내적 방법을 사용하며, 그 식은 다음과 같다.

$$\sum_{i=1}^{t} X_i \cdot Y_i \quad (2)$$

식 (2)에서 t는 인덱스들의 총 수이며, X_i , $1 \le i \le t$, 는 X라는 문서에서 i번째 인덱스의 가중치값이다. 문서 X와 문서 Y간의 내적 유사도는 전체 t개의 인덱스들에 대한 문서 X와 문서 Y내에서의 가중치들을 각각 계산하여, 대응하는 가중치들과의 곱들의 합으로 계산할 수 있다.

3. 정보검색의 문제점

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 추출

하는 연구가 활발히 이루어지고 있다[3, 4, 5, 6, 7, 8 ,9]. 한국어를 대상으로 복합어 인덱스 기법을 제안하고 실험을 수행한 연구로는 김판구의 연구 [7]과 이현아의 연구 [9]를 들 수 있다. [7]은 상호 정보 개념을 바탕으로 복합어의 유용성을 측정하여 복합 인덱스를 추출하였으며, 실험을 통하여 인덱스의 정확률이 향상되었다고 평가하고 있다. 그리고 [9]에서는 단문을 기반으로 한 명사구 색인 방법을 제안하고 인덱스의 평균 재현률과 평균 정확률을 계산하여 기존의 방법들에 비해 개선되었음을 보임으로써 이 방법론의 유용성을 입증하고 있다. 여기서

$$\text{인덱스의 재현률} = \frac{\text{추출된 적합 인덱스들의 수}}{\text{문서에 존재하는 적합 인덱스들의 수}}$$

이며

$$\text{인덱스의 정확률} = \frac{\text{추출된 적합 인덱스들의 수}}{\text{문서에서 추출된 인덱스들의 총 수}}$$

이다. 즉 그들의 연구에서는 각 연구에서 개발한 인덱스 추출 방법으로 추출한 인덱스들의 재현률과 정확률을 계산하였다. 이 연구에서는 키팩트 추출기를 이용하여 추출된 키팩트들의 재현률과 정확률을 계산하고자 하는 것이 아니라, 코퍼스에 있는 각 문서들을 대상으로 키팩드들을 추출하고, 이들에게 가중치를 부여하는 좋은 방법을 개발하기 위해 키팩트 타입에 따른 몇 가지의 주요도 방법을 제안한다. 많은 질의문들을 개발하여 제안하는 각 방법에 따른 질의문들의 정확률과 재현률을 계산하는 실험을 수행하여 그 결과를 비교 분석하고자 한다. 그러므로 기존의 연구들은 인덱스를 추출하는 방법의 유용성을 입증하려고 한 반면, 이 연구에서는 그것의 검증이 아니라 다양한 유형의 키팩트들의 가중치를 부여하는 방법을 개발하고자 하는 점에서 그 목적이 구별될 수 있다.

4. 키팩트

4.1 키팩트의 정의

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[10].

4.2 키팩트 추출기

1997년 전자통신연구원에서는 문서의 주된 내용을 대표하는 키팩트들을 [중심어, 종속어]의 형태로 추출하는 키팩트 추출기를 개발하였다. 중심어는 주로 명사로 구성되고 종속어는 주로 명사, 관형사, 형용사 그리고 동사로 구성된다. 한 문서가 키팩트 추출기에 입력되면 [중심어, 종속어]들의 리스트로 이 문서의 대표군이 생성된다[10].

표 1 키팩트 리스트의 예

감상 鑑賞

감상이란 이해 관계를 떠난 관심을 의미하는 것으로 꽃 자체나 자연풍경 또는 예술 작품 속의 오묘한 아름다움에 사로잡혀 이를 음미함을 말한다.

[감상1,Nil] [이해관계,Nil] [관심,Nil] [이해관계,떠나다] [의미,Nil] [꽃 자체1,Nil] [꽃,Nil] [자체1,Nil] [자연풍경,Nil] [예술작품,Nil] [아름다움,Nil] [예술작품,아름다움] [아름다움,오묘하다] [아름다움,사로잡히다] [음미함,Nil] [음미함,말하다]

이 논문에서 사용하는 키팩트 추출기는 가장 초기 버전이다. [표 1]은 '감상'이라는 문서 중 한 문장을 키팩트 추출기를 이용하여 추출한 키팩트 리스트이다.

[표 1]의 리스트중 [예술작품, Nil]은 종속어가 없는 키팩트인 경우이다. 이러한 형태는 키워드 형태와 유사하다. 즉 키팩트 리스트는 키워드들을 모두 포함한다고 볼 수 있다. '감상'은 여러 가지의 뜻을 가지고 있다. 키팩트 추출기에서 그 의미를 분석하고 분류하여 '감상 1' 혹은 '감상 2'의 형태로 서로 다른 의미의 '감상'을 구별하고 있다. [표 1]의 키팩트 [감상 1, Nil]에서 '감상'은 '감상 1'로 분류된 경우이다.

일반적으로 키팩트는 [S, T]로 표현할 수 있다. 여기서 S는 명사(단순 명사, 복합 명사)를 나타내고 T는 nil, 명사, 관형사, 형용사 혹은 동사를 나타낸다. T가 nil이 아닌 키팩트 [S, T]가 생성되면 항상 [S, nil]의 키팩트는 생성된다. S내에 있는 명사간의 거리, T내에 있는 품사간의 거리, 그리고 S와 T의 거리는 어휘분석 단위로 표현할 때, 그 상한이 6이다. 어휘분석 단위로 그 상한을 6으로 정한 것은 대용량의 말뭉치를 대상으로 실험하여 본 결과, 추출된 대부분의 키팩트의 구성요소의 거리가 통계적으로 6이내에 있음을 근거로 하고 있다.

[표 1]에서 관찰할 수 있듯이 키팩트 추출기는 단순 명사와 복합 명사 형태의 키워드를 생성하며, 또한 명사, 관형사, 형용사 그리고 동사 형태의 종속어가 있는 다양한 키팩트를 생성함을 알 수 있다.

5. 키팩트 유형의 분류 및 키팩트의 가중치

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

5.1 키팩트 유형의 분류

중심어, 종속어] 형태의 키팩트를 구성하는 기본 요소들의 개수와 종속어의 품사에 따라, 즉 중심어에 나타나는 명사들의 수, 종속어의 유무, 종속어가 명사 혹은 서술격 종속어인 경우들로 세분화하여 다음과 같이 9개의 유형으로 분류한다. 기호 N은 명사를 나타내며, Y는 서술격 품사를 나타낸다.

● 유형 1) [N, Nil]

종속어가 없는 형태로 기존의 키워드 형태와 유사한 유형으로 가장 많이 나타난다.

● 유형 2) [N1 N2, Nil]

두 명사 N1과 N2가 복합 명사의 형태로 의미가 있는 경우에 생성된다.

● 유형 3) [N1 N2 N3, Nil]

2개의 복합 단어로 이루어진 중심어에 대해서는 이미 언급하였다. 이 유형은 중심어가 3개의 명사로 이루어져 있다.

● 유형 4) [N1, N2]

두 명사 N1과 N2가 근접하게 나타나고 있으며, 중심어와 종속어 관계가 있다는 사실을 반영한다. N2는 서술격 종속어가 아니고 독립적으로 중심어가 될 수 있는 명사이다.

● 유형 5) [N1 N2, N3]

두 명사 N1과 N2가 서로 대등한 관계에서 복합 명사의 형태가 되고 명사 N3가 종속어가 되는 경우이다.

● 유형 6) [N1 N2 N3, N4]

이 유형은 3개의 명사로 중심어가 되고, 하나의 명사가 종속어로 이루어져 있다.

● 유형 7) [N, Y]

명사 N과 서술격 종속어 Y는 중심어와 종속어의 형태로 어떤 의미있는 관계를 가지고 있을 때 [N, Y]가 생성된다.

● 유형 8) [N1 N2, Y]

두 명사 N1과 N2가 복합 명사의 형태로 서술격 종속어 Y와 함께 생성되는 경우이다.

● 유형 9) [N1 N2 N3, Y]

이 유형은 3개의 명사로 중심어가 되고, 하나의 서술격 종속어로 이루어져있다.

중심어가 4개 이상의 명사로 이루어지는 키팩트는 유형 3, 유형 6 그리고 유형 9 중의 하나로 분류하고, 2개 이상의 종속어로 이루어지는 키팩트는 유형 9로 분류한다. 그러므로 모든 키팩트는 지금까지 분류한 9개의 유형 중 오직 하나의 유형에 속하게 된다.

5.2 키팩트의 가중치

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.

$$s = q \times w = q \times tf \times idf \tag{3}$$

즉,

$$s_{ij} = q_{cj} \times w_{ij} = q_{cj} \times tf_{ij} \times \log \frac{N}{df_j} \tag{4}$$

라고 표현할 수 있다.

식 (4)에서

N : 문서의 총갯수

s_{ij} : i번째 문서에서 j번째 키팩트 t의 가중치

q_{cj} : j번째 키팩트 t의 유형별 주요도

w_{ij} : i번째 문서에서 j번째 키팩트 t의 빈도에 의한 가중치

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

df_j : 코퍼스에서 j번째 키팩트 t가 나타나는 문서들의 수

$\log \frac{N}{df_j}$: j번째 키팩트 t의 문서들의 식별자로서의 값이다.

6. 모듈 개발

키팩트의 각 유형에 다양한 방법으로 유형별 주요도를 부여하고, 정확률과 재현률을 계산하는 실험을 수행하기 위해 C++로 모듈들을 개발하였다[11]. 다음은 개발한 주요 모듈들을 간단히 소개한다.

● 유형별 주요도 부여

제 5절에서 분류한 9가지 유형별로 주요도를 부여하는 모듈이다.

● 특정 키팩트가 문서당 나타나는 빈도수 계산 모듈

식 (4)의 tf_{ij}를 계산하는 모듈이다. 키팩트 리스트 파일을 읽어, 문서별로 각 키팩트의 빈도수를 세어 (키팩트, 빈도 수)의 형태로 출력한다.

● 특정 키팩트가 나타나는 문서의 빈도수 계산 모듈

식 (4)의 df_j를 계산하는 모듈이다.

● 키팩트 가중치 계산 모듈

$s_{ij} = q_{cj} \times tf_{ij} \times \log \frac{N}{df_j}$를 계산하는 모듈로 앞서 설명한 세 모듈의 결과를 이용한다. 역화일의 형태로 출력한다[12]. 키팩트의 가중치에 대한 역화일은 질의에서

추출된 키팩트들이 어느 문서에 얼마의 가중치로 나타나는 지를 효율적으로 탐색하기 위해

[공통 중심어 가중치1 문서번호리스트1
가중치2 문서번호리스트2 …
가중치n 문서번호리스트n
!종속어1 가중치11 문서번호리스트11
가중치12 문서 번호 리스트12 …
가중치1n 문서 번호 리스트1n …
…
!종속어p 가중치p1 문서번호리스트p1
가중치p2 문서번호리스트p2 …
가중치 pn 문서번호리스트pn]

의 형태로 유지한다.

[표 2]는 '지구'를 공통 중심어로 하고 종속어가 nil, '거리' 그리고 '공전' 등일 때의 키팩트의 가중치에 대한 역화일 구조의 예이다.

표 2 키팩트와 가중치에 대한 역화일 구조의 예

지구 0.2213 132 0.0379 187 185 0.0126 189 188 183 177 158 !거리 1.5353 188 183 29 !공전 1.9184 134 !공전 속도 1.9184 134 !껍데기 1.9184 132 !내부 3.8367 132 !대기 1.9184 29 !둘레 3.8367 132 !모습 1.9184 132 !모양 3.8367 132

● 질의와 코퍼스내에 있는 문서들간의 유사도 계산모듈
키팩트화 된 질의문과 역화일 형태로 출력된 키팩트 가중치 자료를 이용하여 질의문과 문서간의 유사도를 계산한다. 키팩트의 가중치 자료에 대한 역화일은 사용하는 코퍼스의 크기가 커지면 이에 비례하여 매우 커지게 되는데, 키팩트를 효율적으로 저장하고 탐색하기 위해 최소 완전 해시함수를 생성하여 이용함으로써 매우 빠른 속도로 유사도를 계산할 수 있도록 하였다[13].

● 유사도의 순위대로 검색하는 모듈
의문과 코퍼스내의 문서들 간의 유사도를 계산한 값을 토대로, 유사도가 높은 순서대로 각 문서의 내용을 검색하는 모듈이다.

7. 실험 및 결과 분석

7.1 사용한 코퍼스와 질의문

실험을 위해 사용한 코퍼스는 계몽사 대백과 사전의 일부인 250개의 문서이다[14]. 전자통신연구원에서 이들을 바탕으로 45 종류의 질의문을 개발하였으며, 각 질의문과 밀접한 관련이 있는 문서들의 목록을 관련 순위별로 작성하여 제공하였다].

코퍼스내에 있는 각 문서의 크기는 매우 다양하다. 예를들면 , '고구려'에 대한 문서는 721개의 단어로 이루어져 있고 문자수로는 2,359개이다. 반면에 '105인 사건'에 대한 문서는 65개의 단어로 이루어져 있고 문자수로는 199개이다. 대부분의 질의문은 짧은 단문의 형태이고 다양한 범주에 속한다고 볼 수 있다. 다음은 사용된 질의문들의 예이다.

- 우리나라의 좋은 낚시터는?
- 바다의 적조 현상이 일어나는 것은 무엇 때문입니까?
- 한지를 만들 때 모두 몇 번의 공정을 거쳐야 하는가?

7.2 키팩트 추출

코퍼스로 사용된 250개의 문서와 45개의 질의문을 대상으로 키팩트를 추출하였다. 유형별 주요도를 부여하는 과정에서 각 유형의 빈도수의 비율을 참고할 수 있도록 코퍼스에서 생성되는 키팩트들을 유형별로 그 빈도수를 세어 보았다. 다음의 [표 3]은 코퍼스에서 나타나는 키팩트들의 유형별 빈도수를 나타낸다. 유형 1의 빈도의 비율이 65.31%로 가장 높고, 유형 2의 비율이 16.61%로 그 다음으로 높으며 유형 9의 비율이 0.13%로 가장 낮음을 알 수 있다. 이 코퍼스에서는 종속어가 없는 유형이 가장 많고, 그 다음으로 종속어가 명사인 유형, 마지막으로 종속어가 서술형인 순서로 분포하고 있다. 또한 유형번호의 역순으로 그 분포가 크게 증가하고 있다.

표 3 키팩트의 유형별 빈도

키팩트 유형		빈도수	빈도수 비율
유형 1	[N, Nil]	12,842	65.31%
유형 2	[N1 N2, Nil]	3,266	16.61%
유형 3	[N1 N2 N3 …, Nil]	1,565	7.96%
유형 4	[N1, N2]	1,073	5.46%
유형 5	[N1 N2, N3]	456	2.32%
유형 6	[N1 N2 N3 …, N4]	186	0.95%
유형 7	[N, Y]	178	0.91%
유형 8	[N1 N2, Y]	70	0.35%
유형 9	[N1 N2 N3 …, Y]	27	0.13%
총계		19,663	100.00%

7.3 키팩트의 유형별 주요도 부여 및 실험

키팩트의 유형별 주요도를 부여하여 정확률과 재현률을 향상하는 하는 실험을 수행하기 위해 다양한 방법으로 주요도 값을 부여한다. 먼저 이 실험 내에서 기존의 키워드 기반 정보 검색과 유사한 검색을 해보기 위해

종속어가 없는 유형에만 주요도값을 부여해 본다. 그리고 앞 절에서 계산한 유형별 빈도를 참고하여 빈도수가 높은 유형에 높은 주요도값을 부여하거나 낮은 주요도값을 부여하여 실험을 수행하고자 한다. 빈도수를 참고로 하여 실험을 수행하는 경우에는 빈도수에 근거하여 수행한 이전의 결과를 토대로 하여 적절히 유형별 주요도 값들을 조정하고자 한다.

[그림 1]은 실험을 수행하기 위해 개발한 각 모듈들의 구동 및 대강의 실험 순서를 나타낸다. 각 질의문으로부터 추출된 키팩트와 기존의 코퍼스 내에 관련이 있는 문서들을 검색하여 그 문서들을 대상으로 정확률과 재현률을 계산한다. 정확률과 재현률의 계산은 질의와 이와 관련한 문서들을 순위별로 나타내고 있는 자료를 활용한다.

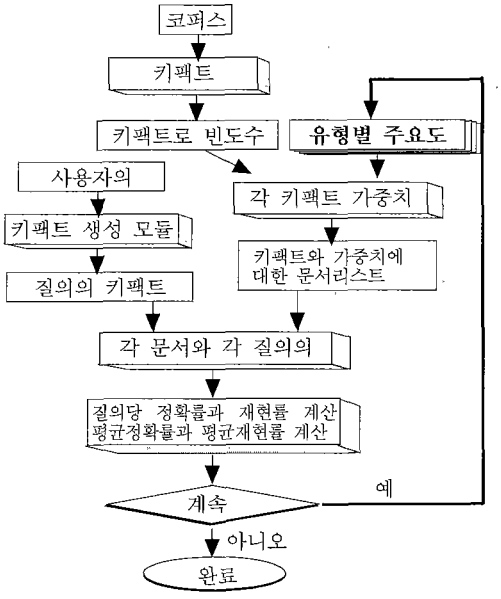

그림 1 키팩트 가중치 부여에 대한 순서도

7.3.1 키팩트의 유형별 주요도 부여

다양한 방법에 의한 정확률과 재현률을 비교해 보기 위해 다음의 9가지의 방법으로 실험을 수행하였다. 기존의 키워드기반 정보 검색과 유사한 검색을 수행해 보기 위해 2가지 방법으로 주요도를 부여하여 실험을 수행하며, 나머지 7가지 방법은 빈도수의 통계에 근거한 방법이다. 앞서 언급한 대로 빈도수를 기초로 한 주요도 부여는 빈도수에 근거하여 수행한 이전의 결과를 토대로 하여 적절히 유형별 주요도 값들을 조정하였다.

● 방법 1

[N, Nil] 유형에만 주요도 1.0을 부여한다. 이 방법은 단순명사로 이루어진 키워드기반 정보 검색과 유사하다.

● 방법 2

[N, Nil], [N1 N1, Nil] 그리고 [N1 N2 N3, Nil] 유형에만 균등한 주요도 1.0을 부여한다. 이 방법은 단순명사와 복합명사를 포함하는 명사구로 이루어진 키워드 기반 정보 검색과 유사하다.

● 방법 3

코퍼스 내에서 나타나는 빈도수가 높은 순서대로 주요도를 0.1씩 차이를 두며 부여한다. 가장 빈도수가 높은 유형 [N, Nil]에 1.0을 부여하고 빈도수가 가장 낮은 [N1 N2 N3, Y]유형에 0.2를 부여한다.

● 방법 4

방법 3과는 반대로 코퍼스 내에서 나타나는 빈도수가 낮은 순서대로 유형별 주요도를 0.1씩 차이를 두며 부여한다. 가장 빈도수가 낮은 유형 [N1 N2 N3, Y]에 1.0을 부여하고 빈도수가 가장 높은 유형 [N, Nil]에 0.2를 부여한다.

● 방법 5

방법 4와 유사하나 각 유형별의 값의 차이를 조금 크게 하여 부여한다.

● 방법 6

코퍼스 내에서 나타나는 빈도수에 따라 5개의 소그룹으로 분류하여, 빈도수가 높은 그룹에 낮은 주요도를 부여한다. 이 방법은 방법 4와 방법 5와 유사하나 그룹간의 빈도수에 따라 부여하는 값의 차이를 크게 한다. 빈도수가 높은 그룹 순서대로 0.01, 0.2, 0.5, 0.9 그리고 1.0의 값들을 각각 부여한다.

● 방법 7

방법 6과 매우 비슷하나 빈도수가 가장 높은 그룹의 주요도를 0.01에서 0.005로, 빈도수가 두 번째로 높은 그룹의 주요도를 0.2에서 0.3으로 변경한다.

● 방법 8

방법 7을 조금 더 세분화하여 주요도를 부여하면서 빈도수가 가장 높은 그룹의 주요도를 0.005에서 0.007로 변경한다.

● 방법 9

방법 8에서 빈도수가 가장 낮은 그룹의 영향을 알아보기 위해 이들의 주요도를 1.0에서 0.0, 0.5, 1.5로 각각 변경하여 실험한다.

표 4 키펙트 유형별 주요도 부여

키펙트 유형		유형 1	유형 2	유형 3	유형 4	유형 5	유형 6	유형 7	유형 8	유형 9
유형별 주요도	방법 1	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	방법 2	1.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
	방법 3	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2
	방법 4	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
	방법 5	0.1	0.2	0.3	0.3	0.4	0.5	0.8	0.9	1.0
	방법 6	0.01	0.2	0.5	0.9	0.9	0.9	1.0	1.0	1.0
	방법 7	0.005	0.3	0.5	0.9	0.9	0.9	1.0	1.0	1.0
	방법 8	0.007	0.35	0.6	0.8	0.9	0.9	1.0	1.0	1.0
	방법 9 1	0.007	0.35	0.6	0.8	0.9	0.9	0.0	0.0	0.0
	방법 9 2	0.007	0.35	0.6	0.8	0.9	0.9	0.5	0.5	0.5
	방법 9 3	0.007	0.35	0.6	0.8	0.9	0.9	1.5	1.5	1.5

[표 4]는 9가지의 방법으로 주요도 값을 부여한 것을 나타낸다.

7.3.2 실험결과 및 분석

앞 절에서 고안한 9가지 방법으로 IBM PC 펜티엄 III-450 MHz상에서 실험을 수행하였다. 유사도가 매우 낮은 문서인 경우에는 검색에서 제외하는 것이 바람직하다. 이 실험에서는 이러한 *임계값(threshold)*을 가장 높은 유사도의 15%이상으로 제한한다. [표 5]는 몇 개의 질의문에 대해 계산한 정확률(P)와 재현률(R)을 나타낸다. 그 질의문들은 다음과 같다.

표 5 질의문에 대한 정확률과 재현률의 예

질의 no. / 유형별 주요도		1		21		37		42	
		P	R	P	R	P	R	P	R
방법 1		0.36	1.00	0.30	0.43	0.54	1.00	0.30	1.00
방법 2		0.36	1.00	0.43	1.00	0.20	0.67	1.00	1.00
방법 3		0.36	1.00	0.43	1.00	0.20	0.67	1.00	1.00
방법 4		0.36	1.00	0.43	1.00	0.20	0.67	1.00	1.00
방법 5		0.36	1.00	0.43	1.00	0.20	0.67	1.00	1.00
방법 6		1.00	0.60	1.00	0.33	0.20	0.67	1.00	1.00
방법 7		1.00	1.00	1.00	0.33	0.20	0.67	1.00	1.00
방법 8		1.00	0.40	1.00	0.33	0.20	0.67	1.00	1.00
방법 9	1	1.00	0.40	1.00	0.33	0.20	0.67	1.00	1.00
	2	1.00	0.40	1.00	0.33	0.20	0.67	1.00	1.00
	3	1.00	0.40	1.00	0.33	0.20	0.67	1.00	1.00

- 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?

표 6 각 방법에 따른 평균 정확률과 평균 재현률

유형별 주요도		평균 정확률	평균 재현률	특징
방법 1		0.4572	0.8486	단순명사로 이루어진 키워드 기반 정보 검색과 유사하다.
방법 2		0.4756	0.8517	명사구로 이루어진 키워드 기반 정보 검색과 유사하다.
방법 3		0.4756	0.8517	빈도수가 높은 순서대로 주요도를 부여한다.
방법 4		0.4830	0.8497	빈도수가 낮은 순서대로 주요도를 부여한다.
방법 5		0.4830	0.8497	방법 4와 유사하나 유형별 값의 차이를 조금 크게 한다.
방법 6		0.5394	0.7697	방법 5와 유사하나 빈도수에 따라 5개의 소그룹으로 분류하고 빈도수가 가장 높은 그룹에 0.01을 부여하고 그룹간의 차이를 크게 한다.
방법 7		0.5487	0.7399	방법 6에서 빈도수가 가장 높은 두 그룹을 0.01->0.005, 0.2->0.3으로 변경한다.
방법 8		0.5542	0.7653	방법 7을 세분화하여 주요도를 부여하면서 빈도수가 가장 높은 두 그룹을 0.005->0.007로 변경한다.
방법 9	1	0.5542	0.7653	방법 8에서 빈도수가 가장 낮은 그룹의 영향을 알아보기 위해 이 이들의 주요도를 1.0에서 0.0, 0.5, 1.5로 각각 변경한다.
	2	0.5542	0.7653	
	3	0.5542	0.7653	

[표 6]은 [표 4]에 있는 주요도 값들을 적용하여 실험을 수행하여 나온 결과를 요약한 것으로 45개의 질의문들의 정확률의 평균과 재현률의 평균을 나타낸다.

앞서 언급하였듯이 이 실험에서는 질의와의 유사도가 매우 낮은 문서를 제외하기 위해 그 임계값을 가장 높은 유사도의 15%이상으로 제한하였다. 임계값을 무시하고 유사도가 0보다 큰 문서들을 모두 검색하였을 때의 평균 정확률과 평균 재현률은 각각 0.3528과 0.8919이다.

각 방법에 따른 주요도 값의 차이와 평균 정확률과 평균 재현률의 증감을 토대로 하여 결과 분석을 하고자 한다.

- 관찰 1)

방법 1은 단순명사로 이루어진 키워드 기반 정보 검색과 유사한 것으로 평균 정확률은 46%이고 평균 재현

률 85%이다. 방법 2는 명사구를 인덱스로 하는 키워드 기반 정보 검색과 유사한 것으로 평균 정확률은 48%이고 평균 재현률 85%이다. 이 두 방법을 비교하여 볼 때 단순 명사로 검색을 하는 것보다 단순 명사를 포함하는 명사구로 검색을 하는 것이 정확률을 향상할 수 있음을 알 수 있다.

● 관찰 2)

방법 2와 방법 3은 그 결과에 차이가 없다. 방법 2에서는 유형 1, 2, 3의 주요도 값은 각각 1.0이고 유형 4부터 유형 9까지는 주요도 값은 모두 0.0이다. 방법 3에서는 유형 1에서 유형 9까지의 주요도 값이 각각 1.0, 0.9, 0.8, …, 0.2이다. 이 두 방법에서 유형 1, 2, 3의 주요도 값은 크게 차이가 없으나, 유형 4부터 유형 9까지는 매우 다르다고 볼 수 있다. 그럼에도 불구하고 평균 정확률과 평균 재현률에 차이가 없는 것은 유형 4부터 유형 9까지의 빈도수의 비율이 약 10%이므로 상대적으로 빈도수의 비율이 큰 유형 1, 2, 3의 주요도 값에 그 결과가 좌우되었다고 짐작할 수 있다.

● 관찰 3)

방법 4와 방법 5는 공통적으로 빈도수가 낮은 순서대로 주요도를 부여하고 방법 5에서는 유형별 값의 차이를 방법 4보다 조금 크게 한 것으로, 그 결과에 차이가 없다. 그러나 방법 2와 방법 3에 비해서는 평균 정확률이 조금 증가하고 평균 재현률은 조금 감소하였다.

● 관찰 4)

방법 6에서는 유형별 순서대로 유형 1에서 유형 9까지 0.01, 0.2, 0.5, 0.9, 0.9, 0.9, 1.0, 1.0, 1.0의 값을 각각 부여하였다. 방법 6의 특징이라면 지금까지의 방법보다는 빈도수가 높은 그룹에 매우 작은 주요도 값을 부여한 것이라고 할 수 있다. 특히 빈도수가 가장 높은 유형 1에 주요도 값 0.01을 부여한 점이 방법 5로부터 크게 다른 점이라 볼 수 있다. 방법 6의 결과는 평균 정확률이 크게 증가하고 평균 재현률 크게 감소하였다. 이 연구에서는 정확률을 향상시키는 데 주안점을 두므로 재현률이 다소 감소하더라도 정확률을 높이는 방향을 모색한다.

● 관찰 5)

방법 7은 방법 6에서 유형 1과 유형 2의 값을 각각 0.01에서 0.005로, 0.2에서 0.3로 각각 변경하였다. 그런데 방법 6에 비해 평균 정확률은 조금 증가하였고 평균 재현률은 많이 감소하였다.

● 관찰 6)

방법 8은 방법 7에서 유형 1과 유형 2의 값을 각각 0.005에서 0.007로, 0.3에서 0.35로 각각 변경한 것이 큰 차이점이다. 그런데 방법 7에 비해 평균 정확률과 평균 재현률이 동시에 다소 증가하여 평균 정확률이 향상되었으며 평균 재현률은 방법 6과 거의 같아졌다.

● 관찰 7)

방법 9는 빈도수가 가장 낮은 그룹의 영향을 알아보기 위해 방법 8에서부터 이들의 주요도를 1.0에서 0.0, 0.5, 1.5로 각각 변경한 방법이다. 실험의 결과로 이들의 값의 변경이 평균 정확률과 평균 재현률에 영향을 주지 않음을 보여준다.

방법 1과 방법 2는 키워드 기반 정보 검색과 유사하며 그 외의 방법들은 키팩트 기반 정보 검색에서 고유하게 수행할 수 있는 방법들이라 볼 수 있다. 방법 1과 방법 2에 의한 정확률보다는 키팩트 기반 정보 검색이라 볼 수 있는 다른 방법들에서 정확률이 더 높음을 알 수 있고, 특히 방법 7과 방법8은 평균 정확률이 높은 방법들이라 할 수 있다. [그림 2]는 임계치를 무시한 경우에 45개의 질의문들의 정확률과 재현률의

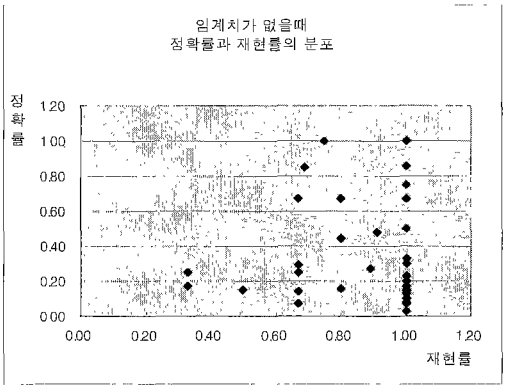

그림 2 임계치가 없을 때의 정확률과 재현률의 분포

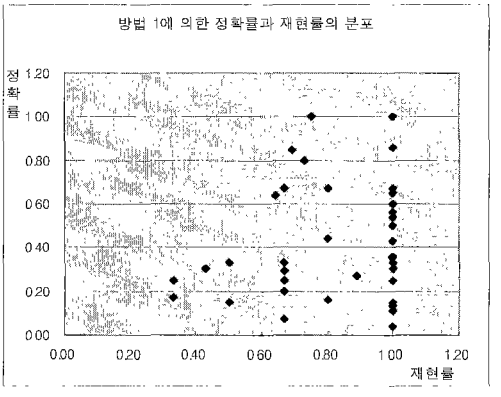

그림 3 방법 1에 의한 정확률과 재현률의 분포

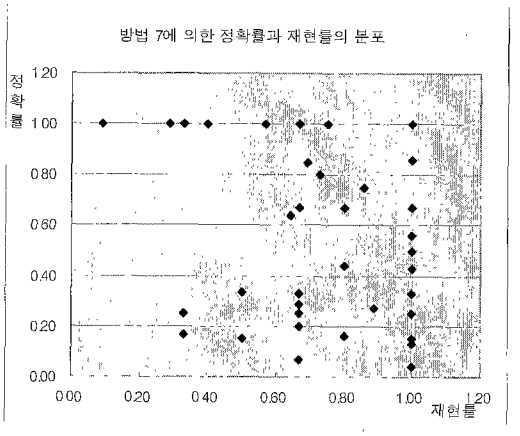

그림 4 방법 7에 의한 정확률과 재현률의 분포

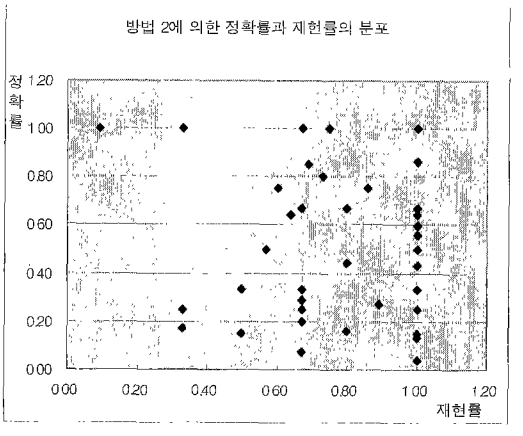

그림 5 방법 2에 의한 정확률과 재현률의 분포

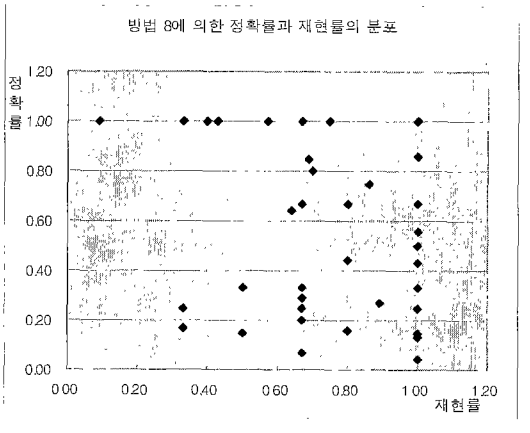

그림 6 방법 8에 의한 정확률과 재현률의 분포

분포를 나타내며, [그림 3], [그림 4], [그림 5] 그리고 [그림 6]은 방법 1, 방법 2, 방법 7과 방법 8에 대한 정확률과 재현률의 분포를 각각 나타낸다.

지금까지 9가지의 방법으로 키팩트의 각 유형에 주요도를 부여하여 실험을 수행하고 그 결과의 변화를 요약하여 보았다. 키워드 형태의 유형들의 빈도수가 높음으로 해서 이들 유형에만 주요도를 부여하는 방법 1과 방법 2로 출발하여 빈도수의 비율이 높은 유형에 상대적으로 낮은 주요도를 점차적으로 부여함으로써 정확률이 점점 향상됨을 알 수 있다.

인접한 다른 방법간에 평균 정확률과 평균 재현률의 미세한 증감으로 어느 방법이 더 우수하다고 단정할 수는 없지만, 단계적인 실험을 통하여 빈도수의 비율이 매우 높은 유형에 매우 작은 주요도 값을 부여하고 빈도수의 비율이 중간 위치에 있는 유형들의 주요도 값을 적절히 변경함으로써 평균 재현률의 감소를 되도록 작게 하면서 평균 정확률을 향상하는 방법을 개발하였다. 수행한 방법들 중 방법 8이 평균 정확률을 높이는 적절한 방법이라고 할 수 있다. 이 방법의 평균 정확률은 임계값이 없을때의 평균 정확률에 비해 약 57% 증가하였으며 평균 재현률은 약 14% 감소하였다. 즉 약 14%의 평균 재현률의 희생으로 평균 정확률을 약 57% 향상하였다고 볼 수 있다.

사용한 질의문들의 수는 충분하다고 판단이 되지만 그들이 모두 짧은 단문이어서 질의문으로부터 추출된 키팩트들의 수가 적고 다양한 형태가 되지 않았다. 이 점이 주요도의 값들의 미세한 변경에 평균 정확률과 평균 재현률이 별로 영향을 받지 않은 요인이 된 것 같다. 평균 정확률을 향상하기 위해 여러 가지 방법을 시도해 보는 것은 사실 실험자가 자발적으로 열심히 하고자 하는 의지에 달려 있다.

8. 결 론

이 논문에서는 키팩트를 이루는 [중심어, 종속어]의 유형을 9가지로 분류하고, 9가지의 유형별 주요도 값을 부여하는 방법을 고안하였다. 그리고 빈도수에 근거를 둔 가중치에 유형별 주요도를 반영하여 키팩트 가중치를 계산하는 모델을 개발하였다.

키팩트의 각 유형에 다양한 방법으로 유형별 주요도를 부여하고, 키팩트 가중치 s와 내적을 이용하여 질의문과 검색대상이 되는 문서들과의 유사도를 계산하고, 정확률과 재현률을 계산하는 실험을 수행하기 위해 필요한 여러 단계의 모듈들을 C++로 개발하였다.

전자통신연구원에서 개발한 초기버전의 키팩트 추출

기를 이용하여, 계몽사 대백과 사전의 일부를 대상으로 전자통신연구원에서 제공한 45개의 질의문을 사용하여 실험을 수행하였다.

실험 결과를 살펴보면 종속어가 없는 유형에만 주요도 값을 부여하고 그 외의 유형에는 주요도 값을 부여하지 않은 방법에서 가장 높은 재현률 85%를 보였다.

종속어가 있고 빈도수의 비율이 중하위그룹에 속하는 유형에 매우 높은 주요도 값을 부여하고, 빈도수의 비율이 매우 높은 유형에 매우 낮은 주요도 값을 부여한 방법들에서 높은 평균 정확률 55%를 보였고 평균 재현률 역시 74% ~ 77% 범위에 분포하고 있다. 비교적 만족스러운 정확률과 재현률의 범위가 약 50% ~ 60%라고 볼 때, 이 결과는 매우 고무적이다[2].

이 실험을 통하여 단순명사에 의한 검색 방법보다는 단순명사를 포함하는 명사구에 의한 검색 방법이 조금 더 향상됨을 볼 수 있었다. 그리고 키워드 기반 검색보다는 키팩트 기반 검색이 전체적으로 우수함을 관찰할 수 있었는 데, 이 사실은 기존의 키워드 기반 정보검색에서 문제시되는 정확률을 키워드 기반 정보검색에서 개선할 수 있는 가능성을 시사하고 있다.

이 연구를 바탕으로 다양한 크기의 질의문들과 대규모의 코퍼스를 이용하여 정확률과 재현률의 향상을 위한 실험을 수행하여 보다 이상적인 유형별 주요도의 부여와 키팩트 가중치 모델의 개발이 필요하다.

참 고 문 헌

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

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

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

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

[5] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어형태소 분석, 서울대학교 공학박사 학위논문, 1993.

[6] Yasushi OGAWA, Ayako 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 an Development in Information Retrieval, 1993.

[7] 김판구, 조유근, "상호 정보에 기반한 한국어 텍스트의 복합어 자동색인", 한국정보과학회논문지, 제21권, 제7호, 1994.

[8] 윤준태, 송만석, "한국어의 대등접속구문 분석", 정보과학회논문지(B), 제24권, 제3호, 1997.

[9] 이현아, 이종혁, 이근배, "단문 분할을 통한 명사구 색인 방법", 정보과학회논문지(B), 제24권, 제3호, 1997.

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

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

[12] 이경호, 파일처리론, 정익사, 1997.

[13] 김수희, 박세영, "대규모의 정보검색을 위한 효율적인 최소 완전 해시함수의 생성", 한국정보처리학회 논문지, 제5권, 제9호, 1998.

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

김 수 희

1979년 부산대학교 과학교육학과 졸업(이학사). 1986년 University of Georgia Dept. of Computer Science (MAMS). 1988년 University of Georgia Dept. of Mathematics (MA). 1993년 University of South Carolina Dept. of Computer Science (Ph.D) 1993년 Benedict College Assistant Professor. 1994년 ~ 현재 호서대학교 컴퓨터공학부 부교수. 관심분야는 정보검색, 데이터베이스, 하이퍼미디어, 정보보안.

남 효 돈

1991년 3월 ~ 1995년 2월 호서대학교 전자계산학과 재학 (이학사). 1995년 3월 ~ 1996년 2월 아주반도체 전산실 근무. 1996년 3월 ~ 1999년 8월 호서대학교 대학원 전자계산학과(이학석사). 2000년 1월 ~ 현재 한국 보쉬 기전 전산실. 관심분야는 데이터베이스, 정보검색.