논문 99-24-11A-13 한국통신학회논문지 '99-11 Vol.24 No.11

한국어 정보 검색에서 엔트로피와 사용자 프로파일을 이용한 질의 확장

정회원 최 준 혁*, 허 준 희**, 이 정 현**

A Query Expansion using User Profile and Entropy in Korean IR

Jun-Hyeog Choi*, Jun-Hui Her**, Jung-Hyun Lee** *Regular Members*

요 약

일반적으로 정보 검색 시스템을 사용하는 이용자들의 관심은 재현률보다는 높은 정확도를 요구하며, 그 중에서도 자신이 검색하고자 하는 문서가 상위에 순서화되길 희망한다. 이러한 사용자의 요구를 만족시키기 위하여 본 논문에서는 엔트로피와 사용자 프로파일을 이용한 정보 검색 시스템을 설계하고 구현하였다. 이때, 사용자 프로파일은 가장 최근 검색에 사용된 키워드들로 구성되었으며 단어의 수는 복잡도를 고려하여 최대 세 단어로 제한하였다. 엔트로피 계산을 위해서 각 확률 정보 값에 상호 정보량을 적용하기 위한 새로운 수식을 유도하였다. 문서 순위를 조정하기 위한 엔트로피 계산은 질의어와 사용자 프로파일로 구성된 질의어 벡터와 질의어에 의해 동적으로 검색된 문서의 색인어 벡터와의 계산 값을 이용하였다. 또한 정확도를 더욱 높이기 위한 방법으로 베이지안 학습에 의해 산출된 가중치 값을 사용자 프로파일에 반영하였다.

ABSTRACT

Generally, IR system users require high precision ratio rather than recall ratio. Furthermore, they hope that the documents to be retrieved will be ranked by priority of there preference. We designed and implemented entropy and user profile based on Korean IR system to satisfy these requirements of users. In our system, user profile consists of keywords retrieved recently and the frequency of keywords should be limited three considering complexity. To calculate entropy, a new formula is derived in order to apply mutual information to each probability information. The entropy calculation for document ranking is achieved with two vector values. One is the query vector which consists of a query and user profile and the other is the index terms vector of the document which is retrieved dynamically by query. To improve the precision, we reflected the weight value derived from Bayesian learning to user profile.

I. 서 론

최근 컴퓨터의 높은 보급률과 인터넷의 발전으로 방대한 양의 정보가 인터넷상에 웹 문서의 형태로 분산되어 있으며 정보 또한 동적으로 변화되는 특성을 갖고 있다. 이렇게 산재해있는 정보들 중에서 자신이 원하는 자료를 찾기는 쉽지 않다. 이러한 이유로 정보검색 시스템이 도입되었다.

정보검색 시스템이란 시스템의 이용자가 필요로

* 김포대학 컴퓨터 계열 (jhlee@dragon.inha.ac.kr)
** 인하대학교 전자계산공학과 (jjun2@nlsun.inha.ac.kr)
논문번호 : 99203-0519, 접수일자 : 1999년 5월 19일
※ 본 연구는 인하대학교 97년도 교내 연구비 지원에 의해 수행되었음

하는 정보를 수집하여 내용을 분석한 뒤, 찾기 쉬운 형태로 조직하여 두었다가 정보에 대한 요구가 발생할 때 적합한 정보를 검색하여 제공하는 시스템이다[15].

정보 검색 시스템의 유형 중에서, 역화일을 이용한 정보 검색 방법은 질의어에 나타난 키워드들이 문서에서 어느 정도의 가중치를 가지고 존재하느냐를 기준으로 문서들을 순서화한다. 이는 단순히 해당하는 질의어에 나타난 단어들에 대한 문서 내의 존재 여부를 반영하므로 관련있는 문서를 찾아내는 능력에는 한계가 있다[12].

이러한 문제를 해결하기 위한 연구의 하나로 클러스터링을 이용하는 방법이 있는데 이는 검색 대상 문서 전체를 탐색하는 대신 정보 요구 주제와 관련된 문헌 클러스터만을 탐색함으로써 탐색 시간의 절약과 검색 효율의 향상을 기대하는 방법이다[4].

정보 검색 시스템의 중요한 목적 중의 하나는 단순히 사용자 질의를 만족하는 문서들의 집합을 검색하는 것이 아니라, 질의를 만족하는 정도에 따라 검색된 문서들에 순위를 부여함으로써 사용자들이 필요한 정보를 얻는데 소요되는 시간을 최소화하는 것이다[12].

본 논문에서는 한국어 정보 검색 시스템의 정확도 향상을 위하여 엔트로피와 사용자 프로파일을 이용한 한국어 정보 검색 시스템을 설계하고 구현하였다. 특히, 역화일 내에 사용자 질의어가 존재하는지의 여부만을 반영하는 기존의 검색 시스템과는 달리 사용자 질의어에 포함되어 있는 의미 정보를 반영하기 위하여 Shannon의 정보 이론을 이용하였다. 이때 확률 벡터 정보량의 계산은 사용자 질의어 및 사용자 프로파일과 각 문서에 포함되어 있는 색인어 사이의 상호 정보량을 기반으로 한다. 이를 바탕으로 엔트로피를 계산하며 엔트로피 값이 적은 문서를 상위에 위치시키는 문서 순위 조정 알고리즘을 구현하였다. 또한, 사용자 프로파일에 존재하는 각각의 단어에 가중치를 부여하기 위하여 Bayesian 추정치와 Sparck Jones의 역문헌 빈도를 이용하여 엔트로피 값을 계산하였다.

실험 대상으로는 KT set95 4,414개의 문서들을 사용하였으며 색인어 추출을 위한 전처리 과정으로 형태소 분석 과정을 거친다. 추출된 색인어들에 대한 가중치 계산은 Sparck Jones의 역문헌 빈도 계산 방법을 이용하였다. 이렇게 계산된 색인어들과 가중치 값은 각각의 문서에 대하여 색인어 후보와 용어 빈도, 장서 빈도, 역문헌 빈도, 용어 가중치 값으로 구성된 n-차원 벡터 공간 모델로 저장된다.

Ⅱ. SMART 시스템에서의 문서 순위화 방법

문서의 순위화를 위한 기존의 방법들 중에는 적합성 피드백을 이용한 방법[1], 데이타 퓨전을 이용한 방법, 불리안 모델에 기반한 방법 등을 들 수 있다. 이러한 방법에서는 사용자 질의어에 나타난 의미보다는 단순히 해당되는 문서내에 검색하고자 하는 색인어가 존재하는지의 여부만을 반영하기 때문에 최종 검색 결과를 살펴보면 사용자의 의도와는 관련없는 문서들이 상위에 순서화되는 것을 볼 수 있다[12].

Salton에 의해 설계된 SMART 시스템은 역화일 구조를 이용하는 대표적 시스템으로 문서의 내용을 식별할 수 있는 식별자를 생성하기 위하여 완전 자동 인덱싱 방법을 사용한다[5, 6, 7].

SMART 시스템은 문서와 질의를 각 문서와 질의가 담고 있는 용어들의 벡터로 표현한다. 문서와 문서 사이, 그리고 문서와 질의 사이의 유사도는 다음의 코사인 상관 계수를 통해 측정한다.

$$Cosine(Doc_i, Query_j) = \frac{\sum_{k=1}^{t}(Term_{ik} \cdot Qterm_{jk})}{\sqrt{\sum_{k=1}^{t}(Term_{ik})^2 \cdot \sum_{k=1}^{t}(Qterm_{jk})^2}}$$

위 식은 문서와 질의 사이의 유사도를 계산하기 위한 것으로, 이 계산식을 통해 산출된 값이 높은 순으로 검색한 문서의 순위를 조정한다.

Ⅲ. 엔트로피와 사용자 프로파일

1. Shannon의 정보 이론과 공기 정보의 결합

문서와 사용자 질의와의 유사도 계산을 위하여 본 논문에서는 정보이론에 기반한 엔트로피를 이용하여 유사도를 계산하였다. 이때 사용하는 확률벡터 정보량은 Shannon의 정보 이론에 근거하여 불확실성의 크기를 엔트로피로 측정하였는데, 이것은 각 사용자 질의어가 갖는 평균 정보량이 된다[9, 14, 16].

불확실성에 대한 양을 나타내는 엔트로피 H는 다음과 같이 정의할 수 있다.

$$H = -\sum_{i=1}^{n} p_i \log_2 p_i$$

여기서, 메시지 i가 갖는 정보량 $\log_2 p_i$는 메시지가 선택될 확률 p_i에 의해 결정된다. 따라서, H는 n개의 메시지가 갖는 평균 정보량이며, p_i는 i번째 메시지가 선택될 확률을 나타낸다. 일반적으로 H는 다음과 같은 특성을 갖는다.

(*i*) $0 \le H(p) \le \log_2 N$

(*ii*) $H(p) = \log_2 N$, if $p_1 = p_2 = \ldots = p_n = 1/N$

(*iii*) $H(p) = 0$, if $p_i = 1$, $p_j = 0$ $(1 \le j \le N, j \ne i)$

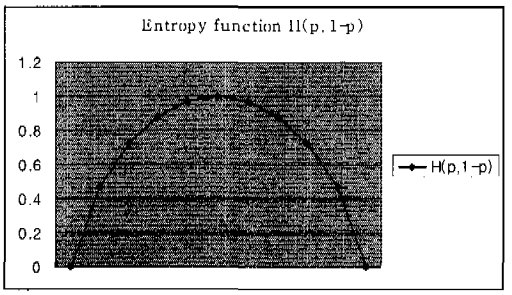


그림 1. H(p, 1-p)의 엔트로피 함수

그림 1은 아래의 식과 같이 정의되는 엔트로피 함수, $H(p_i) = p\log p + (1-p)\log(1-p)$를 설명하고 있다. 여기서, p_i가 균일한 분포일 때, 즉 $p = 1/2$일 때 $H(p_i)$는 최대 값을 갖는다.

2. 상호 정보량을 이용한 엔트로피 계산

본 논문에서는 사용자 질의어에 의해 검색된 문서들에 대한 문서 순위 조정을 위하여 사용자 질의어-프로파일 벡터와 검색된 문서들에 포함되어 있는 주제어들에 대하여 엔트로피를 계산한다. 엔트로피 계산 식에 포함되어 있는 확률 값 p_i를 구하기 위하여 $p_i(d_i \mid q)$ 즉, 사용자 질의에 포함되어 있는 질의어 q가 i번째 문서에서 출현할 확률을 계산하기 위하여 상호 정보량을 이용하였다.

상호 정보량(Mutual Information)이란 확률 변수 x와 y사이의 의존 관계를 정량적으로 나타낸 것으로 x와 y대신에 조사하고자 하는 단어들을 대입시키면 단어와 단어 사이의 의존 관계를 정량적으로 나타낼 수 있다.

문서에서 확률 변수 사이의 관계는 단어와 단어 사이의 관계에 대한 값을 의미하므로 단어의 발생 빈도에 대한 평균을 구하는 것은 의미가 없으므로 다음 관계식을 상호 정보량을 구하는데 이용한다[13].

$$MI(x,y) = \log_2 \frac{p(x,y)}{p(x) \cdot p(y)}$$

윗 식은 다음과 같이 재정의 할 수 있다.

$MI(x,y) = \log_2 p(y \mid x) \cdot p(y)$, 따라서

$MI(x,y) = \log_2 p(y \mid x) + \log_2 p(y)$가 된다.

여기서 $p(y)$는 질의 x를 포함하고 있는 문서에서의 색인어 집합을 의미하므로 항상 1로 간주할 수 있다. 즉, $\log_2 p(y) = \log_2 1$이므로 문서 y에서 x가 출현할 확률은 x와 y의 상호 정보량에 비례함을 알 수 있다.

$$p(y \mid x) \propto MI(x,y) \qquad \text{(식 1)}$$

(식 1)에서 엔트로피 계산을 위한 확률 벡터 정보값은 단어와 단어 사이의 상호 정보량으로 대치하여 사용하였다. 엔트로피 계산에 사용되는 확률정보 값 p_i는 다음과 같은 조건을 만족해야 하기 때문에 정규화 과정을 거치게 된다.

$p_i = P(d_i \mid q)$,

for $i = 1, 2, \ldots, N$, $p_i \ge 0$, $\sum_{i=1}^{N} p_i = 1$

3. 사용자 프로파일

사용자 프로파일(user profile)은 사용자 개인의 관심분야를 기술해 놓은 것을 말한다. 정보검색 시스템을 사용하는 사용자 개개인의 기호나 관심분야를 미리 알고 있다면, 사용자의 관심분야에 속하는 문서만을 해당 사용자에게 우선적으로 검색해 줌으로써 사용자에게 편의를 제공할 수 있다[10].

정보검색 시스템에서 사용자 프로파일은 시스템의 핵심적인 구성요소는 아니다. 사용자 프로파일에 의하지 않고서도 대부분의 시스템은 사용자들이 원하는 검색 결과를 제시해줄 수 있기 때문이다. 하지만 많은 검색 결과가 사용자들이 원하지 않는 결과이기 때문에 사용자는 자신이 원하는 결과를 얻기

위하여 많은 수의 문서들을 일일이 확인하거나 다시 질의를 수정해야 하는 번거로움이 있다. 이러한 문제점은 사용자 프로파일을 도입함으로서 완전하지는 않지만 부분적으로 해결할 수 있다.

정보검색 시스템에서의 사용자 프로파일에 대한 연구로는 베이지안 확률을 이용하는 방법과 NN방법, 결정 트리, Rocchio의 알고리즘, 신경망을 이용하는 방법 등이 있다.

그림 2는 본 논문에서 사용자의 최근 검색 행위를 반영하여 정확도를 향상시키기 위하여 설계된 사용자 프로파일 구조이다.

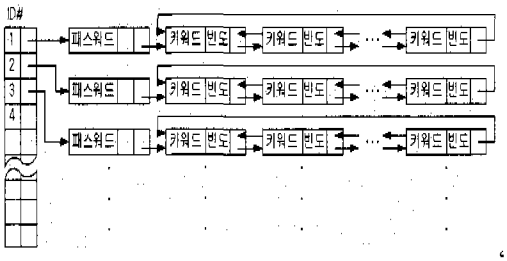


그림 2. 각 키워드의 사용 빈도 정보가 반영된 사용자 프로파일

그림 2에서 사용자 프로파일은 사용자가 최근에 검색에 사용되었던 질의어들 및 사용 빈도의 누적 값으로 구성되어 있으며, 이는 사용 빈도에 따라 순위화되어 있다. 이렇게 사용 빈도에 따라 순위화되어 있는 사용자 프로파일 중에서 최상위 빈도에 대한 세 단어만이 사용자 질의어와 더불어 엔트로피 계산에 이용되며 이렇게 계산된 값은 가중치로 이용된다.

초기 사용자 프로파일이 문서 순위화에 기여하는 역할은 미미하다. 그러나 시간이 경과할수록 사용자의 관심 분야를 반영하고 있는 사용자 프로파일은 사용자 질의어와 더불어 엔트로피 계산에 대한 가중치로서 작용하기 때문에 사용자의 기호도를 반영하여 정확도를 높이는데 많은 영향을 미치게 된다.

Ⅳ. 제안하는 시스템 구조

그림 3은 본 논문에서 사용되는 전체 시스템 구조를 나타낸다. 전체 시스템은 크게 색인어를 추출하기 위한 색인어 구축 모듈과 각 단어들의 상호 정보량을 구축하는 공기 정보 구축 모듈, 그리고 엔트로피 계산 모듈로 분류할 수 있다.

1. 색인어 구축 모듈

먼저, 색인어 후보를 추출하기 위하여 형태소 분석을 수행한다.

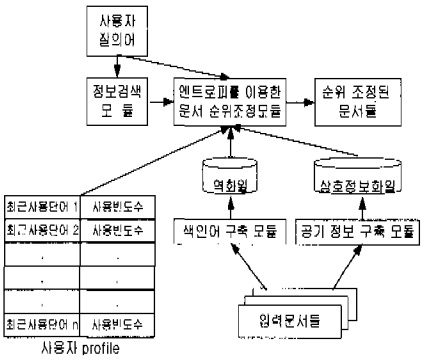


그림 3. 제안하는 시스템 구조

그림 4는 색인어 후보를 추출하기 위하여 본 연구실에서 개발한 형태소 분석기를 적용하기 위한 대상으로 KT set95의 한 문서를 나타낸다.

<id>2860
<title>금성사, 다기능 고화질 팩시밀리 개발
<abstract>
금성사는 화상 인식 기능을 채용해 화상을 문자와 그림으로 구별 인식해 주고 그림의 명암을 32단계로 미세하게 구별해 원본을 더욱 선명하게 재현해 주는 다기능 고화질 팩시밀리를 개발, 24일부터 시판한다고 발표했다. 이 제품은 그림을 강조하면 글자가 흐려지고 글자를 강조하면 그림의 명암을 구별할 수 없었던 기존 제품의 단점을 개선한 것이며 국내 처음으로 기록지 말림 방지시스템을 채용해 기존 감열기록 방식 팩시밀리의 불편함을 없앴다고 금성사는 밝혔다.

그림 4. 형태소 분석 대상 문서 예

그림 5는 그림 4에 대해 형태소 분석기를 적용한 형태소 분석 결과를 나타낸다.

정보검색 시스템에 사용하기 위한 형태소 분석은 명사구에 대해서만 분석을 수행하고, 이에 대한 분석 결과로부터 색인어 후보를 추출한다. 형태소 분석이 끝난 문서들은 Sparck Jones가 제시한 역문헌 빈도를 계산하게 되는데 역문헌 빈도에 의한 가중치 계산은 다음과 같다[6].

```
[금성사 ((금 성사) (N N)) ((금성 사) (N N))]
[다기능 ((다기 능) (N N))]
[고화질 ((고 호 아지 ㅍ) (N N N PCO))]
[팩시밀리 ((팩시밀리) (N))]
[개발 ((개발) (N))]
[금성사는 ((금 성사 는) (N N PS)) ((금성 사 는) (N N PS))]
[화상 ((화상) (N))]
[인식 ((인식) (N))]
                    .
                    .
                    .
[팩시밀리의 ((팩시밀리 의) (N PCD))]
[불편함을 ((불편 하ㅁ 을) (NH AS EN PCO))]
[없앴다고 ((없애 ㅆ 다고) (V EPF EC))]
[금성 ((금성) (N))]
[사는 ((사 는) (N PS)) ((사 는) (NU PS)) ((사 는) (V ED)) ((살 는) (V ED))]
[밝혔다 ]
```

그림 5. 문서(그림 4)에 대한 형태소 분석 결과

$$W_{ik} = \log_2 \frac{n}{DF} + 1 \qquad \text{(식 2)}$$

(식 2)의 역문헌 빈도는 문헌 빈도가 낮은 단어, 즉 적은 수의 문헌에 나타난 단어에 대해 높은 중요도를 부여하는 결과를 가져온다.

Salton은 Sparck Jones의 역문헌 빈도 공식에 문헌 내의 단어빈도 f_{ik}를 곱한 값을 가중치로 줌으로써 문헌빈도가 낮고 단어빈도가 높을수록 역문헌 빈도도 높아지도록 하였다. (식 3)은 Salton의 역문헌 빈도에 의한 가중치 계산을 나타낸다[6, 18].

$$W_{ik} = f_{ik} \cdot [\log(n) - \log_2(DF) + 1] \qquad \text{(식 3)}$$

아래 표 1은 그림 5의 형태소 분석 결과에 따른 색인어 후보들에 대하여 Sparck Jones의 역문헌 빈도와 Salton의 가중치 값을 구한 결과이다.

표 1의 방법으로 모든 문서에서 추출된 색인어들 각각에 대해서는 그것이 포함된 문서들의 리스트, 즉 역 리스트(inverted list)를 구성하였다.

그림 6은 본 연구를 위해 설계한 역화일 자료구조로서 색인어의 첫 음절에 의한 동적 트라이(dynamic trie)로 구성하였다. 이때 색인어는 문서내의 위치와 빈도 정보 외에 추가적인 정보도 지속적으로 포함할 수 있도록 설계되었다.

표 1. 역문헌 빈도에 의한 가중치 계산

문서번호 [2860]				
색인어 후보	용어 빈도	창서 빈도	역문헌빈도	가중치
화상	2	159	4.754888	9.509775
인식	2	308	3.807355	7.614710
문자	2	180	4.584963	9.169925
그림	4	51	6.426265	25.705059
명암	2	11	8.647458	17.294917
용해	2	206	4.392317	8.784635
팩시밀리	1	89	5.614710	5.614710
글자	3	45	6.614710	19.844130
구별	3	18	7.936638	23.809914
기록	3	211	4.321928	12.965784
금성사	2	311	3.807355	7.614710

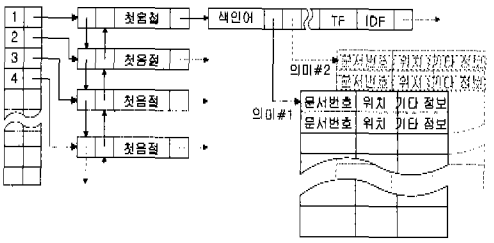


그림 6. 역화일 자료 구조

2. 상호 정보 구축 모듈

상호 정보 수식은 문서에서 단어와 단어 사이의 연관성을 정량적으로 나타내기 위하여 사용되어 왔다. 예를 들어 두 단어 x와 y에 대하여 그들의 확률이 각각 $p(x)$와 $p(y)$로 주어진다면 x와 y간의 상호 정보는 (식 4)와 같다[13].

$$\begin{aligned} MI(x,y) &= \log \frac{p(x,y)}{p(x) \cdot p(y)} \\ &\approx \log \frac{f(x,y)/N}{f(x)/N \cdot f(y)/N} \\ &= \log \frac{N \cdot f(x,y)}{f(x) \cdot f(y)} \end{aligned} \qquad \text{(식 4)}$$

(식 4)에서 $p(x,y)$는 단어 x와 y가 말뭉치 내에서 공기한 확률이다. 이 확률을 말뭉치에서 얻을 수 있는 빈도로 근사시키면 $f(x,y)/N$, 즉 x와 y의 공

기 빈도를 말뭉치의 크기 N으로 나누어 준 값이 될 수 있다. 여기서 N은 말뭉치의 전체 단어 수를 의미한다.

본 연구에서의 말뭉치는 KT set95를 이용하였으며 명사들의 상호 정보량은 본 연구실에서 개발한 공기 정보 구축 모듈에 의해서 구축하였다. 그림 7은 상호 정보량을 구축하는 모듈의 예를 나타내며, 상호 정보 구축 시에 사용하는 윈도 사이즈는 5로 제한하였다.

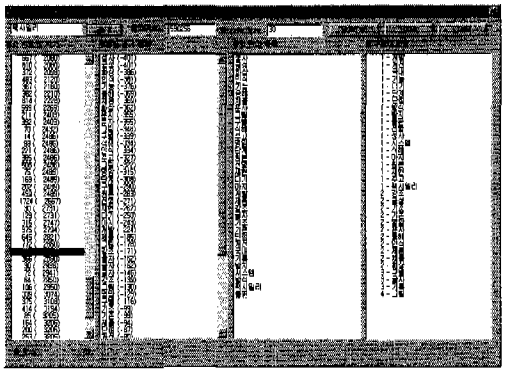


그림 7. 상호 정보량 계산 모듈

3. 엔트로피 계산 및 베이지안 추정치

사용자 질의어에 포함되는 단어들에 대한 문맥 정보를 반영하기 위하여 본 연구에서는 사용자 질의어와 사용자 프로파일 벡터로 구성되는 입력 벡터와 동적으로 검색된 문서들의 색인어들 사이의 엔트로피를 계산한다.

그림 8은 입력 벡터와 동적으로 검색된 문서들의 색인어들 사이의 엔트로피를 계산하기 위한 예를 나타낸다.

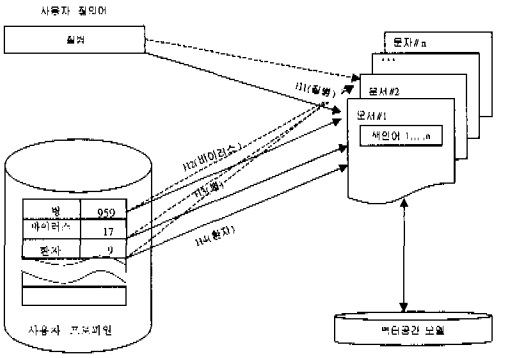


그림 8. 사용자 질의어 및 사용자 프로파일과 각 문서의 색인어들과의 엔트로피 계산

예를 들어, 사용자 질의어가 '질병'이라 하고 [문서 #1]의 주제어들이 [자궁암, 임상, 연구, 환자]라고 한다면 이때의 엔트로피 계산은 (식 5)에 의해 계산할 수 있다.

$$H_1(\text{질병}) = -\sum_{i=1}^{4} p_i \log_2 p_i \qquad \text{(식 5)}$$

(식 5)에서 n의 값 4는 [문서 #1]에 색인어들이 4개 존재함을 의미하며, 각각의 p_i는 사용자 질의어 '질병'과 [문서#1]의 대표 색인어들인 [자궁암, 임상, 연구, 환자]의 상호 정보량에 의해 구할 수 있다.

동일한 작업 과정이 H_2(바이러스), H_3(병), H_4(환자)에 대해서도 수행된다.

질의어 프로파일 벡터와 [문서 #1]의 최종 엔트로피 계산은 (식 6)에 의해서 구할 수 있다.

$$H_{d1} = H_1(\text{질병}) + H_2(\text{바이러스}) + H_3(\text{병}) + H_4(\text{환자}) \qquad \text{(식 6)}$$

본 연구에서는 사용자 질의어-프로파일 벡터와 동적 검색된 문서들의 색인어들 사이에 엔트로피를 계산하는데 있어 역문헌 빈도에 의한 가중치와 베이지안 학습에 의한 추정치를 이용하였다. 이때 사용자 질의어에는 1*역문헌 빈도 값을, 사용자 프로파일의 최상위 빈도 세 단어에는 단어 빈도를 프로파일 단어 전체 빈도의 합으로 나누어 역문헌 빈도 값을 곱한 값을 가중치로 이용하였다. 즉, 본 논문에서 사용하는 사용자 질의어와 사용자 프로파일에 대한 가중치 결정은 베이지안의 사후 확률(posterior probability)을 이용한다. 이때 사용하는 베이지안의 사후 확률은 (식 7)과 같이 정의한다.

$$p(t \mid d_1, \cdots, d_n) \propto p(t) L(d_1, \cdots, d_n \mid t) \qquad \text{(식 7)}$$

여기서 t는 텍스트를 $d_1, \cdots, d_n$은 문서를 나타낸다. 즉, 베이지안의 사후 확률은 문서들에서의 텍스트의 우도함수(likelyhood function)와 사전확률(prior probability)의 곱으로 나타낼 수 있다.

사전확률 $p(t)$에 대한 정보가 없을 때는 일반적으로 1을 사용하지만 본 논문에서는 텍스트의 역문헌 빈도를 사용하여 사전(prior)에 대한 정보를 높혔다. 그리고 우도함수 값은 말뭉치에 나타난 각 단어

의 빈도를 이용하였다.

이렇게 사전(prior)에 대한 구체적 정보인 역문헌 빈도와 말뭉치에서의 각 단어의 빈도를 사용하여 구한 사후 정보를 사용자 질의와 사용자 프로파일 간의 각 단어들에 대한 가중치로 이용한다.

(식 6)에 대한 베이지안 추정치는 사용자 프로파일에 의해서 구할 수 있다. 예를 들어 과거 사용자 프로파일에 [병]이라는 주제어가 959회, [환자]가 9회, [바이러스]가 17회 사용되었다면 베이지안 추정치의 초기 값은 각각 959/1085, 9/1085, 17/1085이 되며, 이 값을 다시 역문헌 빈도의 가중치 값으로 곱하여 최종 가중치를 구할 수 있다. 따라서 (식 6)은 (식 8)로 재계산된다.

$$H_{dl} = (1 * IDF(\text{질병}) * H_1(\text{질병})) + (0.008 * IDF(\text{바이러스}) * H_2(\text{바이러스})) + (0.883 * IDF(\text{병}) * H_3(\text{병})) + (0.015 * IDF(\text{환자}) * H_4(\text{환자})) \quad \text{(식 8)}$$

베이지안 가중치를 사용하는 이유는 사용자 프로파일 내에 있는 단어들의 빈도를 가중치에 반영함에 있어 모집단이 변하는 특성을 반영하기 위한 것으로 실제로 이는 사용자의 행위를 유용하게 모델링하는 데 이용될 수 있을 것으로 기대한다.

그림 9는 베이지안 추정치와 역문헌 빈도를 이용하여 사용자 질의어와 프로파일 사이의 엔트로피를 계산하기 위한 제안 알고리즘이다.

```
Algorithm Calculate_Entropy(Q_U_V[i], D_V[j][k])
//질의어, 사용자 프로파일과 문서들의 색인어들 사이의 엔트로피 계산 알고리즘
 set i, j, k, Global_Entropy to 0;
 Call Calculate_Bayesian_Estimator(Q_U_V[i])
    //베이지안 학습에 의한 추정치 계산
 for i = 1 to Num_of_Doc do
  for j = 1 to Num_of_Query_UserProfile do
   for k = 1 to Num_of_Term do
    Entropy[i] += MI(query(userprofile[j], Term[k]))
          * log2(MI(query(userprofile[j], Term[k]))
    //베이지안 추정치와 역문헌 빈도를 반영하여 엔트로피의 합을 계산
   end k;
  end j;
  for m = 1 to Num_of_Term do
     G_Entropy += (Bayesian[m] * Entropy[m])
  end m;
 set n to 0;
 Call
Reorder_Document(Entropy[Num_of_Query_UserProfile])
 //엔트로피의 합이 적은 문서가 상위에 순서화되도록 문서 순위  조정
 end i;
End Calculate_Entropy
```

그림 9. 베이지안 추정치와 IDF를 이용한 가중치를 계산하기 위하여 설계된 엔트로피 계산 알고리즘

V. 실험 및 평가

실험은 NT 서버인 SMP server 6400Qp상에서 한국통신에서 구축한 시험용 데이터 모음인 KT set95의 4,414개의 문서에 대하여 수행하였다. 실험 대상 문서로부터 10,585개의 색인어 후보를 추출하였다. 이중에서 두 글자 어휘는 6,524개로 전체 색인어 후보 중 61.64%를 구성하고 있다. 또한 한 글자 색인어 후보와 두 글자를 초과하는 색인어 후보는 그 비율이 각각 6.83%와 31.53%를 차지하고 있다. 본 실험에서 한 글자 색인어는 한자어와 접두어 및 접미사가 대부분이고 그 의미 중요도가 상대적으로 낮으므로 실험 및 평가에는 반영하지 않았다.

두 글자 어휘 중 동음 이의어는 324개로 전체 2,787개의 어휘 중 11.62%를 차지하고 있으며, 세 글자 이상 어휘 중 동음 이의어는 2%로 나타났다.

표 2는 사용자 질의어 '질병'과 사용자 프로파일에 존재하는 [병, 환자, 바이러스]와 각각의 문서에 존재하는 색인어들과의 엔트로피 계산에 있어 베이지안 추정치 및 역문헌 빈도를 사용한 엔트로피 합과 가중치를 사용하지 않은 엔트로피의 합을 나타낸다.

이 경우 가중치를 사용하지 않는 엔트로피 계산 결과를 살펴보면 엔트로피 값이 적은 문서 즉, 문서번호 1845, 0387, 1905, 2103의 순으로 문서는 순서화된다. 색인어 내용에서 알 수 있듯이 문서번호 1845번보다는 0387번 문서가 상위에 순서화되어야 바람직하다. 이러한 문제는 베이지안 및 역문헌 빈도를 이용한 가중치를 이용함으로써 재조정될 수 있음을 실험을 통하여 알 수 있었다.

표 2. 질의어에 대한 엔트로피 계산 결과 #1

문서번호	색인어	엔트로피(H) 가중치 사용안함	엔트로피(H) 가중치 사용
0387	[자궁암, 임상, 연구, 환자]	7.16311	6.22314
1845	[북한, 방송, 컴퓨터, 의료]	7.10516	6.34781
1905	[멀티미디어, 프로그램, 그래픽, 교육]	7.53373	6.40312
2103	[정보, 사회, 통신, 산업]	7.60723	6.41195

표 3은 사용자 질의어 '펜티엄'과 프로파일 [컴퓨터(1459), 부속(4), 하드웨어(209)]와의 계산 결과를 나타낸다.

표 3. 질의어에 대한 엔트로피 계산 결과 #2

문서번호	색인어	엔트로피(H) 가중치 사용안함	엔트로피(H) 가중치 사용
1012	[컴퓨터, 통신망, 정보, 멀티미디어]	7.04150	6.28373
1331	[주전산기, 컴퓨터, 통신, 모듈]	7.35789	6.23588
1383	[컴퓨터, 객체지향, 통신, 소프트웨어]	7.38448	6.27997
1353	[전자파장해, 전파법, 검증, 수입]	6.44440	6.39133
1401	[기술, 후지쯔, 전자파, 기준]	6.71769	6.291642

표 4. 질의어에 대한 엔트로피 계산 결과 #3

문서번호	색인어	엔트로피(H) 가중치 사용안함	엔트로피(H) 가중치 사용
2712	[장비, 전자, 전화, 통신]	7.71732	6.53187
2728	[장비, 외국, 홍보, 통신]	7.56886	6.68523
2735	[무선, 호출, 시스템, 교환기]	7.66934	6.55138
2763	[시스템, 전자, 삼성, 이동]	7.91594	6.73411
3354	[기술, 연구소, 시스템, 개발]	7.81129	6.54456
3441	[통신, 전파, 연구, 시스템]	7.58473	6.18019

표 4는 사용자 질의어 '이동통신'과 프로파일 [전화기(49), 무선(137), 통신(1501)]과의 엔트로피 계산 결과를 나타낸다. 이 경우 사용자 질의어는 의미적 유사성을 내포하는 단어가 아닌 관계로 엔트로피의 계산 결과는 큰 차이를 보이지 않고 있다. 이는 의미적 유사성이 없는 사용자 질의어의 경우에는 엔트로피 값이 문서 전체의 순위 조정에 커다란 영향을 미치지 않음을 나타낸다. 즉, 이러한 경우는 대부분의 검색 결과가 문서 순위에 상관없이 사용자를 만족시킨다는 것을 의미한다.

종합적인 재현율과 정확도의 관계를 고찰하기 위하여 KT set95의 질의어 50개 중 10개를 선택하여 평가에 사용하였으며 재현율과 정확도는 (식 9)에 의하여 계산하였다.

$$\text{정확도} = \frac{\text{검색된 적합문서 수 (상위20개중)}}{\text{검색된 문서 총 수}}$$

$$\text{재현율} = \frac{\text{검색된 적합문헌 수}}{\text{적합문서총 수}} \qquad \text{(식 9)}$$

(식 9)의 정확도 계산식에서는 상위 20개의 문서에 대해서만 정확도를 계산하였는데 이는 본 논문이 사용자 질의어에 만족하는 검색 결과를 얻기 위하여 재현율의 저하 없이 정확도를 높이려는 시스템 설계라는 점을 의미한다.

표 5. 검색효율 비교표

키워드	기존검색		개선된검색1		개선된검색2	
	재현율	정확도	재현율	정확도	재현율	정확도
정보검색	0.72	0.39	0.72	0.42	0.72	0.53
전자파장해	0.35	0.69	0.35	0.72	0.35	0.81
자동차	0.23	0.75	0.23	0.77	0.23	0.87
펜티엄	0.45	0.52	0.45	0.53	0.45	0.61
방사능	0.31	0.73	0.31	0.75	0.31	0.85
레이저	0.57	0.48	0.57	0.51	0.57	0.57
냉 매	0.67	0.43	0.67	0.44	0.67	0.55
질 병	0.91	0.31	0.91	0.32	0.91	0.43
이동통신	0.42	0.55	0.42	0.58	0.42	0.68
세탁기	0.83	0.35	0.83	0.41	0.83	0.47

표 5는 KT set95의 50개 질의어 중에서 10개를 추출하여 (식 9)에 의한 정확도-재현율의 결과를 나타낸다. 여기서 '개선된 검색1'은 베이지안 및 역문헌 빈도에 의한 가중치를 사용하지 않는 경우이고 '개선된 검색2'는 가중치를 이용한 결과를 나타낸다.

계산 결과 가중치를 사용하지 않는 검색 시스템은 역화일을 이용하는 SMART 시스템에 비해 평균 4%의 정확도가 증가되었음을 알 수 있었으며, 가중치를 사용하여 엔트로피를 계산하는 '개선된 검색2'의 경우에는 평균 10%의 정확도 향상이 있음을 알 수 있었다. 각각의 해당 경우를 그림 10, 11에 나타내었으며 종합적인 비교 평가 그래프를 그림 12에 나타내었다.

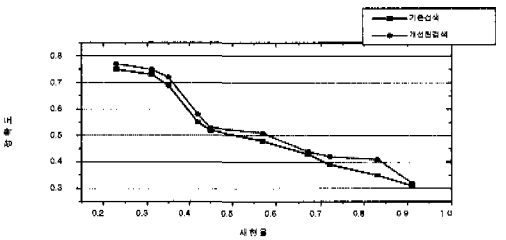


그림 10. 가중치를 사용하지 않은 정확도-재현율 그래프

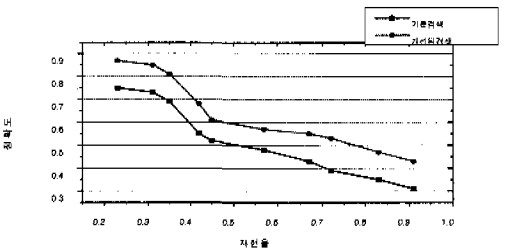


그림 11. 가중치를 사용한 정확도-재현율 그래프

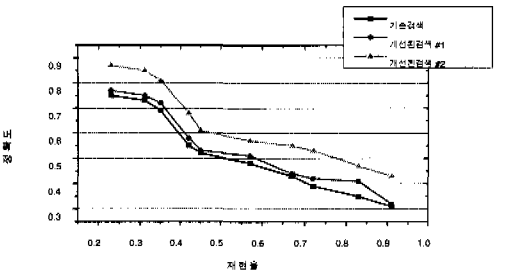


그림 12. 최종 정확도-재현율 그래프

VI. 결 론

본 논문에서는 사용자 질의어의 동적 검색 결과에 대하여 질의어의 의미 정보를 반영하기 위하여 단어들간의 상호 정보량을 이용한 엔트로피 계산을 통하여 문서 순위 조정 알고리즘을 설계하고 구현하였다. 특히, 엔트로피 계산을 위한 확률 벡터 정보량을 계산하기 위하여 상호 정보량을 이용하는 새로운 수식을 유도하였다. 또한, 가장 최근에 사용된 단어 및 사용빈도로 구성된 사용자 프로파일을 설계하여 사용자 질의어와 더불어 동적 검색된 문서와의 엔트로피 계산에 이용하였다. 이때 본 논문에서 설계한 베이지안 학습 및 역문헌 빈도에 의한 가중치를 적용한 결과 기존의 SMART 시스템에 비하여 평균 10%의 향상된 정확도를 얻을 수 있었다.

실험에 사용된 KT set95는 의미적 중의성을 고려하지 않고 설계되었다. 따라서 의미적 중의성이 내포되어 있고 제한된 영역이 아닌 큰 규모의 말뭉치를 구축하여 실험에 사용한다면 정확도는 더욱 향상될 것으로 기대된다.

향후 연구과제로는 KT set95와 같은 한정된 도메인의 공기 정보 대신 언어의 제반 현상을 잘 설명할 수 있는 말뭉치를 구축하여야 할 것이다. 그리고 색인어 구축시 단 음절에 대한 처리 문제와 복합명사의 처리 문제 또한 선결되어야 할 과제 중의 하나이다. 또한, 엔트로피 계산에 따른 복잡도를 해결하기 위한 병렬처리 알고리즘도 고려해야 할 부분이라고 생각된다.

참 고 문 헌

[1] C. Buckley and G. Salton and J. Allan, "The Effect of Adding Relevance Information in a Relevance Feedback Environment," In Proc. 17th ACM SIGIR International Conference on Research and Development in Information Retrieval, pp. 292-298. 1994.

[2] K.W. Church, "Word Assiciation Norms, Mutual Information, and Lexicography," Computational Linguistics, Vol. 16, No. 1, pp. 22-29, 1990.

[3] Scott Deerwester and Susan T. Dumais and Richard Harshman, "Indexing by Latent

Semantic Analysis," Journal of the American Society for Information Science, 41(6): 391-407, 1990.

[4] Iwayama Makoto, Tokunaga Takenobu, "A probabilistic model for text categorization: Based on a single random variable with multiple values," 94-TR0008, 1994.

[5] G. Salton, editor, "The SMART Retrieval System-Experiments in Automatic Document Processing," Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1971.

[6] G. Salton and M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, New York, 1983.

[7] G. Salton, A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing," Communication of ACM, Vol. 18(11), pp. 613-620, 1975.

[8] Tokunaga Takenobu, Iwayama Makoto, "Text categorization based on weighted inverse document frequency," 94-TR0001, Tokyo Institute of Technology, 1994.

[9] S.K.M WONG and Y.Y.YAO, "A Statistical Similarity Measure," In tenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3-12, 1987.

[10] 김남일, 박영찬, 남기춘, 최기선, "정보검색에서 다유전자군 관리에 의한 사용자 프로파일 시스템", 제 9회 한글 및 한국어 정보처리 학술발표 논문집, pp.44-51, 1997.

[11] 강승식, 음절 정보와 복수어 단위 정보를 이용한 한국어 형태소 분석, 서울대학교 대학원 공학 박사 학위 논문, 1992.

[12] 강현규, 박세영, 최기선, "자연언어 정보검색에서 상호정보를 이용한 2단계 문서 순위 결정 방법," 한국 정보과학회 논문지(B), 제 23권, 제 8호, 1996.

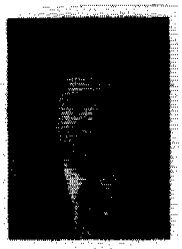
[13] 양재형, 한국어 분석 모호성 해소를 위한 명사의 공기 유사성, 서울대학교 대학원 공학 박사 학위 논문, 1995.

[14] 정영미, "Shannon의 정보이론과 문헌정보," 한국 도서관학, 제 6집, 1979.

[15] 정영미, 정보검색론, 구미무역, 1993.

[16] 한미성, 사용자 프로파일을 이용한 한국어 문서 자동 분류 시스템, 인하대학교 공학 석사 학위 논문, 1998.

최 준 혁(Jun-Hyeog Choi) 정회원



1990년 2월 : 경기대학교 전자계산학과 졸업
1995년 2월 : 인하대학교 대학원 전자 계산공학과 석사
1995년 3월 ~ 현재 : 인하대학교 전자계산공학과 박사과정
1997년 ~ 현재 : 김포대학 컴퓨터계열 조교수
1998년 ~ 현재 : 김포대학 전자계산소장
<주관심 분야> 자연언어처리, 정보검색, 신경망

허 준 희(Jun-Hui Her)

1998년 2월 : 인하대학교 전자계산공학과 졸업
1998년 3월 ~ 현재 : 인하대학교 전자계산공학과 석사과정
<주관심 분야> 자연언어처리, 정보검색, 기계학습

이 정 현(Jung-Hyun Lee)

1977년 2월 : 인하대학교 전자공학과 졸업
1980년 8월 : 인하대학교 전자공학과 석사
1988년 2월 : 인하대학교 전자공학과 박사
1979년~1981년 : 한국전자기술연구소 시스템 연구원
1984년~1989년 : 경기대학교 교수
1989년~현재 : 인하대학교 전자계산공학과 교수
<주관심 분야> 자연언어처리, 인간과 컴퓨터의 상호작용, 정보검색, 음성인식, 고성능 컴퓨터 구조