색인어 연관성을 이용한 의료정보문서 분류에 관한 연구

임 형 근† · 장 덕 성††

요 약

현대사회에서 웹을 통한 정보 제공 서비스가 늘어나면서 병원에서도 홈페이지와 E-mail을 통하여 많은 질문과 상담이 진행되고 있다. 그러나, 이것은 관리자에 대한 업무부담과 답변에 대한 응답시간 지연의 문제가 있다. 본 논문에서는 이런 질의문서에 대한 자동응답시스템의 기초연구로 문서 분류 방법을 연구하였다. 실험방법으로 1200개의 환자질의문서를 대상으로 66%는 학습문서로, 34%는 테스트문서로 활용하여 이것을 NBC(Naive Bayes Classifier), 공통색인어, 연관계수를 이용한 문서분류에 사용하였다. 문서 분류 결과, 기본적인 NBC방법 보다는 본 논문에서 제안한 두 방법이 각각 3%, 5%정도 더 높게 나타났다. 이것은 색인어의 빈도보다, 색인어와 카테고리간의 연관성이 문서 분류에 더 효과적이라는 것을 의미한다.

A Study on Classification of Medical Information Documents using Word Correlation

Hyeong-Geon Lim† · Duk-Sung Jang††

ABSTRACT

As the service of information through web system increases in modern society, many questions and consultations are going on through Home page and E-mail in the hospital. But there are some burdens for the management and postponements for answering the questions. In this paper, we investigate the document classification methods as a primary research of the auto-answering system. On the basis of 1200 documents which are questions of patients, 66% are used for the learning documents and 34% for test documents. All of these are also used for the document classification using NBC (Naive Bayes Classifier), common words and coefficient of correlation. As the result of the experiments, the two methods proposed in this paper, that is, common words and coefficient of correlation are higher as much as 3% and 5% respectively than the basic NBC methods. This result shows that the correlation between indexes and categories is more effective than the word frequency in the document classification.

키워드 : 문서분류(document clssification), 공통색인어(common index), 연관 계수(coefficient of correlation)

1. 서 론

현재 개인 병원이나 종합 병원에서 사용하는 홈페이지는 정적인 정보만을 제공하고 있으며, 의료 정보에 대한 환자들의 질의사항들은 특정 게시판에 게시하거나 e-mail로 전달한다. 일반적인 질의에 대해서는 홈페이지 관리자가 직접 답변하고, 특수한 질의에 대해서는 분류하여 담당 전문의에게 답변을 넘긴다. 그러나 의사도 관리자도 그 많은 답변을 일일이 성의있게 답변하기에는 시간이 충분치 못하며, 질문이 많을 때는 그것을 분류하기도 쉽지 않을 것이다. 이와 같은 질문을 자동 의료정보 분류시스템을 통해 전공별로 구분할 수 있다면 많은 노력을 절약할 수 있을 것이다. 본 논문에서 다루고자 하는 의료정보 분류시스템은 자동 분류(Automatic Classification)에 기반을 두고 있다. 문서의 자동 분류란 기계학습을 이용하여 미리 학습하여 둔 카테고리 중 하나로 문서를 분류해 주는 처리를 일컫는다[8, 16, 18, 19].

문서를 분류하기 위해서는 문서들의 특징을 전혀 모르는 상황에서 각각의 문서의 특징을 추출할 수 있어야 하며, 그들 사이의 공통된 패턴을 발견하고, 목적과 일치되는 판별기준을 수식화하여 문서를 분류할 수 있어야 한다[4]. 문서분류 기법으로 1980년대에는 지식공학에 기반하여 수동으로 문서 분류기를 작성했으며, 1990년대 전반에는 Automated Rule Learning, K-NN Classifier[22], Bayesian Classifier[12], Decision Tree[15]와 같은 기계학습 기법을 이용한 자동 문서분류기 연구가 시작되었다. 1990년대 후반부터는 다양한 기계학습 기법의 적용이

† 정 회 원 : (주)아이큐페드
†† 정 회 원 : 계명대학교 컴퓨터공학과 교수
논문접수 : 2000년 7월 4일, 심사완료 : 2001년 2월 16일

시도되었는데, Entrophy Model[13], Bagging[7], Boosting[10, 20], Stacking[21], EM(Exectation-Maximization)[22] 등이 그것이다.

본 논문에서는 베이시안(Bayesian) 학습법[12]을 이용한 분류 기법을 사용한다. 베이시안 학습법은 확률적 모델을 학습에 적용한 것으로, 타 학습 방법에 비해 경험상 정확도가 높으나, 계산량이 많고 데이터가 잘못되어 임의의 값을 가지거나 값이 생략된 경우에 문제가 발생할 때 보완이 필요하다[14]. 베이시안 학습법을 이용한 대표적인 분류방법으로는 NBC(Naive Bayes Classifier)가 있는데, 평면화(flat)된 카테고리별로 각 단어마다 확률 계산을 하여 분류를 수행한다[22]. 하지만 단어의 빈도수가 높은 것이 꼭 그 문서를 정확하게 대표하는 단어라고는 확신할 수 없다. 예를 들어 의료정보사전에 포함되어 있는 단어 중 "치료"라는 단어는 모든 카테고리에서 쓰이는 단어이며 단어의 학습 문서 양에 따라 확률 계산값은 확연하게 차이가 날 것이다. 실제로 많은 문서에서 그 문서를 대표하는 단어는 그리 높게 발생하지는 않고 있으며, 이러한 문제를 해결하기 위하여 여러 문서에서 높은 빈도를 나타내는 용어는 일반적인 용어로서 문서를 대표할 수 있는 단어로서의 가치는 떨어진다고 볼 수 있다. 이를 보완하기 위한 방법으로는 TF-IDF(Term Frequency-Inverse Document Frequency)알고리즘의 방법이 있다. TF-IDF는 역 문서 빈도수(inverse document frequency)를 단어의 빈도수와 같이 적용함으로서 그 문서를 대표하는 단어들을 효율적으로 찾을 수 있는 알고리즘이다[11, 17].

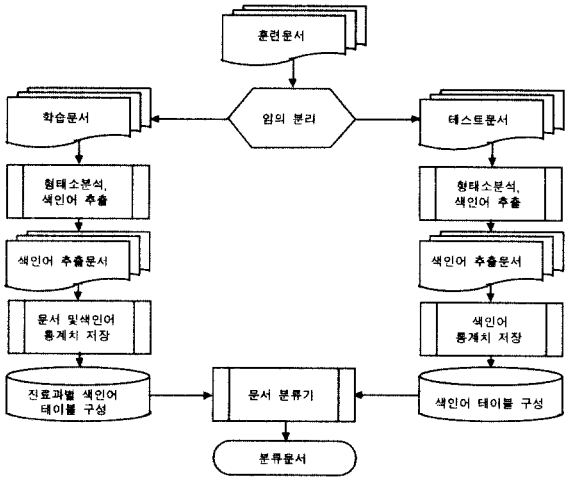
본 논문에서는 학습 문서와 사용될 테스트 문서에 대해서 색인어를 추출하고 추출된 색인어를 대상으로 데이터베이스를 구축하기 위해 한국어 형태소 분석기 HAM[6]을 사용한다. 문서에 포함되어 있는 키워드를 추출하려면 각 단어들을 형태소 분석에 의하여 키워드로서 가치가 있는 명사들을 추출해야 한다. 이 때 형태소 분석 기법의 활용 정도에 따라 색인어휘집을 이용하는 방법, 기능어휘집을 이용하는 방법, 그리고 형태소 분석기를 이용하는 방법이 있다[2]. 그런데 색인어휘집과 기능어휘집을 이용하는 방법은 자동색인 기능으로서 한계가 있기 때문에 최근에는 대부분 형태소 분석을 이용하는 방법을 취하고 있다. 형태소 분석은 문서로부터 각 어절(띄어쓰기 단위)들을 분석하여 명사, 조사, 동사, 어미 등으로 분해하는 작업을 말한다[1, 5]. 한국어 형태소 분석을 기반으로 하여 자동색인 및 철자검사 기능까지 가능한 HAM은 자동색인에 매우 적합한 형태소 분석기를 이용하기 때문에 문서의 종류나 유형에 관계없이 문서에 나타난 키워드를 추출한다.

이러한 이론과 도구들을 바탕으로 본 논문에서는 다음과 같은 두 가지 방법을 연구하였다. 첫 번째 방법은 공통색인어를 추출한 분류방법으로 학습된 진료과에서 중복성을 띠는 색인어들에 대해 확률 계산에 미치는 영향을 최소화 한 것이다. 두 번째 방법은 색인어 연관성을 이용한 분류방법으로 학습된 색인어사전의 각 색인어마다 얼마나 많은 진료과에 영향력을 주었는가를 측정하여 이것을 확률계산에 이용한다. 두 방법 모두 NBC를 기초로 사용한 문서 분류 방법이다. 이 두가지 방법을 기본적인 NBC 분류방식과 비교하기 위해 세가지 실험을 수행한다. 첫 번째는 NBC를 이용한 분류방법이며, 두 번째는 공통색인어를 이용한 방법이며, 마지막으로 색인어 연관성을 이용한 문서분류기를 구현하여 이 세가지 방법에 대해 분류결과를 비교하여 보았다.

2. 의료정보문서 분류시스템의 구조

2.1 전체 시스템 구성도

(그림 1)은 문서 분류를 위한 시스템의 전체적인 구성도이다. 먼저 학습문서로는 환자들의 질의문을 사용하였는데, 병원관련 웹사이트에서 해당 진료과별로 100개씩 전체 1,200개의 질의문을 수집하였다. 여기서, 첫 번째로 학습시킬 문서와 테스트 문서에 대해 임의 분리(Random Split)가 필요한데, 본 논문에서는 학습에 필요한 문서는 진료과별로 66%를 사용하였으며, 나머지 34%는 테스트문서로 분류하는데 사용하였다.



(그림 1) 전체 시스템 구성도

두 번째로 문서에 대한 형태소분석 및 색인어 추출과정을 거쳐, 명사와 복합명사로 구성된 색인어 추출문서를 생성하게 된다. 세 번째로 색인어 추출문서를 이용해 해당 진료과별로 문서, 색인어에 대한 통계치 및 테이블을 구성하여 문서분류기를 통해 분류를 수행하게 된다.

2.2 형태소 분석 및 자동색인

여러 기관에서 형태소 분석기를 구현하였으며 그 중의 일부는 상용 정보검색 시스템에서 자동색인 기능으로 사용

되고 있다. 자동색인에서는 형태소 분석기의 성능이 사용자들에게 직접적으로 드러나지 않으므로, 정보검색 시스템에서 색인어 추출의 정확성은 사용자가 판단하기 어렵다. 왜냐하면 검색된 문서에서 중요한 문서가 빠져 있다 하더라도 사용자는 그 문서가 검색되어야 한다는 사실을 모르는 경우가 대부분이다. 그러므로, 자동색인 시스템의 성능은 정보자료의 중요성만큼 정보검색 시스템에 미치는 영향이 매우 크다. 사용자가 특정한 정보자료의 검색이 불가능함을 알았을 때에야 비로소 자동색인의 중요성을 인식하게 되고, 그제서야 자동색인의 성능을 향상시키려면 저장되어 있는 모든 문서에 대한 색인을 다시 해야 하는 심각한 사태가 발생하게 될 것이다.

본 논문에서는 한국어 형태소 분석 라이브러리 HAM을 사용하여 한글 문서에서 색인어로서 가치가 있는 용어들을 문서의 종류나 유형에 관계없이 추출하게 된다. HAM은 기본적으로 1음절 명사, 품사 불용어(용언, 부사어, 관형어, 독립언어 등), 숫자로 시작되는 용어(날짜, 시간, 번호 등), 기호(문장부호 포함) 등을 불용어로 간주하고 있다. 따라서, 색인어로서 가치가 있는 용어가 추출되지 않는 경우가 발생할 수 있는데, 이를 보완하기 위해 불용어 사전(stopword.dic)에 특수색인어 등록기능이 있으나, 특정 유형들을 모두 추출할때는 option 기능을 적절히 이용하면 된다. 다음은 HAM의 자동색인 기능을 이용하여 한글문서에서 색인어를 추출한 결과이다.

〈 원문 〉

저는 나이는 41세이고 3개월 전부터 오른쪽 엉덩이 부위와 다리에 통증이 와서 MRI촬영을 해 본봐 4,5번 사이의 디스크가 파열이 되어 수술이 불가피 하다고 하며 수술 방법은 뼈 사이에 뼈를 받쳐줄 수 있는 것을 넣고 입원기간은 8~10일을 예상한 답니다. 헌데 본인이 근래 일 관계로 그렇게 시간을 낼 수 없는 처지이고 또한 허리에 수술 받는 것에 선입관이 있어 결정을 못 내리는 것도 사실입니다. 이런 경우 레이저 수술 방법으로 짧은 입원기간이 가능한 지 궁금합니다.

〈 색인어 추출 결과 〉

나이 / 오른쪽 / 엉덩이 / 부위 / 다리 / 통증 / MRI촬영 / MRI / 촬영 / 본봐 사이 / 디스크 / 파열 / 수술 / 불가피 / 수술 / 방법 / 사이 / 입원기간 / 입원 / 기간 / 입원기 / 예상 / 본인 / 근래 / 관계 / 시간 / 처지 / 수술 / 선입관 /결정 / 경우 / 레이저 / 수술 / 방법 / 입원기간 / 입원 / 기간 / 입원기 / 가능 / 궁금

2.3 NBC를 이용한 문서 분류

환자의 질의문서 66%를 이용하여 학습된 각 진료과별 색인어 정보와 34%의 테스트문서로 사용된 질의문서의 색인어 정보를 분석해서 NBC 분류방법을 이용하여 진료과별로 테스트 문서의 확률값을 계산, 비교하여 문서를 분류하게 된다. NBC 학습 알고리즘은 (그림 2)와 같다.

- *Learn_Naive_Bayes_Text(Examples, V)*
1. *Examples* 에서 발생하는 모든 단어와 토큰을 수집한다.
   - *Vocabulary* ← *Examples* 에 있는 모든 서로 다른 단어 및 토큰들
2. 확률 조건들 $P(v_j)$와 $P(w_k|v_j)$의 계산이 필요
   - *For* 각 타겟값 V에서 v_j *Do*
     - $docs_j \leftarrow$ 타겟값 v_j인 *Examples*의 부분 집합
     - $P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$
     - $Text_j \leftarrow docs_j$의 모든 부분을 연결하여 만든 단 하나의 문서
     - $n \leftarrow Text_j$ 에 존재하는 단어 수의 합계 (여러 번 중복된 단어들을 모두 센다.)
     - *For Vocabulary*에 있는 각각의 단어 w_k *do*
       - $n_k \leftarrow Text_j$에 발생하는 단어 w_k 의 회수
       - $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$

(그림 2) NBC 학습 알고리즘

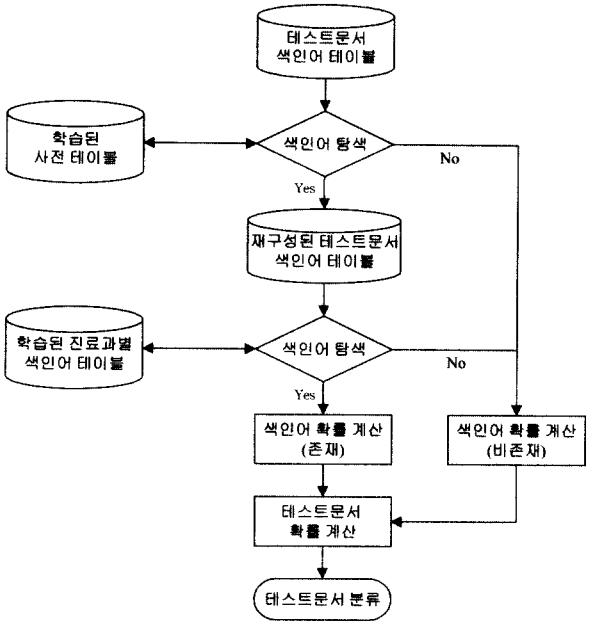
문서 분류할 때는 학습시 수집한 데이터들을 이용하여 새로운 문서에 대한 분류작업을 수행하는데, 이때 사용되는 NBC 분류 알고리즘은 (그림 3)과 같다.

- *Classify_Naive_Bayes_Text(Doc)*
  - *positions* ← *Vocabulary*에서 발견된 토큰들을 포함하는 Doc내의 모든 단어들의 위치
  - *Return* V_{NB},

$$V_{NB} = \underset{v_j \in V}{argmax}\, P(v_j) \prod_{i \in position} P(a_i|v_j)$$

(그림 3) NBC 분류 알고리즘

(그림 4)는 NBC을 이용한 문서분류 구조를 나타내고 있다. 먼저 입력된 테스트문서 색인어 테이블에서 학습된 사전의 색인어 테이블과 비교하여 사전에 존재하지 않는 색인어를 대상으로 곧바로 비존재시 확률계산을 한다. 왜냐하면, 각 진료과별로 중복성을 제거하여 수집한 색인어들로 사전을 구성하였는데 사전에 존재하지 않는다는 것은 어떠한 진료과에도 그 색인어가 존재하지 않는다는 것을 의미하며, 확률계산에서 모든 진료과에 똑같이 중복 계산될 수 있는 시간적 낭비를 줄일 수 있다. 테스트문서에서 사전에 존재치 않는 색인어를 제외한 나머지 색인어를 대상으로 재구성된 색인어 테이블과 학습된 진료과별 색인어 테이블의 비교를 통해서 색인어 존재유무에 따라 확률 계산을 하게 되며, 이를 진료과별로 색인어의 확률 계산값들에 대한 통계치를 종합, 비교하여 문서를 분류한다.



(그림 4) NBC를 이용한 문서분류 구성도

2.4 공통 색인어 추출

많은 문서들 중에서 문서를 대표하는 특징을 찾아내기 위해서 색인어의 빈도수를 많이 이용한다. 하지만 색인어의 빈도수가 높은 것이 그 문서를 대표한다고는 말할 수 없으며, 실제로 많은 문서에서 그 문서를 대표하는 색인어는 빈도가 그리 높게 발생하지 않는다. 또한 테스트문서에서 추출한 색인어들중에는 모든 진료과에 동일하게 쓰이고 있는 공통 색인어들이 많이 존재하게 된다[9].

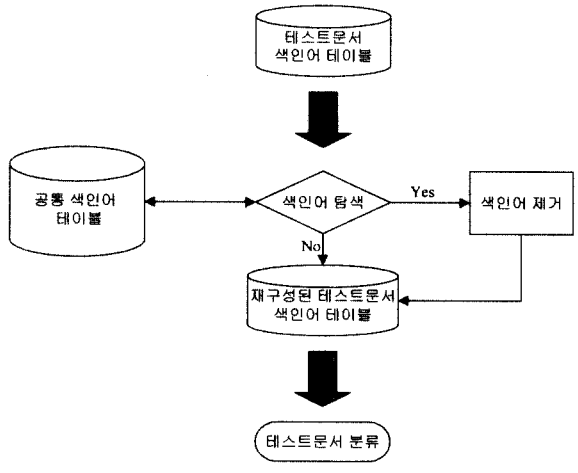
예를 들어 의료 정보사전에 포함되어 있는 색인어 중 "치료"라는 색인어는 모든 진료과에서 공통적으로 사용되는 공통 색인어 이다. 학습문서의 양이 늘어나면 늘어날수록 공통 색인어의 빈도수도 늘어날 것이다. 그러나 이들 공통 색인어는 특별히 어떤 진료과를 대표하지도 않으며, 전체문서에서 일반적으로 사용되는 성격의 색인어로서 빈도수에 의한 확률계산에 많은 영향을 준다. 이와 같은 공통 색인어들을 추출하여 테이블을 구성한 결과, 색인어들이 학습된 사전에서 빈도수가 높은 상위 10%에 모두 속하였다. 본 논문에서는 이러한 공통 색인어들을 이용하여 새로운 테이블을 생성하고, 생성된 테이블을 이용하여 테스트문서에서 공통 색인어들을 제거한 뒤 재구성된 테스트문서로 문서분류를 하였다.

공통 색인어는 진료과별 색인어 테이블을 참조하여 각 색인어 테이블에 공통으로 존재하는 색인어를 찾는다. (알고리즘 : 공통색인어 발견) 처럼 모든 색인어 테이블을 순차검색해서 모든 진료과별 색인어 테이블에 존재하면 공통 색인어 테이블에 등록한다. 이와 같이 하여 파악된 공통색인어는 '치료', '상태', '피곤', '증상', '효과' 등 73개가 된다.

```
NoMatching = False
Do While [사전테이블의 EOF]

    VA = 사전테이블의 색인어
    Do While [진료과이름 테이블의 EOF]

        CA = 진료과 이름
        진료과별 색인어 테이블의 Seek "=", CA, VA
        IF NoMatch = True Then
            공통색인어가 아니므로 다음 진료과의 색인어 테
                이블 검색.
            NoMatching = True로 설정, B루프를 빠져나감
        Else
            EOF가 될 때까지, 다음 진료과에도 색인어가
            존재하는지 검색
        End if

    Loop (B루프)
    IF NoMatching = False Then
        모든 진료과에 색인어가 존재함으로 공통색인어 테이
        블에 색인어를 등록.
    End if

Loop (A루프)
```

(알고리즘 : 공통색인어 발견)

(그림 5)는 공통 색인어를 이용한 문서분류의 구성도이다. 학습된 색인어 테이블은 이미 공통 색인어를 제거한 색인어 테이블이다. 테스트문서 색인어 테이블에 대해서도 마찬가지로 공통 색인어를 찾아 제거하여야 한다.



(그림 5) 공통 색인어를 이용한 문서분류 구성도

2.5 색인어 연관성

문서 내용을 분류 기준으로 적용할려면 자연어처리에서 사용되는 구문 분석, 의미 분석 과정을 거쳐야만 된다. 그러나 지금까지도 이것은 연구 진행 중인 분야이며, 만족할 만한 성능에 이르지 못하고 있다. 또한 실제 구현이 되어도 방대한 처리량에 따른 속도 문제가 남아있다. 그러므로 본 논문에서는 학습된 사전의 색인어를 대상으로 진료과별로 연관성 정도를 조사하여 이를 NBC 학습알고리즘에 적용시켜 보았다[3].

학습된 사전의 색인어 중 모든 진료과에 존재하는 공통 색인어와는 달리, 전체 진료과 중 몇 개의 진료과에만 특수하게 존재하는 색인어도 있다. 만약에 어떤 색인어가 한 개의 진료과에만 속한다면 이는 매우 특수한 경우라고 볼 수 있는데, 다음과 같은 두가지 가정이 가능하다.

(1) "만성위염"과 같이 소화기내과만 존재하는 색인어는, 해당 진료과에서 주요한 역할을 차지하는 색인어이다.
(2) "지속기"와 같이 우연히 소화기내과에 나타났지만, 진료과의 특성을 나타내는 색인어라고도 할 수 없다.

본 논문에서는 가정(1)과 같이 색인어가 해당 진료과에서 주요한 역할을 담당하는 것으로만 국한하기로 한다. 사전에 나타난 각각의 색인어가 얼마나 많은 진료과에 존재하는지 조사하여 이를 관계회수라 한다. 예를 들어, "만성위염"이라는 색인어가 1개의 진료과에만 나타나는 독립적인 색인어라면 진료과와의 관계회수는 1이다. 그러나, 만약 "만성위염"이 전체 12개의 진료과에 모두 속한다면 관계회수는 12가 될 것이다. 이와 같이 색인어와 진료과와의 관계를 연관관계라 말하고, 연관관계에 따른 색인어의 관계회수를 연관계수라고 규정하며 R_k로 나타낸다. 색인어의 연관계수가 높을수록 해당 진료과에 영향력이 낮으며, 연관계수가 낮을수록 색인어가 해당 진료과에 미치는 영향력이 높다고 가정하여 이를 NBC 학습알고리즘에 적용하면, (그림 6)과 같은 수식을 만들어 낼 수 있다.

- 연관계수를 적용한 NBC 학습 알고리즘
  - R_k ← *Vocabulary* 에 나타난 모든 색인어의 연관계수
  - n_k ← $Text_j$에 발생하는 단어 w_k 의 회수
  - E_k ← 연관계수 R_k가 $Text_j$ 에 발생하는 단어 w_k 의 회수에 미치는 영향력
    - $E_k \leftarrow n_k / R_k$
  - For *Vocabulary*에 있는 각각의 단어 w_k *do*
    - $P(w_k | v_j) \leftarrow \dfrac{E_k + 1}{n + |Vocabulary|}$

(그림 6) 연관계수를 적용한 NBC 학습 알고리즘

(그림 6)의 수식은 색인어가 각 진료과에 얼마나 연관되어 있는가를 나타내는데, 색인어의 영향력 E_k는 색인어 빈도수 n_k를 대상으로 연관계수 R_k로 나눈값이다. 영향력 E_k를 사용하면 문서분류는 색인어 빈도수가 아닌 연관계수에 의해 영향을 받게 되며, 만약 같은 연관계수를 가지는 각 진료과별 색인어에 대해서는 빈도수에 영향을 받게된다.

3. 시스템 구현 및 결과

3.1 학습문서

학습 및 질의 문서로 병원관련 웹사이트에서 수집한 환자들의 질의문 1,200개를 대상으로 하였다. 이중 66%의 문서들은 학습에 사용되었으며, 나머지 34%는 시스템 성능평가를 위한 테스트용 질의에 사용하였다. <표 1>은 12개의 진료과를 나타내며, 각각의 진료과명은 실제 학습테이블에서도 필드명으로 구성된다.

〈표 1〉 12개의 진료과

• 산부인과	• 정형외과	• 피부과
• 혈액종양내과	• 흉부외과	• 비뇨기과
• 성형외과	• 소화기내과	• 신경외과
• 신경정신과	• 안과	• 이비인후과

학습예제로 사용된 질의문서는 진료과별로 환자의 질문사항들에 대해 기록된 문서들로 구성되어 있으며, 학습된 진료과별 색인어 수는 대략 4000-6000개 정도이다. (그림 7)은 학습예제로 사용된 환자의 질문사항을 나타낸 질의문서 중 일부분이다.

번호 : 1776
일자 : 2000-05-03
상담 : 디스크 수술 환자인데요...
병력 : 키 : 163cm, 몸무게 : 60kg, 혈액형 : A형

저는 4월 18일에 요추간판탈출(3~4번) 수술(절개해서 디스크를 재거했습니다)을 한 휴학생입니다. 그런데 수술후 앉고 일어설때 마다 허리 부분에서 "우두둑"하는 소리가 계속 납니다. 이상이 있는 것인지. 그리고 허리를 굽히기가 아직은 겁이 나구요. 적당한 운동을 어떻게 해야 될지도 모르겠구요. 그리고 저는 간호대생인데 실습할때나 공부할때 혹은 나중에 취직후 계속 서있는 생활을 해야하는데 무리는 없는지도 알고 싶습니다. 마지막으로 오래 앉아 있는 것이 무리라고 하고 무거운것은 들지 않는 것이 좋다고들 하던데 그 기준이 궁금합니다.

(그림 7) 환자의 질의문서 중 일부분

이 문서에서 번호, 일자, 상담제목, 병력 등을 나타낸 상위부분은 무시하고, 하위부분의 문서내용만 이용하는데, 이를 한국어 분석 라이브러리 HAM을 사용하여 색인어를 추출하게 되고, 추출한 색인어를 대상으로 학습된 해당 진료과별 색인어와 비교, 분석하여 문서분류를 하게된다.

3.2 분류결과

주어진 학습예제들을 이용한 분류결과는 (그림 8), (그림 9), (그림 10)과 같으며 이 그림들은 실제 데이터베이스에 저장된 결과값들을 엑셀의 매크로 언어인 VBA(Visual Basic for Application)을 이용하여 출력을 한 것이다.

〈표 2〉 Class에 해당되는 진료과 정보

A	비뇨기과	G	안 과
B	산부인과	H	이비인후과
C	성형외과	I	정형외과
D	소화기내과	J	피 부 과
E	신경외과	K	혈액종양내과
F	신경정신과	L	흉부외과

* Test document : 408 (- Row : 테스트문서 진료과 - Column : 학습된 진료과)

class	A	B	C	D	E	F	G	H	I	J	K	L	정확성
A	23	1	0	0	4	1	0	0	0	0	5	0	67.65%
B	2	28	0	0	0	1	1	0	0	0	2	0	82.35%
C	1	0	25	0	3	2	2	0	0	1	0	0	73.53%
D	1	1	0	26	0	1	0	0	1	0	3	1	76.47%
E	0	0	0	0	31	0	1	0	1	0	1	0	91.18%
F	1	0	0	0	1	30	0	0	0	1	1	0	88.24%
G	1	3	1	0	1	0	26	1	1	0	0	0	76.47%
H	0	1	0	1	2	4	3	17	0	1	3	2	50.00%
I	1	0	0	0	12	1	0	0	13	0	6	1	38.24%
J	2	4	2	0	3	2	0	1	0	17	3	0	50.00%
K	0	1	0	1	1	0	0	0	1	0	30	0	88.24%
L	1	4	2	1	7	1	0	0	0	0	0	18	52.94%

* 정확문서 : 284 / 408 (- 정확성 평균 : 69.61%)

(그림 8) NBC를 이용한 분류 결과

* Test document : 408 (- Row : 테스트문서 진료과 - Column : 학습된 진료과)

class	A	B	C	D	E	F	G	H	I	J	K	L	정확성
A	23	2	0	0	3	1	0	0	1	0	4	0	67.65%
B	2	27	0	0	0	1	1	0	0	0	2	1	79.41%
C	1	2	26	0	1	1	2	0	0	1	0	0	76.47%
D	1	1	0	28	0	1	0	0	0	0	2	1	82.35%
E	0	0	0	0	29	1	1	0	3	0	0	0	85.29%
F	1	0	0	0	1	31	0	0	0	1	0	0	91.18%
G	1	0	1	0	0	1	25	2	1	1	1	1	73.53%
H	0	1	0	2	0	2	3	20	0	2	1	3	58.82%
I	1	0	1	0	10	2	0	0	15	0	4	1	44.12%
J	0	3	4	1	1	2	1	1	0	20	1	0	58.82%
K	0	2	0	2	1	0	0	0	1	0	28	0	82.35%
L	1	2	1	1	2	1	1	0	0	0	1	24	70.59%

* 정확문서 : 296 / 408 (- 정확성 평균 : 72.55%)

(그림 9) 공통 색인어를 이용한 분류 결과

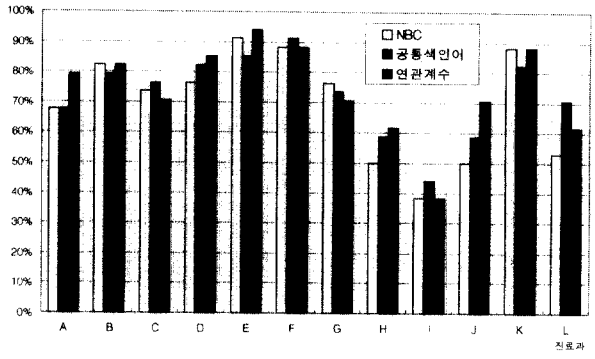
* Test document : 408 (- Row : 테스트문서 진료과 - Column : 학습된 진료과)

class	A	B	C	D	E	F	G	H	I	J	K	L	정확성
A	27	0	0	0	2	1	0	0	0	0	4	0	79.41%
B	2	28	0	0	0	0	2	0	0	0	2	0	82.35%
C	1	0	24	0	1	1	3	2	0	2	0	0	70.59%
D	1	1	0	29	0	0	0	0	0	0	2	1	85.29%
E	0	0	0	0	32	0	1	0	1	0	0	0	94.12%
F	1	0	0	0	1	30	1	0	0	1	0	0	88.24%
G	1	3	1	0	2	0	24	1	1	1	0	0	70.59%
H	1	1	0	1	2	2	3	21	0	0	1	2	61.76%
I	0	1	0	0	10	2	2	0	13	0	5	1	38.24%
J	0	3	1	0	2	2	0	0	1	24	0	1	70.59%
K	0	1	0	1	0	0	1	0	1	0	30	0	88.24%
L	2	1	0	0	6	1	3	0	0	0	0	21	61.76%

* 정확문서 : 303 / 408 (- 정확성 평균 : 74.27%)

(그림 10) 연관계수를 이용한 분류 결과

(그림 11)은 NBC, 공통색인어, 연관계수의 분류 결과에 대한 비교그래프이다. 3가지의 분류법의 평균향상률을 <표 3>을 통해 살펴보면, NBC에 색인어의 연관관계를 적절히 이용한 분류법에서 정확율이 높아진 것을 알 수 있다.



(그림 11) NBC, 공통색인어, 연관계수의 분류 결과 비교그래프

〈표 3〉 분류 결과 비교

class	진 료 과	N B C	공통색인어	연 관 계 수
A	비 뇨 기 과	67.65%	67.65%	79.41%
B	산 부 인 과	82.35%	79.41%	82.35%
C	성 형 외 과	73.53%	76.47%	70.59%
D	소 화 기 내 과	76.47%	82.35%	85.29%
E	신 경 외 과	91.18%	85.29%	94.12%
F	신 경 정 신 과	88.24%	91.18%	88.24%
G	안 과	76.47%	73.53%	70.59%
H	이 비 인 후 과	50.00%	58.82%	61.76%
I	정 형 외 과	38.24%	44.12%	38.24%
J	피 부 과	50.00%	58.82%	70.59%
K	혈 액 종 양 내 과	88.24%	82.35%	88.24%
L	흉 부 외 과	52.94%	70.59%	61.76%
정 확 성 평 균		69.61%	72.55%	74.27%
평균향상률(NBC 기준)		-	+2.94%	+4.66%

그 이유는 색인어가 다른 외부의 진료과와는 별개로, 독립성을 유지할수록 해당 진료과의 특징을 잘 나타내는 것으로 볼 수 있으며, 분류 결과에서 더 좋은 결과를 얻을 수 있는 것으로 볼 수 있다. 그리고, 진료과중 "정형외과"와 같은 경우에는 다른 진료과와는 달리 정확도가 현저하게 떨어지는데, 그 이유를 조사하기 위하여 해당 문서들을 살펴본 결과, "정형외과"의 색인어들은 환자들이 주로 피부손상에 따른 신경계에 대한 질의사항이 많아서 연관성이 높은 "신경외과"쪽으로 많이 분류되는 것으로 나타났다.

<표 4>은 학습된 질의문서중에서 240개의 질의문서를 추출하여 이것을 테스트문서로 NBC알고리즘, 공통색인어, 연관계수에 적용한 결과이다. 추출방법은 무작위로 해당 진료과별로 학습된 질의문서 20개씩을 추출하여 이것을 테스트 문서로 사용하여 보았다. <표 4>의 분류 결과들을 보면 모두 90% 이상의 정확성을 보이며, 평균정확율은 대체로 95%~99%의 높은 정확율을 보였다. 이것은 해당 진료과의 학습문서로 사용된 질의문서를 임의 추출하여 다시 테스트 문서로 사용하게 됨으로써 비교적 높은 정확성을 보였다.

〈표 4〉 학습된 질의 문서를 이용한 문서 분류 결과 비교

class	진 료 과	N B C	공통색인어	연관계수
A	비 뇨 기 과	18	19	19
B	산 부 인 과	20	20	20
C	성 형 외 과	14	19	19
D	소 화 기 내 과	19	19	20
E	신 경 외 과	20	20	20
F	신 경 정 신 과	20	20	20
G	안 과	20	20	20
H	이 비 인 후 과	20	20	20
I	정 형 외 과	20	20	20
J	피 부 과	20	20	20
K	혈 액 종 양 내 과	20	20	20
L	흉 부 외 과	19	20	20
전 체 정 확 성 (240)		230	237	238
정확성향상 (NBC 기준)		-	+ 7	+ 8

4. 결 론

현대사회에서 인터넷은 없어서는 안될 하나의 문화가 되어 가고 있으며, 이러한 인터넷의 확산으로 웹을 통한 정보 제공 서비스가 늘어 가고 있다. 병원에서도 홈페이지를 통하여 환자와 상담하며 질의사항에 응답을 해주고 있는데, 이것은 관리자에 많은 업무를 주게 되어 질의사항에 대한 효과가 많이 떨어진다고 볼 수 있다. 환자의 질의사항을 특정 게시판을 통하여 질의병명에 해당되는 담당 의사에게 문서가 직접 전달된다면 환자는 자신의 병에 대한 정확한 이해와 의사에 대한 신뢰를 가지게 될 것이며, 병원으로서는 환자들의 질의사항들에 대해 빠르고 정확하게 답변할 수 있을 것이다.

본 논문에서는 NBC알고리즘을 기초로 공통 색인어를 이용한 분류 방법과 색인어 연관성을 이용한 분류 방법을 적용하여 문서분류 시스템을 구현하였다. 12개의 진료과를 대상으로 질의문서들을 분류시킨 결과, 첫 번째로 NBC를 이용한 분류방식의 평균 정확율은 70% 정도로 나타났으며, 두 번째로 공통색인어를 이용한 방법에서는 약 73% 정도의 평균 정확율을 보였다. 마지막으로 진료과별 색인어가 미치는 연관성을 이용한 방법에서는 74% 정도의 평균 정확율을 보여 기존에 이용한 NBC방법에 비해 3%~5%정도 높게 나타났다. 이와 같이 일반적인 색인어의 빈도수가 문서 분류에 미치는 영향력을 최대한 줄이며, 색인어의 연관성을 이용하는 것이 기존의 NBC 분류방식만을 이용한 것보다 더 효율적인 방법으로 나타났다.

이 논문의 한계는 색인어의 통계적 정보만을 이용하여 질의문서의 색인어들을 조정하고 가중치를 부여하는데 있다. 따라서 차후 과제로 색인어의 의미 정보를 이용한 문서 분류 방법과 학습 문서의 특성에 따라 변할 수 있는 분류 방법에 대한 연구가 필요할 것이다. 또한 지식 정보로서 사용하고 있는 문서의 양이 늘어나게 되는데 따른 효과적인 저장 방법이 있어야 할 것이며, 자주 갱신되는 정보의 경우에는 상호 정보를 관리할 수 있는 방법이나 도구가 필요할 것으로 보인다.

참 고 문 헌

[1] 강승헌, 유재수, "문자열 부분검색을 위한 색인기법의 설계 및 성능평가", 한국정보처리학회논문지, 제6권 제6호, pp.1458-1467, 1999.

[2] 강호관, "한국어 의존관계 파싱에 적합한 구문단위의 정의", 포항공과대학교 석사학위논문, 1998.

[3] 신진섭, 이창훈, "단어의 연관성을 이용한 문서의 자동분류", 한국정보처리학회 논문지, 제6권 제9호, pp.2422-2430, 1999.

[4] 우종원, 윤승현, 유재수, "문서관리시스템을 위한 질의처리기 설계 및 구현", 한국정보처리학회논문지, 제6권 제6호, pp. 1419-1432, 1999.

[5] 이운재, "한국어 문서 태깅 시스템의 설계 및 구현", 한국과학기술원 석사 학위 논문, 1993.

[6] 한국어 분석 라이브러리, http : //ham.hansung.ac.kr/ham/ham.html/.

[7] L. Breiman, "Bagging predicators," Machine Learning 24(2), pp.123-140, 1996.

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

[9] W. Cohen and Y. Singer, "Context-sensitive learning methods for text categorization," SIGIR-96, 1996.

[10] H. Drucker, "Improving regressors using boosting techniques," Proc. 14th Conf. on Machine Learning, Nashville TN, pp.107-115, 1997.

[11] T.Joachims, "A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization" Proc. 14th Conf. on Machine Learning, Nashville TN, pp.143-146, 1997.

[12] G. H. John, and P. Langley, "Estimating continuous distributions in Bayesian classifiers," Proc. 11th Conf. on Uncertainty in Artificial Intelligence, Montreal Canada, pp.338-345, 1995.

[13] R. Kohavi, and M. Sahami, "Error-based and entrophy-based discretization of continuous features," Proc. 2nd Conf. on Knowledge Discovery and Data Mining, Portland OR, AAAI Press, pp.114-119, 1996.

[14] T. Mitchell, Machine Learning, McGraw-Hill, 1997.

[15] J. R. Quinlan, "Induction of decision trees," Machine Learning 1(1), pp.81-106, 1986.

[16] C. J. van Rijsbergen, Information Retrieval, Butterworths,

London. 2nd Edition, 1979.

[17] G. Salton, and C. Buckely, "Improving Retrieval Performance by Relevance Feedback," Journal of the American Society for Information Scienc, 41(4), pp.288-297, 1990.

[18] _____, "Term weighting approaches in automatic text retrieval," Technical Report pp.87-881, Cornell University, Department of Computer Science 1987.

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

[20] R. E. Schapire, Y. Freund, P. Bartlett, and W.S Lee, "Boosting the margin : A New explanation for the effectiveness of voting methods," Proc. 14th Conf. on Machine Learning, Nashville TN, pp.322-330, 1997.

[21] K. M. Ting, and I.H. Witten, "Stacked generalization : When does it work?," Proc. 15th Joint Conf. on Artificial Intelligence, Nagoya Japan, pp.866-871, 1997.

[22] I. H. Witten, and E. Frank, Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations, Academic Press, 2000.

임 형 근

e-mail : ltomcat@jinri.keimyung.ac.kr
1998년 경산대학교 정보처리학과(학사)
2000년 계명대학교 컴퓨터공학과(공학석사)
2000년~현재 (주)아이큐패드 근무
관심분야 : 기계학습, 정보검색, 시스템 관리 등

장 덕 성

e-mail : dsjang@kmu.ac.kr
1979년 경북대학교 컴퓨터공학과 졸업(학사)
1981년 서울대학교 전산과학과(이학석사)
1988년 서울대학교 컴퓨터공학과(공학박사)
1982년~1985년 동아대학교 전산공학과 조교수
1985년~현재 계명대학교 컴퓨터공학과 교수
1992년~1993년 University of Colorado 방문연구교수
1998년~현재 계명대학교 기획정보처 전산원장
관심분야 : 컴파일러, 시각프로그래밍, 자연어처리, 정보검색, 에이전트 등