(제 10회 한글 및 한국어 정보처리 학술대회)

문서 구조 정보에 기반한 웹 페이지 범주화 모델*

정성화, 이종혁

포항공과대학교 정보통신대학원

A Web Page Categorization Model Based on Document Structural Information

Sung-Hwa Jung, Jong-Hyeok Lee

Graduate School for Information Technology, PIRL, POSTECH

{ojsh, jhlee}@madonna.postech.ac.kr

요약

본 논문에서는 주제범주 체계를 이용한 웹 검색이 가지는 장점을 이용 할 수 있도록 인터넷 웹 페이지들을 주제범주 체계에 따라 자동으로 분류하는 모델을 제시한다. 특히 웹 페이지 작성자들의 의도를 범주화에 반영할 수 있는 방법으로 HTML 태그를 이용한다. 즉 웹 페이지의 표현에 있어서 벡터 스페이스 모델에서의 색인어 빈도 가중치에 태그 가중치를 추가 하여 보다 좋은 성능을 얻도록 하였다. 그리고 주제범주를 표현하는데 사용되는 자질의 선정에는 기대상호정보, 상호정보 척도를, 문서간 유사도 비교에는 최근린법을 사용하였다. 전북대에서 정보탐정용으로 분류한 웹 페이지를 대상으로 실험하였으며, 기본 모델 대비 약 7%의 정확도 향상을 얻을 수 있었다.

1 서론

웹 페이지 범주화(Web Page Categorization)는 웹 페이지에 미리 정해진 주제 범주(Category)를 할당하는 작업이며, 텍스트를 대상으로 한다는 점에서 기본적으로 문서 범주화(Text Categorization)와 동일한 문제가 된다[2]. 따라서 문서 범주화를 위해 개발된 방법론들은 웹 페이지 범주화에도 적용이 가능하다. 그러나 웹 페이지와 전통적인 문서 사이에는 여러가지 차이점이 있다. 먼저 웹 페이지는 상호 연결구조에 의해 비순차적(Non-sequential)으로 내용에 접근한다. 이것은 하나의 웹 페이지가 가지는 주제의 범위가 상위 웹 페이지, 즉 홈페이지에 접근하면 할수록 넓어지고, 하위 웹 페이지로 갈수록 좁아지는 것을 의미한다. 문서의 경우 하나의 처리 단위가 비교적 일관된 주제를 가지는 것과는 다른 상황이다. 따라서 웹 페이지 범주화는 기존의 문서 범주화와는 다른 전략이 필요하다[2]. 또 다른 고려 사항으로 웹 페이지 범주화에 있어서 처리 대상이 되는 텍스트는 모두 HTML 태그에 의해 관리되고 표현된다는 것이다. 즉 웹 페이지를 작성하는 사람은 문서의 내용을 쉽게 이해될 수 있는 형태로 표현하고자 노력하며, 이 과정에서 그러한 의도는 태그의 형태로 웹 페이지에 담기게 된다. 그래서 일반 텍스트에서는 파악하기 힘든 내부 구조정보와 추출된 주제어간의 상대적인 중요도에 있어서의 우열 관계를 HTML 태그를 활용하면 비교적 쉽게 얻을 수 있다. 이는 문서의 표현에 있어서 중요한 정보가 되며, 본 논문에서는 이점에 착안하여 범주화를 위한 문서의 표현에 HTML 태그를 이용하는 방법을 제안한다.

일반적으로 문서 범주화의 성능은 자질 추출 방법과 문서 범주화 모델에 크게 영향을 받는다[1]. 본 논문에서의 자질 추출은 벡터 공간 모델에서 기대 상호 정보(EMIM : Expected Mutual Information)[4][7]와 상호 정보(MIM : Mutual

* 이 논문은 '98년도 한국통신(과제제목 : 웹검색 서비스용 자동 문서분류 시스템 연구)의 지원비 지원에 의한 결과임.

Information)를 이용하여 주제범주와 일정값 이상의 관련도를 가지는 색인어를 자질로 선택하고 [2], 색인어 가중치는 TF/IDF에 태그 가중치와 태그 정규화 가중치를 추가하였다. 범주화 모델로는 최근린법(KNN : K-Nearest Neighbor) [3][8]을 사용하였다.

앞서 언급한 바와 같이 본 논문의 핵심은 HTML 문서와 같이 구조화 된 문서에 대한 표현에 있다. 이는 문서 작성자는 문서의 내용이 쉽게 이해될 수 있도록 여러가지 배려를 할 것이라는 것을 전제로 한 것이다. 문서 작성자가 문서의 이해를 돕기 위해 의도적으로 만들어 둔 문서 내부 구조 정보를 태그로부터 얻어, 색인 작업시에 가중치로 반영하는 것이다. 따라서 문서 작성자의 의도를 가장 잘 반영할 수 있도록 태그별로 가중치를 적절하게 조정해 주는 것이 관건이 되며, 태그의 순기능적인 측면이 된다.

반면에 무분별하게 사용된 태그에 의해 문서의 주제가 훼손되는 경우도 예상이 되므로 이를 방지하기 위해서 태그의 내부 발생 빈도에 따라 정규화 작업을 실시하였다. 이것은 평균이상으로 사용된 태그를 그 비율 만큼 가중치를 줄여주는 처리이며, 평균 이하이면 최초의 값을 그대로 사용한다.

이상의 모델로 전북대에서 분류한 정보 탐정용 웹 사이트 주제분류 체계[2]에 대한 실험을 실시하여 성능 향상을 얻을 수 있었다.

2. 웹 페이지 범주화 모델

2.1 제안하는 웹 페이지 범주화 모델

문서 범주화 문제를 해결하기 위한 기존의 연구들은 크게 아래 3가지 접근 방법으로 분류된다 [1]. 먼저 문서가 특정 범주에 포함되는지 여부를 판단할 수 있는 규칙을 전문가가 작성하거나 학습 문서들에게서 추출하는 규칙에 기반한 방법(Rule-based Method)[9][5]이 있다. 그리고 학습 문서에서 추출한 범주 자질(Category feature)을 이용하는 베이지언 확률모델[6][7]이 있으며 마지막으로 문서간 유사도(Similarity) 값을 이용하는 최근린법(K-nearest neighbor method)[3][8]을 들 수 있다. 이러한 방법론들은 대부분 비슷한 결과를 보이는 것으로 알려져 있다[1].

최근린법(K-nearest neighbor method)에서는 일반적으로 역색인 파일(Inverted Index File)로 문서를 표현한다[1]. 본 논문에서는 이때 사용되는 문서의 자질 선정에 기존의 불용어 처리 이외에 기대 상호 정보(EMIM: Expected Mutual Information Measure)와 상호 정보 척도(MIM : Mutual Information Measure)를 사용하였다. 이는 문서 범주화에서는 학습 문서를 문서 중심으로 표현하기 보다는 문서가 가지는 범주 중심으로 표현하는 것이 바람직하기 때문이다[1]. 또한 성능의 향상을 위해 HTML 태그로 부터 파악 가능한 문서 구조 정보를 이용하여 색인어 가중치를 조절하였다.

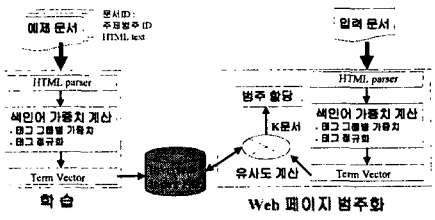

그림 1. 시스템 구성도

최근린법은 두 단계로 이루어진다. 첫 단계에서는 입력 문서와 가장 가까운 k개의 문서를 선택한다. 두번째 단계에서는 추출된 k개의 예제 문서에 할당된 범주들을 근거로 하여 입력 문서에 할당할 범주들을 선택한다. 본 연구의 웹 페이지 범주화 시스템의 기본 구성도는 그림 1과 같다. 그림에서 좌측 부분은 범주화를 위해 예제문서로 학습을 하는 과정이며, 우측 부분은 새로운 문서에 주제 범주를 할당하는 과정이다. 학습과정은 예제 문서를 입력 문서와 비교가 가능하도록 문서의 자질을 추출하고, 색인어에 적절한 가중치를 부여하여 역색인 파일을 구성하는 과정으로 다음과 같다. 먼저 HTML 문서가 들어오면 불용어를 제거하고 텍스트를 어절 단위로 색인기로 보내 명사와 태그만을 구분 해내는 파싱(parsing)을 실시한다. 그리고 문서 빈도(Document Frequency)로 EMIM/MIM 값을 계산하여 주제 범주와 관련이 있는 단어를 문서의 자질로 추출한다. 다음 색인어 빈도(tf, df), 태그 가중치, 태그 정규화 가중치로 문서 내의 특정 색인어의 가중치를 계산하여 역색인 파일을 구성하는 것으로 학습을 완료한다.

학습이 완료되면 그림 2의 우측부분과 같은 과정을 거치면서 새로운 문서에 대한 주제 범주가 할당 된다. HTML 파싱, 색인과정은 학습과정과 동일하며 이때 만들어진 새로운 문서의 벡터 파일과 학습 문서를 사용하여 만든 역색인 파일내의 각 문서들 간의 유사도 계산을 한다. 본 논문에서 사용하는 유사도 계산 방법은 코사인 계수 방법을 이용한다[3].유사도 계산 결과에 따라

상위 k 개의 문서가 추출되면 이 k 개의 문서가 가지는 주제범주들을 이용하여 새로운 문서에 할당한다. 새로운 문서에 할당할 범주를 찾기 위해 아래와 같은 식을 사용한다[12].

$$P(C_k \mid D_x) \approx \sum_{D_j \in \{k \text{ top ranking documents}\}} sim(D_j, D_x) \times P(C_k \mid D_j) \quad \text{----- (1)}$$

식 (1)에서 $sim(D_j,D_x)$는 문서 D_j와 D_x의 코사인 유사도 값이며, $P(C_k|D_j) = 1$ 은 범주 C_k가 D_j에 할당된 경우를, $P(C_k|D_j) = 0$ 은 범주 C_k가 D_j에 할당되지 않은 경우를 나타낸다.

2.2 자질 추출

먼저 색인 작업은 CYK 알고리즘과 비터비 알고리즘을 통합한 포항공대 한국어 색인 시스템[2]을 사용한다. 이 시스템은 하나의 알고리즘으로 띄어쓰기 오류를 수정하면서 복합명사를 단일어로 하고, 처리하는 어절에 미등록어가 있을 경우 이 미등록어를 처리 할 수 있도록 구성되어 있다.

색인어에서 주제 범주를 나타내는 자질을 추출하기 위해 두가지 척도를 사용한다[2]. 먼저 EMIM 을 사용하여 어떠한 범주에서도 같은 역할을 하는 단어들을 제거한다. EMIM 을 척도로 주제 범주와 단어 간의 관계를 학습문서로부터 추출하게 되면, 범주화에 긍정적(positive)으로 도움을 주는 단어와 함께 그 범주와는 전혀 관련이 없어서 부정적인 (negative) 역할을 하는 단어들을 같은 분류로 채택하게 된다. 그런데 최근 린법에서는 부정적인 자질은 사용하지 않으므로 MIM 으로 주제 범주에 긍정적인 단어만을 자질로서 추출한다. 이때 MIM 이 임계치를 넘는 자질만이 색인의 대상이 되며, MIM 을 계산할 때 TF 와 태그 가중치는 고려하지 않는다.

2.3 색인어 가중치 조절

색인어 가중치 부여에는 기존의 TF/IDF 의 빈도 정보에 더하여 HTML 태그 가중치 개념을 도입하였다. 이는 문서를 작성하는 사람이 내용을 이해하기 쉽도록 구성하려고 노력 한다는 것을 전제로 한 것이다. 즉 문서의 내용이 보다 쉽게 이해되도록 하기 위해서 문서의 내용을 큰제목, 소제목 등으로 내부를 구조화하여 배치하고, 이러한 제목에는 문서의 주제를 보다 잘 나타내는 중요한 용어를 사용하리라는 것이다. 그리고 내용 중에 나타나는 중요한 단어는 글자를 크게 하거나 글자체를 다르게 하거나, 글자의 색을 다르게 하는 것도 예상할 수 있다. 또한 주제의 흐름을 훼손하지 않으면서 특정 관심사항에 대한 불연속적인 참조가 가능하도록 문서간 연결 구조를 이용하는데 이때 사용되는 단어도 연결되는 주제를 대표하는 성격을 가지므로 중요한 자질이 될 수 있다. 따라서 태그가 문서의 주제어를 부각시킬 것이라는 점을 고려하면 문서를 나타내는 자질을 보다 잘 표현할 수 있을 것으로 생각된다. 그림 1 은 이러한 개념을 나타낸 것이다. 즉 문서를 대표하는 정도면에서 각 단어가 동일하다는 개념에서 태그 가중치라는 증폭 체계를 사용 하여 중요하다고 생각되는 단어들이 문서를 대표하는 데 더 큰 역할을 하도록 하자는 것으로 문서의 주제어를 부각시키는 효과가 있다.

• 문서의 주제어 부각

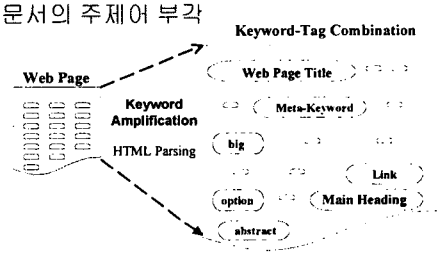

그림 2. 태그 가중치 개념도

태그가중치는 문서의 내용을 대표하는 정도와 일반 용어와 차별화된 정도를 감안하여 전체 38 개의 HTML 태그를 표 1 과 같이 3 개의 그룹으로 나누었다. 개개의 태그에 별도의 가중치를 부여하지 않은 것은 태그별로 성능에 영향을 미치는 정도가 규명되지 않았으며, 처리 속도 측면도 고려한 것이다.

표 1.태그 가중치 테이블

구분	가중치	태그
그룹 1	4	<title> 등 4 개
그룹 2	3	<a><b> 등 34 개
그룹 3	1	기타

<title> : HTML 문서의 제목
<a> : 문서 내외부 연결관계 표시
<b> : bold 글자체라는 의미

한편 이렇게 적절하게 태그를 사용하는 경우 이외에 태그를 무분별하게 사용하는 경우도 충분히 예상할 수 있다. 그래서 본 연구에서는 태그를 적절하게 사용하는 문서 작성자를 위해

태그별로 가중치를 부여하고, 남용되는 경우에 대비하여 태그 출현 빈도를 정규화하였다. 정규화 기준은 문서별 태그 분포와 전체 학습문서의 태그 분포를 비교하여 태그 가중치를 적절하게 조절하는 것으로 식 (2)와 같다.

$$N(tag_i, D_j) = \frac{\text{태그}(tag_i)\text{평균빈도}(\%)}{\text{문서}D_j\text{에서 태그}(tag_i)\text{빈도}(\%)} \quad (2)$$

이 식은 문서 내 특정 태그가 전체 학습문서 집합에서의 평균 빈도보다 많이 사용된 경우에 사용되며 반대인 경우에는 적용하지 않는다. 예를 들면 <a> 태그의 가중치가 3 이고, 학습 문서에서는 전체 태그 집합에서 <a> 태그의 사용 비율이 10% 인데, 범주화를 위해 입력된 문서에서는 20%가 사용되었을 경우, 10/20 = 0.5 가 되어 <a> 태그가 가지는 가중치는 처음 설정된 값의 1/2, 즉 3*0.5=1.5 가 된다. 반대인 경우에는 20/10=2 가 되어 1 보다 큰 값을 가지지만 2 배로 하면 증폭되는 정도가 너무 심한 것으로 생각되어 최초에 설정된 가중치 3 을 그대로 사용한다. 즉 계산 결과가 1 보다 크게 되면 그대로 1 로 두고, 1 보다 적으면 계산 값을 해당 태그의 가중치에 곱하는 것으로 한다.

최근린법에서는 새로운 문서에 주제범주를 할당하기 위해 학습 문서와 입력 문서간의 유사도를 비교한다. 이때 색인어 가중치 부여 방법이 문서간 유사도 계산에 중요한 역할을 한다. 본 논문에서는 문서 D_j에 나타난 색인어 t_k의 가중치 $W_{Dj,tk}$는 단어의 빈도와 함께 식 (3)과 같이 태그 가중치와 태그 정규화 가중치를 추가하여 식 (4)와 같이 정의한다

$$tf_{t_k} = \sum_{i=1}^{n} \alpha(tag_{t_{k,i}}) \times N(tag_{t_{k,i}}, D_j) \quad \text{---- (3)}$$

$$W_{D_j,t_k} = \frac{\log(0.5 + tf_{t_k})}{\log(1.0 + \max \text{tf})} \times \frac{\log N/df_{t_k}}{\log N} \quad \text{--- (4)}$$

$\alpha(tag_{t_{k,i}})$: i번째나타난색인어t_k에붙은태그의 가중치

$N(tag_{t_{k,i}}, D_j)$: 문서D_j에서 i번째 나타난 색인어 t_k에 붙은 태그의 정규화 가중치

n : 색인대상 단어의전체 갯수

N : 학습 문서의 수

df : 색인어 문서 빈도

$\max_{tf}$:특정문서에서 가장 많이 나타나는 색인어의빈도수

4. 실험 및 평가

실험에는 그림 3 과 같이 전북대에서 정보탐정에서 사용하도록 만든 웹 사이트 분류 데이터를 사용하였다. 웹 페이지는 각 웹 사이트의 홈페이지를 대상으로 하였다.

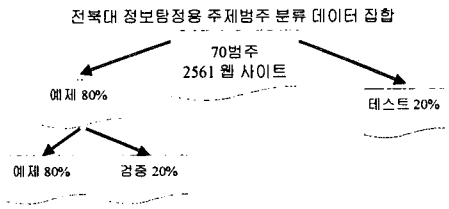

그림 3. 실험용 데이터 구성

실험을 위한 모델 구성은 표 2 와 같다. 즉 모델 1 은 색인어 TF/IDF 만을 사용한 기존의 모델이며, 모델 2 는 태그 가중치를, 모델 3 은 EMIM+MIM 척도에 의한 자질 추출 과정을 거친 것이며, 모델 4 는 자질 추출 과정을 거치고 태그 가중치 부여한 것이다.

표 2 실험을 위한 모델 구성

구 분	자질추출	태그
모델 1	X	X
모델 2	X	O
모델 3	O	X
모델 4	O	O

평가 기준으로는 정보 검색에서 자주 사용하는 11 point recall 에서의 precision 을 평균한 Average Precision 을 사용하였다.

실험 순서는 최적 계수 도출을 위한 검증과 테스트로 나누어 실시하였다. 실험 데이터 배분은 그림 3 과 같이 전체 데이터의 20%를 실험용으로 하였으며, 80% 학습 데이터중에서 다시 20 %를 검증에 사용하였다. 여기에서는 웹 사이트를 구성하는 여러 웹 페이지들의 구조에 대한 고려는 하지 않고 하나의 웹 페이지만을 실험의 대상으로 하였으나 홈 페이지에서 추출된 자질의 수가 25 개를 초과하면 그대로 사용하였고, 25 개 이하이면 최대 20 개 까지 하부 페이지를 추가하여 사용하였다. 이것은 문서를 표현하는데 필요한 최소한의 자질을 확보하기 위한 것이다.

자질 추출을 위한 MIM 의 임계치는 주제범주-자질-MIM 분포 그래프를 참고로 하여 0.0022 로 설정하였다.

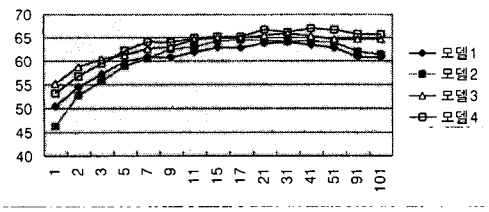

그림 4. K 값 결정을 위한 검증 실험 결과

그림 4는 모델별로 최적의 K값을 찾기 위한 실험 결과이다. 모델 1 : 31, 모델 2 : 17, 모델 3 : 31, 모델 4 : 41 에서 각각 가장 좋은 성능을 나타내었다.

표 3은 위의 실험에서 결정된 K값으로 학습문서를 100% 사용하여 실험한 결과이다. 학습문서의 수가 약 16 % 증가했는데 결과는 모델 4를 기준으로 약 3% 향상되었다. 실험 결과는 새로운 요소를 추가한 모델들이 점점 나아지는 결과를 보여 주었지만 차이가 2% 내외로 기대에 미치는 것은 아니었다. 몇 가지 요인을 생각해 볼 수 있다.

표 3. 모델별 실험 결과

구분	Avr-Precision	K 값
모델 1	65.91	31
모델 2	66.22	17
모델 3	68.71	31
모델 4	70.47	41

먼저 위의 실험에서 사용한 태그 가중치와 MIM 임계치는 모두 직관에 의한 것이다. 따라서 가장 최적의 값이 선정되었다고 볼 수 없다. 실제로 <title>의 가중치를 변화시켜 가면서 실험한 결과 그림 5와 같은 결과를 얻을 수 있었다. 태그 가중치로 인해 성능이 변화하는 모델 2와 모델 4에서 <title> 태그가 가져야 할 가장 적절한 가중치는 7이 된다. 태그가 성능에 미치는 영향이 독립적이라고 가정하면 태그별로 위의 실험과 같은 최적 계수 도출 과정을 거치면 더 좋은 결과를 얻을 수 있으리라 생각 된다.

그리고 태그의 사용이 문서의 표현에 있어서 특정 부분의 단어 영향력을 증폭시키는 측면을 고려하면 불용어에 대한 배려가 필요하리라 생각된다. '*디지털 조선일보*'라는 문장에 <title> 태그가 붙어 있는 경우를 생각 해보자. 이 문장의 색인 결과는 '*디지털*', '*조선*', '*일보*'라는 세 가지 단어가 될 것이다. 먼저 '*조선*'이라는 고유명사는 자질 추출 과정에서 색인 대상에서 사라지지만 '*디지털*'은 '*신문*'이라는 주제범주와 같이 증폭되어 도리어 문서 표현을 왜곡시키는 결과를 가져오게 된다. '*디지털*'이라는 자질이 사용되는 주제범주에서 '*디지털*'이라는 자질을 포기하더라도 크게 문제가 되지 않는다면 불용어 처리과정에서 삭제하여도 무방하다 할 것이다. 물론 이러한 처리의 대상이 되는 단어는 인터넷이라는 특수한 환경에서만 예상 가능한 것이어야 할 것이다. 상용 시스템을 지향 한다면 이러한 방법으로 지속적인 성능 향상을 꾀하는 것도 하나의 대안이 될 수 있으리라 생각된다.

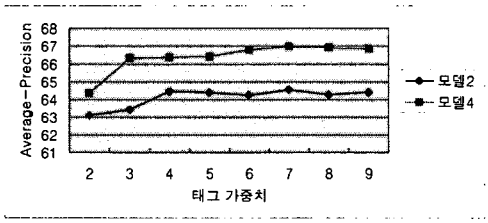

그림 5. 태그 가중치 변화에 따른 결과

5 결론 및 향후 계획

본 논문에서는 문서 작성자의 입장에서 중요하다고 생각되는 순서로 태그 가중치를 부여하여 문서 표현이 범주화에 미치는 영향을 알아 보고자 하였다. 실험 결과 태그 가중치가 웹 페이지 범주화에서 성능에 영향을 미친다는 것을 확인하였으며, 태그별 최적 가중치를 실험을 통해 구하여 적용한다면 더 좋은 결과를 얻을 수 있을 것으로 생각된다. 그리고 무분별한 태그 사용에 대비하여 태그 정규화 처리는 반드시 해야 한다는 결론을 얻었다. 시스템의 성능에 학습문서의 양도 중요한 인자가 된다는 사실도 확인하였다.

보다 더 문서를 잘 표현하는 방법으로 현재 많이 이야기되고 있는 주제 범주와 관련이 없는 부정적 자질(Negative Feature)을 범주화에 활용하는 방법과 여기에 문서 구조정보를 결합하는 방법에 대한 연구도 필요할 것으로 생각된다.

참고문헌

[1]권오욱, 이종혁, 이근배 "Nearest Neighbor 방법을 이용한 문서 범주화에서 범주 자질의 평가", 제9회 한글 및 한국어 정보처리 학술대회, pp7-14, 1997.

[2]권오욱, 이종혁, "웹 검색 서비스용 자동 문서분류 시스템", 한국통신 '97정보통신 기초 연구과제 최종 보고서, 포항공과대학교 정보통신연구소, 1997.

[3] Brij Masand, Gordon Linoff and David Waltz, "Classifying News Stories using Memory Based Reasoning", In Proceedings of the 15th Annual International Conference on Research and Development in Information Retrieval(SIGIR'92), pp.59-65, 1992.

[4] C. J. Van Rijsbergen, "A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval" , pp.106-119, Journal of Documentation, Vol.33, No2, June 1977.

[5] Childanand Apte, Fres Damerau and Sholom M. Weiss, "Towards Language Independent Automated Learning of Text Categorization Models", In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp23-30, 1994.

[6] David D. Lewis and Marc Ringuette, "A Comparison of Two Learning Algorithms for Text Categorization" ,In Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, pp.81-93, 1994.

[7] David D. Lewis, "Representation and Learning in Information Retrieval", PhD thesis, Department of Computer Science;Univ. of Massachusetts; Amherst, MA01003, 1992.

[8] Makoto Iwayama and Takenodu Tokunaga, "Cluster-Based Text Categorization:A Comparison of Category Search Steategies", In Proceedings of the 18th Annual International Conference on Research Development in Information Retrieval (SIGIR'95), pp.273-280, 1995.

[9] Philip J. Hayes, Laura E. Knecht and Monica J. Cellio, "A News Story Categorization System", In Proceedings of the 2nd Conference on Applied Natural Language Processing, pp.9-17, 1988

[10] Philip J. Hayes, "Intelligent High-Volume Text Processing Using Shallow, Domain-specific Technique", In Paul S. Jacobs, editor, Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval, pp.227-241, Hillsdale, New Jersey, 1992.

[11] Philip J. Hayes, P.M. Andersen, I.B. Nirenburg and L.M. Schmandt, "TCS: A Shell for Content-Based Text Categorization", In Proceedings of the 6th IEEE AI Applications Conference, 1990.

[12]Yiming Yang, "Expert Network : Effective and Efficient Learning from Human Decision in Text Categorization and Retrieval", In Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp13-22, 1994.