한국산학기술학회논문지
Vol. 11, No. 2, pp. 517-524, 2010

ETOM+RPost기반의 문서분류시스템의 설계 및 구현

최윤정[1*]
[1]서일대학 정보통신과

Design and Implementation of Text Classification System based on ETOM+RPost

Yun-Jeong Choi[1*]
[1]**Department of Information Communication, Seoil University**

요 약 최근의 컴퓨터 기술과 인터넷 기술의 발달로 인해 분석 데이터가 급속도로 증가함에 따라 이들을 다루기 위한 자동분류시스템에 대한 요구가 높다. 문서분류시스템은 감독학습이 필수적이기 때문에 최소한의 전문가의 개입만으로도 높은 정확도가 보장되는 자동화 시스템에 대한 요구가 크다. 반면, 분류할 데이터들은 형식이나 내용상으로 그 복잡도가 높아지고 있어서, 일반적인 분류방법으로는 좋은 분석결과를 얻기 어려운 양상을 보인다. 특히 스팸성 데이터와 같이 어떠한 의도가 반영되어 가공되거나 변형되는 데이터는 분석의 어려움을 가중시킨다. 본 논문에서는 분류알고리즘의 성능향상을 위해 제안한 ETOM과 RPost방법을 구현하였다. 분류의 경계선 상에 있는 스팸문서들에 구현시스템을 적용하여 그 과정을 분석하였다. 실험결과 제안방법에 의한 정확도가 0.795에서 0.93으로 약 16%의 증가하였음을 확인하였다.

Abstract Recently, the size of online texts and textual information is increasing explosively, and the automated classification has a great potential for handling data such as news materials and images. Text classification system is based on supervised learning which needs laborous work by human expert. The main goal of this paper is to reduce the manual intervention, required for the task. The other goal is to increase accuracy to be high. Most of the documents have high complexity in contents and the high similarities in their described style. So, the classification results are not satisfactory. This paper shows the implementation of classification system based on ETOM+RPost algorithm and classification progress using SPAM data. In experiments, we verified our system with right-training documents and wrong-training documents. The experimental results show that our system has high accuracy and stability in all situation as 16% improvement in accuracy.

Key Words : Machine Learning, Text Classification System, Learning Algorithm, Feedback System

1. 서론

최근 정보의 양이 급증함에 따라 효율적인 정보관리 및 검색기능이 요구되고 있다. 문서들의 성격은 과학논문이나 의료정보문서와 같이 복잡하고 쉽게 이해하기 힘든 특성을 띄고 있어서 분석의 어려움을 가중시킨다. 이러한 추세는 한층 더 높아진 사용자의 요구사항으로 반영되는데 최근에는 전문가 개입을 최소화하면서도 더 지능적이고 높은 수준의 정확도와 성능을 보장할 수 있는 시스템에 대한 관심이 높다[1,2]. 본 논문에서는 이전 연구에서 제안한 확장된 분류체계를 이용한 학습알고리즘(ETOM, Expanded Training Set Organization Method)과 강화된 후처리분석(RPost, Reinforcement Post-Processing)방법을 이용하여 자동문서분류시스템을 구현한다[3]. 본 연구에서 향상시키려는 성능의 대상은 목표항목으로 정확히 분류되는 정확성(accuracy)과 학습과정의 오류에서도 옳게 분류되는 안정성(stability)이다. 구현시스템의 성능평가를 위해 정확히 가려내기 힘든 스팸문서에 적용하여 정확성

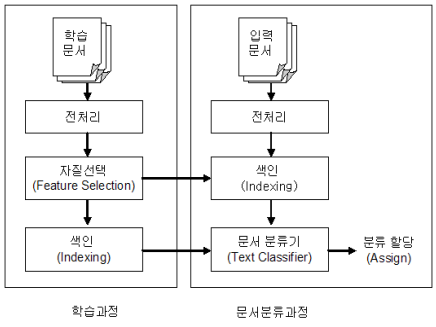
*교신저자 : 최윤정(cris@seoil.ac.kr)
접수일 09년 10월 23일 수정일 10년 01월 30일 게재확정일 10년 02월 24일

과 안정성을 확인하였다.

본 논문은 다음과 같이 구성된다. 2장에서는 일반적인 자동분류시스템에 대해 살펴본다. 3장에서는 ETOM+RPost 알고리즘에 대해 간략히 설명하고, 이를 기반으로 구현한 시스템의 설계와 구현방법을 보인다. 4장에서는 구현시스템을 스팸문서들에 적용한 분류과정을 보인 후, 5장에서 결론을 맺는다.

2. 자동문서분류시스템

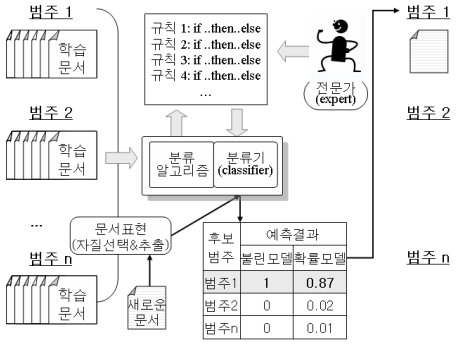
자동문서분류란 문서의 내용에 기반하여 정의된 범주에 문서를 자동으로 할당하는 기법과 관련된 연구분야로 효율적으로 문서를 관리하고 검색할 수 있게 하는 동시에 방대한 양의 수작업을 감소시키는 것이 목적이다[3,4]. 그림 1은 전통적인 문서분류시스템을 도식화한 것으로 주요작업을 요약하면 다음과 같다.



[그림 1] 전통적인 문서 분류 시스템

- 전처리(preprocessing) : 수집된 문서를 기계가 처리할 수 있도록 변환하고 문서의 내용이나 특징을 반영하는 내용어(content word)를 추출한다.
- 자질선택(feature selection) : 전처리로 얻은 주요 용어들 중, 학습에 유용하게 사용될 용어(feature)만을 선택한다.
- 문서색인(indexing) : 선택된 자질을 통해 어떻게 문서를 표현(document representation)할 것인가에 대한 색인 작업이 이루어진다.
- 문서분류(classification) : 학습문서와 분류알고리즘을 이용하여 얻은 분류규칙으로 새로운 문서를 분류한다.

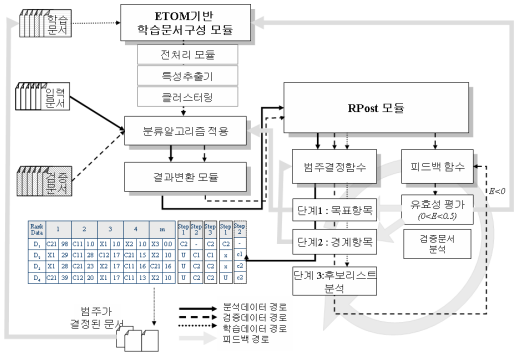
그림 2는 자동문서분류시스템의 기본 개요를 나타낸 것으로 기본과정인 학습과정과 범주지정과정을 자동화시켜 구현한다. 문서분류시스템은 감독학습(supervised learning)에 의한 학습과정을 전제로 하며 학습과정과 분류과정으로 나뉜다. 학습과정에서 각각의 범주에 속하는 학습문서들로부터 주요 자질들을 구별해내면, 분류기는 새로 입력된 문서와 범주간의 유사성을 계산한다[5,6]. 문서분류를 위해 사용되는 분류기는 대부분 정보검색모델에 기반하고 있는데, 문서표현에 사용된 자질들의 가중치 계산방법에 따라 불리언모델, 벡터모델과 확률모델로 구분된다[4,5,7]. 불리언모델에서는 문헌과 질의가 색인어의 집합으로 표현되며 집합이론에 근거한다. 벡터모델은 t차원공간의 벡터로 표시되어 대수적 모델이라 불리며 확률모델은 확률론에 근거한다.



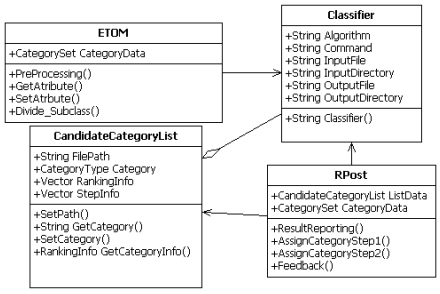
[그림 2] 자동문서분류시스템의 기본 개요

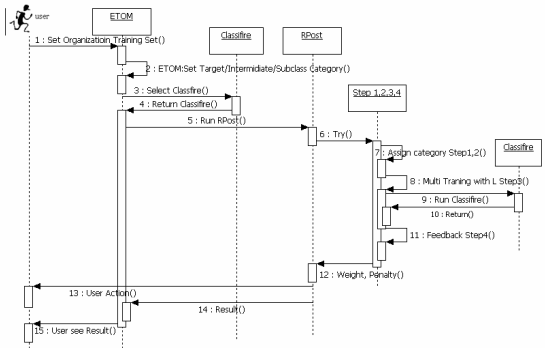
3. ETOM+RPost 기반 문서분류시스템의 설계 및 구현

ETOM+RPost 기반 자동문서분류시스템은 그림 3과 같이 동작한다. 경계범주를 자동으로 탐색하여 분류체계를 확장시키는 알고리즘인 ETOM 시스템은 분류기 적용을 위한 전처리 역할을 하며, RPost 시스템은 분류기의 분류결과를 후처리하는 역할을 한다[3]. 그림 4는 ETOM+RPost의 클래스 다이어그램과 시퀀스 다이어그램을 나타낸다. 본 논문에서는 다양한 분류알고리즘이 잘 구현되어있는 BOW와 SVM-light를 연계하였다[6,8]. BOW는 CMU(Univ. of Carnage Mellon)에서 개발되었으며 SVM-light은 Joachims에 의해 개발된 것으로 텍스트 분류 실험에 주로 사용된다[9-11]. 경계범주와 세부범주 탐색을 위한 클러스터링 분석을 위해 미네소타 대학에서 개발된 CLUTO-클러스터링 알고리즘을 연계하였다[13]. ETOM과 RPost 시스템은 JAVA로 구현하였다.



[그림 3] ETOM + RPost 기반 자동분류시스템의 흐름도





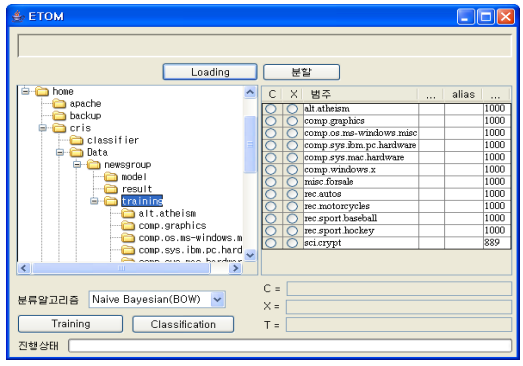
[그림 4] ETOM + RPost 시스템 설계를 위한 클래스 다이어그램과 시퀀스 다이어그램

▪ **ETOM 시스템 : 확장된 분류체계에 의한 학습문서 집합 구성**

ETOM의 사용자인터페이스는 그림 5와 같다. 사용자는 미리 정의되어 있는 분류체계와 학습문서를 로딩하여 경계범주를 탐색한 후 분류알고리즘을 선택하여 분류모델을 만든다. ETOM 시스템은 문서의 헤더와 태그를 제거할 수 있는 전처리모듈과 계층분류를 위한 목표범주 분할 모듈, 분류알고리즘 선택 모듈 등으로 구성되어있다. 사용자는 각 목표범주를 디렉토리로 구분해야 하며, 각각의 학습문서들은 해당 디렉토리에 개별 파일로 저장되어 있어야 한다.

▪ **RPost 시스템 : 강화된 후처리분석에 의한 범주결정 및 피드백**

RPost 시스템은 초기 분류결과들의 개요정보를 보여주는 리포팅 모듈과 범주결정(단계 1~단계 3) 모듈, 그리고 피드백(단계 4)모듈로 구성된다. 이 과정에서는 초기 결과에 대해 후처리 분석을 진행하여 분류결과들의 변화 상태를 요약한다. RPost의 피드백과정의 목적은 학습체계와 범주결정방법이 유효한지 평가하고 어떠한 방법으로 피드백 할 것인지 제시하는 것이다.



[그림 5] ETOM의 사용자인터페이스

이와 같이 피드포워드 제어를 통해 각 단계별 수치의 범위를 변경하여 적용하면 범주의 재결정 결과를 즉시 얻을 수 있다. 높은 신뢰도를 가지고 결정된 문서들은 새로운 학습문서로 사용되도록 사용자에게 제시하거나 자동으로 추가할 수 있도록 재학습-재분류 모듈로 연결된다. RPost 시스템의 주요모듈은 4장의 실험과정과 함께 설명하기로 한다.

4. 실험 및 성능형가

4.1 실험계획

제안방법의 성능을 잘 나타낼 수 있도록 분류가 어려운 문서집합을 사용하였다. 실험 데이터는 아파치社의 스팸프로젝트(Apache Spam Assassin Project)의 문서들이다[13]. 스팸메일을 분류하다 보면, 업무상 관련된 메일도 스팸처리 되는 일이 발생한다. 본 실험에서는 스팸스타일의 일반문서와 필터링 규칙을 통과하는 스팸문서가 올바로 분류되는지를 확인한다.

1) 실험 데이터

아파치의 스팸문서 집합은 다음과 같이 구성된다. 여

기서 'ham'은 'spam'이 아닌 일반문서를 뜻한다. 그림 6은 서버의 필터링 시스템을 통과한 문서의 예를 보인다. 이처럼 시스템에서 필터링 되는 용어를 피해 교묘히 작성된 문서들은 올바로 분류되기 힘들다.

- spam : 필터링 규칙을 통과한 스팸문서(non-spam-trap, 약 500개).
- easy_ham : 스팸문서와 매우 쉽게 구별되는 일반문서들(약 2500개).
- hard_ham : 스팸문서 스타일의 일반 문서(약 250개).
- easy_ham_2: non-spam 메시지(약 1400개)
- spam_2 : 스팸 메시지(약 1400개)

4.2 실험방법

1) 학습문서집합 구성

확장된 목표범주에 따라 학습문서의 구성과 수를 표 1과 같이 구성한다. 목표범주 C= {spam, ham}가 주어졌을 때 경계범주는 X={hard_spam, hard_ham}로 구성된다.

2) 실험 조건

기존방법과 제안방법의 성능비교를 위한 실험 조건은 표 2와 같다. 자질들을 교란시키기 위해 일반문서와 스팸문서를 교차지정 하여 오류상황을 만들었고, 분류알고리즘으로 Naïve Bayesian(NB)을 적용하였다

Title : still waiting for your reply
Hi! I am Ekaterina. I am a kind, sociable woman.
I have a lot of tenderness inside my soul which waits for somebody to be given to. I dream someone to steal my heart one day - I want to love and to be loved! ...
1/ Are you interested in serious relations with Russian woman?
2/ Are you planning to visit Russia?
3/ Would you like to correspond or to talk by phone?
4/ Why are you interested in Russian lady?
5/ Have you ever been to Russia?
6/ What is important for you in relations and am I right for you?
I will be waiting for your reply to admin@1WIFEFORU.INFO
Would you like to know me better and to meet me?
Waiting for your reply, Ekaterina.

[그림 6] 필터링 시스템을 통과한 스팸문서의 예

[표 1] 확장된 분류체계에 의한 학습문서집합의 구성

확장된 분류체계			학습문서 개수		
목표범주 (C)	하위항목 (subclass)	경계범주 (X)	학습 문서	오류 (약 10%)	총 문서 수 (500, 550)
Spam	S1		100	(10)	200(20)
	S2		50	(5)	
	S3		50	(5)	
	X1	hard_spam	100	(10)	100(10)
	X2	hard_ham	100	(10)	100(10)
Ham	H(일반)		100	(10)	100(10)

[표 2] 정확성/안정성 비교를 위한 실험조건

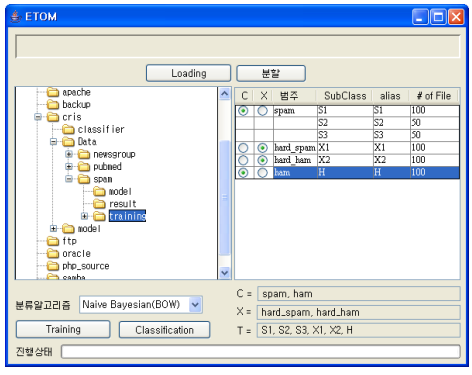
	실험	분류체계		실험조건		
		목표범주(C)	경계범주(X)	분류기	오류포함	후처리 분석
기존 방법	E1	spam= {S1,S2,S3,X1}, ham= {H,X2}	-	NB	X	X
	E2			NB	O	X
제안 방법	E3	spam={S1,S2,S3},ham ={H}	hard_spam={X1}, hard_ham={X2}	NB	X	O
	E4			NB	O	O

4.3 구현시스템을 이용한 분류실험

ETOM+RPost 기반의 분류시스템을 이용한 스팸문서의 분류과정을 보인다. 결과확인을 위한 검증문서는 학습에 사용되지 않은 문서들로 경계범주와 목표범주의 비율을 7:3으로 구성하였고, 경계범주에 포함된 문서의 분류과정을 관찰하였다.

1) 학습과정의 오류가 없는 환경에서의 분류

그림 7은 ETOM 시스템의 사용자 인터페이스로써 학습문서집합을 구성하고 분류알고리즘을 적용하여 분류모델을 만든 후, 검증문서를 분류기에 적용시켜 분류 결과값을 얻는다. 그림 8은 분류의 결과값으로 문서의 순위별 범주와 순위값들의 리스트이다. 이 때 순위값은 정규화시켜 표현도록 하였다.



[그림 7] 스팸문서 분류를 위한 학습문서집합구성

Data	1		2		3		4		5	
269046.txt	H	0.79	X2	0.07	S2	0.06	X1	0.04	S3	0.04
1922737.txt	H	0.78	S3	0.09	X1	0.07	X2	0.03	S1	0.03
954946.txt	H	0.71	X2	0.11	S1	0.08	X1	0.07	S3	0.03
1611373.txt	H	0.69	S1	0.11	S3	0.08	X2	0.06	X1	0.06
214392.txt	H	0.69	S3	0.09	S2	0.08	S1	0.07	X2	0.06
134338.txt	H	0.67	X2	0.11	S3	0.08	S2	0.07	X1	0.07
637102.txt	H	0.62	X1	0.11	S2	0.10	X2	0.09	S3	0.08
689287.txt	H	0.61	S2	0.16	X2	0.09	S3	0.07	X1	0.06
902771.txt	H	0.61	S2	0.12	S1	0.11	X1	0.09	S3	0.06
1580627.txt	H	0.56	S1	0.16	X1	0.11	S2	0.09	S3	0.08
165874.txt	S3	0.53	X2	0.23	S1	0.11	X1	0.09	H	0.04
4636.txt	S3	0.53	S2	0.16	S1	0.13	X1	0.10	X2	0.08
1568271.txt	S3	0.53	H	0.14	X2	0.13	S1	0.10	S2	0.10
1049798.txt	H	0.52	S2	0.17	X1	0.12	S3	0.09	X2	0.09
23156.txt	S3	0.49	X1	0.20	S2	0.16	S1	0.08	H	0.07
253431.txt	S3	0.49	X2	0.18	S2	0.12	H	0.11	S1	0.10
1309353.txt	S3	0.49	H	0.14	X1	0.13	S2	0.13	X2	0.11
1006816.txt	X1	0.48	H	0.16	X2	0.15	S1	0.13	S2	0.09
1920985.txt	X1	0.47	S3	0.22	H	0.13	S2	0.11	X2	0.07
1960798.txt	S1	0.47	H	0.21	X1	0.14	S2	0.12	S3	0.06
603672.txt	S2	0.47	X1	0.21	H	0.15	X2	0.09	S1	0.08
1436672.txt	X1	0.47	S2	0.17	H	0.16	X2	0.10	S3	0.10
1047127.txt	S3	0.47	H	0.16	X1	0.15	X2	0.13	S2	0.09
1944233.txt	S3	0.47	S1	0.16	X2	0.15	H	0.12	S2	0.11

[그림 8] 검증문서를 입력으로 한 분류기의 결과값

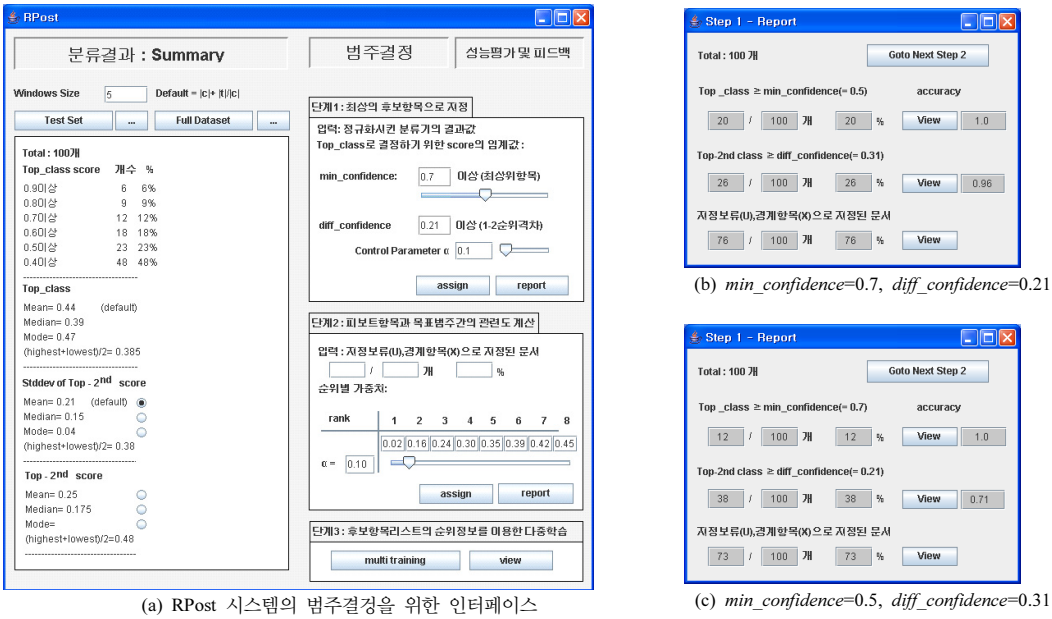
■ 단계 1의 분석과정

그림 9는 후처리분석을 위한 RPost 시스템의 초기화면으로 분류 결과값에 대한 개요를 나타낸다. 그림 9의 (a)는 최상위항목들의 분포와 대표값, 1순위와 2순위 수치값에 대한 편차와 대표값을 보임으로써, 높은 신뢰도를 갖는 문서를 우선 분류하기 위한 임계값의 가이드라인을 제시한다. 사용자는 이 개요정보를 통해 현 분류상황을 파악하여 각 임계값을 정한다. 이 때의최상위 후보항목이 0.5이상의 확신을 갖는 경우는 23개로 나머지 70%의 문서들은 목표범주로 분류되기 불확실한 상태이다. 검증문서들의 각 범주별 편차의 평균은 약 0.2로 각 범주들과의 관련이 작으므로 오분류 될 가능성이 매우 높은 문서들임을 의미한다.

RPost 시스템의 범주결정과정 중 단계 1에서는 분류결과의 신뢰도가 높은 문서를 우선 지정할 수 있도록 최상위항목의 수치값의 하한값인 *min_confidence*와 1순위와 2순위 격차의 하한값인 *diff_confidence*의 값을 정하고 있다[2]. 본 시스템에서 *min_confidence*의 값은 임의로 0.7, *diff_confidence*는 0.21로 설정되어있다. 이는 1순위와 2순위의 관련도 차이를 0.21 이상으로 설정해야 함을 의미한다. 그림 9의 (b)는 위의 값을 만족하는 문서들에 대해 범주가 결정된 상태를 나타낸다. 여기서 최상위항목이 0.7이상의 신뢰도를 갖는 문서는 12건이며 정확도는 1로써 이 문서들은 모두 정확한 범주로 지정되었다. 한편, 1순위와 2순위 격차값이 0.21이상인 문서 26 건의 범주가 추가적으로 결정될 수 있으나 정확도는 약 0.7로서 38개의 문서 중 약 10개 이상의 문서가 오분류되거나 미분류 되었음을 보인다. 그림 9의 (c)는 *min_confidence*를 0.5, *diff_confidence*를 0.31로 정한 이후의 범주가 결정된 문서들의 상황을 나타내고 있다. *diff_confidence*를 상향조정한 후 범주가 결정된 문서의 수는 줄었지만 오분류율은 약 3% 감소되었다.

■ 단계 2의 분석과정

단계 1에서 지정이 보류된 문서들과 경계항목으로 결정된 문서들은 단계 2의 입력이 된다. 이 과정에서는 경계항목을 피봇으로 하고 각 목표항목과의 관련도를 계산하고 보다 가까운 쪽의 항목으로 지정되도록 한다. X1이 피봇이 되고 이와 가까운 범주인 스팸범주가 선택되거나 X2가 피봇이 되고 일반문서로 범주결정이 되는 상황은 제안방법이 추구하고 있는 방향이다. 한편, X2이 피봇일 때 스팸문서로 결정되거나 X1이 피봇일 때 일반문서로 결정되는 상황은 현재의 학습문서가 범주를 잘 대표하고 있지 않음을 나타낸다. 그림 10은 단계 2를 수행한 후의 분류결과를 나타낸다. 선택된 피봇항목과 각 목표범주와의 거리계산을 통해 범주가 결정된 문서들의 개수와 정확도를 표시한다. 여기서 정확도는 0.91로써 76개의 문서 중 약 7개의 문서가 오분류 되어있음을 보인다.



(a) RPost 시스템의 범주결정을 위한 인터페이스

(b) *min_confidence*=0.7, *diff_confidence*=0.21

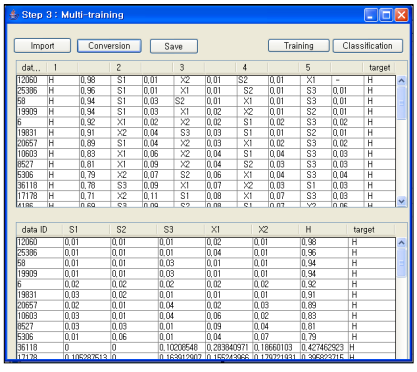
(c) *min_confidence*=0.5, *diff_confidence*=0.31

[그림 9] RPost 시스템의 인터페이스와 단계 1의 임계값 조율에 따른 결과 리포팅

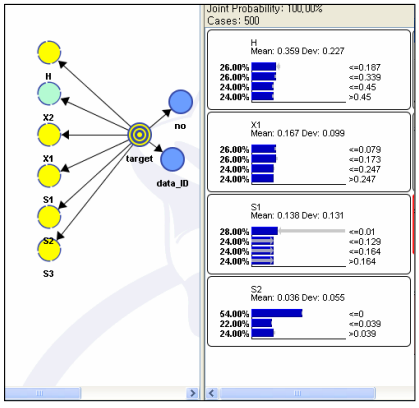
[그림 10] RPost 시스템의 단계 2의 임계값 조율에 따른 결과 리포팅(*min_confidence*=0.5, *diff_confidence*=0.31)

■ 단계 3의 분석과정

그림 11은 본 시스템의 단계 3으로써, (a)의 다중학습 인터페이스를 이용하여 후보항목리스트의 형태를 학습에 적당한 양식으로 변환한 후, (b)의 분류알고리즘을 적용하게 된다. 다중학습에 의한 결과는 앞 장에서 설명한 바와 같이 시스템의 성능평가에서 안정성을 측정하는 요소로 사용된다.



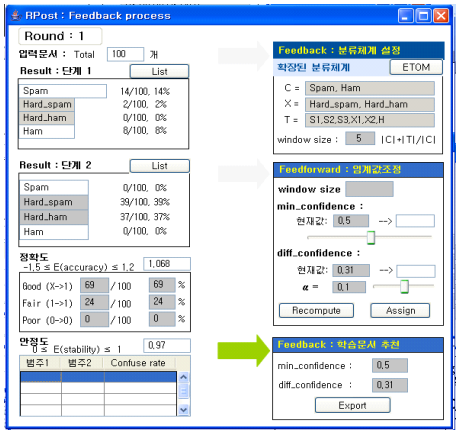
(a) 단계 3의 사용자 인터페이스



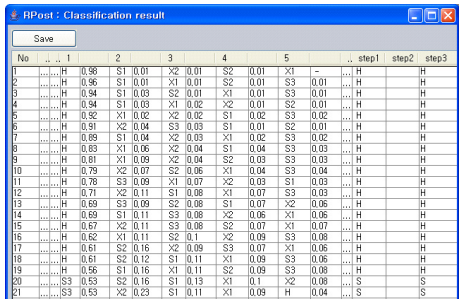
(b) 문서의 후보항목리스트를 이용한 다중학습과 분류

[그림 11] RPost 시스템의 단계 3의 인터페이스&분석

그림 12의 (a)는 후처리분석의 피드백과정으로 범주가 지정된 문서들의 최종적인 보고를 작성하고 성능평가 결과에 따른 피드백방법을 제시한다. 오류가 없는 학습체계 하에서 범주가 결정된 문서들의 정확도는 매우 높은 수치를 보인다. 높은 신뢰도를 갖는 문서에 대한 피드백방법으로 새로운 학습문서로 지정하도록 하는 사용자의 소극적 개입이 제시된다. 구현시스템에서는 학습문서 설정을 위하여 그림 12의 (b)의 인터페이스를 제공한다. 피봇항목과 각 목표항목간의 계산결과에 따른 리스트를 연결시킴으로써 사용자가 편리하게 확인할 수 있도록 하였다.



(a) 시스템 성능평가에 다른 피드백



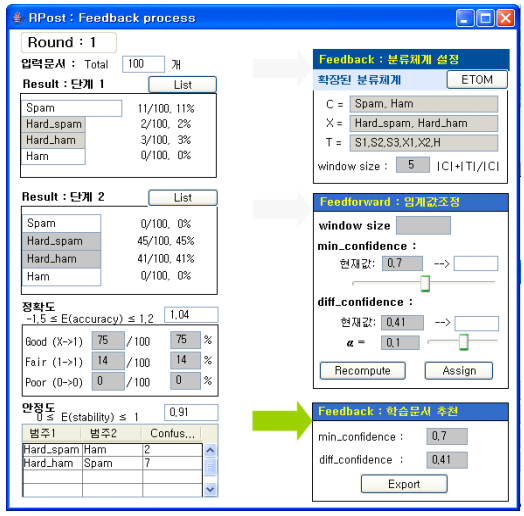
(b) 결정된 범주에 높은 확신을 갖는 문서들을 이용한 학습 예

[그림 12] 피드백과 재학습을 위한 인터페이스

2) 학습과정의 오류가 있는 환경에서의 분류

다음은 약 10%의 학습 오류를 포함하고 있을 때의 분류과정을 보인다.

(a) 학습오류 하에서의 분류과정

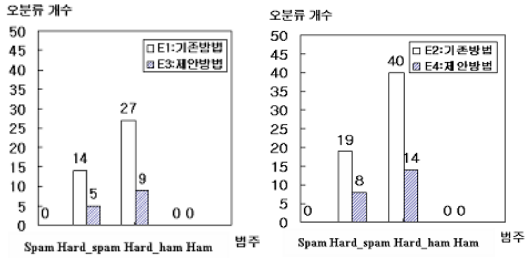


(b) 시스템 성능평가에 다른 피드백

[그림 13] 학습문서에 오류 하에서의 범주지정결과 및 피드백 과정

그림 13은 범주결정과정 단계 1과 단계 2에서 범주가 지정된 문서들에 대한 요약 정보를 나타낸다. 단계 1에서 임계값은 *min_confidence* = 0.7, *diff_confidence* = 0.4 로, 결정되는 문서의 개수는 총 14개에 불과하였다. 이전의 오류가 없는 환경과 비교하여 각 범주간 관련도의 편차가 극명히 나타나거나 높은 확신을 갖는 문서도 드물다. 단계 2에서는 86개 중 약 75개의 문서에 대해서만 옳게 분류되었고 11개는 오분류되었다. 그림 13의 (b)는 입력문서에 대한 성능평가 결과와 피드백 방법을 제시한다. 유효성 평가함수에서 분류결과의 정확도와 학습체계의 안정성을 측정하기 위해 각 문서의 단계별 결과의 일치도를 계산하였다. 정확도는 결과값을 알고 있는 검증문서에 대하여 표시되며 -1.5 ~1.2의 범위를 가진다. 본 논문에서는 단순문서와 불확실한 문서가 같은 비율로 구성되었을 때 불확실한 문서가 모두 오분류되면 0.5의 값을 가지도록 설정되어 있다. 0 이하의 값을 갖는 경우는 경계영역에 속하지 않는 단순한 문서들의 오분류율이 더 높을 경우이다.

이 실험에서 성능평가 결과는 E(accuracy)=1.04, E(consistency)= 0.91 로 나타난다. 단계 1에서 미분류되거나 경계범주로 지정된 문서 중 9건의 단계별 결과가 일치하지 않음을 나타낸다. 이 때 상반된 결과값을 나타내는 범주들은 'hard_ham'/'spam' 이 7건이고 'hard_spam'/'ham'이 2건임을 표시하고 있다. 학습오류로 인해 경계범주와 목표항목을 연결하는 자질의 구분력이 감소되었음을 나타낸다. 이러한 정보를 통해 오분류의 원인을 추측하고 사용자는 각 범주간 구분력을 강화하기 위해 단계 2에서 결정된 문서들을 학습문서로 선택하여 사용한다.



(a) 학습오류 0% (b) 학습오류 10%

[그림 14] 기존방법과 제안방법에서의 오분류 된 문서

[표 3] 기존방법과 제안방법에서의 정확도와 오분류율

실험조건	정확도 (precision)	spam (recall)	ham (recall)	오분류율 (misclassification rate)
E1	0.795	0.86	0.73	20.5%
E3	0.93	0.95	0.91	7%
E2(noisy)	0.705	0.81	0.60	29.5%
E4(noisy)	0.89	0.92	0.86	11%

4.4 성능평가 결과

본 시스템의 검증을 위해 학습에 사용되지 않은 문서로서 4개의 범주마다 50개의 문서들로 구성하여 200개의 문서들로 결과를 확인하였다.

그림 14는 각 범주별 50개의 검증문서에 대한 오분류 문서의 개수를 나타낸다. 표 3은 기존방법과 제안방법에

서의 정확도와 오분류율을 비교하고 있다. ‘spam’ 범주보다 일반문서들의 집합인 ‘ham’ 범주에서 오류 차이가 좀 더 크게 발생하였고 이 경우의 재현율도 함께 낮아졌음을 알 수 있다. 기존방법이 0.73에서 0.60으로 감소율이 큰 것에 비하여 제안방법은 0.91에서 0.86으로 소폭 감소되었음을 보인다.

5. 결론 및 향후연구

일반적으로 자동 분류시스템의 정확도는 학습과정과 적용하는 분류알고리즘에 의존하는 편이다. 본 논문에서는 일률적으로 최상위항목으로 지정하는 방식을 개선함으로써 오분류가 일어나는 상황을 줄이고 있다. 본 논문에서 제안된 피드백과정에 따라 재학습이 필요한 시기와 오류의 원인을 찾아 수정할 수 있으며,재학습이 요구되는 범주에 새로운 학습문서를 제시하고 있다. 오분류율이 높은 스팸문서들을 대상으로 실험한 결과 제안방법은 기존의 방법보다 16%의 정확도 향상 효과를 보였다.

참고문헌

[1] O.Dekel,J.Keshet, "Large Margin Hierarchical Classification.", In Proc. of the ICML'04, pp.209~216, 2004

[2] M.Ruiz and P.Srinivasan, "Hierarchical Text Categorization Using Neural Networks", Information Retrieval, Vol.5, No.1, pp.87~118, 2002.

[3] 최윤정, 지정규, 박승수, "경계범주 자동탐색에 의한 확장된 학습체계 구성방법", 한국정보처리학회 논문지B , 제16-B권, 제6호, pp.0479-0488, 2009.12.

[4] 김재준, 김한구, “베이지언 문서분류시스템을 위한 능동적 학습기반의 학습문서집합 구성방법”, 한국정보과학회 논문지, 제29권, 제12호,2002.12

[5] 윤성희, “자연어 질의유형 판별과 응답 추출을 위한 어휘 의미 체계에 관한 연구”, 한국산학기술학회 논문지, 제5권,제6호,pp.539-545, 2004.12

[6] 김수희, “XML 문서의 구조기반 검색성능 평가”, 한국산학기술학회 논문지, 제10권, 제2호, pp.396-406, 2009.2

[7] M.Lan, C.Tan, H.-B. Low, and S.Y. Sung, "A Comprehensive Comparative Study On Term Weighting Schemes For Text Categorization With Support Vector Machines", In Proc. of 14th International World Wide Web Conference, pp.1032~1033, 2005.

[8] Y.Zhao and G.Karypis, "Hierarchical Clustering Algorithms for Document Datasets", Data Mining and Knowledge Discovery, Vol.10,No.2,pp. 141~168, 2005.

[9] Rainbow(BOW),http://www.cs.cmu.edu/~mccallum/bow

[10] Bayesian Classifier,http://www.bayesia.com/GB/ home/

[11] SVM-light, http://www.cs.cornell.edu/People/tj/svm_light/

[12] CLUTO-Clustering Algorithms, http://glaros.dtc.umn.edu/gkhome/views/cluto

[13] Apache Assassin Project, http://spamassassin.apache.org/

최 윤 정(Yun-Jeong Choi) **[정회원]**

- 2001년 9월 : 이화여자대학교 대학원 컴퓨터학과 (공학석사)
- 2007년 2월 : 이화여자대학교 대학원 컴퓨터학과 (공학박사)
- 2007년 8월 ~ 2008년 2월 : 서강대학교 컴퓨터학과 Post.Doc
- 2009년 3월 ~ 현재 : 서일대학 정보통신과 강의전담 교수

<관심분야>
인공지능, 기계학습, 온톨로지, 유비쿼터스 센서네트워크