Journal of the Research Institute of Industrial Technology. Vol. 23, 2006

사용자 행동과 점진적 기계학습을 이용한 쓰레기 편지 여과 시스템의 설계 및 구현

김강민* · 김재훈**

*한국해양대학교 컴퓨터 공학과 대학원, **한국해양대학교 IT공학부 부교수

Design and Implementation of spam Mail Filtering System Using User Reaction and Incremental Machine Learning

*K. M. Kim** · *J. H. Kim***

**Graduate school of National Korea Maritime University, Pusan 606-791, Korea*

***Division of IT Engineering, National Korea Maritime University, Pusan 606-791, Korea*

요 약 : 인터넷의 급속한 성장에 의해 전자 편지 서비스는 빠르고 비용이 거의 들지 않는 편리한 의사교환 수단이 되었다. 그러나 쓰레기 편지 발신자는 이러한 전자 편지의 편리함을 악용하여 사용자가 원하지 않음에도 불구하고 다양한 상업성, 음란성 편지를 무분별하게 발송하는 실정이다. 사용자는 일반적으로 정보의 유용한 정도에 따라 매우 다른 행동을 보이므로 정보에 대한 사용자의 행동 차이는 문서 분류 시스템에서 유용하게 활용할 수 있다. 이 논문은 사용자의 행동 정보를 점진적으로 학습하여 쓰레기 편지를 여과 하는 시스템을 제안한다. 제안하는 시스템은 편지 학습 데이터와 그에 대한 사용자의 행동을 수집하고 이 데이터를 점진적으로 학습하여 편지 여과 작업에 활용하는 분류 모델을 만든다. 분류 모델은 점진적으로 갱신되어 사용자에게 더 정확한 분류 결과를 제공한다. 실험 결과 사용자에 따라 81%~93%의 분류 정확도를 보였다. 그리고 분류 정확도는 학습 말뭉치의 양이 900개~1,000개일 때, 가장 높았다. 쓰레기 편지와 정보성 편지에 대해 명확하게 행동 차이를 보인 사용자의 분류 결과가 높았다. 사용자의 행동 정보를 포함하는 편지 분류 결과는 그렇지 않은 결과에 비해 평균 14%의 분류 정확도가 향상되었다는 사실은 매우 유용한 정보이다. 결과적으로 제안하는 시스템은 쓰레기 편지 여과 작업에서 사용자의 행동 정보를 유용하게 활용할 수 있다는 사실을 보였다.

핵심용어 : 스팸메일 필터링, 묵시적 피드백, 사례기반 기계학습, 점진적 기계학습

ABSTRACT : *With rapidly developing Internet applications, an e-mail has been considered as one of the most popular methods for exchanging information because of easy usage and low cost. The e-mail, however, has a serious problem that users can receive a lot of unwanted e-mails, what we called, spam mails, and then the user's mailbox can grow exponentially. On the other hand, in e-mail client systems, users do different actions according to usefulness of information, and some classification and recommendation systems like GroupLens utilize the actions to improve performance. This paper presents a mail filtering system using user actions and incremental machine learning. E-mail data and user actions are collected through some user's interface implemented in CGI/Perl. The proposed system consists of two models: one is an action inference model to draw a user action from an e-mail and the other is a mail classification model to decide if an e-mail is spam or not. All the two models are derived using incremental learning, of which an algorithm is IB2 of TiMBL. The accuracy is 81 ~ 93% according to each person. Our proposed system outperforms a system that does not use any information about user actions. Consequently, we have shown that information about user actions is useful for e-mail filtering.*

KEY WORDS : spam filtering, implicit feedback, instance-based machine learning, incremental machine learning

* kkangmin@gmail.com 051)410-4896
** jhoon@mail.hhu.ac.kr 051)410-4574

1. 서 론

월드 와이드 웹(WWW : World Wide Web)을 중심으로 하는 인터넷의 급속한 성장과 더불어 전자 편지 서비스(e-mail service)는 이제 의사교환의 필수적인 매체로 사용되고 있다. 그러나 전자 편지 서비스를 악용한 쓰레기 편지(spam mail)'는 사회적인 문제점으로 대두되고 있다. 여기서 쓰레기 편지는'수신자가 원하지 않는 전자 편지'이므로 수신자에 따라 그 정의가 달라질 수 있다. 그럼에도 불구하고 쓰레기 편지에는 다음과 같은 공통적인 특성이 있다. 일반적으로 쓰레기 편지는 수신자가 원하거나 요청하지 않았다는 성질, 영리 목적의 상업성, 그리고 대량성을 공통적으로 지닌다[1]. 쓰레기 편지는 우선 수신자를 성가시고 짜증나게 하고, 전자 편지 서비스 제공 업체에게는 저장 장치의 용량을 부족하게 만드는 문제를 일으킨다. 경제적 측면에서 쓰레기 편지의 폐해는 다음과 같다. ITU의 조사보고서에 따르면, 2003년 한 해 전 세계적으로 쓰레기 편지로 인해 발생한 경제 손실 비용이 약 25조 원에 달하는 것으로 추정되었다[2]. 같은 해 쓰레기 편지로 인한 국내 피해액 역시 약 1조 3천억 원에 이르는 것으로 추정되었다[3].

쓰레기 편지가 심각한 문제로 부각되자 각국 정부에서는 법적, 제도적 대책을 수립하고, 쓰레기 편지 규제에 적극적으로 나서기 시작하였다. 국내의 경우 정보통신부의 관장 하에 '정보통신망이용촉진및정보보호등에관한법률'을 제정하고, 정보통신부 산하기관인 한국정보보호진흥원에서는 2003년부터 '불법스팸대응센터'를 운영하고 있다. 또한 각 연구기관 및 정보통신 기업체들은 쓰레기 편지에 대한 기술적인 대처 방안을 연구하고 관련 시스템을 제안하였다[4].

이와 같은 문제를 보완하기 위해 이 논문에서는 쓰레기 편지와 정보성 편지에 대한 사용자들의 행동(reaction)이 각각 상이하다는 점에 착안하여 사용자의 행동을 쓰레기 편지 분류를 위한 특징값으로 사용하는 쓰레기 편지 여과 시스템을 제안한다. 사용자 행동의 예를 들면 쓰레기 편지로 간주되는 편지는 읽지 않고 편지 제목만 확인한 후 바로 삭제할 수 있다. 반대로 중요한 정보성 편지일 경우 따로 보관함을 만들어서 보관하거나, 다른 사람에게 전달 또는 답장을 보내는 행동을 할 수 있다. 제안하는 시스템은 이러한 사용자의 행동을 저장하여 쓰레기 편지/정보성 편지 분류를 위한 사례 기반 학습[4]의 특징으로 활용하는 것이다. 이는 편지 내에서 추출할 수 있는 정보만을 이용하여 편지를 분류하는 방법에 비해 더 나은 성능을 기대할 수 있다.

2. 관련 연구

2.1 쓰레기 편지 여과 기술

쓰레기 편지 여과 기술은 분류 대상인 편지를 분석하여 쓰레기 편지의 여부를 판단하는 것이다. 이 기술은 수신자가 동의하지 않은 편지를 사전에 차단하는 것인지, 수신된 편지가 쓰레기 편지인지 아닌지의 여부를 식별하여 사후 차단하는 것인지에 따라 'opt-in'과 'opt-out'방식으로 나눌 수 있다[5]. 'Opt-in' 방식은 수신자가 수신을 원하는 편지 목록을 편지 서비스 제공업체나 편지 관리 프로그램에 등록하여 자신이 등록하지 않은 곳에서 전송되는 모든 편지를 원천 차단하는 방식이다. 'Opt-out' 방식은 쓰레기 편지의 일반적인 특성을 추출하고 이를 이용하여 쓰레기 편지를 차단하는 작업이다. 이 논문이 제안할 시스템은 'opt-out' 방식을 사용하므로 'opt-out' 방식에 대해 구체적으로 설명한다. 'Opt-out' 방식에서 사용하는 편지의 특성은 편지가 포함하는 단어 정보, 발송자 정보, 편지에 포함된 이미지 정보, 쓰레기 편지의 패턴 정보가 있다.

'편지가 포함하는 단어 정보'를 특성으로 하는 여과 작업은 사용자가 수신하는 쓰레기 편지에서 반복 사용되는 특정 단어를 찾아 이후 이 단어를 포함한 편지를 차단하는 작업이다. '발송자 정보'를 이용한 차단 작업은 쓰레기 편지를 자주 보내는 전송자 편지 주소나 해당 편지 서버의 IP 주소, URL등의 정보(black list)를 저장하여 여과하는 방법이다. 하지만 전송자 정보를 위·변조하는 것이 용이한 현실에서 크게 실효성은 없다. '편지가 포함하는 이미지 정보'를 이용한 여과 작업은 특히 음란물 등을 걸러내는데 많이 활용될 수 있으나, 상대적으로 분류의 정확도가 낮고 분류 작업을 위한 속도가 현저히 떨어진다는 단점이 있다. '쓰레기 편지의 패턴 학습 정보'를 이용한 여과작업은 쓰레기 편지의 패턴을 지속적으로 인식하여 차단하는 작업을 말한다. 쓰레기 편지의 패턴으로 단어의 조합, 문장의 조합이나 출현 횟수의 누적치 및 가중치 등을 기반으로 쓰레기 편지의 여부를 판별하는 것인데, 많은 계산량을 요구한다는 단점이 있지만 비교적 분류 정확도가 높다[6].

2.2 사례 기반 학습

사례 기반 학습(instance-based or case-based learning)은 기억기반 학습(memory-based learning)이라고도 하는데, 기존의 사례를 근거로 하여 새로운 사례를 판단하여 범주화하거나 분류하는 학습방법이다. 다시 말해서 유사한 사례로 새로운 문제의 사례를 범주화하는 알고리즘이다. 처음 훈련 데이터로 학습시킬 때 시간이 오래 걸리지만 일단 사례학습이 끝나면 그 다음부터 주어지는 새로운 사례에 대해서는 매우 우수한 학습 효율을 보여준다. 일반적인 사례기반 시스템은 학습부와 실행부로 구성된다. 학습부에서는 유사한 사례를 군집화하거나, 빠른 검색을 위해서 색인하여 적절한 형태로 사례를 저장한다. 실행부에서는 주어진 입력에 대해서 학습부에서 저장된 사례와 가장 비슷한 사례를 추출하고, 주어진 입력과 저장된 사례들의 유사도를 계산하여 여과 작업에 사용되는 분류를 수행한다[7]. Fig. 1은 사례기반 학습의 흐름도이다.

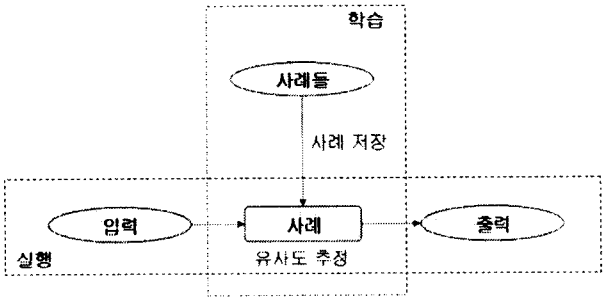

Fig. 1 A flow diagram of instance-based classification systems

2.3 묵시적 피드백

묵시적 피드백(implicit feedback)은 사용자에 대한 지식 습득과정에서 사용자의 어떠한 직접적인 참여도 요구하지 않는 대신 입력되는 데이터에 대한 사용자의 행위(reaction)을 기록, 저장하여 사용자와 데이터간의 연관성(쓰레기 편지의 여부)을 정의하는 작업을 말한다[8].

Morita와 Shinoda는 그들의 연구에서 사용자가 문서를 읽는데 소비하는 시간과 정보 요구 사이의 상관관계를 발견하였으며[9], Konstan이 제안한 협력 여과 시스템(collaborative filtering system)인 'GroupLens'는 문서를 읽는데 투자하는 시간을 관찰하여 이것을 사용자의 문서에 대한 관련 정도로 판단하였다. 그리고 Goecks과 Shavlik은 대상 웹 페이지에 대한 사용자의 하이퍼링크(hyperlink) 클릭 동작과 마우스 스크롤 행위를 활용하였다[10]. 사용자의 문서 데이터에 대한 저장, 삭제, 출력 등의 행동 역시 관심사를 표현하는 증거로 활용할 수 있는데, 특히 Kim과 Oard는 최근 사용자의 행동을 검토(examination), 저장(retainment), 참조(reference), 주석 첨가(annotation)로 분류하고 이를 묵시적 피드백의 증거 데이터로 활용하는 방법을 제안하고 Table 1과 같이 분류하였다[11].

이 표의 세로축은 사용자의 행동의 종류를 나타내고, 가로축은 분류 대상의 최소 범위를 나타낸다. 예를 들어 사용자의 복사/붙이기 행위는 문서의 일부분을 최소 범위로 하는 참조 행위로 해석할 수 있다. 그러나 현재까지 Kim과 Oard의 분류법을 활용한 시스템이 구현된 사례가 없다. 이 논문이 제안하는 시스템은 Kim과 Oard이 제안한 분류법을 참고하여 사용자 행동을 모델링하고, 이를 이용한 쓰레기 편지 여과시스템을 구현하였다.

2.4 편지 학습 말뭉치

객관적인 편지 여과 시스템의 성능을 평가하기 위해서는 공

Table 1 Classification of Kim's user reaction

최소범위 / 행위의 종류	문서 일부분	한 개의 문서	문서 묶음
검토	보기 듣기	선택	
저장	출력	북마크 저장 획득 삭제	동의
참조	복사/붙이기 인용	이동 답변 링크 걸기 참조하기	
주석 달기	기호로 표시	문서 평가 출판	재배치

개된 편지 말뭉치가 필요하다. 영어로 구성된 편지 말뭉치 구축 작업은 매우 활발하게 이루어지고 있다. 그 예로 'SpamArchive'라는 편지 말뭉치 구축 사이트에서는 하루에 평균 5,000여 개의 쓰레기 편지 데이터를 모으고 있다. 그리고 대표적인 편지 말뭉치로 'Enron e-mail corpus'가 있는데, 영문으로 구성된 뉴스그룹 데이터와 편지로 구성되어 있다. 이 편지 말뭉치에서 주목할 만한 점은 다른 편지 말뭉치와 달리 정보성 편지의 집합으로 구성되어 있다는 것이다.

영어 편지 말뭉치 구축작업이 활발하게 이루어지고 있는 반면 한글 편지를 대상으로 하는 공개된 편지 말뭉치는 현재 없다. 따라서 쓰레기 편지 여과시스템을 개발하는 연구자들은 자신의 시스템 실험을 위해 각자 평가 데이터를 마련해야 하는 실정이다. 쓰레기 여과 시스템의 평가를 위한 공개 한글 말뭉치 구축 작업이 필요한 상황이다.

3. 사용자 행동과 점진적 기계학습을 이용한 쓰레기 편지 여과 시스템의 설계 및 구현

제안된 시스템은 크게 편지 말뭉치와 사용자 행동정보를 이용한 학습 말뭉치 구축 과정, 사례기반 기계 학습을 이용하여 행동 추론 모델과 분류 모델을 만드는 학습 과정, 행동 추론 모델과 분류 모델을 이용한 편지 분류 과정으로 나눌 수 있다(Fig. 2 참조).

3.1 학습 말뭉치 구축 과정

제안하는 시스템은 학습을 통한 분류 모델 구축을 위해서

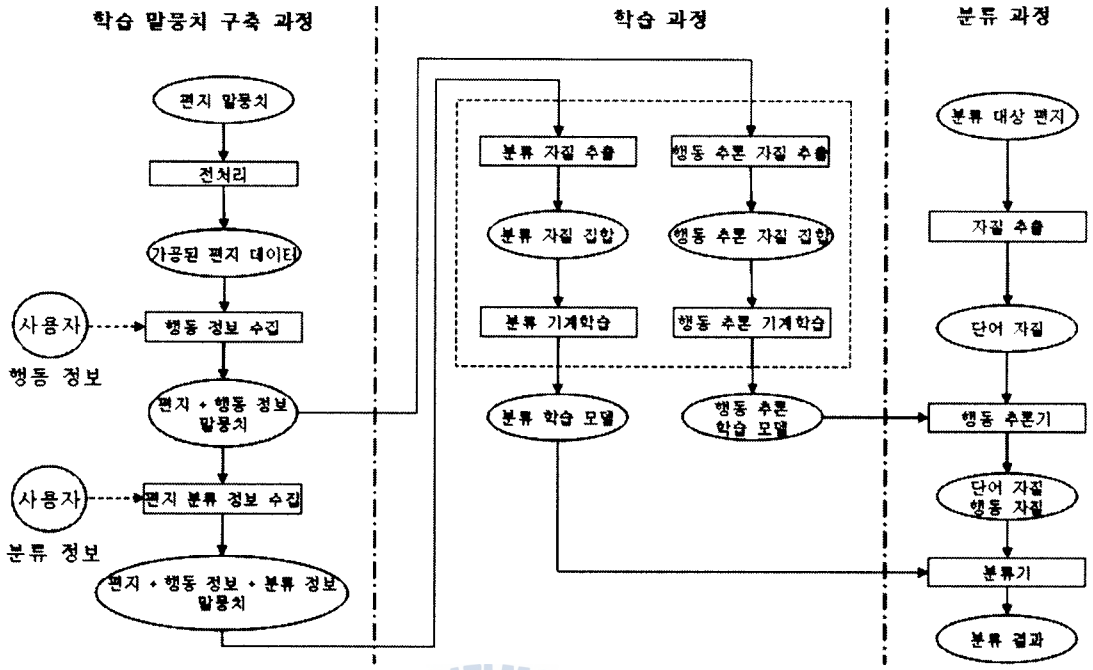

Fig. 2 Overview of the proposed spam mail filtering system

우선 편지 말뭉치와 편지에 대한 사용자의 행동을 가공하여 학습 말뭉치로 구축하는 작업을 수행한다. 학습 말뭉치 구축 과정은 편지 말뭉치를 전처리(preprocessing)하는 과정과, 전처리를 통해 출력된 데이터를 이용한 사전 구축작업, 사용자 인터페이스를 통한 사용자 정보 수집과정으로 이루어져 있다.

전처리 과정은 HTML문서로 구성되어 있는 전자 편지를 쓰레기 편지 여과작업에 활용할 수 있는 정보들의 집합으로 가공하는 과정을 말한다. 편지 내용에서 헤더(header) 정보를 추출하고 본문 중에서 HTML 태그를 제거하게 된다. 태그가 제거된 문자열중 명사 데이터를 추출해 내기 위해 이 논문에서는 명사추출 모듈[12]을 활용한다. 전처리 과정은 Fig. 3와 같다. 전처리 과정을 통해 가공된 편지 말뭉치를 이용하여 사전을 구축한다. 사전은 이후 학습 과정에서 자질을 추출하기 위해 사용된다. 처리할 데이터 양을 줄이기 위해 편지로부터 추출한 명사정보는 사전에서 바이그램(Bi-gram) 형태로 표현되는데, 이는 한자 문화권의 영향을 받은 한국어 명사가 보통 두 개의 음절로 구성되는 경우가 많다는 사실에 기인하였다[13]. 그리고 사전에서는 고빈도어(high frequency term)와 저빈도어(low frequency term)를 제거하는데, 이들은 여과작업에 의미 없는 데이터일 경우가 대부분이기 때문이다.

제안하는 시스템은 사용자 인터페이스를 통해 편지에 대한 사용자의 행동 정보를 수집한다. 사용자 인터페이스는 각 편지에 대한 사용자의 행동을 미리 정의된 형태로 가공한 후 시스템에게 전달하는 역할을 한다. 제안하는 시스템이 사용자

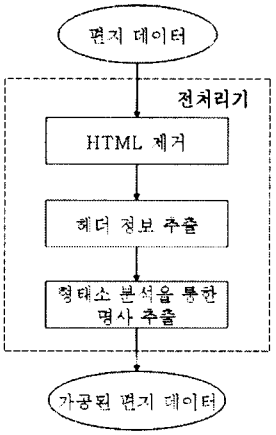

Fig. 3 Preprocessing step

인터페이스를 통해 수집하는 사용자 행동 패턴은 '읽기(read)', '삭제하기(delete)', '분류하기(classify)', '전달하기(forward)', '답장 보내기(reply)'가 있다. 그리고 제안하는 쓰레기 편지 여과 시스템은 사용자의 대상 편지에 대한 분류 결과를 '쓰레기 편지로 분류(I2T: Information to Trash)', '정보성 편지로 분류(T2I: Trash to Information)'로 모델링 하였다.

3.2 학습 과정

학습 과정은 말뭉치 구축 과정을 통해 구축된 학습 말뭉치

로부터 행동 추론 학습기와 분류 학습기의 입력 데이터인 자질 벡터를 추출하여 행동 추론 모델과 분류 모델을 구축하는 과정이다. 제안하는 시스템은 각 모델을 구축하기 위한 기계학습기로 TiMBL[14]을 사용한다.

말뭉치 구축 과정을 통해 구축된 학습 말뭉치는 사용자가 제공한 행동 데이터를 포함하여 행동 추론 모델과 분류 모델의 생성을 위한 벡터 형태의 학습 자질이 된다. 일반적으로 한 개의 편지에서 추출한 학습 자질은 그 데이터양이 매우 적다. 따라서 학습 자질은 희소벡터(sparse vector)로 표현한다. 학습 자질은 편지로부터 추출한 정보들로 구성된 편지부와 사용자 행동정보에서 추출한 행동부로 이루어져 있다. 편지부는 해당 편지에서 추출한 단어의 사전 색인(index)번호와 출현 빈도수를 쌍으로 하는 정보, 본문 내의 이미지의 개수 정보를 저장한다. 행동부는 사용자 인터페이스로부터 얻은 사용자 행동정보를 이진수 형태로 저장하고 있다. 학습 자질을 추출하는 과정은 Fig. 4와 같다.

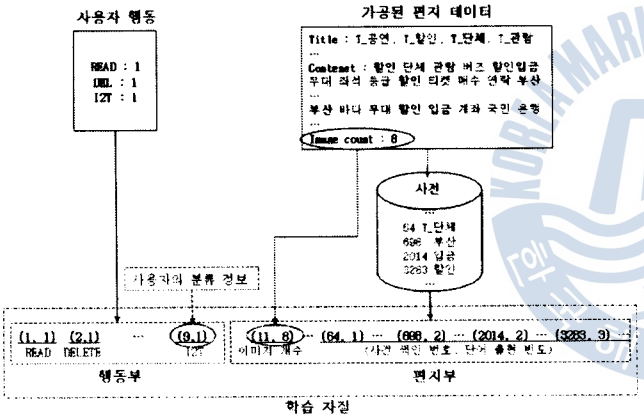

Fig. 4 An example of the feature extraction

이 논문에서는 두 종류의 학습 모델이 사용된다. 하나는 기본 편지로부터 사용자의 행동을 추론하기 위한 학습이고 하나는 추론된 사용자의 행동을 이용하여 편지를 분류하기 위한 학습 모델이다. 이 두 가지 모델을 생성하기 위한 기계학습 도구는 TiMBL이며, TiMBL에서 제공되는 IB2는 점진적 기계학습 방법이므로 이 논문에서는 IB2 알고리즘을 이용한다.

제안하는 시스템은 입력 편지로부터 사용자의 행동을 추론해 내기 위한 학습 모델과, 이를 이용하여 편지를 분류하기 위한 학습 모델을 만들기 위한 두 개의 학습 프로세스를 가진다. 우선 사례기반 학습은 지도학습(supervised learning)의 일종으로 사례를 학습하기 위해 학습 자질이 정답 데이터를 포함해야 한다. 따라서 행동 추론 모델은 사용자의 행동 정보를 추론하는 것이 목적이므로 사용자 인터페이스를 통해 편지를 제공하고, 각 편지에 대한 사용자의 행동 정보 즉 정답 데이터를 입력받아 일정량의 학습 데이터를 구축하고 학습을 통해 모델(사례, 자질별 가중치 정보)을 만들어 내는 작업을 하는 것이다. 분류 모델 구축 작업은 사용자의 행동 정보와 함께 사용자가 제공하는 분류 결과(쓰레기 편지/정보성 편지)를 포함하는 자질을 구축한 후 이를 학습하여 분류결과를 추론하기 위한 사례를 만드는 것이다.

제안하는 시스템은 분류대상이 되는 편지를 자질로 삼아서 사례기반 기계학습을 통해 만들어진 행동 추론 모델을 통해 사용자의 행동을 추론해 내고, 추론된 행동 정보를 포함하는 자질을 구성한 후 분류 모델에 적용하여 대상 편지가 쓰레기 편지인지를 결정하는 작업을 수행하게 된다. 학습과정과 마찬가지로 분류 작업은 사례기반 학습 도구인 TiMBL 테스트 기능을 사용하였으며, 학습을 통해 추출한 모델(사례, 가중치)과 분류 대상 편지의 유사도는 IB2 알고리즘을 이용하였다.

4. 실험 및 평가

4.1 실험 말뭉치

학습 및 실험에 말뭉치로 사용한 편지는 한메일(www.hanmail.net)에서 제공하는 편지 백업 기능을 이용하여 12명의 사용자로부터 추출한 10,000개의 편지 데이터와 해당 편지에 대한 각 사용자의 행동 정보이다. 편지 데이터의 수집 기간은 2005년 3월부터 6월까지 3개월이었으며, 각 사용자는 한메일에서 제공하는 쓰레기 편지 여과 기능을 사용하지 않고 편지 데이터를 수집하였다. 사례 기반 기계 학습 시스템의 학습 말뭉치 구축을 위해 각 사용자는 1,000개의 편지 말뭉치를 대상으로 행동 정보를 제공한다. 수집된 각 사용자별 1,000개의 데이터는 900개의 학습 데이터와 100개의 실험 데이터로 구분한 후 테스트를 수행하였다. 행동 정보는 제안하는 시스템이 제공하는 사용자 인터페이스를 통하여 수집한 것이며, 수집된 행동 데이터의 통계는 Fig. 5와 같다.

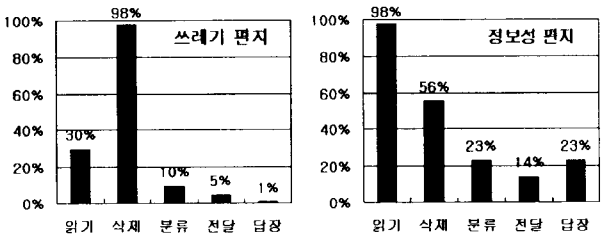

Fig. 5 The ratio of reaction in spam/non-spam in learning corpus

4.2 성능 평가 방법

이 논문에서는 제안한 여과 시스템의 정확도를 평가하기 위하여 다음과 같이 정확도 A를 계산하였다.

$$A = \frac{E}{N} \qquad (1)$$

수식 1에서 N은 실험 데이터의 개수, 는 E사용자의 분류결과와 일치하는 시스템의 분류결과의 개수이다.

4.3 분류 정확도 평가와 분석

제안하는 시스템은 사례기반 기계학습을 이용하여 추출한 분류 모델을 활용하여 여과 작업을 수행하는데, 우선 학습 데이터 양에 따라 분류 정확도의 변화 추이를 조사하였다. 실험은 각 사용자의 학습 데이터가 100개, 300개, 500개, 900개일 때 분류 정확도를 측정하는 방식으로 진행하였고, 기본적으로 학습 데이터의 양이 늘어남에 따라 분류 정확도가 향상되었음을 Fig. 6에서 확인할 수 있다.

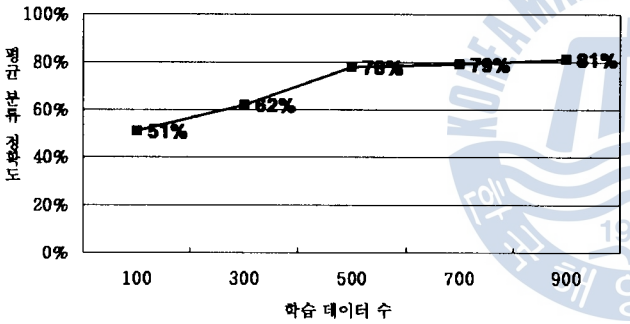

Fig. 6 A mail classification precision according to learning data

이 논문에서는 편지를 분류하기 위해 일반적으로 사용하는 편지에서 추출한 단어뿐만이 아니라 대상 편지에 대한 사용자의 행동 정보를 이용하는 기법을 제안하였다. 따라서 사용자 행동 정보를 제외한 여과 성능과 사용자 행동 정보를 포함한 여과 성능을 비교할 필요가 있는데, 결과는 Fig. 7에서 확인할 수 있듯이 학습 데이터 양에 따라 최소 6%에서 최대 14%의 분류 정확도가 향상되었음을 확인할 수 있다. 결론적으로 이 연구는 편지에 대한 사용자의 행동 정보가 쓰레기 편지 분류 작업에 효과적인 요소가 되었다는 사실을 보였다.

5. 결론 및 앞으로의 연구 과제

이 논문에서는 사용자의 행동 정보와 점진적 기계학습을 이용한 쓰레기 편지 여과 시스템을 제안하였다. 학습 단계에서는 사례기반 학습기를 이용하여 행동 추론 모델과 분류 모델을 만들고, 분류 단계에서는 대상 편지를 입력으로 행동 추론 모델과 분류 모델을 거쳐 쓰레기 편지 여부를 판단하였다.

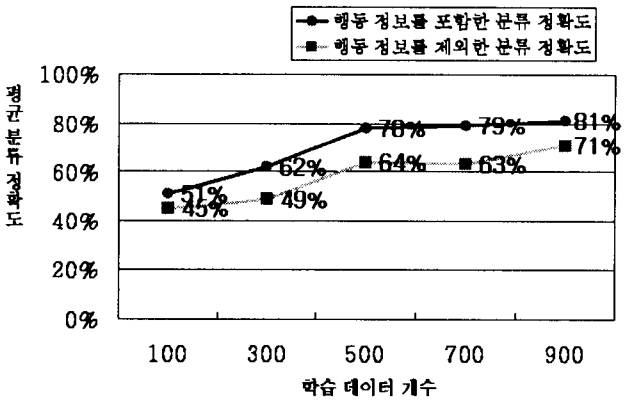

Fig. 7 Usefulness of user action in mail classification

이 시스템은 학습 데이터가 900개일 때, 평균 약 81%의 정확도를 보였으며, 학습 데이터 양이 늘어감에 따라 과적합(over fitting) 문제로 정확도가 약간 떨어지는 구간도 발견되었지만 일반적으로 분류 성능이 좋아지는 사실을 발견할 수 있었다. 실험을 통해 얻은 중요한 사실은 사용자의 행동 정보를 포함한 편지 분류 작업이 사용자의 행동 정보를 제외한 편지 분류 작업에 비해 6%~14% 정도의 분류 정확도가 향상되었다는 것이다. 결과적으로 사용자의 행동은 편지 분류 작업에 있어서 유용한 정보가 된다는 것을 보였다.

향후에는 제안한 시스템에서 정의한 사용자의 행동 외의 다양한 사용자의 행동 패턴을 연구하여 쓰레기 편지 분류 작업에 활용하고 사례기반 학습 이외의 다양한 쓰레기 편지 분류 기법을 이용하여 사용자의 행동을 이용한 여과 성능 향상을 꾀할 수 있다.

참고 문헌

[1] Sorkin, D. E. (2001), "Technical and legal approaches to unsolicited electronic mail", San Francisco University Raw Review, vol. 35, pp. 334.

[2] ITU, (2004), "spam in the information society: Building frameworks for international cooperation".

[3] 한국정보보호진흥원 (2002), 이메일 추출 방지 프로그램의 원리 및 기능분석, http://www.kisa.or.kr/index.jsp.

[4] 한국정보보호진흥원 (2004), 알기쉬운 스팸 대응 현황 자료집, http://www.kisa.or.kr/index.jsp.

[5] Mertz, D. (2002), spam filtering techniques: Six approaches to eliminating unwanted e-mail,

[6] Graham, P. (2002), "A plan for spam".

[7] Mitchell, T. M. (1997), Machine Learning, McGraw-Hill Companies Inc.

[8] Hanani, U., Shapira, B. and Shoval, P. (2001), "Information filtering: Overview of issues, research and systems", User Modeling and User-Adapted Interaction, vol. 11, no. 3, pp. 203-259.

[9] Morita, M. and Shinoda, Y. (1994), "Information filtering based on user behavior: Analysis and best match text retrieval", Proceedings of SIGIR, pp. 272-281.

[10] Goecks, J. and Shavlik, J. (2000), "Learning user's interests by unobtrusive observing their normal behavior", Proceedings of The 5th International Conference on Intelligent User Interfaces, pp. 129-132.

[11] Kim, J. and Oard, D. W. (2001), "Observable behavior for implicit user modeling: A framework and user studies", 한국문헌정보학회지, vol. 35, no. 3, pp. 173-189.

[12] 김재훈, 김준홍 (2001), "도합유사도를 이용한 한국어 문서요약 시스템", 한국인지과학회 논문지, vol. 12, no. 2, pp. 35-42.

[13] 강승식 (2003), "음절 바이그램 단순화 기법에 의한 한국어 자동 띄어쓰기 시스템의 성능 개선", 제15회 한글 및 한국어 정보처리 학술발표 논문지, pp. 227-231.

[14] Daelemans, W., Zavrel, J. and Ko, van der S. (2004), TiMBL: Tilburg Memory-Based Learner Version 5.1 reference guide, Tilburg University, ILK Technical Report, ILK-0104, http://ilk.kub.nl/download/pub/papers/ilk0402.pdf.

원고접수일 : 2005년 12월 30일
원고채택일 : 2006년 1월 10일