사용자 패턴 기반의 부정오류(FN) 수준 평가를 활용한 스팸메일 분류

남명국*, 이상훈*
*국방대학교 전산정보학과
e-mail : nmk10068@naver.com, hoony@kndu.ac.kr

Spam Filtering by False Negative(FN) Value Analysis based on User Pattern

Myoung-Kuk Nam*, Sang-Hoon Lee*
*Dept of Computer Science & Information, Korea National Defense University

요 약

전자 메일의 사용이 급증함에 따라 스팸메일의 양도 함께 증가하고 있다. 증가되는 스팸으로 인한 피해를 줄이기 위하여 여러 가지 기법들이 사용되고 있지만, 지능화되어가는 스패머들의 기술에 완전한 스팸메일의 차단은 불가능하며, 수신된 메일에 대해 사용자는 자신만의 기준으로 스팸메일 여부를 판단하고 있다. 본 논문에서는 스팸메일임에도 불구하고 수신되는 메일(FN)에 대해, 사용자의 반응 패턴을 통하여 이를 판단하고자 한다. 수신된 메일의 송신자와 제목, 보관 편지함 등에서 형태소 추출을 하고 이를 PN_DB(Positive형태소와 Negative형태소로 구성된 DB, 이하 PN_DB)로 구축한 뒤, Negative 형태소들을 Balcklist로 사용하여 FN 메일을 판단한다. FN 메일로 판단된 경우에 PN_DB에서 계산된 각각의 가중치 값을 적용하여 사용자의 과거 스팸 판단 성향이 반영된 FN_value를 시각적으로 표현함으로써 사용자의 판단을 용이하게 하는 시스템을 제안한다.

1. 서론

네트워크 속도의 향상과 사용자 친화적인 인터넷 환경의 발달은 좀 더 빨리, 좀 더 많은 양의 정보를 전달하고 교환하기를 원한다. 특히 e-mail은 기존의 data 전송방식이나 전통적인 방법에 비해 매우 빠르고,한 번에 많은 data를 다수의 수신자에게 보낼 수 있는 기능 등 효과적인 특성으로 인해 많은 사람들이 사용하는 보편적인 정보 통신 수단이 되어 왔다. 그러나 이와 같은 특성을 이용하여 상업적인 목적, 웜 및 바이러스의 전파, 해킹 등의 목적으로 전파되는 e-mail의 양도 급증하였다. 웹사이트 MessageLabs에 따르면, 2008년 2월 발신자 주소를 통해 확인된 스팸메일은 전 세계적으로 72.7%를 차지했으며, 특히 미국 68.9%, 캐나다 74%, 영국 60.4%, 독일 69%, 네덜란드 64.6%, 중국 70.9%, 일본 65.4%의 분포를 보였다[1]. 스팸메일의 폐단이 사회 각 분야에서 두드러지자 정보통신부는 '정보통신망 이용촉진 및 정보 보호 등에 관한 법률 시행령 및 시행규칙 개정안'을 마련하여 모든 광고 메일에는 제목 처음에 [광고]라는 문구를 넣고 제목의 끝에는 '@'을 첨가하도록 하였다. 이에 따라 스팸메일들은 광고 메일에 주로 나타나는 단어들을 'Blacklist'로 등록하여 자동적으로 필터링 되는 방법에 의해 주로 식별되었다[2]. 그러나 스팸메일이 마케팅 수단으로 사용되면서 이 법률을 준수하지 않는 스팸메일들이 범람하고, 이에 따라 메일 서비스 업체부터 개인 사용자에게 이르기까지 많은 피해자가 속출하게 되었고, 그 피해 규모 또한 적지 않다.

스팸메일이 인터넷 전 분야에 걸쳐 다양한 피해를 일으키고, 업무 효율성을 저해하는 요인으로 인식됨에 따라, 스팸메일을 효과적으로 차단하기 위한 다양한 기법들이 사용되고 있다. 가장 많이 사용되는 기법으로는 Listing, DNS lookup, 그리고 단어 필터링 등이 있다[3]. Listing 기법은 이미 알려진 스패머의 정보를 등록시킴으로써 스팸메일의 차단을 시도하는 기법으로, 수신된 e-mail이 등록된 정보와 부합할 경우, 이를 스팸메일로 간주하는 기법이다. DNS lookup 기법은 송신자의 인증을 통해 스팸메일을 차단하는 기법이다. 단어 필터링 기법은 수신된 e-mail의 단어 또는 문장을 통해 스팸메일의 여부를 판단하는 기법으로, 규칙 기반 필터링 기법과 학습 기반 필터링 기법으로 세분화된다. 그 중 학습 기반 필터링 기법은 현재, 안티 스팸메일 분야에서 가장 널리 사용되고 있는 스팸메일 차단 기법 가운데 하나이다. 이처럼 e-mail 서비스 업체에서 다양한 filtering 기능을 제공하고 있지만, 기존의 스팸메일 차단 연구는 분류과정에서 FN(False Negative, 스팸메일임에도 일반메일로 분류)과 FP(False Positive, 일반메일임에도 스팸메일로 분류하여 차단)를 완전히 제거하지 못하고 그 수치를 감소하는 방향으로 연구가 진행되었다. 특히 FN 메일의 경우는 반드시 사용자의 판단에

의해서 분류가 될 수밖에 없다. FN 메일에 대해 사용자들은 제목과 송신자만을 보고 '삭제'하는 경우와 내용을 확인하고 '스팸으로 등록'하는 부정적인 판단 방법, 긍정적인 메일에 대해서는 그것이 상업적인 것이라 할지라도 내용을 확인하고 편지함에 저장하거나 답장을 보내는 등의 긍정적인 행동을 한다. 본 논문에서는 부정적인 FN 메일에 대한 제목과 송신자 이름에서 형태소를 추출하여 Negative 형태소로 PN_DB에 등록하고, 긍정적인 메일과 보관된 메일, 편지함 이름, 보낸 편지함, 주소록에 등록된 이름 등의 제목과 송·수신자 이름에서 형태소를 추출하여 Positive 형태소로 PN_DB에 등록하여 사용자 Profile을 구축한다. 새로운 메일에 대해 제목과 송신자 이름에서 형태소를 추출하여 Blacklist filtering 방법을 통해 스팸 여부를 판단하고, FN 메일의 경우 사용자 Profile의 PN_DB에서 계산된 가중치 값을 적용하여 스팸메일의 정도(FN_value)를 평가할 수 있다.

본 논문의 구성은 2장에서 형태소 분석 방법, 가중치 부여 방법, SVM에 따른 classification 방법, Blacklist에 의한 분류 방법에 관해 기술하고, 3장에서는 사용자 패턴에 따른 Profile을 디자인하고, 4장에서는 FN 메일에 대한 적용방법의 아키텍쳐를 설계하고, 5장에서는 결론 및 향후 연구과제 제시로 실험 및 성능 평가를 위한 데이터의 수집과 적용방법에 대해 기술한다.

2. 관련 연구

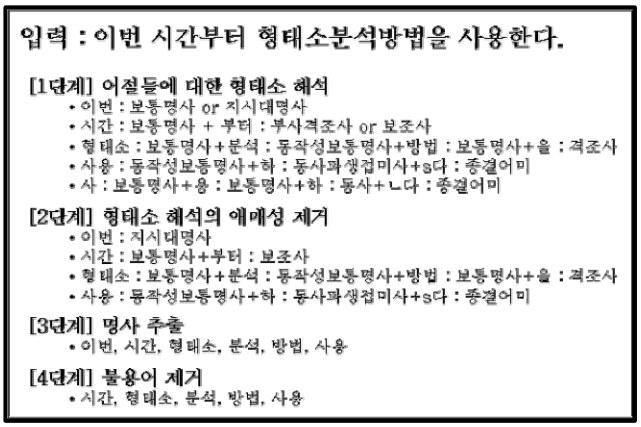
2.1 일반적인 텍스트 전처리

문헌 전처리 과정은 주로 다음과 같이 텍스트 연산(즉, 변형)으로 구분할 수 있다[4].

- 숫자, 하이픈, 문자 부호, 대소문자 처리를 다루는 것이 목적인 텍스트 어휘 분석(lexical analysis)
- 검색에서 변별력이 매우 낮은 단어들을 걸러내는 것이 목적인 불용어 제거
- 불용어를 제외한 나머지 단어에 대해 접사(즉, 접두사와 접미사)를 제거하고, 질의어의 구문적 변형(예를 들어, 'connect', 'connecting', 'connected')을 포함하는 문헌의 검색을 허용하는 것이 목적인 스테밍
- 색인어로 사용할 단어 / 스템(stem)(혹은, 단어 집합)의 선정. 특정 단어를 색인어로 사용할지의 여부는 보통 그 단어의 구문적 성질과 관련이 있으며, 실제로 명사는 형용사, 부사, 동사보다 종종 더 많은 의미를 지니고 있다.
- 본래의 질의를 관련 용어로 확장할 수 있도록(이는 일반적으로 유용한 과정임) 시소러스와 같은 용어 분류 구조를 구축하거나 텍스트에 표현된 구조를 직접 추출하는 작업의 단계를 거친다.

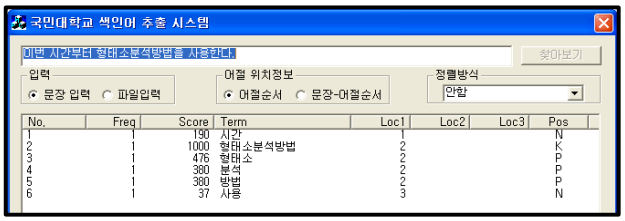
2.2 형태소 분석기

형태소 분석기의 목표는 입력된 문장을 형태소로 분리하는 것이며, 그들 간의 결합 관계를 밝히는 것이다. 입력 문장이 정형화되어 있지 않으므로 접속 관계 처리, 불규칙 및 음운 현상 처리, 복합어 처리, 미등록어 처리 과정 등을 거친다. 이러한 처리를 위하여 형태소 분석기들은 경험규칙, 사전, 규칙 명세들을 이용한다. 또한 형태소 결합 테이블, 불규칙 변형 테이블, 준말 변형 테이블 등을 사용하여 주어진 입력과 순차적으로 비교하면서 형태소를 분석한다[5].



(그림 1) 형태소 분석 단계

본 논문에서 사용되는 형태소 분석기는 '국민대학교 색인어 추출기'를 활용한다[6].



(그림 2) 국민대학교 색인어 추출 시스템

2.3 가중치 부여 방법

문서에서 각 term의 가중치는 해당 term의 빈도와 역문헌빈도 등을 이용하여 다양한 방법으로 결정할 수 있으며, 가장 보편적인 방법으로는 이진 가중치, 단어 빈도 가중치(word frequency weighting), TF-IDF 가중치 등이 있다.

이진 가중치는 가장 단순한 방법으로, 만약 문서에 해당 term이 출현하면 '1'로 표현하고 그렇지 않다면 '0'으로 표현하는 방법이다. i번째 문서에서 term k의 가중치는 식 (1)과 같이 나타낸다. 이 방법은 다음

에 설명하는 방법들에 비해 계산량이 적은 장점이 있으며, 영역에 따라서는 다음에 설명하는 좀 더 복잡한 방법들보다 더 좋은 성능을 나타낸다는 연구 결과도 있다[7]. (f_{ik}는 문서 d_i에서 출현한 term k의 빈도)

$$a_{ik} = \begin{cases} 1 & (\text{if } f_{ik} > 0) \\ 0 & (otherwise) \end{cases} \quad (1)$$

단어 빈도 가중치 방법은 문서에 출현한 term의 빈도를 그대로 문서 표현에 적용하는 방법이다. i번째 문서에서 term k의 가중치는 식 (2)와 같이 나타낸다.

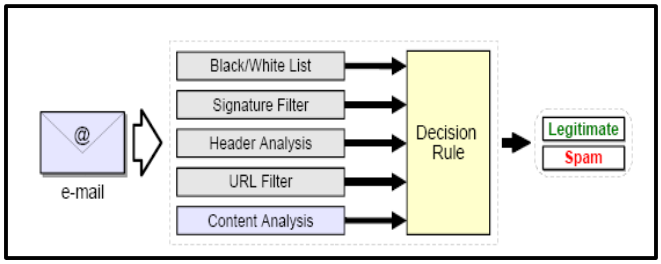
$$a_{ik} = f_{ik} \quad (2)$$

TF-IDF 가중치 방법은 가장 많이 알려진 방법 중 하나로, 문서에서 각 term의 가중치는 해당 문서에서 각 term의 빈도와 역문헌빈도의 곱으로 나타낸다[8]. i번째 문서에서 term k의 가중치는 식 (3)과 같이 나타낸다.(N은 전체 문서의 수이며, n_k는 term k가 출현한 문서의 수)

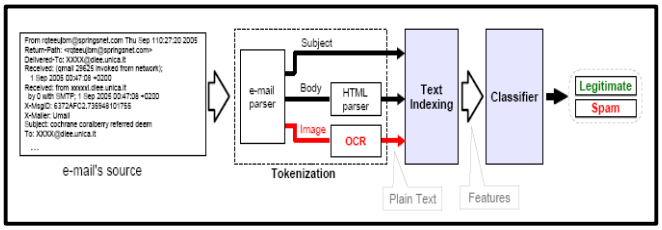
$$a_{ik} = f_{ik} \times \log\left(\frac{N}{n_k}\right) \quad (3)$$

사용자가 FN 메일의 판단 기준으로 많이 사용한 내용의 term들에 대해 상대적으로 높은 가중치를 부여함으로써 사용자에게 높은 수준의 FN_value 값을 표현할 수 있다.

2.4 일반적인 스팸메일 차단 방법



(그림 3) 텍스트 내용의 스팸메일 차단 방법



(그림 4) 이미지가 포함된 스팸메일 차단 방법

일반적인 스팸메일 차단 방법으로는 크게 두 가지로 나눌 수 있는데, 첫 번째가 (그림 3)과 같이 텍스트 내용에 대한 차단 방법이며, 두 번째가 (그림 4)와 같이 텍스트와 그림이 포함된 메일에 대한 차단 방법이다. 두 방법 모두 텍스트 정보를 분석하는 기법으로 스팸메일과 긍정적인 메일의 리스트를 통한 방법, 서명에 의한 방법, 메일 헤더부분에 포함된 발신자 및 URL을 활용한 방법, 내용에 대한 의미해석을 통한 방법, 이미지에 포함된 텍스트를 추출하여 기존의 텍스트 분석방법을 적용하는 필터링 기법이 사용되었다[9].

2.5 SVM(support vector machine)

SVM은 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년 Vapnik에 의해 소개된 학습 기법으로 두 개의 클래스의 구성 데이터들을 잘 분리할 수 있는 결정면(Decision surface, Hyperplane)을 찾는 모델이다[10]. 지지벡터들의 관계를 규명하는 함수를 찾음으로써 주어진 데이터들을 명확하게 긍정과 부정의 두 개 부분으로 분류할 수 있다. SVM은 다차원 공간에서도 계산이 용이하여 가장 많이 사용되는 분류방법 중의 하나이다. 그러나 사용자 행동 패턴을 볼 때, 스팸의 요소들이 조금이라도 존재할 경우 다른 요소들은 사용자가 스팸메일로 판단하는데 결정적인 영향을 미치지 않는다. 따라서 SVM의 분류방법과 같이 긍정과 부정의 요소들의 영향력을 모두 적용하는 방법으로는 본 논문의 스팸메일 판단 방법에 적절치 않은 것으로 판단된다.

2.6 필터링(Filtering)

스팸메일의 제목이나 내용, 보낸 사람의 정보를 활용하여 그 내용이 스팸으로 등록된 경우에 스팸메일로 분류하는 방법이 필터링 방법이다. 더 향상된 방법으로는 스팸메일 헤더나 보낸 사람, 내용에 존재하는 단어의 빈도를 조사하는 방법, 발송 서버의 IP를 신고하여 이를 차단하는 방법 등 다양한 방법이 사용되고 있다. 그러나 필터링의 방법도 사용자가 선택한 경우에만 적용이 가능하여 변화되는 다양한 형태(일반메일과 유사한 형태의 메일)의 스팸메일에 대한 효과는 크지 않다.

본 논문에서는 일반적인 스팸메일 flitering 방법을 회피하여 수신된 스팸메일(FN 메일)에 대해 식별/분석을 하는 것으로 사용자에게 보여지는 정보(제목과 송신자 정보)만을 활용하여 스팸 여부를 판단한다. 또한 사용자 행동 패턴을 볼 때, 스팸의 요소들이 존재할 경우 다른 요소들은 사용자의 판단에 거의 영향을 미치지 않으며, 주어지는 정보들에 대해서는 항상 사용자의 관심사항에 따라 그 평가 방법이 변화가 되어야 한다. 따라서 주어진 정보가 사용자 행동 패턴에 의해 구축된 사용자 Profile의 PN_DB에 등록된 Negative 형태소와 일치하는 형태소가 있을 경우에만 FN 메일로 판단할 수 있으므로, FN 메일의 판단 방법은 'Black List

filtering'방법과 유사하게 형태소 일치 식별 방법을 적용한다.

3. 사용자 Profile 설계

3.1 FN 메일에 대한 사용자 행동 패턴

e-mail 사용자들은 수신된 메일들에 대해 송신자와 제목만을 보고 다음의 방법으로 메일을 판단 및 분류한다.

- 제목만 보고 스팸이라고 생각되는 것은 그냥 "삭제" 한다.(악성코드나 바이러스가 의심될 경우 반드시 내용을 보지 않고 삭제한다.)
- 나에게 관심이 있는 메일일 경우, 내용을 확인하고 그 결과 스팸인 경우 스팸메일 리스트에 등록하고, 불필요한 경우 "삭제" 한다.
- 유용한 정보나 필요한 사항은 "삭제" 하지 않는다.
- 기존에 단순 "삭제" 했던 내용의 메일도 시간이 지남에 따라 사용자의 관심분야가 변하여 내용을 확인하고 "저장" 또는 "편지함 이동" 으로 분류하는 경우가 있다.(기존의 스팸 가능성을 줄임)

즉, 사용자의 관심 여하에 따라 광고성 메일이라 할지라도 유용한 메일이 될 수 있으며, 스팸의 판단은 전적으로 사용자의 성향에 달려 있다고 할 수 있다. 이에 따라 수신된 메일에 대한 사용자의 판단 패턴을 긍정적인 요소들과 부정적인 요소들로 구분하여 적용한다면 새로운 메일에 대한 판단이 가능할 것이다.

3.2 사용자 행동 패턴에 따른 개인 Profile 생성

사용자 행동 패턴은 전적으로 송신자와 제목에서 판단이 되므로 각각의 단어와 형태소를 분석하여 그것을 긍정적인 요소들과 부정적인 요소들로 구분하여 그 정도를 판단하는 방법이 적용되어야 한다. 각각의 DB에 대한 빈도와 메일 건수에 대한 가중치 적용방법은 TF-IDF 가중치 방법을 사용한다.

- 부정적인 형태소 : 제목만 보고 "삭제" 하는 메일과 확인 후 "스팸 등록" 을 하는 메일의 송신자와 제목에서 단어와 형태소를 식별하고, 사회적 이슈가 되면서 사용자의 호기심을 자극하는 단어들(야동, 연예인 X파일 등), 광고성 문구에 대한 단어들(이율, 대박, % 등)에서 그 빈도수를 취하여 PN_DB에 Negative 형태소로 등록하며 그 예는 (표 1)과 같다.
- 긍정적인 요소 : 보관된 메일들의 송신자와 제목, 편지함의 이름, 보낸 편지함 또는 주소록의 수신자 등에서 긍정적인 요소들의 단어와 형태소를 추출하여 이를 PB_DB에 Positive 형태소로 등록하며 그 예는 (표 1)와 같다.
- DB upgrade 방법 : DB 내용이 무한정으로 확장되어서는 안되며, 사용자의 관심변화에 따라 형태소들이 수시로 첨가 및 삭제가 이루어져야 한다. 이를 위해서 첫째, 메일 내용을 확인하고 '스팸 등록' 을 할 경우, PN_DB에서 검사 대상 메일의 모든 형태소들과 일치하는 형태소들의 형태를 Positive에서 Negative로 변경하고 '출현빈도' 와 '메일 건수' 를 '1' 로 초기화 한다. 둘째, 스팸으로 판단되어 단순히 '삭제' 되는 메일에서 추출된 형태소가 PN_DB에서 Positive 형태소로 등록되어 있는 경우는 Negative 형태로 변경하지 않는다. 즉, Positive 형태소로 등록되지 않은 형태소만 Negative 형태소로 PN_DB에 등록한다. 셋째, Positive 형태소로 등록되는 메일(메일 저장, 편지함 생성, 답장 등)의 형태소가 Negative 형태소에 포함되어 있을 경우 Negative 형태소에서 Positive 형태소로 형태를 변경하고 '출현빈도' 와 '메일 건수' 를 '1' 로 초기화 한다(기존에 FN 메일로 판단되었던 형태소일지라도 User의 관심 변화에 따라 일반메일로 재분류 가능함을 반영).

일련번호	형태소	출현빈도	출현메일건수	가중치(w)	PN
1	오빠	4	4	3.11	N
2	할인	3	3	2.71	N
3	돈	5	3	4.52	N
4	자료	10	9	4.26	P
5	아버지	2	2	2.16	P
6	교회	5	3	4.52	P

(표 1) Negative_DB 테이블 형태

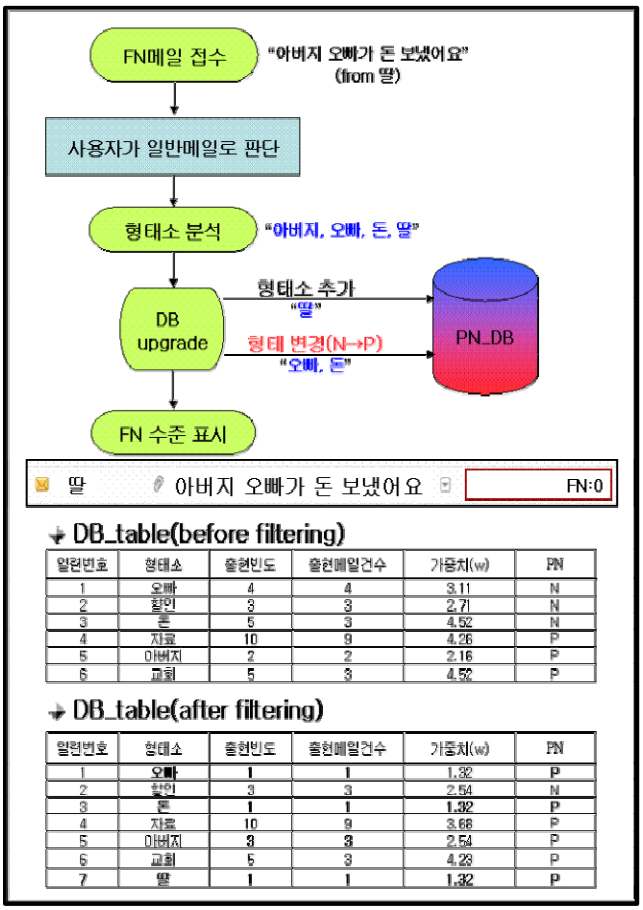
4. FN_value 적용 방안

사용자의 수신된 메일에 대해서는 먼저 모든 메일이 FN 메일 가능성이 존재한다는 가정을 한다. 수신된 메일에 대해 송신자와 제목에 대하여 형태소 분석을 한 뒤, 추출된 형태소는 사전에 구축되어진 PN_DB에 등록된 형태소들과 비교하여 Positive, Negative, Nothing의 세 가지 형태로 분류한다. 이후 Negative 형태소와 일치하는 형태소가 존재할 경우 해당 메일을 FN 메일로 판단한다. 일반메일로 판단되는 경우에는 FN_value를 '0' 으로 표시하며, FN 메일로 판단이 되는 경우에만 가중치 값이 적용된 수준 판단을 표시한다. 이를 표현한 아키텍쳐는 수신된 메일을 사용자가 필요한 메일로 판단하는 경우(그림 5)와 스팸메일로 분류하여 그 FN_value를 계산하는 경우(그림 6)의 두 가지로 살펴볼

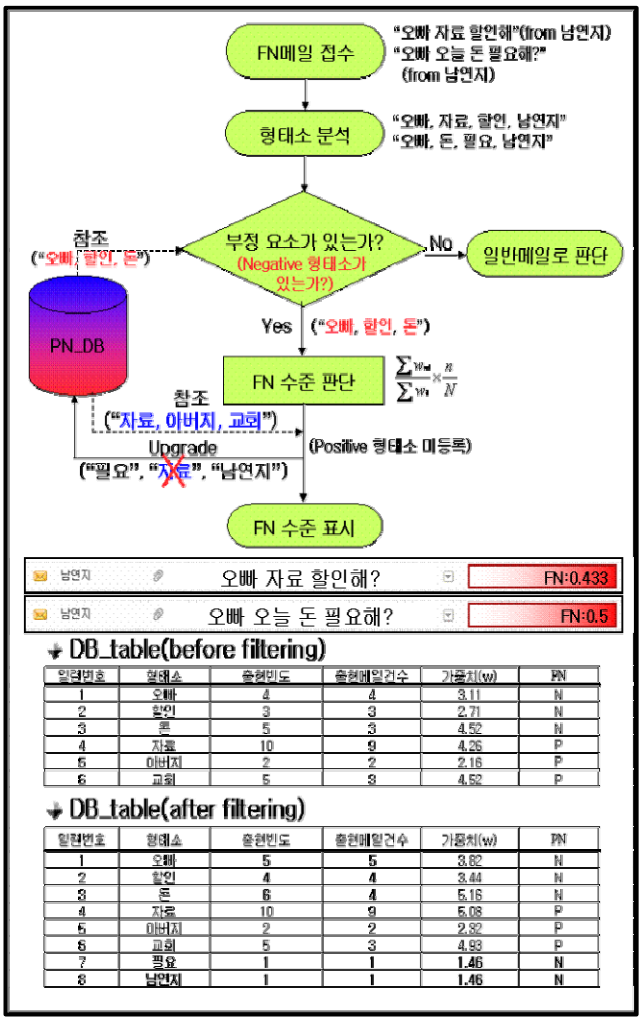
수 있다. FN_value 계산은 가중치 값(형태소들의 빈도와 메일 수에 따른 가중치)을 활용하는데, 그 값은 사용자의 과거 경험에 비추어 볼 때 스팸 성향이 어느 정도 있는지를 나타내는 척도가 된다. 따라서 FN_value는 식별된 모든 형태소에 대한 Negative 형태소의 비율로써 나타낼 수 있다. FN_value 계산은 FN 메일 판단의 경우와 다르게 Positive 형태소의 가중치 값을 적용하는데, 이는 사용자가 스팸이라고 판단하는 형태소들과 적절한 메일이라고 판단하는 형태소들이 혼재되어 있을 경우 긍정과 부정의 판단 정도를 가늠해야 하므로 긍정의 형태소에 대한 가중치 값을 적용하게 된다. 또한 기존의 패턴이 어느 정도로 수신 메일에 영향을 미치는지 알아보기 위해서 수신 메일의 전체 형태소 대비 PN_DB에 포함된 형태소의 수를 고려한다. 계산 방법은 먼저 수신된 메일에서 추출된 모든 형태소에 대하여 가중치 값을 계산하고, Negative 형태소 가중치의 합을 전체 형태소 가중치 값의 합으로 나누고, 전체 형태소 중에 PN_DB 내 존재하는 형태소가 차지하는 비율을 곱하여 그 값을 구할 수 있다. 이를 적용한 FN_value 계산식은 식 (5)와 같으며, 적용 예시는 (표 2)와 같다.

$$FN\ value = \frac{\sum w_{ni}}{\sum w_i} \times \frac{n}{N} \qquad (5)$$

(w_i : FN 메일의 형태소 가중치, w_{ni} :FN 메일의 Negative 형태소 가중치, n : PN_DB에 등록된 FN 메일의 형태소 개수, N : FN 메일 형태소 전체 개수)



(그림 5) FN Analysis architecture(legitimate 메일)



(그림 6) FN Analysis architecture(spam 메일)

제목	형태소	계산	FN value
오빠 자료 할인해	**오빠**, 자료, **할인**, 남연지	(3.11+2.71) / (3.11+2.71+4.26) ×3/4	0.433
오빠 오늘 돈 필요해?	**오빠**, **돈**, 필요, 남연지	(3.11+4.52) / (3.11+4.52) ×2/4	0.5

(표 2) Negative 형태소 FN_value 적용 예시

5. 결론 및 향후 연구

제안된 시스템의 실험은 '국민대학교 색인어 추출 시스템'을 사용하여 형태소를 추출하고, Mysql로 PN_DB를 구현하여 시스템을 구축한다. 사용될 dataset은 한글 data는 개인적으로 수신된 e-mail dataset을, 영문 data는 'Data Mining Cup 2003'의 dataset을 사용한다[11].

기존의 스팸 차단 시스템은 e-mail 제공업체의 이익과 스패머들의 기술 발전으로 인해 사용자가 요구하는 스팸 차단 기능을 만족시키지 못했던 것이 사실이다. 제안된 방법을 사용할 경우, 수신된 모든 메일에 대해서 사용자의 성향과 특성에 따른 스팸메일의 판단이 이루어질 수 있다. 본 논문에서 제안한 방법은 기존의 스팸 차단 시스템이 처리하지 못한 스팸메일을 대상으로 하는 것이지만, 두 방법이 동시에 병행해서 사용될 경우 그 효과가 더욱 높을 것이다.

향후 연구계획으로는 가중치 적용 방법의 변화에 따른 FN_value 정도 변화를 통해 더욱 민감한 FN_value를 도출할 수 있을 것이며, FN_value가 어느 정도 수준일 때 사용자들이 스팸으로 판단하는지에 대한 실험치를 획득하여 분석한다면 보편적인 기준치를 제시할 수 있을 것으로 예측된다.

참고문헌

[1] MessageLabs, "MessageLabs Intelligence: February 2008 "Spam from Gmail Hooks CAPTCHA 'Breaker'," available at http://www.messagelabs.com

[2] C. O' Brien and C. Vogel. "Spam Filter: Bayes vs. Chi-Squared; Letter vs. Words." In Proc. of the 1st Intl. Sym on Information & Comm. Technologies, pages 24-26, 2003.

[3] 김범배, 최형기, "신경망과 유전자 알고리즘을 이용한 스팸메일 필터링 기법의 구현과 성능평가," 정보처리학회논문지 C 제13-C권 제2호, 2006.

[4] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, "Modern Information Retrieval," 1999.

[5] 강승식, "한국어 형태소 분석과 정보검색," 홍릉과학출판사, 2002.

[6] "국민대학교 언어공학, 정보검색 연구실," http://nlp.kookmin.ac.kr

[7] S. T. Dumais, J. Platt, D. Heckerman and M. Sahami, "Inductive learning algorithms and representations for text categorization," In Proceedings of ACM-CIKM98, pp.148-155, Nov. 1988.

[8] G. Salton and M. J. McGill, "An Introduction to Modern Information Retrieval," McGraw-Hill, 1983.

[9] Giorgio Fumera, Ignazio Pillai, Fabio Roli, "Spam Filtering Based On The Analysis Of Text Information Embedded Into Images," 2006.

[10] V. Vapnik. "The nature of statistical learning theory," 1995.

[11] DATA MINING CUP 2003, available at http://www.data-mining-cup.com/2003/Wettbewerb