http://dx.doi.org/10.3745/KIPSTB.2011.18B.5.315

단어 반복 특징을 이용한 스팸 문서 분류 방법에 관한 연구

이 성 진[†] · 백 종 범[†] · 한 정 석[†] · 이 수 원[††]

요 약

인터넷 환경에서 스팸의 범람은 개인 정보의 유출, 피싱에 의한 금전적 손해, 무분별한 유해 콘텐츠의 유통 등 심각한 사회 문제를 야기하고 있다. 또한 사회적 통제를 필요로 하는 유해 정보를 무차별적으로 유통시키는 스팸의 형태와 기술이 갈수록 다양해지고 있다. Bag-of-Words 모델을 이용한 학습 기반 스팸 분류 방법은 현재까지의 연구 중에서 가장 일반적으로 사용되는 방법이다. 그러나 이 방법은 분류 모델 학습 과정에서 사용된 키워드의 출현 정보만으로 스팸 문서를 분류하기 때문에 최근 흔히 발견할 수 있는 스팸 차단 회피 방법에 대한 대처 능력이 부족하다.

본 논문에서는 이러한 문제를 해결하기 위해 문서에서 등장하는 반복 단어의 특징을 이용한 스팸 문서 탐지 방법을 제안한다. 최근 대부분의 스팸 문서에서는 노출하고자 하는 스팸 문구를 반복하는 경향이 있으며, 이는 스팸 문서를 판별하는 기준으로 사용될 수 있다. 본 논문에서는 단어 반복의 특징을 표현할 수 있는 6개의 변수를 정의하고 이를 분류 모델 생성을 위한 속성으로 사용한다. 본 논문에서 제안하는 스팸 탐지 방법의 성능 평가를 위해 블로그 포스트 데이터와 이메일 데이터를 이용하여 기존 방법들과의 비교 실험을 진행하였고, 결과 분석을 통해 제안 방법이 우수함을 확인하였다.

키워드 : 스팸 차단, 스팸, 스팸덱싱, 단어 스패밍, 단어 반복

A Study on Spam Document Classification Method using Characteristics of Keyword Repetition

Seongjin Lee[†] · Jongbum Baik[†] · Chung-Seok Han[†] · Soowon Lee[††]

ABSTRACT

In Web environment, a flood of spam causes serious social problems such as personal information leak, monetary loss from fishing and distribution of harmful contents. Moreover, types and techniques of spam distribution which must be controlled are varying as days go by. The learning based spam classification method using Bag-of-Words model is the most widely used method until now. However, this method is vulnerable to anti-spam avoidance techniques, which recent spams commonly have, because it classifies spam documents utilizing only keyword occurrence information from classification model training process.

In this paper, we propose a spam document detection method using a characteristic of repeating words occurring in spam documents as a solution of anti-spam avoidance techniques. Recently, most spam documents have a trend of repeating key phrases that are designed to spread, and this trend can be used as a measure in classifying spam documents. In this paper, we define six variables, which represent a characteristic of word repetition, and use those variables as a feature set for constructing a classification model. The effectiveness of proposed method is evaluated by an experiment with blog posts and E-mail data. The result of experiment shows that the proposed method outperforms other approaches.

Keywords : Spam Filtering, Spam, Spamdexing, Term Spamming, Word Repetition

1. 서 론

1994년 국내에 처음 인터넷이 등장한 이후 2010년 5월 현재 만3세 이상 인구의 인터넷 이용률이 무려 77.8%에 이를 정도로 꾸준히 성장해 왔다[1]. 그러나 인터넷 이용의 증가는 스팸의 범람이라는 역기능을 함께 수반하고 있다. 스팸이란 정보통신망을 통해 이용자가 원하지 않음에도 불구하고 일방적으로 제공되는 영리목적의 광고성 정보를 의미하

※ 이 논문은 2009년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2009-0075552).
† 준 회 원 : 숭실대학교 컴퓨터학과 박사과정
†† 정 회 원 : 숭실대학교 컴퓨터학부 교수(교신저자)
논문접수 : 2011년 5월 2일
심사완료 : 2011년 7월 8일

며, 점차 그 규모가 커지고 있다[2]. 특히 최근 검색 문화가 활성화되면서 Spamdexing(Spam+Indexing)을 이용한 다양한 방식의 스팸이 등장하고 있다[3][4]. Spamdexing이란 검색 엔진의 문서 랭킹 알고리즘에 영향을 미쳐 사용자의 의도와는 상관없이 스팸 문서의 노출 순위를 부당하게 높이는 방법을 말하며, 크게 Boosting과 Hiding으로 구분할 수 있다.

이 중 Boosting은 TF-IDF 알고리즘에서의 노출 순위를 높이기 위해 단어를 조작하는 Term Spamming과 HITS[5]나 PageRank[6] 알고리즘에서의 노출 순위를 높이기 위해 링크를 조작하는 Link Spamming으로 구분된다. <표 1>을 통해 이에 대한 구체적인 예를 살펴 볼 수 있다.

〈표 1〉 Boosting 구분과 종류

구분	Spam Technique	방 식
Term	Body Spam	문서 본문에 스팸 문구 삽입
	Title Spam	제목에 스팸 문구 삽입
	Meta Tag Spam	메타 태그에 스팸 문구 삽입
	Anchor Text spam	링크 텍스트에 스팸 문구 삽입
	URL Spam	URL에 스팸 문구 삽입
	Repetition	특정 단어를 반복하여 특정 쿼리에 대한 노출 순위를 높임
	Dumping	연관성 없는 단어를 나열하여 광범위한 쿼리에 대한 노출 유도
	Weaving	정상 문서에 스팸 문구를 삽입
	Phrase stitching	여러 정상 문서를 섞어서 나열함
Link	Honey Pot	정상 페이지를 이용, 스팸 페이지의 검색 랭킹 조작
	Link Farm	다수의 Junk Page를 생성, 서로간의 링크를 통해 검색 랭킹 조작
	Comment Spam	댓글 등에 스팸 링크 삽입
	Expired Domain	사용 중지된 도메인을 구입, 기존 사용자들을 스팸 페이지로 유도

한편 Hiding은 <표 2>와 같이 구분된다.

〈표 2〉 Hiding의 종류

Spam Technique	방 식
Contents Hiding	스팸 문구를 제외한 문구는 흰색으로 처리, 사용자는 볼 수 없지만 검색 랭킹은 높음
Cloaking	검색 엔진의 크롤러에게는 정상 페이지를 반환해주는 방법
Redirection	자동으로 스팸 페이지로 이동


시설비/시설비 2010/11/08 14:56

카지노사이트♦바카라사이트

본중 예산에서 할카지노사이트♦바카라사이트▸www.VSVS.KR◂피하고 공교육의 질을 높이기 위해 교육사업비를 대폭 증액했으며, 기관별 중카지노사이트♦바카라사이트▸www.VSVS.KR◂특사업과 1회성 행사성, 낭비성 사업은 과감히 정리했다"고 말했다.

서울 남부지방검찰카지노사이트♦바카라사이트▸www.VSVS.KR◂라이브샘방송인터넷생중계실시간온라인게임사이트청 형사 5부(부장검사 김주원)는 카지노사이트♦바카라사이트▸www.VSVS.KR◂8일 코스닥 업체에서 돈을 받고 해당 기업의 사업 전망을 밝게 보도한 혐의(배임수재)로 카지노사이트♦바카라사이트▸www.VSVS.KR◂전(前) 인터넷 언론사의 증권담당 기자 장모(32)씨를 구속기소했다.

검찰에 따르면 장씨는 2008년 10월부터 2010년 3월까지 모 코스닥 상장 건설업체의 IR담당자로부터 주가 상승에 도움이 되는 호재성 기사를 써달라는 부탁과 함께 4천만원을 받고 13차례에 걸쳐 우호적인 내용의 기사를 쓴 혐의를 받고 있다.

조사결과 장씨는 해당업체가

(그림 1) Term spamming의 혼합 사용

스팸의 형태와 기술은 날로 고도화되고 있지만 이를 효과적으로 차단할 수 있는 기술의 발전은 더딘 상황이다. 현재 스팸 차단 분야에서 가장 널리 사용되고 있는 Bag-of-Words 모델 방식은 스팸 문서 분류시 학습에 사용된 단어의 출현 정도만 고려하기 때문에 Term Spamming에 취약하다. (그림 1)에서는 Weaving과 Repetition을 혼합하여 사용하고 있으며, 정상적인 문서로 인식할 가능성이 크다.

Repetition은 Term Spamming 중에서 가장 사용 빈도가 높은 방법으로, Repetition을 사용한 스팸 문서는 정상 문서와 대비되는 특징을 띄게 된다. 첫째, 반복 단어의 출현 빈도(Term Frequency, *TF*)가 비정상적으로 높게 나타나며, 둘째, 이로 인해 문서에 출현한 단어들의 TF 분포가 일반 문서와는 구분이 된다.

본 논문에서는 문서별 TF 분포로부터 반복 단어의 특징을 추출하고 이를 학습하여 스팸 문서 분류에 활용하는 방법을 제안한다. 이를 위해 단어들의 TF 분포를 분석하여 단어 반복의 패턴을 수치화한 문서 벡터를 생성하고, 기계학습 알고리즘을 이용하여 스팸 분서 분류 모델을 학습한다. 제안하는 방법은 미리 정의된 소수의 변수들로만 학습 예제를 구성하기 때문에 분류의 정확도 향상뿐 만 아니라 학습 소요 시간의 감소라는 추가적인 효과를 기대할 수 있다.

본 논문의 구성은 다음과 같다. 2장에서 본 연구에 대한 관련 연구를 살펴보고, 3장에서는 단어 반복의 특징을 이용한 스팸 문서 분류 방법을 설명한다. 4장에서는 블로그 포스트 데이터와 이메일 데이터를 이용한 검증 실험을 기술하고, 5장에서는 결론과 향후 연구 과제를 제시한다.

2. 관련 연구

2.1 Bag-of-Words 모델 기반 스팸 필터링

Bag-of-Words 모델 기반 스팸 필터링은 일반적으로 사용되는 학습 기반 스팸 필터링 방법으로 토큰(Token)들의 집합을 특징으로 추출하고 이를 이용하여 분류 모델을 학습한다[7-12]. 스팸 분류 모델을 학습하는 방법으로는 나이브

베이지안 분류기를 이용하는 것이 가장 일반적이다. Pantel[8]과 Sahami[9]는 1998년도에 처음으로 나이브 베이지안 분류기(Naive Bayesian)를 스팸 필터링 방법에 활용하였다. 기본적으로 베이지안 분류기는 다른 분류 알고리즘들에 비하여 상대적으로 빠른 속도와 정확한 성능을 보이는 것으로 알려져 있다. Li[10]의 경우 나이브 베이지안 알고리즘을 개선한 새로운 알고리즘을 제안하였다. Li가 제안한 방법은 스팸 필터링 속도를 향상시키면서도 정확도의 손실을 최소화한 것으로 평가된다. 나이브 베이지안 분류기 외에도 k-Nearest Neighbor[11], SVM(Support Vector Machine)[12] 등이 스팸 분류 모델 학습 기법으로 이용되고 있다.

최근 [13]의 연구에 따르면 SVM을 이용한 스팸 분류 모델의 정확도가 가장 높은 것으로 나타났다. 그러나 [7]에서 상용/비상용 스팸 필터링 솔루션들을 조사한 결과에 따르면 대부분의 스팸 필터링 솔루션들은 나이브 베이지안 분류기를 이용하고 있는 것으로 나타났다.

2.2 콘텐츠 스팸 필터링

콘텐츠 스팸 필터링은 Bag-of-Words 모델 기반 스팸 필터링을 무력화시키기 위한 스패머들의 다양한 스팸 콘텐츠 생성 방법에 대응하기 위한 법이다. 스패머들은 Bag-of-Words 모델 방법이 특정 단어의 출현 여부에 의존한다는 점을 이용하여 '카_지_노'와 같이 단어 사이에 특수문자를 삽입하는 방법, 정상적인 문서 하단에 특정 스팸 문구를 삽입하는 방법 등의 다양한 변형 사례를 이용하여 스팸 콘텐츠를 생성한다.

최근에는 다양한 유형의 스팸 사례들에 대응하기 위하여 N-Gram을 이용한 스팸성 자질 추출 방법[14], 스팸 문서의 구조적 특징을 추출하는 방법[15] 등이 연구되고 있다. N-Gram을 이용한 스팸성 자질을 추출하는 연구[14]는 학습 데이터를 생성하는 과정에서 토큰 대신 N-Gram을 이용하여 색인된 자질들을 특징으로 이용한다. 그러나 N-Gram 색인을 이용하는 방법은 새로운 패턴의 단어 구성이 등장하는 문서에 대한 적응력이 부족하다는 문제점이 있으며, [14]에 따르면 N-Gram 색인의 결과로 최소 5,000여 개에서 최대 58,000여 개의 특징이 추출되므로 분류 속도에서 문제가 발생하는 것으로 나타났다.

이러한 문제를 해결하기 위하여 스팸 문서의 구조적 특징을 추출하여 스팸 분류 모델을 학습하는 방법들[15]이 연구되고 있다. 이와 같은 방법들은 일반적으로 반복 출현한 단어의 비율, 링크의 수, 명사의 비율, 불용어의 비율, 문장의 길이 등 정상적인 문서와 구분되는 스팸 문서의 구조적 특징을 추출하여 학습 데이터 셋을 구축한다. 스팸 문서의 구조적 특징을 이용하여 학습 데이터 셋을 구축할 경우, 특징 추출 시간 및 학습/예측 시간이 크게 단축되며 보다 다양한 스팸 유형에 대한 대응력이 좋다는 장점을 지닌다.

2.3 링크 스팸 필터링

링크 스팸은 페이지 간의 링크 관계를 이용하여 PageRank와 같은 링크 기반 랭킹 알고리즘을 이용하는 검색엔진에서 특정 스팸 페이지의 순위를 높이는 스팸 기법을 의미한다.

링크를 이용한 스팸 필터링 방법들[16][17]은 대부분 특정 페이지와 연결된 주변 페이지들(인공적으로 생성된 페이지)이 특정 페이지의 검색 랭킹을 향상시키는지 여부를 검증하기 위하여 링크 분석을 수행한다. Baeza-Yates[18]는 Link Farm에 포함된 페이지가 많은 in-link를 지니고 있음으로 인하여 높은 PageRank 값을 지니는 것을 방지하기 위하여 Damping Function을 제안하였다. Damping Function은 PageRank 계산 시, 첫 번째 레벨에 속하는 링크들의 직접적 공헌도를 무시하는 함수이다. 또한 [18]은 Damping Function과 각 페이지의 링크구조를 분석하여 각 페이지를 링크하고 있는 Supporter들의 개수를 측정하는 알고리즘을 이용하여 웹 스팸을 탐지하는 방법을 제안하였다.

3. 연구 내용

3장에서는 본 논문에서 스팸 문서 분류 모델을 학습하기 위해 필요한 변수들의 의미와 계산 방법에 대해 기술한다.

3.1 문서별 최다 출현 단어 비율

Repetition을 사용하는 스팸 문서에서는 반복되는 단어의 *TF*가 해당 문서에서의 *MaxTF*가 될 가능성이 높다. 또한 문서 길이 대비 *MaxTF*의 비율 즉, *MaxTF* / *SumTF*의 값이 크게 나타난다. 반복 횟수가 많을수록 이 값이 커지며, 스팸 문서일 가능성이 크다고 할 수 있다. 이를 본 논문에서는 최다 출현 단어의 빈도 비율이라 정의하며 (식 1)과 같이 계산한다. (식 1)에서 d_i는 문서 집합 D에서 i번째 문서를 의미한다.

$$MaxTFRatio(d_i) = \frac{MaxTF(d_i)}{TFSum(d_i)} \qquad \text{(식 1)}$$

3.2 문서별 단어 반복 지수

*MaxTFRatio*는 문서 내에서 가장 많이 출현한 단어에 대한 반복 강도를 설명할 수 있지만 여러 단어를 반복하는 경우를 설명하지는 못한다. 예를 들어 2-3개의 단어를 동일한 빈도로 반복할 경우, *MaxTFRatio*는 이 중 한 단어의 빈도만 반영된다는 문제점이 있다. 이럴 경우 TF가 큰 상위 N개의 단어를 사용하는 방법이 있다. 하지만 N을 고정하면 문서별 특징이 제대로 반영되지 않는다. 본 연구에서는 이를 해결하기 위해 *TF*의 표준편차를 이용하여 *MaxTFRatio*를 보완한다.

Repetition을 사용하는 문서에서 반복 패턴을 가진 단어는 나머지 정상적인 단어와의 빈도 편차가 크게 나타난다. 따라서 스팸 문서와 정상 문서 비교시 TF 평균이 비슷하더라도 스팸 문서의 표준편차에서 큰 차이를 보이며, 표준편

차가 클수록 스팸 문서일 가능성이 크다. 표준편차는 평균을 중심으로 봤을 때 측정값의 분포 정도를 나타내는 척도로, 표준편차가 클수록 값이 넓게 분포되어 있다는 것을 의미한다. 경우에 따라 평균±표준편차의 범위를 벗어나는 값은 이상치라고 볼 수 있다.

본 연구에서는 문서 i에서 등장한 단어 w의 빈도가 +1σ(평균+표준편차)의 범위를 벗어나면 이상치, 즉 Repetition을 이용한 반복 단어일 수 있다고 간주하고 이 조건에 해당되는 단어들의 빈도 비율을 이용하여 문서 i의 단어 반복 지수로 사용한다. (식 1)은 (식 2)와 같이 수정되어 문서별 단어 반복 지수를 계산한다. (식 2)에서 $w_{i,k}$는 문서 i에서 k번째 등장한 단어를 뜻한다. *WordRepetitionIndex*는 0~1 사이의 정규화된 값을 가진다.

$$WordRepetitionIndex(d_i) = \frac{\sum_{k \in R} TF(w_{i,k})}{TFSum(d_i)},$$
$$R = \{r \mid TF(w_{i,r}) > AvgTF(d_i) + StdTF(d_i)\}$$

(식 2)

3.3 문서별 반복 단어 비율

*WordRepetitionIndex*는 문서의 반복정도를 설명할 수 있는 유효한 변수이기는 하지만 반복지수에 영향을 주는 단어가 얼마나 되는지를 설명하지 못한다. 또한 단어별 출현 빈도가 거의 비슷한 경우라면 +1σ의 범위를 벗어나는 단어가 많아지기 때문에 *WordRepetitionIndex*의 값이 왜곡될 수 있다.

따라서 WordRepetitionIndex에 영향을 주는 단어의 수 역시 반복 특징을 설명하는 유효 변수가 된다. 본 연구에서는 이를 반복 단어 비율로 정의하고 (식 3)에 의해 계산하며 0~1사이의 값을 갖는다.

$$WordRepetitionRatio(d_i) = \frac{NumOfWord(w_{i,k} \mid TF(w_{i,k}) \geq AvgTF(d_i) + StdTF(d_i))}{NumOfWord(d_i)}$$

(식 3)

3.4 문서별 단어 스팸 지수

문서에서 등장하는 반복 단어의 특징은 Term Spamming을 적용한 스팸 문서에 대한 변별력이 뛰어나다. 하지만 이러한 구조적 특징 외에도 각 단어들의 스팸 지수 역시 스팸 판별에 대한 변별력을 지니고 있음을 부인할 수 없다. 여기서 각 단어의 스팸 지수란 스팸 문서에서 출현할 확률(이하 스팸 확률)을 의미하며 문서별 단어 스팸 지수는 단어별 확률의 합으로 정의할 수 있다. 이 때 단어별 확률만 합하게 되면 단어별 출현 빈도를 무시하게 되므로, 출현 빈도에 스팸 확률을 곱한 값으로 각 단어의 스팸 지수를 계산한다. (식 4)은 단어별 스팸 지수의 합으로 문서별 단어 스팸 지수를 계산하는 식이다. (식 4)에서 PrS($w_{i,j}$)는 단어 $w_{i,j}$가 스팸 문서에서 등장할 확률이다.

$$WordSpamIndex(d_i) = \sum_j TF(w_{i,j}) \times \Pr S(w_{i,j})$$

(식 4)

(식 4)에 의해 계산한 문서별 단어 스팸 지수는 문서의 길이에 큰 영향을 받는다. 즉, 단어가 많이 등장한 문서일수록 스팸 지수가 높아진다는 단점이 있다. 이를 해결하기 위해서 문서의 길이, 즉 문서에서 등장한 모든 단어의 빈도합을 이용하여 정규화한다. (식 5)는 정규화된 문서별 단어 스팸 지수의 계산식이다.

$$WordSpamIndex(d_i) = \frac{\sum_j TF(w_{i,j}) \times \Pr S(w_{i,j})}{TFSum(d_i)}$$

(식 5)

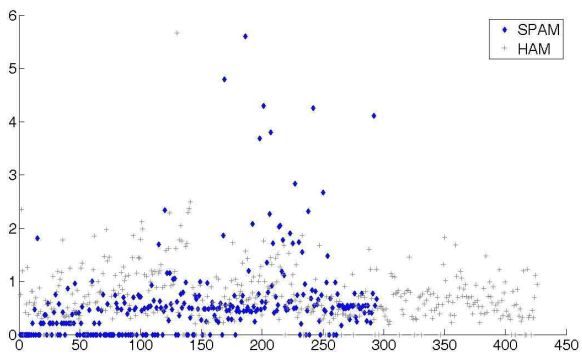
스팸 지수를 계산할 때 추가적으로 고려해야 하는 점이 있다. 스팸 확률이 높은 단어 중 어떤 단어는 정상문서에서도 출현할 확률이 높을 수도 있다는 것이다. 즉, 모든 문서에서 자주 등장하는 단어들은 스팸 판별에 대한 변별력이 없기 때문에 문서별 스팸 지수 계산 시 반영 비율을 낮춰줄 필요가 있다. 본 논문에서는 이러한 문제를 해결하기 위해 스팸 확률을 전체 문서에서의 출현 확률로 나눈 상대적 확률값을 계산하고 log를 취한다. 최종적인 문서별 단어 스팸 지수는 (식 6)과 같이 계산한다.

$$WordSpamIndex(d_i) = \frac{\sum_j TF(w_{i,j}) \times \log \frac{\Pr S(w_{i,j})}{\Pr(w_{i,j})}}{TFSum(d_i)}$$

(식 6)

3.5 문서별 단어 빈도의 표준편차

본 논문에서 스팸 문서 분류 모델 학습을 위한 특징 변수로 *WordRepetitionIndex*와 *WordRepetitionRatio*를 사용하는 근거는 *TF*의 표준편차가 스팸 문서와 정상 문서에서 차이가 난다는 것이다. 특히 *TF*의 표준편차는 그 자체가 빈

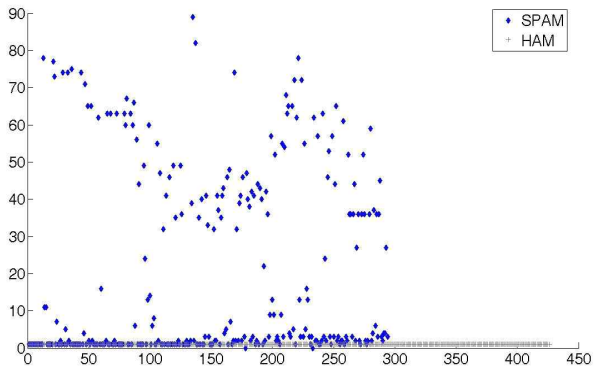


(그림 2) 이메일 데이터의 TF 표준편차 분포

도 분포의 특징을 설명하는 변수이며, 따라서 스팸 문서를 판별할 수 있는 변별력이 있다고 볼 수 있다. 본 논문에서는 학습을 위한 다섯 번째 특징 변수로 *TF*의 표준 편차를 사용한다. (그림 2)는 4장의 성능 평가 실험에서 사용한 이메일 데이터에 대한 *TF*의 표준편차의 분포표이다. (그림 2)에서 x축은 문서 번호이며, y축은 문서에 대한 *TF*의 표준편차이다.

3.6 문서별 링크의 수

스팸 문서는 스팸 사이트로의 방문 유도를 목적으로 하고 있기 때문에 다수의 링크를 포함하고 있는 경우가 많다. 정상적인 문서에도 링크가 발견되기는 하지만 스팸 문서에 비하면 극히 적은 수를 포함한다. 본 논문에서는 학습을 위한 마지막 특징 변수로 문서별 링크 수를 사용한다. (그림 3)은 이메일 데이터에 포함된 링크 수의 분포이다. +로 표시되는 정상 문서의 링크 수는 그래프의 하단에 집중되어 있음을 알 수 있다.



(그림 3) 이메일 데이터의 링크 수 분포

지금까지 본 논문에서 제안하는 스팸 문서 분류 방법의 핵심 내용인 단어 반복 특징을 표현할 수 있는 특징 변수에 대해 정의하고, 이를 계산하는 방법에 대해 살펴 보았다. 이를 정리한 내용은 <표 3>에서 확인 할 수 있다.

〈표 3〉 특징 변수 리스트

특징 변수	의미
MaxTFRatio	최다 출현 단어의 반복 비율
WordRepetitionIndex	반복 단어들의 반복 지수
WordRepetitionIRatio	반복 단어들의 반복 비율
WordSpamIndex	단어들의 스팸 지수의 합
TFStdDev	단어 출현 빈도의 표준편차
LinkCount	링크의 수

4. 실험 및 결과

4.1 실험 조건 및 방법

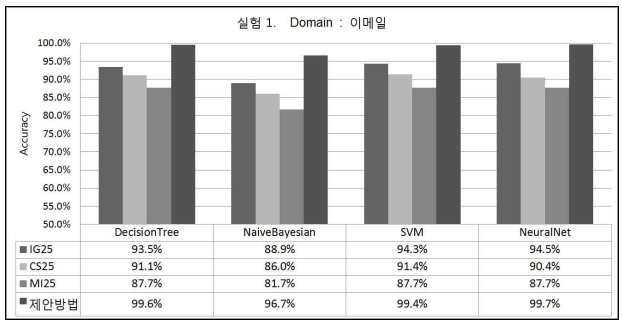
〈표 4〉 비교 실험 정보

항목	내용
실험 환경	인텔 Core2 Quad CPU 2.83GHz 8GB Memory, Window-7 64Bit OS Oracle11g, Java SDK 1.6
실험 데이터	블로그 포스트 (총 5,225건 / 스팸 1,391건) 이메일 데이터 (총 720건 / 스팸 294건)
학습 알고리즘	Decision Tree, Naive Bayesian, SVM, 신경망 (Weka 3.6.4의 오픈 소스를 활용)
비교 대상	1. Bag-Of-Words 모델 2. [Archana, 2009] [15] 3. 도메인 교차 실험
실험 방법	10-fold Cross Validation
평가 척도	Accuracy, Precision, Recall, F-measure, FP-Rate

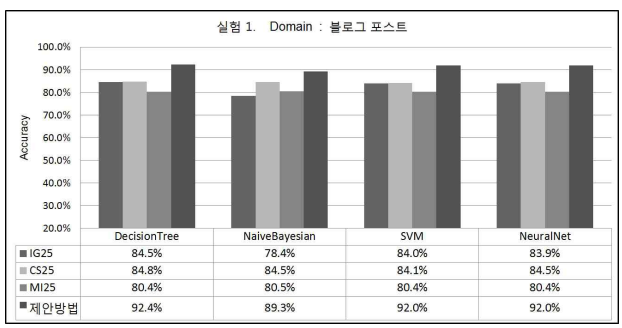
본 논문에서 제안하는 방법의 성능 평가를 위해 <표 4>와 같은 조건에서 비교 실험을 진행하였다.

4.2 Bag-of-Words 모델과의 비교

Bag-of-words 모델을 적용할 때는 효과적인 자질어 선택 방법과 학습 알고리즘이 중요하다. 이를 위해 일반적으로 사용하는 자질어 선택 방법인 정보획득량(Information Gain, IG), 카이제곱통계량(Chi-square, CS), 상호정보량(Mutual Information, MI)을 사용한 경우와 각각 비교하였으며 자질어의 수는 25개로 하였다.



(그림 4) 실험1. 이메일 데이터에 대한 Accuracy 비교



(그림 5) 포스트 데이터에 대한 Accuracy 비교

(그림 4)와 (그림 5)는 각각 이메일 데이터와 포스트 데이터를 사용한 실험 결과이며, 평가 척도는 Accuracy이다.

실험 결과 이메일이나 포스트 데이터 모두 본 논문에서 제안하는 방법의 성능이 가장 우수하게 나왔다. Bag-of-Words 모델의 경우 일반적으로 알려진 바와 같이 정보획득량(IG)을 사용할 때 성능이 가장 좋게 나온 것을 확인 할 수 있다.

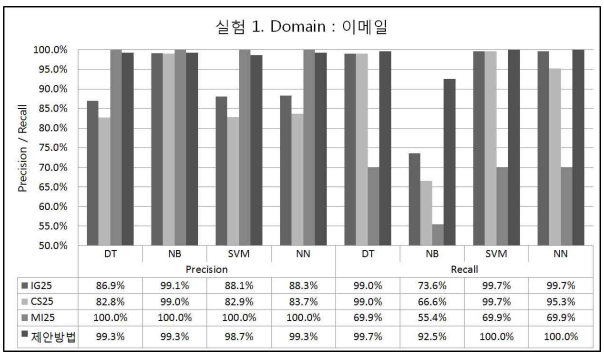
(그림 6)과 (그림 7)은 동일한 실험에 대해 Precision과 Recall을 이용한 평가 결과이다. 스팸 문서 분류 분야에서 Precision이란 스팸으로 분류한 결과가 얼마나 정확한지를 나타내는 척도이며, Recall은 실제 스팸인 문서를 얼마나 잘 찾아내는가를 나타내는 척도이다. 클래스별 분류 결과를 보여주는 Confusion Matrix가 <표 5>와 같을 때 Precision과 Recall은 각각 (식 7)과 (식 8)에 의해 평가한다.

〈표 5〉 스팸 문서 분류 결과에 대한 Confusion Matrix

분류결과 / 실제스팸여부	Spam	Ham
Spam	TP	FP
Ham	FN	TN

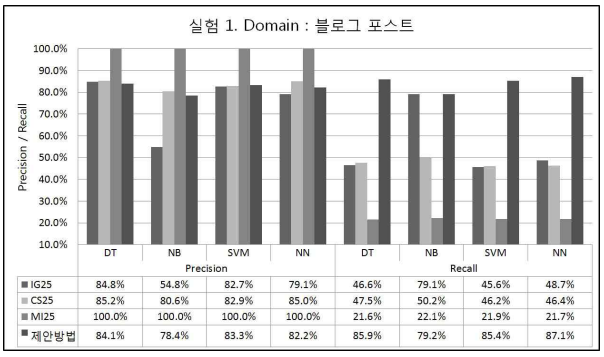
$$Precision = \frac{TP}{TP+FN} \quad \text{(식 7)}$$

$$Recall = \frac{TP}{TP+FP} \quad \text{(식 8)}$$



	Precision				Recall			
IG25	86.9%	99.1%	88.1%	88.3%	99.0%	73.6%	99.7%	99.7%
CS25	82.8%	99.0%	82.9%	83.7%	99.0%	66.6%	99.7%	95.3%
MI25	100.0%	100.0%	100.0%	100.0%	69.9%	55.4%	69.9%	69.9%
제안방법	99.3%	99.3%	98.7%	99.3%	99.7%	92.5%	100.0%	100.0%

(그림 6) 실험1. 이메일 데이터에 대한 Precision/Recall 비교



	Precision				Recall			
IG25	84.8%	54.8%	82.7%	79.1%	46.6%	79.1%	45.6%	48.7%
CS25	85.2%	80.6%	82.9%	85.0%	47.5%	50.2%	46.2%	46.4%
MI25	100.0%	100.0%	100.0%	100.0%	21.6%	22.1%	21.9%	21.7%
제안방법	84.1%	78.4%	83.3%	82.2%	85.9%	79.2%	85.4%	87.1%

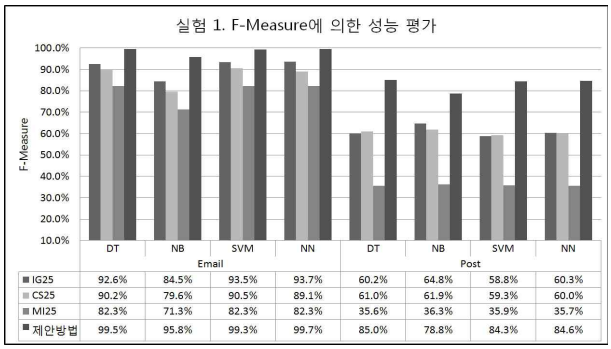
(그림 7) 실험1. 포스트데이터에 대한 Precision/Recall 비교

실험 결과 Recall은 모든 경우에 제안 방법의 성능이 우수하게 나왔다. 반면 Precision은 MI의 성능이 가장 좋게 나왔으며, 제안 방법은 두 번째로 우수한 성능을 보였다. 그러나 MI는 Recall의 성능이 현저히 떨어진다.

Precision과 Recall은 어느 정도 반비례 관계가 있으며, 하나의 척도로만 성능을 평가할 수는 없다. F-Measure는 Precision과 Recall을 동시에 고려한 평가 척도로 (식 9)와 같이 계산한다. (식 9)에서 β를 이용하여 Precision과 Recall의 반영 비율을 조절할 수 있으며, β가 1이면 두 척도를 동등하게 반영하겠다는 의미이다.

$$F-Measure = \frac{(\beta^2+1)\times Precision \times Recall}{\beta^2 \times Precision + Recall} \quad \text{(식 9)}$$

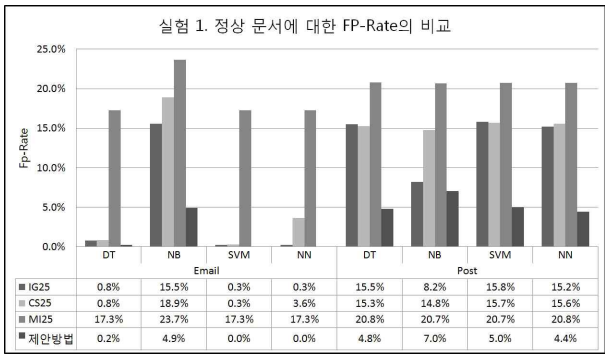
(그림 8)은 F-Measure에 의한 평가 결과로, 모든 경우에서 제안하는 방법의 성능이 우수한 것을 확인할 수 있다. 특히, 블로그 포스트 도메인의 경우 다른 방법에 비해 그 성능이 월등히 우수하다.



	Email				Post			
IG25	92.6%	84.5%	93.5%	93.7%	60.2%	64.8%	58.8%	60.3%
CS25	90.2%	79.6%	90.5%	89.1%	61.0%	61.9%	59.3%	60.0%
MI25	82.3%	71.3%	82.3%	82.3%	35.6%	36.3%	35.9%	35.7%
제안방법	99.5%	95.8%	99.3%	99.7%	85.0%	78.8%	84.3%	84.6%

(그림 8) 실험1. F-Measure 비교

비교 실험1에서 마지막으로 살펴볼 평가 척도는 FP-Rate이다. 스팸 문서 분류에서는 일반적으로 정상 문서를 스팸으로 오분류하는 것이 반대의 경우보다 risk가 크다고 알려져 있다. FP-Rate는 문서의 오분류율을 나타내는 척도이며, (그림 9)는 정상 문서에 대한 FP-Rate이다. 도메인과 학습 알고리즘을 비교한 모든 경우에 제안 방법의 성능이 가장 우수하게 나타났으며 오분류율이 0에 가까운 것을 알 수 있다.



	Email				Post			
IG25	0.8%	15.5%	0.3%	0.3%	15.5%	8.2%	15.8%	15.2%
CS25	0.8%	18.9%	0.3%	3.6%	15.3%	14.8%	15.7%	15.6%
MI25	17.3%	23.7%	17.3%	17.3%	20.8%	20.7%	20.7%	20.8%
제안방법	0.2%	4.9%	0.0%	0.0%	4.8%	7.0%	5.0%	4.4%

(그림 9) 실험1. FP-Rate의 비교

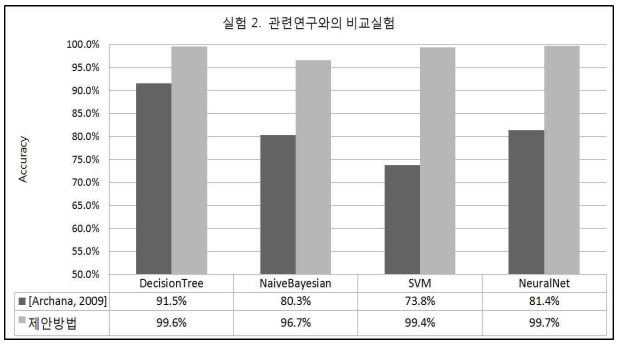
4.3 [Archana, 2009]와의 비교 실험

2장에서 살펴 본 관련 연구 중 [15]는 블로그 포스트의 스팸 댓글을 차단하는 방법에 관한 연구로, <표 6>과 같은 특징 변수를 사용한다.

〈표 6〉 [Archana Bhattarai, 2009]의 특징 변수

항목	내용
Post-Comment Similarity	Post와 Comment에서 출현한 색인어 벡터간의 코사인 유사도
Word Duplication	$1-\frac{No.\ of\ unique\ words}{Total\ No.\ of\ words}$
Number of Anchor Texts	Comment에 포함된 링크의 수
Noun concentration	$\frac{No.\ of\ noun\ phrases\ present}{Total\ No.\ of\ words}$
Stop ratio	$\frac{No.\ of\ stopwords\ present}{Total\ No.\ of\ words}$

[15]의 특징 변수를 살펴보면 단어의 중복 출현 비율, 명사의 집중도, 불용어 비율 등을 사용하고 있으며, 단어의 출현 특징을 분석한다는 점에서 본 연구와 비교될 만 하다. [15]의 특징 변수 중 ***Post-Comment Similarity***는 포스트 또는 이메일 분류 시 사용할 수 없기 때문에 나머지 4개의 변수를 사용하여 비교 실험을 진행 하였다. (그림 10)은 이메일 데이터와 포스트 데이터에 대한 평가 실험 결과이며, 척도는 Accuracy이다.

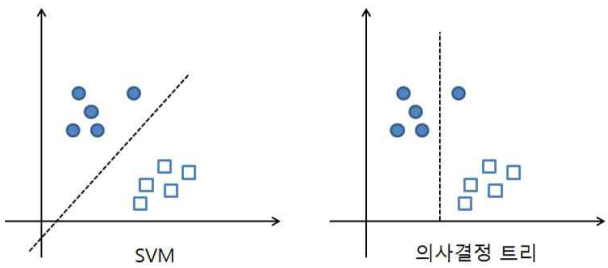


(그림 10) 실험 2. 관련 연구와의 Accuracy 비교

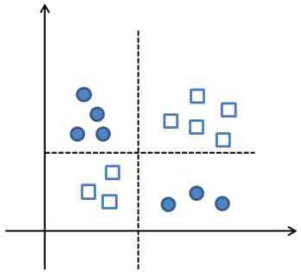
실험 결과를 보면 모든 비교 조건에서 제안 방법의 성능이 우수하게 나왔다. [15]의 결과를 보면 의사결정트리를 사용했을 때의 성능이 가장 좋게 나오며 이는 [15]의 실험 결과와도 일치한다. 제안방법 역시 미세한 차이긴 하지만 의사결정트리의 성능이 좋다고 할 수 있다. 흥미로운 사실은 실험에 사용한 4가지 학습 알고리즘 중 의사결정트리가 수치형 데이터에 대한 학습 능력이 가장 떨어진다는 점이다. 그럼에도 불구하고 성능이 가장 좋게 나온 것은 다음과 같은 이유라고 해석된다.

우선 학습에 사용하는 특징 변수들은 서로 독립이라고 보기는 어려우며, 서로간의 상관관계를 가지고 있다. 또한 학습에 사용하는 변수들은 스팸 여부를 결정짓는 변별력에서 있어 모두 동등한 역할을 수행한다고는 할 수 없다. 학습 알고리즘 중에서 Naive Bayesian의 경우 학습 성능을 보장하기 위해서는 특징 변수가 서로 독립적이어야 한다. 그러나 본 연구에서 제안한 특징 변수들은 서로 보완적인 관계를 가지고 있어 변수간의 독립성을 보장하지 못한다. 이는 비교 연구에서도 마찬가지이며, Naive Bayesian의 성능 저하의 원인이라고 할 수 있다.

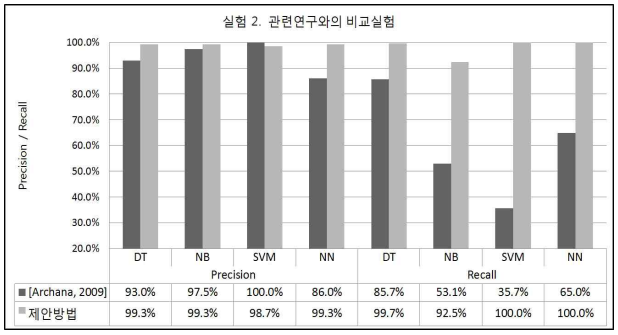
일반적으로 문서 분류 분야에서 가장 좋은 성능을 보인다고 하고 SVM의 경우 특징 변수들이 클래스에 미치는 영향이 어느 정도 동등해야 한다. 자질어 기반의 내용 기반 스팸 차단에서는 각 자질어의 출현 정도가 스팸 여부를 결정하는데 있어 비슷한 정도의 영향을 미친다고 할 수 있다. 하지만 본 연구에서는 특수 문자의 출현 특징을 설명하는 변수의 중요도가 상대적으로 낮기 때문에 이러한 가정에 위배된다. 또한 SVM은 하나의 결정면으로 클래스를 분할하기 때문에 (그림 11)과 같은 예제는 효과적으로 분류할 수 있지만, (그림 12)과 같은 예제는 하나의 결정면으로는 효과적인 분리가 어렵다. 반면 특징 변수의 교차에 의해 여러 구간으로 결정면을 분할하는 의사결정트리의 성능이 더 좋게 나타날 수 있다. 특히 의사결정 트리는 클래스를 결정하는데 영향이 큰 변수부터 결정면을 분할하고, 다른 변수는 분할된 결정면을 다시 분할하는 보조적인 역할을 수행한다. 따라서 이러한 특징을 지닌 의사결정 트리는 본 연구에서 제안하는 특징 변수들을 이용한 학습에 더 적합하다고 할 수 있다.



(그림 11) 하나의 결정면을 사용하는 SVM과 의사결정 트리의 비교



(그림 12) 의사 결정 트리에 의한 효과적인 결정면 분할의 예



	Precision				Recall			
	DT	NB	SVM	NN	DT	NB	SVM	NN
[Archana, 2009]	93.0%	97.5%	100.0%	86.0%	85.7%	53.1%	35.7%	65.0%
제안방법	99.3%	99.3%	98.7%	99.3%	99.7%	92.5%	100.0%	100.0%

(그림 13) 실험 2. 관련 연구와의 Precision / Recall 비교

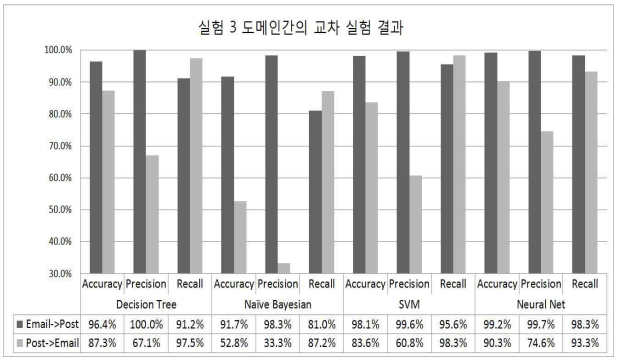
[15]와의 비교 실험에서 Accuracy외의 다른 척도들 모두 제안 방법의 성능이 우수한 것으로 나타났으며, 이 중 Recall에 대한 결과를 검토하고자 한다. (그림 13)은 [15]와의 비교 실험 결과 중 Recall에 대한 비교이다.

[15]은 Recall의 성능이 크게 떨어지게 나온다. Recall이 실제로 스팸인 문서를 얼마나 잘 찾아내느냐에 대한 척도인 것을 고려할 때, 제안 방법은 스팸 문서의 구조적 특징을 좀 더 세분화했기 때문에 [15]에 비해 안정적인 성능을 보이는 것으로 판단된다.

4.4 도메인 간의 교차 실험

스팸 문서 차단에 관한 연구들이 가지는 공통적인 제약 중 하나는 도메인에 특화된 분류 모델을 만들어야 한다는 점이다. 특히 Bag-of-words 모델을 사용하는 경우 도메인별로 나타나는 자질어가 다를 가능성이 매우 높다. 따라서 도메인별로 분류 모델을 학습하기 위한 특징 변수가 다르므로, 다른 도메인에서 학습된 분류 모델을 사용하는 것은 불가능하다. 그러나 본 연구에서 제안하는 방법은 도메인에 상관없이 동일한 특징 변수들을 사용하기 때문에 도메인간의 교차 사용에 대한 가능성이 열려 있다.

실험 3은 이런 가능성을 확인해 보기 위한 실험으로 이메일과 포스트 데이터 간의 교차 적용 실험을 진행하였다. 즉, 이메일 데이터로 학습된 분류 모델을 이용해 포스트 데이터에 대한 테스트를 진행하였다고, 반대의 경우에 대한 실험도 진행하였다.



	Decision Tree			Naïve Bayesian			SVM			Neural Net		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Email->Post	96.4%	100.0%	91.2%	91.7%	98.3%	81.0%	98.1%	99.6%	95.6%	99.2%	99.7%	98.3%
Post->Email	87.3%	67.1%	97.5%	52.8%	33.3%	87.2%	83.6%	60.8%	98.3%	90.3%	74.6%	93.3%

(그림 14) 실험 3. 도메인간의 교차 실험

실험 결과 본 논문에서 제안하는 방법은 도메인간의 교차 적용이 가능할 정도로 우수한 성능을 보이고 있다. 특이한 점은 이메일 데이터로 분류 모델을 학습하고 포스트 데이터에 적용하는 경우의 성능이 반대의 경우보다 월등히 뛰어나다는 점이다. 이는 도메인간의 교차 적용 시 학습 도메인의 학습 예제가 성능에 영향을 미칠 수 있다는 의미이다.

또한 이번 실험의 경우 학습 알고리즘에 따른 성능 차이가 확연하게 드러나고 있다. SVM이나 NN의 경우 Accuracy는 높지만 Precision이나 Recall에 취약점을 보이고 있다. 하지만 의사 결정 트리는 모든 척도에서 안정적인 성능을 보이고 있어 도메인간의 교차 적용 시 의사결정트리를 이용하는 것이 합리적으로 보인다.

5. 결론 및 향후 연구

사회적으로 심각한 문제를 야기하고 있는 스팸 문서를 차단하기 위한 다양한 연구들이 이루어지고 있다. 기존의 연구는 주로 Bag-of-Words 모델을 이용한 기계학습 기반의 스팸 차단 방법이 주를 이루고 있다. 하지만 이 방법은 최근 스팸 문서에서 자주 발견되는 Term Spamming에 대해 취약하다는 문제점을 가지고 있다. Term Spamming은 스팸 문서에 포함된 단어를 조작하여 검색 결과의 노출 순위를 부당하게 높이는 방법을 말하며, 가장 많이 사용되는 방식은 Weaving과 Repetition을 혼합하여 사용하는 것이다. 스팸 문서에서 Repetition을 사용하는 경우 특이한 특징이 나타나는데 반복 단어의 영향으로 문서 출현 빈도(Term Frequency : TF)가 비정상적으로 높은 단어들이 출현한다는 것이다. 이는 TF의 빈도 분포에 있어 이상치라고 할 수 있을 것이다.

본 논문에서는 Term Spamming을 사용하는 스팸 문제를 해결하기 위해 단어 반복 특징을 이용한 스팸 분서 분류 방법을 제안하였다. 이는 Repetition을 이용한 스팸 문서에서 나타나는 TF 분포의 이상치를 이용하는 방법이다. 구체적으로는 반복 단어의 특징을 표현할 수 있는 6가지의 특징 변수와 TF분포로부터 이를 추출하는 방법을 제시하였다. 본 논문에서 제안하는 6가지 특징 변수는 최다 출현 단어의 반복 비율, 반복 단어들의 반복 지수, 반복 단어들의 반복 비율, 단어들의 스팸 지수의 합, 단어 출현 빈도의 표준편차, 링크의 수이다.

본 논문에서 제안한 스팸 분류 방법의 타당성을 검증하기 위해 블로그 포스트와 이메일 데이터를 이용하여 Bag-of-Words 모델을 사용하는 방법과 본 연구와 가장 유사한 [Archana,2009]의 연구와의 비교 실험을 진행하였다. 비교 실험 결과 본 논문에서 제안하는 방법이 다른 방법에 비해 월등히 우수함을 확인할 수 있었다.

또한 본 논문에서 제안하는 방법은 도메인에 상관없이 동일한 특징 변수를 사용하기 때문에 도메인 간의 교차 적용에 대한 가능성이 열려 있다고 할 수있다. 이런 가능성을 검증하기 위하여 비교 실험에서 사용했던 데이터를 이용하

여 상호 교차 실험을 진행하였다. 실험 결과 의사결정트리는 분류 정확도의 평균이 91.85%로 두 경우 모두 안정적인 성능을 보였다.

향후에는 자음/모음 띄어쓰기, 특수 문자 끼워넣기 등의 단어 파괴를 통한 Term Spamming에 대한 특징을 반영할 수 있는 연구가 추가적으로 필요하며, Link 조작에 의한 Abusing이나 Splog에 대한 탐지 방법에 관한 연구가 필요하다.

참 고 문 헌

[1] "2010년 인터넷이용실태조사", *방송통신위원회,한국인터넷진흥원*, 2010. 9.

[2] "2008 불법스팸방지 가이드라인", *방송통신위원회, 한국정보보호진흥원*, 2008. 11.

[3] Zoltan Gyongyi, Hector Garcia-Molina, "Web Spam Taxonomy", *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, 2005.

[4] Hassan Najadat1, Ismail Hmeidi, "Web Spam Detection Using Machine Learning in Specific Domain Features", *Journal of Information Assurance and Security 3 (2008) 220-229*, 2009.

[5] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", *Journal of ACM*, 1999.

[6] Amy Langville and Carl Meyer. "Deeper inside PageRank", *Technical report, North Carolina State University*, 2003.

[7] Enrico Blanzieri and Anton Bryl, "A survey of Learning-based Techniques of Email Spam Filtering", *Artificial Intelligence Review, Springer*, 2008.

[8] Pantel P and Lin D, "Spamcop:a spam classification & organization program", *In AAAI'98 Workshop, Learning for Text Categorization*, 1998.

[9] Sahami M, Dumais S, Heckerman D and Horvitz E, "A bayesian approach to filtering junk e-mail", *In AAAI'98 Workshop, Learning for Text Categorization*, 1998.

[10] Li K and Zhong Z, "Fast statistical spam filter by approximate classifications", *In SIGMETRICS 2006*, 2006.

[11] Androutsopoulos I, Paliouras G, Karkaletsis V, Sakkis G, Spyropoulos C and Stamatopoulos P, "Learning to filter spam e-mail: a comparison of a naive bayesian and a memory-based approach". *In workshop on machine learning and textual information access, 4th European conference on principles and practice of knowledge discovery in databases, PKDD 2000*, 2000.

[12] Drucker H, Wu D and Vapnik V, "Support vector machines for spam categorization", *IEEE Transactions on Neural Networks, Vol.10, No.5, pp.1048 - 1054*, 1999.

[13] 이신영, 길아라, 김명원, "링크구조분석을 이용한 스팸 메일 분류", *정보과학회논문지:소프트웨어 및 응용 제34권 제1호*, 2007. 01.

[14] 이호섭, 조재익, 정만현, 문종섭, "비정상 문자로 조합으로 구성된 스팸 메일 탐지 방법", *정보보호학회논문지, 제18권 제6(A)호, 2008.* 12.

[15] Archana Bhattarai, Vasile Rus, Dipankar Dasgupta, "Characterizing Comment Spam in the Blogsphere through Content Analysis", *ACM Transactions on the Web, Vol.2, No.1, Article 2*, 2009

[16] Yitong Wang, Xiaofei Chen and Xiaojun Feng, "Combating Link Spam by Noisy Link Analysis", *Advanced Data Mining and Applications:Lecture Notes in Computer Science, Vol.6440/2010, pp.453-464*, 2010.

[17] Luca Becchetti, Carlos Castillo, Debora Donato, Ricardo Baeza-YATES, Stefano Leonardi, "Link Analysis for Web Spam Detection", *Journal of ACM Transactions on the Web, Vol.2, No.1*, 2008.

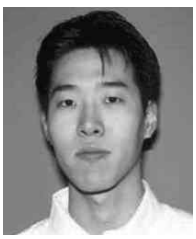
[18] BAEZA-YATES R, BOLDI P, AND CASTILLO C, "Generalizing pagerank:Damping functions for link-based ranking algorithms", *In Proceedings of ACM SIGIR*, 2006.

이 성 진

e-mail : ptnrev93@mining.ssu.ac.kr
2002년 숭실대학교 컴퓨터학부(학사)
2004년 숭실대학교 컴퓨터학과(석사)
2004년~현 재 숭실대학교 컴퓨터학과 박사과정
관심분야 : 인공지능, 기계학습, 개인화

백 종 범

e-mail : jbb100@ssu.ac.kr
2006년 한국IT전문학교 전자계산학(학사)
2009년 숭실대학교 컴퓨터학과(석사)
2009년~현 재 숭실대학교 컴퓨터학과 박사과정
관심분야 : 데이터 마이닝, 정보 검색, 패턴 인식, AI



한 정 석

e-mail : jshan97@mining.ssu.ac.kr
2004년 관동대학교 컴퓨터공학과(학사)
2007년 숭실대학교 컴퓨터학과(석사)
2007년~현 재 숭실대학교 컴퓨터학과 박사과정
관심분야 : 데이터마이닝, 인공지능, 기계학습

이 수 원

e-mail : swlee@ssu.ac.kr

1982년 서울대학교 계산통계학과(학사)

1984년 한국과학기술원 전산학과(석사)

1994년 University of Southern California 전산학과(박사)

1995년~현 재 숭실대학교 컴퓨터학부 교수

관심분야 : 데이터마이닝, 인공지능