제 18 회 한국정보처리학회 추계학술발표대회 논문집 제 9 권 제 2 호 (2002.11)

발신지 추적기법과 사례기반학습을 이용한 한국어 스팸메일 필터의 설계 및 구현

하홍준*, 원일용*, 박호준*, 송두헌**, 이창훈*
*건국대학교 컴퓨터공학과
**용인송담대학 컴퓨터소프트웨어학과
e-mail : greatsk@konkuk.ac.kr

Design and Implementation of Korean Spam mail Filter using the Place of Dispatch Tracking and IBL

Hong-Joon Ha*, Ill-Young Weon*, Ho-Joon Park*, Doo-Heon Song**, Chang-Hoon Lee*
*Dept. of Computer Science, Kon-Kuk University
**Dept of Computer Software, Yong-In Songdam College

요 약

스팸메일이 급증함에 따라 신뢰할 수 있는 전자메일 필터의 요구가 늘어나는 추세다. 스팸메일을 보내는 스패머(spammer)의 거의 대부분은 광고가 주요 목적이다. 멀티미디어(multimedia)기반의 전자메일은 정보전달 및 시각효과가 뛰어나 스패머가 선호하는 전자메일의 한 형태이다. 이런 종류의 전자메일은 텍스트 기반(基盤) 스팸메일 필터의 성능을 떨어뜨리거나 필터링을 아예 불가능하게 한다. 본 연구에서 발신지(發信地) 추적기법과 사례기반학습을 이용해 신뢰할 수 있는 한국어 스팸메일필터를 설계 및 구현하였다.

1. 서론

인터넷의 폭발적 사용증가로 기존의 전화(電話), 서신(書信), 팩스(Fax)에서 비용효과 및 전파효과가 뛰어난 전자메일(E-mail)의 사용이 급증하고 있다. 특히, 최근 들어 메시지 내용은 텍스트기반(Monomedia)의 데이터에서 정보전달 및 시가적 효과가 뛰어난 멀티미디어(Multimedia)기반으로 전환되고 있다. 그러나, 잘못된 사용으로 인한 피해 사례도 잇따르고 있다. 전자메일 서비스 제공업체 상당수가 스팸메일 때문에 업무부담과 비용부담을 겪고 있고, 이용자 역시 스팸메일에 의한 시간낭비와, 정보수신의 방해, 사용요금의 낭비 등의 피해를 입고 있다. 또한 스패머(불법메일 발신자)가 수신대상을 고려하지 않고 전자메일을 무차별하게 발송함으로써 포르노그라피 사이트로부터 온 유해한 메일이 어린이에게 전달될 수 있다.

원하지 않는 전자메일을 차단하는 시스템을 스팸메일필터(이하 "필터"라 함)라고 한다. 필터는 키워드 패턴에 기반(基盤)을 두는 것이 일반적인 형태(예, MS 사의 outlook)인데, 이런 형태의 필터는 사용자가 직접 키워드 패턴을 설정해야 하며 필터의 적중률이 비교적 낮다. 게다가 필터에서 필터링하는 단어가 이미 지화되어 있어 멀티미디어 데이터만을 포함한 스팸메일은 필터링 자체가 불가능하다. 따라서, 메일의 형식에 관계없이 필터링할 수 있는 스팸메일필터의 개발이 시급하다.

적중률이 높은 필터를 개발하기 위해서는 전자메일의 발신지(예, HTML 태그 속성에 포함된 URL)를 추적해서 더 많은 텍스트 데이터를 확보하는 발신지 추적기법과 필터를 자동으로 생성하는 학습 알고리즘이 필요하다.

감독자가 자료를 집단별로 구분해 놓고 분류기준은 컴퓨터 프로그램이 학습에 의하여 발견하도록 하는 방식을 감독학습이라 하며, 이런 학습은 문서의 분류기준을 찾을 때 유용하다. 학습알고리즘에서 데이터 표현 방식은 매우 중요하다. 텍스트 데이터에 학습알고리즘을 적용할 경우 단어를 분리한 다음 단어의 출현유무나 단어의 발생빈도를 구하는 것이 일반적이다. 텍스트에 출현하는 모든 단어를 추출하려면 자연어

처리의 상위단계인 의미분석 및 형태소 분석 과정을 거쳐야 한다. 그러나, 본 연구의 실험에서는 추출이 용이한 명사로 한정하였다.

본 논문은 발신지 추적기법과 명사추출기법, 사례기반학습 알고리즘을 이용해 키워드 패턴기반의 필터보다 적중률이 높은 한국어 스팸메일필터의 설계 및 구현에 관해서 논한다.

2. 전자메일의 전송형태

MIME(Multipurpose Internet Mail Extensions)은 아스키 데이터만을 처리할 수 있는 SMTP(Simple Mail Treansfer Protocol)를 확장하여 오디오, 비디오, 이미지, 응용프로그램, 등 여러가지 종류의 데이터파일들을 주고받을 수 있도록 기능이 확장된 프로토콜이다. HTML 파일 또한 전자메일에 의해 전송되는 데이터파일 중 하나이며, 그 특징은 다음과 같다.

- 0 개 이상의 속성을 가지는 태그로 구성된다.
- <a>태그, <img>태그 등에 기술된 속성은 표현하고자 하는 멀티미디어 자료의 위치정보(URL) 값을 가진다.
  예) **<a href=http://www.yahoo.co.kr>**
- 다른 HTML 문서로 이동할 수 있는 하이퍼 텍스트 기능을 가지고 있다.

위에서 살펴본 바와 같이 태그는 스팸메일의 발신지를 나타내는 위치정보 값을 가진다.

ICNA(http://www.iana.org/assignments/media-types 에서 인터넷 메일의 형태를 확인 할 수 있다.

3. 전자메일의 전송형태별 사례분석

연구실의 연구원을 대상으로 전자메일의 전송형태별 사례를 분석한 결과 다음과 같은 결과를 얻을 수 있었다.

<표 1> 전송형태별 전자메일 수신비율

전자메일의 유형 / 전송형태	일반메일	스팸메일
텍스트	99.8%	63.7%
HTML	0.2%	36.3%

위 <표 1>에 의하면 스팸메일의 대부분 HTML 형태로 전송된다고 보아도 타당할 것이다. 물론, 사례분석이 연구실의 연구원을 대상으로 이루어져 100%신뢰할 수는 없지만, 본 논문에서는 일반인들이 수신하는 메일도 이와 비슷할 것으로 가정한다.

4. 발신지 추적 기법

3 장과 4 장에서 스팸메일의 전송유형과 사례를 분석한 결과 대부분의 스팸메일은 HTML 형태로 전송되며, 전송 데이터에는 발신지에 대한 위치정보(URL)가 들어 있었다.

발신지 추적은 HTML 파일로부터 추출된 명사의 개수가 T 개 보다 적을 때 동작한다. 다음은 발신지 추적 알고리즘의 동작과정을 보인다.

```
i = 1;
html = Email;
currentPage = "";

while( true )
{
        t = ExtractNoun( html );

        if ( t < T )
        {
                if ( i > L ) break;
                url = ParseURL( html );
                ReadHtml( url );
        }
        else
        {
                break;
        }

        i++;
}
```

[설명]

```
L : 추적 깊이의 최대값
i : 현재 추적 깊이
T : 발신지 추적 트리거(triger)를 위한 명사추출 개수의 기준치
t : 추출한 명사수
currentPage : 현재 웹페이지 url

t ExtrantNoun( html ) : Html 파일로 부터 t 개의 명사를 추출
html ReadHtml( URL ) : URL 로 부터 Html 파일을 읽음
url ParsingURL( html ) :
        html 파일로부터 다음을 수행
        <a>태크의 href 속성 값(URL)을 파싱
        <frame>태그의 src 속성 값(URL)을 파싱
        가장 발생 빈도 높은 URL, 같을 경우 랜덤하게 선택
        url 이 상대주소일 경우, currentPage 값 갱신
        currentPage + url 값을 돌려줌
```

(그림 1)발신지 추적 알고리즘

5. 전자메일의 전처리

모든 전자메일은 명사추출과정을 거쳐 벡터공간 안에서 다음과 같은 형태로 표현된다.

$$Msg = < \vec{n}_1, \vec{n}_2, \vec{n}_3, ..., \vec{n}_n >$$

n_x 는 명사가 출현한다면 1, 출현하지 않는 다면 0 의 값을 가진다. N_x 는 에드리뷰트이며 각각은 하나의 명사에 대응한다.

(그림 2) 벡터공간 모델

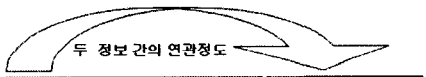
감독학습알고리즘을 적용하기 위해서는 분류정보 에트리뷰트가 추가되어야 한다. 스팸메일이면 1, 일반 메일이면 0 의 값을 가진다.

1	1	0	0	0	1
1	0	0	1	0	1
0	0	1	0	1	0
0	1	0	0	1	0

(그림 3) 분류정보가 추가된 벡터공간 모델

상호정보(MI)는 두 단어간의 연관정도를 나타내는 측도(測度) 중의 하나이다. 본 연구에서는 에트류트와 분류정보 사이의 연관정도를 측정하기 상호정보를 사용한다. 그리고, MI 값이 큰 m 개의 에트리뷰트를 분류기준이 되는 단어 즉, 특징이라고 간주한다.

$$MI(N;C) = \sum_{n\in\{0,1\},C\in\{0,1\}} P(N=n,C=c)\cdot\log\frac{P(N=n,C=c)}{P(N=n)\cdot P(C=c)}$$



1	1	0	0	0	1
1	0	0	1	0	1
0	0	1	0	1	0
0	1	0	0	1	0

(그림 4) 상호정보(MI)

6. 학습

원시적인 사례기반학습(IBL)은 주어진 벡터공간과 새로운 사례 사이에서 유사도가 가장 큰 k 개의 사례를 찾는 알고리즘이다. 본 논문은 k 값이 1 인 원시 사례기반학습에 6.1 절에서 제안한 가중치 결정알고리즘을 적용하였다.

6.1. 가중치 결정

```
for (i=1 ; i<size(T);i++)
   for(j=1; j<size(T);j++)
     Power(Ai)=|Acc(IBL(TA-Ai ,ej,1) – Acc(IBL(TA,ej,1)|
Weight(Ai)=Power(Ai) / sum(Power(Ai)

 * IBL(TA,e,k) : 에트리뷰트 집합 A 를 기반으로 훈련할 집합 T 에 대하여 새로운 사례 e 와 가장 가까운 k 개의 사례를 출력
 * Acc(e) = 사례 e 의 분류 정확도
```

(그림 5) 가중치 결정 알고리즘

6.2. 유사도 계산

두 에트리뷰트간 거리계산 방법은 다음과 같다.

$$D(x,y) = |x-y|$$

- 0 일 때 : 두 값이 유사함, 어떤 단어가 두 사례 모두에 나타나거나, 모두 나타나지 않음
- 1 일 때 : 두 값이 유사하지 않음, 어떤 단어가 하나의 사례에는 나타나고 다른 하나의 사례에는 나타나지 않음

가중치를 적용한 두 에트리뷰트간 거리계산 방법

$$D'(x,y) = D(x,y)\times w$$

두 사례간 유사도 계산 방법

$$E(x,y) = \sum_{a=1}^{m} D'(x_a,y_a)\times w$$

7. 필터링

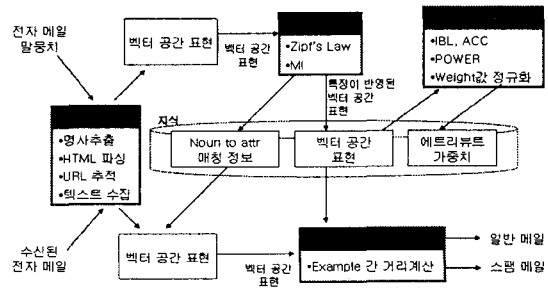
사례기반학습에 의해 생성된 필터는 다음과 같은 과정을 거쳐 새로 수신된 전자메일을 제거한다.

- 새로 수신된 전자메일에서 명사를 추출한다.
- 추출한 명사 개수 t 가 T 보다 적을 경우 발신지 추적 및 텍스트 수집
- 수신된 전자메일에서 MI 에 의해 결정된 단어의 출현유무를 따져 새로운 사례 i 를 생성한다.
- 벡터공간 안에 있는 모든 사례 I_x 와 i 사이의 유사도를 계산한다.
- 가장 유사도가 높은 I_x 의 분류 정보를 i 의 부류(class)라고 판단한다.
- 부류가 스팸메일일 경우 해당메시지를 제거한다.

8. 스팸메일 필터의 설계

스팸메일필터 시스템은 4 장에서 7 장까지의 연구결과를 바탕으로 특징추출, 가중치 결정, 발신자 추적, 메일필터로 구성된다.

명사와 에트리뷰트간의 매칭정보, 각 사례의 축약된 정보인 벡터 공간 표현, 에트리뷰트 가중치는 사례기반학습에 의해 생성된 지식에 해당한다.



(그림 6) 스팸메일 필터링 시스템의 구조

9. 실험결과 및 결론

발신지 추적기법의 효과를 알아보기 위해 다음과 같은 실험을 하였다.

컴퓨터공학 관련 한국어 전자메일에서 100 건, 컴퓨터공학 관련 국내 웹사이트에서 50 건의 HTML 또는 텍스트 형태의 말뭉치를 수집하고 일반메일로 분류하였다. 한국어를 사용한 광고성 전자메일에서 150 건의 말뭉치를 수집하고 스팸메일로 분류하였다.

n 개의 사례가 있을 때, n-1 개의 사례로 학습하고 나머지 하나로 테스트를 하는 절차를 n 번 반복하는 성능평가 방법을 Leave One Out 이라고 한다. 본 실험에서 Leave One Out 을 성능평가방법으로 사용했으며, 상호정보(MI)에 의해 선택될 에튜리뷰트 개수는 300, 추적의 깊이(L)는 3 으로 고정하였다.

결과는 <표 2>와 같다.

	발신지 추적 트리거(Triger)를 위한 명사추출 개수의 기준 T	
	5	10
적중률	94.6%	96.7%

<표 2> 스팸메일필터의 성능평가

트리거 T 값을 5 와 10 으로 설정했을 때 2.1%정도로 다소 적은 성능향상을 보였지만, 멀티미디어에 기반한 스팸메일이 늘어날수록 이 수치는 높아질 것으로 기대된다.

10. 향후 과제

본 연구에서 발신지 추적으로 인한 처리비용 문제를 고려하지 않았다. 발신지 추적 알고리즘을 이용해서 추출한 URL 과 문서분류정보 사이의 연관정도를 분석하고, 그 통계적 지식을 바탕으로 블랙리스트를 만들고 관리하는 알고리즘에 대한 연구가 필요하다.

참고문헌

[1] Aha. D.W., "A Study of Instance-Base Algorithms for Supervised Learning Task: Mathematical, Empirical, and Psychological Evaluations", Technical Report 90-42, University of California, Irvine(1990)

[2] http://www.aic.nrl.navy.mil/~aha/software/

[3] Ion Androutsopoulos, Georgios Paliouras, Vangelis Karkaletsis, Georgios Sakkis, Constantine D. Spyropoulos and Panagiotis Stamatopoulos, "Learning to Filter Spam E-Mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach", Proceedings of the "Machine Learning and Textual Information Access" Workshop of the 4th European Conference on Principles and Practice of Knowledge Discovery in Databases PKDD 2000, 2000.

[4] Mehran Sahami, Susan Dumais, David Heckerman, Eric Horvitz, "A Bayesian Approach to Filtering Junk E-Mail", Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence, August 2001

[5] 임희석, 윤보현, 임해창, "배제 정보를 이용한 효율적인 한국어 형태소 분석기", 한국정보과학회 논문지, 제 22 권 제 6 호, pp.957-964, 1995.

[6] http://www.ietf.org/rfc/rfc2046.txt, Multipurpose Internet Mail Extensions(MIME) Part Two: Media Types

[7] http://www.ietf.org/rfc/rfc2854.txt, The 'text/html' Media Type