스팸메일 필터링을 위한 한글 변칙어 인식 방법

Recognition Method of Korean Abnormal Language for Spam Mail Filtering

안희국*, 한욱표*, 신승호*, 양동일**, 노희영*

Hee-Kook Ahn*, Uk-Pyo Han*, Seung-Ho Shin*, Dong-Il Yang** and Hee-Young Roh*

요 약

전자메일은 사용의 편리성과 정보전달의 신속성 때문에 널리 사용되고 있지만, 광고목적이나 악의성을 갖는 스팸메일의 양도 증가하여 사회적 경제적으로 큰 문제를 야기한다. 스팸메일을 필터링하기 위한 방법은 수용 전 단계와 수용 후 단계로 나누어서 접근할 수 있는데, 수용 후 접근의 경우는 메시지로부터 단어나 문장 단위로 자질을 추출하고 그로부터 학습이나 매칭방법을 통하여 필터링을 하는 과정을 포함한다. 하지만, 필터링을 우회하기위해 스패머는 계속적으로 단어를 변형시켜 메일을 발송시키고 있다. 특히 한국어의 경우는 특성상 한 음절을 이루는 음소의 변화로부터 변형이 가능하기 때문에 그 변칙적 사용이 더 다양하다고 할 수 있다. 따라서, 기존의 정규식이나 학습알고리즘은 대처에 한계를 갖게 된다. 이에 본 논문에서는 한글의 변칙어를 인식할 수 있는 방법을 제안함으로서 스팸메일분류 시스템의 성능을 향상시키고자 한다. 이를 위해, 자소접근방법을 사용하고, Smith-Waterman알고리즘을 적용하였다. 메일서버로부터 추출한 필터키워드와 메일로부터 제안한 방법을 실험한 결과 유사도 수준에 따라 한글 변칙어들을 정확히 인지해 낼 수 있었다. 실험을 통해 소요 공간 및 시간은 허용될 수 있는 수준임을 확인하였다.

Abstract

As electronic mails are being widely used for facility and speedness of information communication, as the amount of spam mails which have malice and advertisement increase and cause lots of social and economic problem. A number of approaches have been proposed to alleviate the impact of spam. These approaches can be categorized into pre-acceptance and post-acceptance methods. Post-acceptance methods include bayesian filters, collaborative filtering and e-mail prioritization which are based on words or sentances. But, spammers are changing those characteristics and sending to avoid filtering system. In the case of Korean, the abnormal usages can be much more than other languages because syllable is composed of chosung, jungsung, and jongsung. Existing formal expressions and learning algorithms have the limits to meet with those changes promptly and efficiently. So, we present an methods for recognizing Korean abnormal language(Koral) to improve accuracy and efficiency of filtering system. The method is based on syllabic than word and Smith-waterman algorithm. Through the experiment on filter keyword and e-mail extracted from mail server, we confirmed that Koral is recognized exactly according to similarity level. The required time and space costs are within the permitted limit.

Key words : Spam Mail Filtering, Korean Abnormal Language, Smith- Waterman Algorithm, Keyword Similarity

I. 서 론

전자메일은 정보 교환의 신속성과 비용의 저렴성

* 강원대학교
** 한림성심대학
· 제1저자 (First Author) : 안희국　　교신 저자 : 노희영
· 투고일자 : 2011년 3월 23일
· 심사(수정)일자 : 2011년 3월 24일 (수정일자 : 2011년 4월 23일)
· 게재일자 : 2011년 4월 30일

때문에 이미 사용자 간의 주된 통신매체로서 인식되고 이용되어져 왔다. 하지만, 사용자가 원하지 않거나, 요청하지 않았는데 수신되는 상업/비상업 목적의 전자메일인 스팸메일은 여전히 사용자에게 전달되고 있다. 국내의 경우, 2003년 이후로 이메일 스팸 수신량은 해마다 감소하여 왔으나, 2009년에는 2.16통으로, 2008년의 2.12통에 비하여 0.04통이 증가하는 추세를 보이고 있다. [1].

스팸메일은 사회적 경제적으로 커다란 피해를 유발하고 있으며 문제해결을 위해 수신된 메일을 스팸(spam)과 비스팸(nonspam)으로 분류하는 필터링작업을 수행하게 된다[2]. 메일필터링은 메일을 받아들이기 전에 결정하는 수용 전(pre-acceptance)접근 방법과 받아들인 후에 결정하는 수용 후(post-acceptance) 접근방법으로 구분될 수 있다[3]. 전자의 경우는 송수신 허용리스트(Domain, IP, E-mail address)를 이용하며, 후자의 경우는 대표적으로 나이브 베이지안 분류기를 기반으로 하고 있으며[4-8], 지지벡터기계(Support Vector Machine), 유전자 알고리즘, 신경망 등의 학습 알고리즘을 이용한다[9, 10]. 이러한 방법들은 스팸메일의 97%까지 방어하는 것이 가능하다고는 하지만[11], 적법한 메일이 스팸으로 분류되는 오탐침(false positive)율이 15%까지 높을 수 있고, 스패머에 의한 변형적 접근(예. 메일주소 변경, 메시지상에 오탈자삽입, 등)으로 주기적인 갱신과 새로운 접근방법이 필요한 상황이다. 이에 최근에는 카이제곱 통계량과 SVM을 이용하여 스팸메일을 필터링하는 방법을 통하여 98.9%의 정확도를 얻을수 있는 방법까지 제안되었다[12]. 하지만, 이러한 수용 후 접근방법의 경우, 단어나 문장을 기본 단위로 사용하여 동작하기기 때문에 자소단위로 변형이 일어나는 한국어의 경우 전혀 인식을 못하거나 인식을 위해 추가적으로 값을 유지하고 학습해야하는 문제점이 발생하게 된다. 이는 재학습의 과정을 거치기 이전까지는 부정확한 필터링을 야기시키고, 재학습을 하더라도 동일키워드의 변칙어들이 분산되어 서로 독립적으로 학습되기 때문에 정확한 학습을 어렵게 한다. 메일정보서비스 제공자(ISP)는 개별 사용자에게 키워드 프로파일을 작성할 수 있는 공간을 제공함으로서 최종수신자에게 도착하는 메일을 필터링하도록 서비스하고 있지만, 이는 사용자와 제공자 모두 시간과 공간비용의 소모를 발생시키며, 향후 발생될 수 있는 변칙어에 대해서는 능동적으로 대처하지 못한다는 문제점을 안고 있다. 특히 한국어의 경우, 특성상 다양한 변칙어를 갖을 수 있어서 한글 메일의 경우, 변칙어를 인식하기위한 필요성이 더 심하다고 할 수 있다.

이에 본 논문에서는 스팸메일 시스템의 후처리단계에서 사용되는 키워드 필터의 변칙어 인식의 문제를 한국어메일을 중심으로 해결하고자 한다. 접근 방법은 단어 단위의 비교를 음절을 구성하는 자음과 모음 단위로 변환하고, 동적프로그래밍 알고리즘(Dynamic Programmi- ng Algorithm. DPA)을 적용함으로서 한글 변칙어를 인식하도록 한다. 이를 위해 한글 변칙어의 특성을 분석한다. DPA로부터 필터키워드와 비교하고자하는 메일키워드간의 유사도 거리를 측정하고 임계값을 통하여 스팸키워드를 인식해내는 방법을 사용한다. 실험을 위하여 메일서버에 수동으로 입력되거나 학습을 통해 자동 갱신된 필터 키워드들 중에서 동일한 것으로 판단되는 키워드들을 수집하고 이를 제안하는 방법을 통해 적용한 결과, 유사도 수준에 따라 정확히 변칙어를 인식할 수 있음을 확인하였다.

본 논문의 구성은 II장에서 관련연구로 본 연구에 사용한 동적프로그래밍 알고리즘과 변형된 형태인 smith-watherman 알고리즘에 대해 알아보고, III장에서는 본 논문의 주제어인 변칙어를 정의하고, 변칙어 인식과 관련된 기존 해결방법 및 한계에 대하여 대학의 메일서버에서 추출한 스팸메일 및 필터키워드들을 중심으로 설명한다. 또한, 한글변칙어의 패턴분석으로 통해 한글 변칙어 인식의 접근방법을 도출한다. IV장에서는 본 논문에서 제안하는 변칙어 인식방법을 상세히 서술한다. V장에서는 실험을 통한 결과를 설명하고, VI장에서는 결론 및 의의, 향후 연구 과제를 기술한다.

II. 관련연구

2-1 Dynamic Programming Algorithm(DPA)

1970년 Needleman & Wunsch에 의해 소개된 DPA는 문제를 분할하여 작은 문제를 해결해 나감으로서 전체문제를 해결할 수 있는 divide & conquer 알고리즘이다. 최소의 비용을 갖는 경로를 찾거나, 작업스케줄링, 스트링 매칭에서의 최장일치서열 탐색 등에 이용될 수 있는 최적화 알고리즘으로서 편집거리를 중심으로 해당 서열들 간의 유사도를 측정하는 데 이용될 수 있다[13, 14].

비교하고자하는 두 문자열 S와 T가 영문알파벳(a,b,c,...z)으로 구성된다고 할 때, 서열간의 유사도는 각 원소들을 정렬한 후, 각 원소들에 대한 점수함수(scoring function: σ)의 합으로 나타낼 수 있다. 여기서 정렬이라 함은 비교하는 두 서열의 심벌순서는 유지하되 편집연산(삽입, 삭제, 치환)을 허용하면서 나열하는 경우를 말한다.

정의 1. x와 y가 각각 한 문자나 공백이면, σ(x, y)는 x와 y를 정렬(align)하는 함수로 표기하고, 점수함수(scoring function: σ)라 한다.

점수함수는 비교하는 심벌이 일치할 경우, +2값을, 일치하지 않을 경우나 공백을 포함하고 있을 경우 -1을 할당함으로서 정렬값에 대한 벌점(penalty)를 부과한다.

두개의 다른 두 문자 a와 c에 대해 σ(a, a)=σ(c, c)=+2, σ(a, c)=σ(a, -)=σ(-, c)=-1를 갖는다. 그림 1은 두 문자열 x(acbcdb)와 y(cadbd)의 서열정렬의 한 예로서 정렬값 6-5=1을 보여준다.

```
x :  a c - - b c d b
y :  - c a d b - d -
```

그림 1. 서열 정렬의 예
Fig. 1. An example of sequence alignment.

정의 2. S와 T의 최적정렬(optimal alignment)은 이러한 두 strings들에 대한 최대값을 갖는 정렬이다.

즉, 두 서열간의 최적정렬을 찾는 것이 두 서열간의 유사도를 측정하는 방법이 되며 대표적으로 DPA를 통하여 효율적으로 찾아낼 수 있다. DPA를 Needleman- Wunsch 알고리즘을 중심으로 알아보면, 크게 행렬생성단계와 역추적단계로 구성된다.

행렬생성단계

문제는 |S|=n, |T|=m을 갖는 string S, T가 주어졌을 때, S와 T의 최적정렬을 찾는 것이다. V(i, j)를 S[1]....S[i]와 T[1].....T[j]의 최적 정렬값으로 정의한다. 따라서, S와 T의 최적 정렬값은 V(n, m)이다.

S의 i번째 문자가 T의 0번째 문자(공백)와 정렬되는 경우 A(i,0)를 고려하기위해 반복과정에 앞서 다음의 초기화과정을 시행한다.

초기화과정 :

$$A(0,0)=0$$
$$A(i,0)=A(i-1,0) + \sigma(S[i],-),\ \text{for } i>0$$
$$A(0,j)=A(0,\ j-1) + \sigma(,-,\ T[j]),\ \text{for } j>0$$

그리고, 0<i, 0<j에 대하여 다음을 반복함으로서 행렬을 생성을 완성한다.

반복과정 :

$$A(i,j)=\max(\ A(i-1,j-1) + \sigma(S[i],T[j]),\ A(i-1,j) + \sigma(S[i],-),\ A(i,j-1) + \sigma(-,T[j])\ )$$

진행방향은 상→하, 좌→우로 진행한다.

S=BASEBALL과 T=BASKETBALL에 대한 행렬생성의 예는 그림 2와 같다.

		B	A	S	K	E	T	B	A	L	L
	0	−1	−2	−3	−4	−5	−6	−7	−8	−9	−10
B	−1	2	1	0	−1	−2	−3	−4	−5	−6	−7
A	−2	1	4	3	2	1	0	−1	−2	−3	−4
S	−3	0	3	6	5	4	3	2	1	0	−1
E	−4	−1	2	5	5	7	6	5	4	3	2
B	−5	−2	1	4	4	6	6	8	7	6	5
A	−6	−3	0	3	3	5	5	7	10	9	8
L	−7	−4	−1	2	2	4	4	6	9	12	11
L	−8	−5	−2	1	1	3	3	5	8	11	14

그림 2. DPA에 의한 S, T 매트릭스
Fig. 2. S, T matrix of DPA.

역추적단계

역추적단계는 최적정렬을 찾아내는 과정으로서 방법은 생성된 행렬의 n, m 엔트리로부터 현재 값에 영향을 준 엔트리를 따라 역으로 추적함으로서 그림 3처럼 최적정렬을 찾아낼 수가 있다.

		B	A	S	K	E	T	B	A	L	L
	0	−1	−2	−3	−4	−5	−6	−7	−8	−9	−10
B	−1	2	1	0	−1	−2	−3	−4	−5	−6	−7
A	−2	1	4	3	2	1	0	−1	−2	−3	−4
S	−3	0	3	6	5	4	3	2	1	0	−1
E	−4	−1	2	5	5	7	6	5	4	3	2
B	−5	−2	1	4	4	6	6	8	7	6	5
A	−6	−3	0	3	3	5	5	7	10	9	8
L	−7	−4	−1	2	2	4	4	6	9	13	11
L	−8	−5	−2	1	1	3	3	5	8	11	14

그림 3. DPA에 의한 최적정렬 탐색과정
Fig. 3. An optimal search of DPA.

따라서, 그림4와 같은 최적정렬을 찾을 수 있다.

```
B A S K E T B A L L
| | |   |   | | | |
B A S - E - B A L L

8 matches : 8*2=16
0 mismatches : 0*-1=0
2 gaps : 2*-1=-2
Total : 14
```

그림 4. DPA에 의한 최적정렬 탐색결과
Fig. 4. Result of optimal alignment with DPA.

본 논문에서는 두 서열의 유사도를 얻어내기 위하여 최적 정렬값을 사용한다. 즉, 스팸필터링을 빠져나가는 한국어의 변칙어들을 스팸키워드와 유사도측정을 통해 인식하고 이를 스팸필터링 및 학습알고리즘에 제공하고자 하는 것이다.

2-2 Smith-waterman algorithm

DPA가 각각의 길이를 m과 n으로 하는 두 개의 서열을 비교대상으로 하고, 전체서열에 대한 유사도를 측정하는데 비하여, Smith-Waterman에 의해 제시된 방법은 지역적 유사도를 측정하기위해 확장된 알고리즘이다[15]. 방법은 행렬생성과정에서 모든 시작가능한 유사점을 고려하기위해 유사도점수가 0이하가 될 경우, 0으로 치환하고, 행렬의 어느 지점에서라도 정렬을 끝낼 수 있도록 최대값을 각 엔트리에 저장하면서 진행하게 된다. 이를 기본과정과 반복과정으로 나누어 기술하면 다음과 같다.

초기화과정 : (0<i, 0<j에 대하여)

$$A(i,0)=0$$

$$A(0,j)=0$$

반복과정 :

$$A(i,j)=\max(\ 0,$$

$$A(i-1,j-1)\ +\ \sigma(S[i],T[j]),$$

$$A(i-1,j)\ +\ \sigma(S[i],-),$$

$$A(i,j-1)\ +\ \sigma(-,T[j])\ )$$

S=abcxdex, T=xxxcde라하고, match: +2, mismatch: -1, space: -1이라할 때, 전체 생성매트릭스와 최적부분정렬을 탐색하는 과정은 다음과 같다.

그림 5에서 최적 정렬값은 A(6,6)=5이다. 따라서, 해당엔트리로부터 엔트리 값이 0일 때까지 역추적해 가면서 서열을 추출한다.

i \ j	0	1 x	2 x	3 x	4 c	5 d	6 e
0	0	0	0	0	0	0	0
1 a	0	0	0	0	0	0	0
2 b	0	0	0	0	0	0	0
3 c	0	0	0	0	↖ 2	1	0
4 x	0	2	2	↖ 2	←↑ 1	1	0
5 d	0	1	1	1	1	↖ 3	2
6 e	0	0	0	0	0	2	↖ 5
7 x	0	2	2	2	1	1	4

그림 5. smith-waterman알고리즘의 회복과정
Fig. 5. Recovering of smith-waterman algorithm.

역추적과정의 경로에 대응하는 최적 부분정렬은 다음과 같다.

c x d e	x - d e
c - d e	x c d e

그림 6. smith-waterman알고리즘의 부분정렬
Fig. 6. Partial alignment of smith-waterman algorithm.

그림 6의 두 결과는 3개의 일치와 1개의 공백을 가지므로, 3·(2)+1·(-1)=5값을 갖는다.

III. 스팸메일의 유형분석

본 장에서는 대학교의 메일서버로부터 스팸메일의 유형을 파악하고, 한글 변칙어의 패턴을 분석한다.

3-1 스팸메일의 유형분석

조사대상 대학 서버의 2009년도 하반기 스팸메일 수를 보면, 전체 수신메일 459,885개의 수신 메일 중

에서 정상메일은 44,678개로서 9.5%의 정상메일과 89%에 해당하는 419,126개의 스팸메일, 1.2%에 해당하는 6,081개의 바이러스메일을 확인할 수 있었다. 이들의 내용을 보면, 대출관련 스팸메일이 40%로서 가장 큰 비율을 차지하고, 다음으로 성기능약품과 광고메일이 각각 20%, 자격증 또는 창업관련메일이 15%, 성인광고메일이 5%의 비율을 차지하였다.

스팸메일들이 스팸필터를 우회하기위해 사용하는 유형을 살펴보면, 크게 5가지로 분류될 수 있었다.

① 각종 랜덤 텍스트를 메일본문에 덧붙여 필터를 우회하게끔 메시지를 구성

② 메일에 포함되어 있는 HTML코드를 각종 방법으로 변조함으로서 필터를 우회

③ 메일에 포함되어 있는 URL스트링을 조작해서 필터를 우회

④ 메일 내용물에 신문기사의 일부 문단이나 명언 등을 랜덤하게 삽입

⑤ 메일 제목이나 본문 하단에 의미없는 쓰레기 문자를 삽입

즉, 이러한 방법들을 통해서 필터를 지능적으로 피하고 있었다. 각각의 유형에 대해서는 정규식 필터나, URL필터, 유해사이트 차단필터, HTML주석을 이용한 정규식 필터, Meta-character를 이용한 정규식 필터 등으로 효과적으로 차단 할 수가 있다.

문구 : (광고), [광 고], (광^^고)
차단방법 : 정규식 이용.
(\[| \<| \(| \ {|#| \$|%| \^|&| *| [|<| (| {|
|- 【).*광[,-~!@#$%^&*♥() \+|<>/.]*고.*(】
-| |} |) |>|] |!|@|#| \$|%| \^|&| *| }| \)|
\>| \])

그림 7. 한국어 스팸메일 예와 대응방법
Fig. 7. Response method and example for Korean spam mail.

하지만, 한국어의 경우 한 단어를 중심으로 사용자가 문법에 맞지 않는 변칙어를 생성 후 사용하는 다음과 같은 경우에는 정규식으로 접근이 불가능하게 된다.

▷문자를 경음화, 또는 격음화 시키는 경우
광고 → 꽝고, 섹스→쎅스
▷문자대신 비슷한 숫자로 대치하는 경우
비디오 → ㅂ1디오

그림 8. 일반 정규식으로 해결 불가능한 경우
Fig. 8. Impracticable case of regular expression.

이러한 경우는 정규식을 만들기 어렵기 때문에 일반적인 단어매칭을 하거나, 베이지안과 같은 학습알고리즘을 통하여 필터링하게 된다. 이는 필터링키워드와 일치할 경우에만 스팸으로 인식하므로, 가능한 변칙어들을 모두 등록시키거나 학습시켜야하며, 그만큼 공간과 시간비용을 소모해야만 하는 문제점을 갖게 된다.

본 연구의 서술을 위하여 우리는 변칙어를 다음과 같이 정의하고 진행하도록 한다.

정의 3. 한글 변칙어란, 교육적, 문화적인 통일을 위해 표준이 되게 정한 말인 표준말의 범위에 들지 않는 글자로서 문법규칙에 어긋나더라도 일시적으로 의미 전달만을위해 개인 사용자에 의해 임의로 생성되고 사용되어질 수 있는 말로 정의한다.

스패머는 기존의 스팸필터링을 통과하기위해 고의적으로 기존 표준어를 교묘하게 변형하여 사용하고 있으며, 지능적인 변형을 하고 있는 상황이다.

본 논문에서는 한글 단어의 유사도 측정방법을 도입함으로서 변칙어를 인식하는 방법을 제안하고자 한다. 이를 위하여 먼저, 한글 변칙어의 패턴을 다음과 같이 분석한다.

3-2 한글 변칙어의 패턴 분석

영어 변칙어의 경우, 단어를 구성하고 있는 철자 중에서 발음상 제한된 범위, 즉, 특정 철자범위 내에서만 변형이 발생하며, 이는 비슷한 숫자나 특수문자로 대체되어 사용되고 있었다.

정상어	변칙어
girl	gir1
drag	dr@g

그림 9. 영어의 변칙어 예
Fig. 9. An example of abnormal English.

다음은 한글의 변칙어의 예이다.

정상어	변칙어
섹스	쎅스, 색스, 세엑스, 섹쓰
포토샵	포토삽, 뽀토샵, 포오토샵, 포샵
비아그라	B아그라, ㅂ1아그라

그림 10. 한국어의 변칙어 예
Fig. 10. An example of abnormal Korean.

이처럼, 한글의 경우, 변칙어가 영어보다는 좀 더 다양하게 나타나게 되는데, 그 이유는 한글의 변형이 영어와는 달리, 음(음절)을 구성하는 음소(초성, 중성, 종성)의 미미한 변화로부터 제각기 변화될 수 있기 때문이다.

본 논문에서는 인터넷을 통한 전자문서(웹문서, 전자메일, 메신저)상에서 사용되고 있는 다양한 한글 변칙어의 특성을 크게 다음과 같이 분석하였다.

① 한글 변칙어는 음절이 아닌, 음절을 구성하는 자모의 변화로부터 발생된다. 하지만, 기본음은 전달되도록 변형이 된다.

예) 대출 : ㄷH출, 대츌

② 한글 변칙어는 기본 순서는 유지하면서 변형이 일어난다. 즉, 늘어뜨리거나 줄이더라도 자모의 기본 순서는 유지된다.

예) 인터넷 : 인터네앳, 인텟

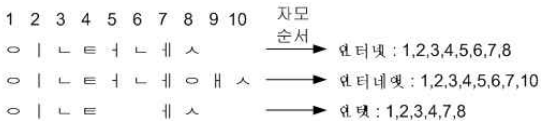

그림 11. 한글변칙어의 특성
Fig. 11. Characteristics of abnormal Korean.

즉, 표준어를 중심으로 발생되는 변칙어들을 보면, 자모의 순서가 유지되면서 변형이 일어남을 확인할 수 있다.

③ 어간과 어미 중 어미의 변화가 더 다양하다. 예외사항으로 경음화나 격음화, 특정자음이나 모음을 유사한 숫자나 영문으로 치환하는 경우가 많이 발생한다.

예) 어미를 변화시키는 경우

대출가능합니다 : 대출가능함니다. 대출가능하셈,

대출가능함.

예) 경음화나 격음화를 시키는 경우

신용불량 : 씬용불량, 비아그라 : 피아그라

④ 단어내의 임의의 위치에서 띄어쓰기를 함으로서 변형을 유도한다.

예) 법정금리 : 법 정금리

이러한 네 가지 특성으로부터 스팸 필터키워드로부터 변칙적으로 사용될 수 있는 변칙어들을 인식하기 위한 기준을 다음과 같이 설정한다.

① 변칙어가 스팸키워드와 동일한 단어인지 아닌지를 판단하기위해서는 단어의 음절비교가 아닌, 음소비교방법을 사용해야 한다.

② 발음상 늘어뜨리거나 줄이더라도 자모의 순서는 유지되므로, 순서에 입각한 비교를 해야 한다.

③ 앞의 음소와 뒤의 음소에 가중치를 두는 비중을 동일하게 하고 비교한다. 하지만 앞에서부터 가중치를 누적시킴으로서 스트링의 유사도를 계산한다.

④ 단어구분의 기분이 되는 공백이 임의의 위치에 사용될 수 있으므로, 비교를 위해서는 전처리과정을 통해서 추출된 명사, 형용사와 같은 독립된 품사별로 진행하는 것이 아니라, 문장전체를 비교하는 방법을 사용한다.

Ⅳ. Koral(Korean Abnormal Language)인식 모듈

본 논문에서는 웹메일에서 스팸키워드를 우회하기위해 사용될 수 있는 변칙어를 인식하기 위한 방법으로서 smith-waterman알고리즘을 적용한다. 본 장에서는 변칙어인식과 관련한 적용알고리즘의 특성을 분석하고, 이를 적용하기위한 방법인 Koral인식 모듈에 대해 설명한다.

4-1 한글변칙어 인식을위한 DPA의 특성

동적프로그래밍 알고리즘은 두 스트링간의 유사도를 결정하기위해서 사용될 수 있다. 기존의 단어비교법을 통해서는 음소의 미미한 변화가 있을 경우, 스팸키워드로 인식을 못하지만, 음소단위 비교를 통해서는 임계값에 따라 변칙어가 스팸키워드와 동일한 단어인지 아닌지를 판단할 수 있다.

스팸키워드 ↓ 변칙어 →

		ㄷ	ㅐ	ㅊ	ㅜ	ㄹ
	0	-1	-2	-3	-4	-5
ㄷ	-1	2	1	0	-1	-2
ㅐ	-2	1	1	0	-1	-2
ㅊ	-3	0	0	3	2	1
ㅜ	-4	-1	-1	2	5	4
ㄹ	-5	-2	-2	1	4	7

그림 12. DPA의 특성분석 I

Fig. 12. Analysis of DPA characteristics I.

위 그림은 점수함수로 match: +2, mismatch:-1, space:-1을 적용하여 “대출”과 "ㄷH출"의 최적 정렬값을 구한 것이다. 정렬값은 행과열의 마지막 엔트리 값인 7이 된다. 완전일치를 가정한다면, 대각선이 모두 +2씩 증가하므로, 10이 되고, 완전 불일치일 때는, -5가 된다. 두 단어간의 유사도를 구하기 위해 다음의 공식을 사용한다.

$$sim = \frac{applied\,agreement}{complete\,agreement} \quad 식 \qquad (1)$$

식 (1)로부터 유사도값(sim)의 범위는 $-0.5 \le sim \le 1$이 된다. 즉, 위 예제는 일치도가 7/10=0.7로서 임계값에 따라 동일한 단어로 인식될 수 있다. 하지만, 기존의 단어비교법을 통해서는 불일치하므로, 스팸키워드로 인식되지 못하고, 필터링을 통과하게 된다.

		ㅇ	ㅣ	ㄴ	ㅌ	ㅓ	ㄴ	ㅖ	ㅇ	ㅐ	ㅅ
	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
ㅇ	-1	2	1	0	-1	-2	-3	-4	-5	-6	-7
ㅣ	-2	1	4	3	2	1	0	-1	-2	-3	-4
ㄴ	-3	0	3	6	5	4	3	2	1	0	-1
ㅌ	-4	-1	2	5	8	7	6	5	4	3	2
ㅓ	-5	-2	1	4	7	10	9	8	7	6	5
ㄴ	-6	-3	0	3	6	9	12	11	10	9	8
ㅖ	-6	-4	-1	2	5	8	11	14	13	12	11
ㅅ	-6	-5	-2	1	4	7	10	13	13	12	14

그림 13. DPA특성분석 II.

Fig. 13. Analysis of DPA characteristics II.

그림 13은 “인터넷”과 “인터네앳”에 대한 최적정렬값을 구한 것으로 유사도는 0.875이다. 압축되어 변형이 일어난, "인텟“과는 0.555의 유사도를 갖는다. 그림15의 경우, 자모의 순서를 바꿔 ”인터앳네“으로 계산하면, 0.437로서 자모의 순서가 바뀔 경우, 완전일치도 값이 떨어지게 된다. 즉, 전체 사용된 자모의 개수가 일치하더라도 순서가 바뀔 경우는 유사도 값이 떨어짐을 알 수 있다.

변칙어의 경우, 앞부분과 뒷부분에서 동일한 심벌 개수 별로 변화가 일어난다면, 동일한 유사도 값을 갖도록 해야 하므로, 처음부터 끝까지 동일한 점수함수를 적용시킴으로서 해결된다.

예) “대출가능“ & ”대출카능“ : 0.85

“대출가능” & “ㄷH출가능” : 0.85

임의의 위치에서 띄어쓰기가 일어날 경우, 메일로부터 메일정보를 추출할 때, 정확하게 품사를 추출하고 이를 비교하는 것이 의미가 없어진다. 따라서, 비교 시에는 Needleman-Wunsch 알고리즘 보다는 Smith-Waterman알고리즘을 적용함으로서 이러한 문제를 해결 할 수 있다. 즉, “법정금리”와 “법 정금리”의 경우, 각각을 분리하여 비교하면 의미가 없어지지만, 분리하지 않고, 전체를 비교함으로서 0.954의 유사도를 얻을 수 있다. 따라서, 기존의 단어비교법을 통해서 스팸으로 분류되지 않던 것이 스팸으로 분류될 수 있다.

4-2 알고리즘의 상세

본 논문의 목적은 기존 스팸메일 필터링에서 발생될 수 있는 변칙어로인한 필터링의 정확도 및 효율성 저하의 문제를 해결하고자 변칙어를 인식할 수 있는 방법을 제안하는데 있다. 즉, 일반적인 필터링시스템 중에서 본 연구의 초점은 다음 그림으로부터 구분될 수 있다.

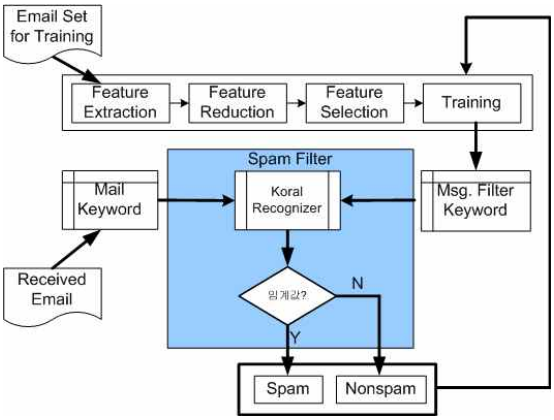

그림 14. 스팸필터링 시스템의 구조
Fig. 14. Structure of spam mail filtering system.

그림 14는 전체 메일 필터링시스템의 개념도로서 학습을 통해 추출된 필터키워드와 수신된 메일로부터 전처리과정을 거쳐 얻어진 메일 키워드를 본 논문에서 제안하는 Koral인식기를 통해 비교함으로서 스팸/비스팸을 결정하고 있다. 결정된 자질들은 다시 학습모듈로 순환되고 있다.

한국어 변칙어 인식모듈은 부분최적정렬을 구하기 위해 사용될 수 있는 Smith-Waterman알고리즘을 사용한다. 이는 3.2절의 변칙어의 특성상 임의의 위치에서 띄어쓰기를 함으로서 스팸필터를 우회하는 것을 막기 위한 방법으로 사용한다. 즉, 전처리과정을 거친 정제된 품사에 대해 비교하는 것이 아니라, 한 문장 전체와 필터키워드를 비교함으로서 가장 유사한 부분을 찾아, 그 부분의 유사도 값을 필터링에 사용하고자 하는 것이다.

제안하는 Koral인식 모듈의 작동순서 다음과 같다.

1. 메시지필터키워드와 임계값설정.
2. 메시지필터키워드와 메일키워드의 자소분리.
3. 그림 6 실행.→부분 최적정렬값 추출.
4. 유사도 계산.
5. 임계값 비교.

참, 메일을 spam으로 인지.
학습알고리즘 전달(필터키워드, 메일키워드)
거짓, 메일을 nonspam으로 인지.
학습알고리즘 전달(메일키워드).

V. 실 험

본 실험에서는 Koral모듈의 임계수준에 따른 한글 변칙어 인식을 파악한다. 사용되는 동적프로그래밍 알고리즘이 서열의 크기가 커질 경우, 시간과 공간비용이 문제가 될 수 있으므로, 이와 관련한 요구 공간 및 수행시간을 분석한다.

5-1 실험환경

제안하는 인식모듈은 다음의 일반적인 소프트웨어와 하드웨어 환경 하에서 구현하고 결과값을 비교한다. 즉, Intel Pentium Processor 1.73GHz, 1G RAM의 하드웨어와 Windows XP Professional.운영체제 환경하에서 Microsoft C++로 구현하고 실험한다.

테스트를 위한 필터링키워드는 대학의 메일서버에서 6개월 동안 메일의 제목과 본문에서 수집하였고, 상위 3019개의 키워드들 중에서 한글 변칙어로 판단되는 키워드들을 50개의 그룹으로 분류하였다. 그리고, 키워드와 변칙어들을 그룹별로 Koral모듈을 통해 인식하도록 하였다. 그로부터 임계값에 따른 인식정도를 파악한다. 시간과 공간분석을 위한 테스트 메일은 subject와 body에 포함된 문자의 개수를 달리하는 메일을 선택한다. DPA를 시행하기위하여 선택된 메일을 음절단위에서 자소단위로 분해 후, 생성된 엔트리의 개수는 각각, 100, 204, 393, 807, 1875, 3794이다. 본 실험에서는 서버에 등록된 2081개의 필터키워드를 조사한 결과 총 44192개의 문자를 갖고 있었고, 한 키워드 당 평균 28개의 자소를 갖고 있었다. 이에, 필터키워드와 subject와 body메시지를 포함하는 메일을 1:1로 비교할 때를 가상하여, 28개의 문자열과 100, 204, 393, 807, 1875, 3794개의 문자열에 대해 테스트하고, 소요 시간 및 공간을 확인한다.

실험에는 점수함수로 match : +2, mismatch : -1, space : -1을 적용하였다.

5-2 실험결과

필터키워드	지역최적정렬값	유사도
대출상담후 10분	7	0.7
대-출	6	0.6
대i출	9	0.9
은행권대출	10	1
대출 최저금리	10	1
대출 긴급자금	10	1
대..출	8	0.8
대/출/	9	0.9
대i출	9	0.9
대^*출	8	0.8
대출한도 상향 조정되었습니다	10	1
대^출	9	0.9
대^^출	8	0.8
ㄷ에출	6	0.6
ㄷH출	7	0.7
대출..이젠 망설이지마세요	10	1
(회사원I공무원),+?최저금리대출	10	1
100% 대출가능	10	1
대납대출전문	10	1

그림 15. “대출”의 유사도
Fig. 15. Similarity value for "대출“.

위 그림은 필터시스템에 수동 또는 학습에 의해 등록되어있는 필터 키워드들 중에서 “대출”그룹에 속하는 단어들 중 일부에 대하여 유사도 비교를 한 것이다. “대출”이라는 글자를 완전히 포함하고 있는 문자열은 유사도 1로서 모두 인식되고, 어둡게 표시된 부분이 변칙어를 사용한 경우에 해당되며, 모두 60%이상의유사도 범위 안에서 인지될 수 있음을 확인할 수 있다. 즉, 키워드에 따라서, 임계값을 다르게 함으로서 변칙어들을 세분화하여 분류할 수 있다. 예를 들어, 임계값을 0.7로 하였을 때, “대출”, “대츨”, “대i출”, “대..출”, “대^*출”, “대^출”, “대^^출”, “ㄷH출”을 인식한다. 나머지 49개의 그룹에 대해서도 동일하게 유사도를 추출할 수 있었다. 특히, 변칙어가 많이 나타나는 육두문자의 경우, 더욱 세분화하여 분류할 수 있었다.

이러한 결과로부터 학습알고리즘에 제공되는 키워드는 기존의 동일한의미의 변칙어를 제공하는 것이 아니고, “대출”이라는 대표키워드만을 제공한다.

자소길이		시행 횟수										평균
Filter Keywo	Mail	1	2	3	4	5	6	7	8	9	10	
28	6493	0.047	0.031	0.047	0.031	0.047	0.047	0.031	0.031	0.031	0.031	0.037
28	3749	0.031	0.016	0.031	0.031	0.016	0.031	0.016	0.031	0.016	0.016	0.024
28	1875	0.016	0	0.016	0.016	0.016	0.016	0.016	0	0.016	0.016	0.013
28	807	0.016	0.016	0	0.016	0	0.016	0	0.016	0.016	0	0.012
28	393	0.016	0	0.015	0.015	0	0.016	0.015	0	0.016	0	0.009
28	204	0.016	0	0.016	0	0	0	0	0.016	0	0	0.005
28	100	0	0	0	0	0.015	0	0	0	0.015	0	0.003

그림 16. 메일 크기에 따른 소요시간
Fig. 16. Required time by mail size.

그림 16은 “직장인I.*?대출”이라는 키워드와 자소길이가 100~ 6493개를 갖는 메일과의 유사도 측정 시 소요되는 시간을 실험한 것으로서 10회 시행 후 평균값을 대표값으로 하였다. 단일 필터키워드로 적용할 때 메일의 크기가 커지더라도 처리시간이 제한요소로 작용하지 않는다는 것을 알 수 있다.

unit : byte 2

Filter Keyword	28	28	28	28	28	28	28
Mail	6,493	3,749	1,875	807	393	204	100
Required Memory	363,608	209,944	105,000	45,192	22,008	11,424	5,600

그림 17. 메일의 크기에 따른 요구공간
Fig. 17. Required space by mail size.

그림 17은 위 실험에 대한 요구공간을 나타낸 것으로 엔트리의 단위크기를 2Byte로 하였을 때, 요구공간이다. 메일의 자소크기가 6,493일 때에, 363KB를 소요한다.

VI. 결 론

본 논문에서는 광고목적이나 악의를 갖고 접근하는 스패머들에 의해 사용되는 이러한 유형의 단어들을 변칙어로 정의하고, 이를 인식하기 위한 방법을 제안하였다. 방법의 적용을 위해 기존 필터링 시스템에서 유지하고 있는 필터링 키워드들을 분석하고, 특히 한글 변칙어의 유형및 특성을 분석하였다. 그로부터 본 논문에서는 두 서열간의 유사도를 측정하여 일치하는 부분만을 찾아낼 수 있는 Smith-Waterman알고리즘을 사용한 결과 제안한 접근방법은 특정 키워드를 A라 할 때, 메일의 subject나 body에 사용된 A의 변칙어들을 유사도 수준에 따라 정확히 인식해 낼 수 있음을 보였다. 또한 인식에 요구된 시간과 공간비용은 단일 비교만을 시행했을 때, 충분히 허용될 수 있음을 확인하였다.

제안한 방법은 학습과정을 수행하기 전 단계에서 작동하며 학습이전에 필터키워드와 유사한 단어들을 검색함으로서 독립적으로 스팸필터링을 수행할 수 있으며, 학습시스템에 정확한 스팸키워드를 제공할 수 있다. 그로인해 학습시스템의 효율도 향상시킬 수 있을 것으로 기대한다. 특히 한국어의 경우 특성상 다른 언어에 비해 더 많은 변칙어가 출현할 수 있으므로, 효과적일 것으로 기대한다.

본 한글 변칙어 인식모듈은 단일키워드와 한 개의 메일문서를 크기만 변화시키면서 실험하였는데, 향후, 다중키워드와 여러 개의 메일을 시스템 환경에 적용시킴으로서 필터링의 효과를 정확도 및 공간, 시간비용의 관점에서 분석할 필요성이 있다.

참 고 문 헌

[1] 한국인터넷진흥원, "2010 국가정보보호백서(National Informatization Protection White Paper)", pp. 107-109, 2010.

[2] 이우권. "사이버공간의 스팸메일 규제정책에 관한 연구", *규제연구 제 13권 2호*, 12월, 2004.

[3] L. H. Gomes and C. Cazita, "Characterizing a Spam Traffic.," in Proc. 2004 Internet Measurement Conference, *Taormina, Sicily*, Italy. Oct. 2004.

[4] V. Keselj, E. Milios, A. Tuttle, S. Wang, and R. Zhang. "TREC 2005 Spam Track: Spam Filtering Using N-gram-based Techniques", *Proceedings of Text REtrieval Conference*, 2005.

[5] 김현준, 정재은, 조근식, "가중치가 부여된 베이지안 분류자를 이용한 스팸메일 필터링 시스템 " *정보과학회논문지, 31권 8호*, pp.1092-1100, 2004.

[6] R. Segal. "IBM SpamGuru on the TREC 2005 Spam Track," *Proceedings of Text REtrieval Conference,* 2005.

[7] Al Brakto, B. Filipic. "Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track," *Proceedings fo Text REtrieval Conference*, 2005.

[8] L. A. Breyer. "DBACL at the TREC 2005," *Proceedings of Text REtrieval Conference*, 2005.

[9] http://www.csie.ntu.edu.tw/~cjlin/libsvm

[10] 공미경, 이경순, "스팸성 자질과 URL 자질의 공동학습을 이용한 최대 엔트로피 기반 스팸메일 필터 시스템," *정보처리학회 논문지B, 15-B권 1호*, pp.61-68, 2008.

[11] F. Zhou, L. Zhuang, B. Zhao, L. Huang, A. Joseph, and J. Kubiatozicz, "Approximate object location and spam filtering on peer-to-peer systems," in Proc. Middleware, *Rio de Janeiro, Brazil*, June 2003.

[12] 이성욱, "카이제곱 통계량과 지지벡터기계를 이용한 스팸메일 필터," *정보과학회 논문지B, 17-B권 3호*, pp.249-254, 2010.

[13] S. B. Needleman and C. D. Wunsch. "A general method applicable to the search for similarities in the amino acid sequences of two proteins," *Journal of Molecular Biology*. vol. 48: 443-453, 1970.

[14] Wagner, R. A. and Fischer, M. J. "The string-to-string correction problem," *J. ACM 21,* 168-173, Jan. 1974.

[15] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *Journal of Molecular Biology,* vol. 147(1): 195-197, Mar. 1981.

안 희 국 (安熙國))

2003년 2월 : 강원대학교 컴퓨터과학과 이학석사
2007년 2월 : 강원대학교 컴퓨터과학과 이학박사
2011년 현재 : 강원대학교 강사

관심분야 : 텍스트마이닝, 생물정보학, 알고리즘

한 욱 표 (韓旭彪)

1993년 2월 : 강원대학교 전자계산학과 학사
1996년 2월 : 강원대학교 전자계산학과 석사
2003년 : 강원인터넷대학 실장
2007년 8월 : 강원대학교 컴퓨터과학과 박사
2011년 현재 : 강원대학교 강사

관심분야 : 유비쿼터스컴퓨팅, 센서네트워크, 네트워크보안

신 승 호 (辛承浩)

1998년 : 수학과 조교
2000년 : 철도경영연수원 사이버운영팀
2001년 : 강원대학교 컴퓨터과학과 석사
2005년 : 강원대학교 컴퓨터과학과 박사 수료
2011년 현재 : 강원대학교 강사

관심분야 : 수치해석, 알고리즘, 암호학, 웹프로그래밍

양 동 일 (梁東一)

2004년 2월 : 강원대학교 컴퓨터과학과 이학석사
2007년 8월 : 강원대학교 컴퓨터과학과 이학박사
2011년 현재 : 한림성심대학 인터넷비즈니스과 교수

관심분야 : 소프트웨어공학, 유비쿼터스, 포렌식

노 희 영 (盧熙瑩)

1972년 고려대(문학사-독문학)
1978년 독일 도르트문대
(Vordiploma-전자계산학)
1982년 독일 도르트문대
(Diploma-전자계산학)
1983년 한국표준연구소 전자계산실
선임연구원
1984년 2월~현재 : 강원대학교 교수
관심분야 : 컴파일러, 자연어처리, 소프트웨어공학