김일성종합대학학보
(자연과학)
주체104(2015)년 제61권 제7호

JOURNAL OF **KIM IL SUNG** UNIVERSITY
(NATURAL SCIENCE)
Vol. 61 No. 7 JUCHE104(2015).

스팸전자우편려과를 위한 확장쏘프트웨어 구성방식의 한가지 방법

윤희광, 손금철

위대한 령도자 **김정일**동지께서는 다음과 같이 교시하시였다.

《우리는 정보기술, 나노기술, 생물공학을 발전시키는데 선차적으로 힘을 넣어야 하며 그 중에서도 정보기술 특히 프로그람기술을 빨리 발전시켜야 합니다.》(《**김정일**선집》 증보판 제22권 21페지)

현재 전자우편이 광범히 보급되여 리용되는것과 관련하여 전자우편에 대한 연구 특히 스팸우편에 대한 연구가 더욱 심화되고있다.

그 처리방법에는 whitelist, blacklist, 호평순위를 리용하는 방법, 우편료금을 제한하는 방법, 내용려과기를 리용하는 방법 등이 있다.[1－3]

스팸려과에서 내용에 기초한 스팸려과처리는 세계적으로 제일 많이 리용되고있으며 지금까지도 다른 모든 방법보다 더 우월한 스팸처리기술로 리용되고있다.

론문에서는 스팸려과기의 전처리로서 동적다중표준기(dynamic multiple normalizer)를 제기하였다. 즉 전자우편으로부터 평문형식의 본문으로의 본문표준화를 진행하여 평문형식의 본문을 처리하게 함으로써 스팸려과기의 성능을 높일수 있는 구성방식을 제기하였다.

1. 스팸우편처리방법에 대한 론의

현재 세계적으로 널리 리용되고있는 스팸처리방법에는 blacklist(whitelist)를 리용하는 방법, 료금제한방법(Postage approach), 내용려과방법 등이 있다.[3]

blacklist방법은 blacklist(감시대상자명부)에 등록된 스팸사용자로부터 오는 전자우편을 모두 스팸으로 간주하여 무시하는 방법이며 whitelist방법은 그와 반대로 whitelist에 등록된 사용자만이 우편을 전송하게 하는 방법이다. 료금제한방법은 전송비를 제한하여 비용이 많이 들어 보낼수 없는 대량우편과 같은 스팸우편들을 차단하게 하는 방법이다.

지금까지 가장 많이 리용되는 방법은 내용에 기초한 스팸전자우편려과방법이다. 이 방법은 려과엔진이 식별된 스팸전자우편과 비스팸전자우편들에 대한 학습을 진행하여 스팸전자우편을 검출할수 있게 한다.

따라서 내용에 기초한 스팸전자우편려과기들이 널리 리용되고있으며 상업용제품으로 출하되고있다.

내용에 기초한 스팸려과기는 크게 두가지 문제점을 가지고있다.

첫째로, 스팸려과기가 다종다양한 스팸우편내용에 대한 심도있는 학습을 진행하기가 어려운것이다.

둘째로, 스팸려과기의 학습자료가 어떤 특정한 형식으로 갖추어지지 않음으로써 려과기가 학습을 할수 없게 하는것이다.

내용에 기초한 스팸려과를 진행하는 내용스팸려과기들의 거의 모두는 전자우편본문에 대한것이며 일부 학습에 리용되는 화상과 HTML꼬리표들, 첨부화일들에서는 스팸의 성질이 잘 나타나지 않는다.

실지로 스팸작성자들은 우편이 스팸이라는것을 나타내는 본문을 숨기기 위하여 할수 있는 모든 방법을 리용한다. 아래에 스팸려과기들을 속여넘기는 일부 방법들을 보여주었다.

① 고의적으로 잘못된 단어를 리용하는것이다.

② 전자우편첨부화일인 화상에 본문을 숨기는것이다.

③ HTML꼬리표를 리용하는것이다.

이 방법들의 결함은 스팸려과기들이 전자우편을 정확히 식별할수 없게 하는것이다.

우리는 스팸전자우편려과를 보다 효과적으로 진행하기 위한 우편봉사기구성방식을 제기하였다. 즉 내용에 기초한 스팸전자우편려과기의 전처리기로서 다중동적표준기(dynamic multiple normalizer)를 정의하였다. 이 표준기는 봉사기에 우편이 도착한 다음 려과기가 스팸식별을 하기 전에 전자우편을 이루는 내용들을 한단계 한단계 평문형식으로 변환한다. 려과기는 표준기를 거친 평문형식의 본문만을 처리하여 스팸식별을 보다 정확히 할수 있다.

2. 스팸전자우편려과를 위한 확장쏘프트웨어구성방식

확장구성방식에서는 표준기를 리용하여 많은 방법으로 은폐된 스팸통보문들을 처리하는 능력을 향상시켜준다.

제안한 확장쏘프트웨어구성방식은 그림과 같다.

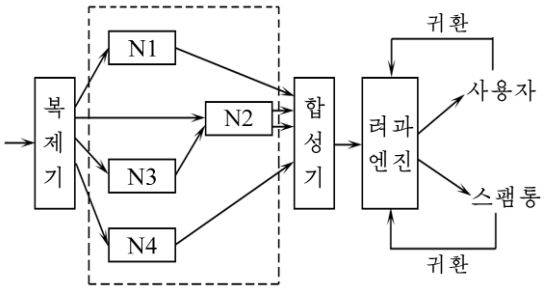

그림. 확장쏘프트웨어구성방식

그림에서 들어오는 전자우편은 먼저 복제기에 의해서 복제되고 표준기들(N1부터 N4)에 전송된다. 개개의 표준기는 전자우편의 원래 본문을 회복시킨다. 마지막에 합성기는 표준기들로부터 복귀된 본문들을 하나의 본문으로 합치고 그것을 스팸려과엔진에로 보낸다. 다음 엔진은 이 본문을 려과엔진에로 보내여 스팸인가 아닌가를 결정한다.

모든 표준기들은 특정한 입력형식을 접수하며 그것을 다른 형태로 변환하여 출력한다. 실례로 HTML표준기는 HTML형식을 접수하여 HTML꼬리표들을 떼여버리고 평문형식으로 출력한다. OCR표준기는 jpeg나 gif와 같은 화상형식을 접수하고 OCR기술을 리용하여 화상으로부터 본문을 추출하며 추출한 본문을 평문형식으로 출력한다. Trigram이나 Markov표준기는 잡음이 있는 본문 즉 철자가 잘못된 단어들을 접수하고 정확한 철자로 회복한다.

이때 모든 표준기들은 다 려과엔진이 요구하는 평문형식을 생성하지는 않는다. 실례로 OCR표준기는 전자우편화상들로부터 본문을 추출하여 평문형식으로 출력하기로 되여있다. 하지만 OCR표준기에 의해 생성된 본문은 많은 OCR잡음들을 포함하고있다. 즉 잘못 인식한 철자오유와 같은것들을 포함하고있다. 대신에 그것을 다른 표준기 실례로 Trigram이나 Markov 표준기로 전송하여 OCR본문의 잡음을 제거하고 정확한 철자를 복귀한다.

표준기는 (입력형식, 출력형식)의 접수가능한 입출력형식으로 정의한다. 즉 HTML표준

기는 (HTML, 본문)으로, OCR표준기는 (화상, 잡음섞인 본문)으로, Trigram표준기는 (잡음섞인 본문, 본문)으로 정의된다.

한편 입출력형식은 간단한 형식으로 할수 있다. 먼저 복제기는 들어오는 우편을 식별하는 유일한 번호를 생성하고 우편을 가능한 모든 형식 실례로 본문과 HTML, 화상 등으로 다중흐름복제를 한다. 다음 표준기는 합성기에 이 유일번호와 흐름번호를 통지한다. 매개 흐름들은 우편의 유일식별값을 가지며 표준기의 입력형식에 기초하여 합성기로 전송된다. 표준기나 표준기망은 전자우편을 스팸려과기에서 리용할 평문형식으로 만든다. 합성기는 복제기로부터 유일번호를 받아 여러 흐름들로부터 들어오는 전자우편을 식별한다. 그리고 흐름들로부터 들어오는 본문부분들을 하나의 본문으로 련결시키고 그것을 려과엔진으로 보낸다.

들어오는 우편이 표준기들을 통과하는 순서는 뒤바뀌여진다. 그러나 매개 표준기들의 입력형식과 출력형식을 명백하게 정의한 후 입출력형식을 사슬지음으로써 표준기들을 통과하는 통로를 동적으로 찾을수 있다. 즉 표준기는 입력형식과 출력형식에 의해서만 정의되며 표준기들은 그 형식에 기초하여 동적으로 사슬지어진다. 최종목적은 려과엔진을 위한 평문형식을 만들어내는것이다. 그림에서 점선으로 된 둥근4각형안의 경로들은 들어오는 전자우편의 내용에 기초하여 동적으로 구성된것이다. 표준기는 체계에 쉽게 추가되거나 제거될수 있다. 체계안에서 필요하다면 한 표준기는 다른 표준기와 사슬로 련결될수 있다.

맺 는 말

스팸우편봉사기를 실현하기 위하여 스팸우편봉사기 postfix의 견지에서 론의한 확장구성방식과 표준기를 제기하였으며 그것으로 하여 스팸우편려과기의 성능을 향상시키였다.

참 고 문 헌

[1] J. Jung et al.; ACM Sigcomm Internet Measurement Conference, 370, 2004.

[2] A. Kumar et al.; ACM Sigcomm Internet Measurement Conference, 3, 2005.

[3] B. Mobasher et al.; Effective Attack Models for Shilling Item-Based Collaborative Filtering System, Springer, 10～23, 2005.

주체104(2015)년 3월 5일 원고접수

A Method of Extendable Software Architecture for Spam Email Filtering System

Yun Hui Gwang, Son Kum Chol

We proposed a method to improve the performance of spam email filtering system.

There are many content-based spam email filters but these filters are not flexible enough to adapt the new development of spam techniques.

To solve this problem we proposed extendable software architecture consist of dynamic multiple normalizers that converts an email to its plain text format as the preprocessors for spam email filters.

Key words: spam, spam filter, spam detection, normalizer