링크구조분석을 이용한 스팸메일 분류
(A Spam Mail Classification Using Link Structure Analysis)

이 신 영 † 길 아 라 †† 김 명 원 ††
(Shin Young Rhee) (Ara Khil) (Myung Won Kim)

요 약 기존의 내용기반 스팸메일 분류는 전자메일이 이미지를 많이 가지고 있고 텍스트는 적게 가지고 있을 경우에는 내용을 분석하기 어려우므로 스팸메일을 분류하는 데 한계가 있다. 이와 같은 문제를 해결하기 위하여 본 논문에서는 전자메일의 구조를 분석하는 링크구조분석 스팸메일 분류 알고리즘을 제안한다. 이것은 전자메일 안의 하이퍼링크의 개수와 하이퍼링크가 가리키는 웹 문서들이 다른 웹 문서에 의해 링크된 수를 측정하여 전자메일의 중요도를 계산한 후 의사결정트리를 학습하여 스팸메일과 정상메일을 분류한다. 또한 위의 링크구조분석 알고리즘과 하이퍼링크의 서버 주소만을 이용한 변형된 링크구조분석 알고리즘, 그리고 SVM(support vector machine)을 이용한 내용기반 방법을 다수결 원칙으로 결합한 통합 스팸메일 분류 시스템을 제안한다. 실험 결과, 제안한 링크구조분석 알고리즘은 기존의 내용기반 방법 보다 스팸메일 분류 정확도가 94.8%로 약간 향상되었으며 또한 통합 스팸메일 분류 시스템도 내용기반 방법과 비교하여 향상된 97.7%를 나타냈다.

키워드 : 스팸메일 분류, 링크구조분석, 의사결정트리, SVM

Abstract The existing content-based spam mail filtering algorithms have difficulties in filtering spam mails when e-mails contain images but little text. In this thesis we propose an efficient spam mail classification algorithm that utilizes the link structure of e-mails. We compute the number of hyperlinks in an e-mail and the in-link frequencies of the web pages hyperlinked in the e-mail. Using these two features we classify spam mails and legitimate mails based on the decision tree trained for spam mail classification. We also suggest a hybrid system combining three different algorithms by majority voting: the link structure analysis algorithm, a modified link structure analysis algorithm, in which only the host part of the hyperlinked pages of an e-mail is used for link structure analysis, and the content-based method using SVM (support vector machines). The experimental results show that the link structure analysis algorithm slightly outperforms the existing content-based method with the accuracy of 94.8%. Moreover, the hybrid system achieves the accuracy of 97.6%, which is a significant performance improvement over the existing method.

Key words : Spam Mail Classification, Link Structure Analysis, Decision Tree, SVM

1. 서 론

스팸메일(spam mail)은 unsolicited bulk e-mail 또는 junk mail이라고도 불리는데, 요청하지 않았음에도 다수의 수신자 집단에게 발송되는 광고성 인터넷 메일이다[1]. 스팸메일은 특성상 발송 비용이 거의 들지 않아 대량 발송이 쉬우며, 이런 대량의 스팸메일은 개인의 업무 시간을 빼앗고 스트레스를 주며 개인, 기업, 국가적으로도 커다란 피해를 주는 정보화 사회의 대표적인 역기능이다.

이러한 스팸메일은 스팸메일을 보내어 공격하는 측과 이를 차단하여 방어하는 측의 공격과 방어로 볼 수 있는데 이런 최근의 스팸메일의 공격 특징과 그에 대한 방어 방법들의 문제점을 분석하면 크게 다음과 같이 나눌 수 있다.

• 스팸메일 안의 단어 변조를 통한 공격

이 공격에 대한 방어 방법은 스팸메일 방어 방법의 가장 기초적인 방법인 특정 단어 정합(matching)을 통해 스팸메일을 분류하는데 스팸메일 공격자가 매우 다

· 본 연구는 서울시 산학연 협력사업(10581cooperateOrg93111)의 연구결과로 수행되었음
† 학생회원 : (주)시뮬레이션연구소
bearrhee@naver.com
†† 정 회 원 : 숭실대학교 컴퓨터학부 교수
ara@comp.ssu.ac.kr
mkim@comp.ssu.ac.kr
논문접수 : 2006년 8월 16일
심사완료 : 2006년 11월 20일

양하고 빠르게 단어 변조를 하고 방어하는 측은 변조한 단어에 대한 규칙을 뒤늦게 정의해야 하므로 스팸메일 분류율이 매우 낮으며 방어하는데 소요되는 비용도 많은 편이다.

- 중국 등 국내가 아닌 외국의 서버를 통한 스팸메일 발송과 잦은 서버 및 발신 주소 변경

이 공격에 대해서 스팸메일 방어자는 스팸메일 발신 서버나 발신 주소를 블랙리스트에 추가하여 방어를 하는데, 스팸메일 공격자가 발신 서버나 발신 주소를 변경하는데 비용이 적게 소모되고 스팸메일 방어자는 공격자가 공격한 후에 블랙리스트에 추가하는 작업 밖에 못하므로 역시 스팸메일 분류율이 매우 낮고 비용도 많이 소요된다.

- 본문을 조금씩 바꾸거나 본문 안에 임의의 코드를 삽입하여 공격

스팸메일 방어자가 스팸메일의 본문의 정합을 통해 스팸메일을 블랙리스트에 추가하면 스팸메일 공격자가 발신 서버나 발신 주소를 변경하더라도 스팸메일 본문의 내용 때문에 계속 스팸메일을 방어할 수 있다. 이에 스팸메일 공격자는 본문의 정합을 피하기 위해 본문을 조금씩 바꾸거나 임의의 코드를 무작위로 삽입하여 공격한다.

- 내용기반 스팸메일 차단을 피하기 위해서 스팸메일 내용을 텍스트 대신에 이미지로 제작하여 공격

특히 이 공격은 국내 스팸메일의 거의 대부분을 차지하는데, 스팸메일에서 제목 부분만 텍스트로 이루어져 있고 본문은 모두 이미지로 이루어진 경우 전자메일의 본문에는 HTML 태그와 하이퍼링크에 대한 정보만 있을 뿐이고 스팸메일이라고 판단할 텍스트가 거의 없기 때문에 이미지의 내용을 문자인식을 통해서 텍스트를 추출해 내지 않는 이상 기존의 내용기반 스팸메일 분류 방식으로는 분류에 많은 어려움을 겪는다.

이와 같이 스팸메일의 공격형태가 빠르고 다양하게 변하며 텍스트 대신 이미지를 사용하기 때문에 기존의 스팸메일 분류 방법은 많은 한계를 가지게 된다. 이에 본 논문에서는 기존의 스팸메일 분류 방법인 스팸메일의 내용을 기반으로 분석하는 것이 아닌 스팸메일의 전체 웹(web)상에서의 링크 구조를 분석하여 스팸메일을 분류하는 방법을 제안한다.

본 논문에서는 웹 문서들 간의 링크 구조를 이용하여 웹 문서의 중요도에 의해 스팸메일을 분류하는 링크구조분석이라는 새로운 방식의 스팸메일 분류 알고리즘을 제안하며 또한 기존의 내용기반 스팸메일 분류와 통합하여 더욱 스팸메일 분류율을 높인 통합 시스템도 제안한다.

본 논문의 구성은 2장에서 기존의 스팸메일 분류 방법과 관련 연구를 기술하고, 3장에서는 본 논문에서 제안하는 링크구조분석 스팸메일 분류 알고리즘을 기술하고 내용기반 방법과의 통합 시스템을 기술한다. 4장에서는 실험 결과를 기술하고 마지막으로 5장에서는 결론 및 향후 연구 방향을 기술한다.

2. 관련 연구

2.1 기존의 스팸메일 분류 방법

스팸메일을 차단하는 기존의 방법으로는 크게 다음과 같은 것들이 있다.

• 목록기반 분류

전자메일 서버 차원에서 발신자 전자메일의 블랙리스트와 화이트리스트를 만들어 스팸메일을 차단한다. 그러나 이 방법은 스팸메일 공격자가 전자메일 주소를 계속 바꾸어 공격하고 리스트를 갱신하는데 비용이 많이 들며 최신의 스팸메일에 대해 느리게 반응한다는 단점이 있다.

• 규칙기반 분류

전자메일의 헤더, 제목, 본문을 분석하여 분류하고자 하는 특정 단어가 정합되었을 때에 전자메일을 차단하는 방식이다. 그러나 특정 규칙을 잘 정의하기 힘들고 스팸메일 공격자 또한 이러한 규칙에 빠르게 대응한다. 특히 스팸메일 공격자가 "광고"를 "광xx고", "광◇고"등과 같이 전자메일의 특정 단어에 다른 문자를 섞어서 단어를 변조하면 규칙기반 분류로는 분류하기 어려운 단점이 있다.

• 내용기반 분류

이는 스팸메일과 스팸메일이 아닌 정상메일(legitimate mail)을 샘플로 하여 불용어(stop-word)를 제거하고 스테밍(stemming) 등의 전처리 작업을 한 후, 의사결정 트리(decision tree), 나이브 베이지안(naive Bayesian), SVM(support vector machine), 인공신경망(artificial neural networks) 등의 기계 학습 방법으로 학습한 후 전자메일을 분류하는 방식이다. 최근에는 나이브 베이지안에 의한 스팸메일 분류[2]나 SVM을 이용한 스팸메일 분류가 연구되었다[1,3,4]. 그러나 최근의 스팸메일은 본문의 내용을 텍스트가 아닌 이미지로 보내는 경우가 대부분이어서 내용기반 분류에 한계가 있다.

• 협업적 분류

이는 한 전자메일 서버안의 사용자들이 스팸메일이라고 신고한 전자메일들은 다른 사용자에게도 스팸메일이라고 판단하여 차단하는 방법이다. 이 방법은 전자메일 포털 서비스 업체를 중심으로 사용되고 있고 스팸메일 분류율은 좋은 편이나 서로 다른 전자메일 서버 간에 스팸메일에 대한 정보를 공유하지 않아서 한 전자메일 서버 안에서만 사용해야 하는 단점이 있다. 또한 같은

표 1 형태소 색인을 거친 후 전자메일 데이타의 정규화 과정

단어 사전	299차원	(대출, 연체, 성인, 광고, 누드, ..., 카드)
전자메일 데이타	N차원	{ 한국, 정부, 최초, 카드, 대출, ..., 인터넷 }
전자메일 데이타 정규화	299차원	(1, 0, 0, 0, 0, ..., 1)

내용의 스팸메일에 대해서 스팸메일 공격자가 스팸메일에 임의의 코드를 삽입함으로써 같은 내용의 스팸메일을 서로 다른 전자메일로 인식하게 하여 분류율을 떨어뜨린다. 그리고 스팸메일 공격자가 스팸메일의 공격자 주소나 발신 서버의 주소를 계속 바꾸기 때문에 역시 분류율이 떨어지게 된다.

- **사회 연결망(Social Network) 분석을 통한 스팸메일 분류**

[5]에 의해 제안된 최근의 방법으로 사회 연결망 분석을 기초로 하고 있다. 이는 전자메일을 보낸 사람을 노드로, 전자메일을 보낸 관계를 링크로 표현하여 전자메일을 서로 자주 주고받는 사람들을 그래프에서 군집으로 표현하는 방법으로, 결국 그래프에서 군집은 서로를 아는 친구관계를 의미하게 된다. 따라서 그래프의 군집이 아닌 노드들은 스팸메일 공격자라고 판단하여 분류할 수 있다. 그러나 이 방법은 전자메일을 보낸 관계를 추적해야 하므로 한 전자메일 서버 안에서만 사용해야 하는 단점과 새로운 발송자 메일을 스팸메일로 분류하는 문제가 있을 수 있다.

2.2 페이지랭크(PageRank) 알고리즘

[6]에서는 모든 웹 문서를 보편적 중요도 순으로 순위를 매기기 위해 웹의 링크 구조를 이용하였다. 이 랭킹 알고리즘은 페이지랭크라고 불리며 현재 검색엔진 구글[1]에서 쓰이는 알고리즘이다.

대략적으로 현재 크롤링(crawling) 가능한 웹 그래프는 약 1억 5천만 개의 노드(웹 문서)와 17억 개의 간선(링크)이 있다고 알려져 있다. 각각의 웹 문서는 그 웹 문서로부터 밖으로 나가는 아웃링크(out-link 또는 forward link)와 그 웹 문서를 가리키는 인링크(in-link 또는 backward link)를 갖는다. 어떤 웹 문서의 모든 인링크를 다 찾아내는 것은 불가능하지만 웹 문서를 다 운로드하고 나면, 아웃링크가 무엇인지 알 수 있다. 일반적으로 링크가 많이 된 웹 문서일수록 그렇지 못한 웹 문서보다 더 중요하다.

2.3 내용기반 스팸메일 분류

내용기반 스팸메일 분류에는 나이브 베이지안, SVM 등이 주로 사용되나 최근에는 특히 SVM이 많이 사용되고 있는데, SVM의 장점으로는 명백한 이론적 근거에 기반 하므로 결과 해석이 용이하고 적은 학습 샘플만으로도 신속하고 높은 분류율을 나타낸다.

1) http://www.google.com

SVM을 사용한 내용기반 스팸메일 분류를 위해서는 전자메일의 제목과 본문에서 HTML 태그를 제외한 텍스트를 추출하여 형태소 단위 해석을 통하여 명사를 추출하며 형태소 단위 해석을 통한 색인과정은 그림 1과 같다.

그 뒤 표 1과 같이 전자메일로부터 추출 한 명사를 스팸메일에 사용된 단어의 빈도수를 기준으로 단어 사전을 생성하고, 단어 사전 벡터와 학습/테스트 데이타의 단어 집합이 정합되면 1로, 정합되지 않으면 0으로 표시하여 SVM에 사용하기 위해 정규화를 한다.

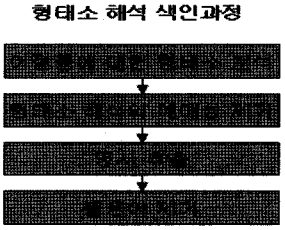


그림 1 SVM 학습의 전처리를 위한 형태소 해석 색인 과정

이렇게 스팸메일과 정상메일의 학습 샘플을 정규화한 후 SVM을 학습하여 모델을 생성한다.

스팸메일 분류는 그림 2와 같이 입력에 해당하는 전자메일을 형태소 해석 색인 과정을 거쳐 전처리하고 단어 사전을 통해 정규화한 후 학습한 모델에 적용하여 분류를 수행하게 된다.

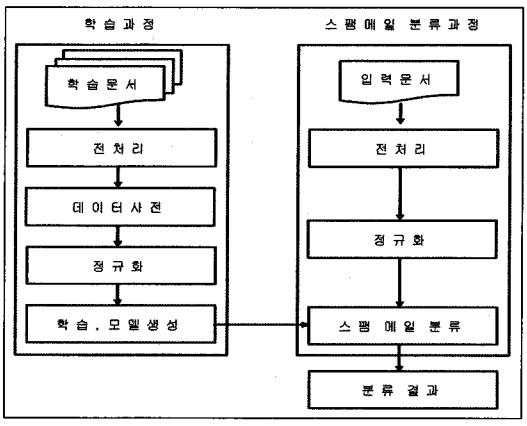


그림 2 SVM 스팸메일 학습, 분류 과정

3. 링크구조분석 스팸메일 분류 알고리즘

전자메일은 일종의 웹 문서로 볼 수 있으며 그 안에는 다른 웹 문서로의 하이퍼링크를 가지고 있다. 특히 스팸메일의 경우에는 대부분 광고하려는 사이트로의 링크를 가지고 있다. 왜냐하면 광고 목적의 전자메일 마케팅을 실현하려면 반드시 사용자를 광고하려는 사이트로 유도해야 하기 때문이다. 반면에 스팸메일이 아닌 정상메일의 경우에는 뉴스레터 등과 같이 다른 사이트로의 링크를 가지고 있는 경우도 있고, 개인 간의 주고받는 전자메일 같은 경우에는 링크가 없고 텍스트만 있는 경우로 나눌 수 있다.

3.1 웹 문서의 중요도

검색엔진 구글은 페이지랭크 알고리즘을 사용하여 모든 웹 문서의 랭킹을 계산하여 색인(indexing)한 결과를 가지고 있다. 여기서 웹 문서의 랭킹이란 어떤 웹 문서가 다른 웹 문서에 의해 얼마나 많은 링크로 참조되고 있는 정도를 수치화한 것이다. 다시 말하면 해당 웹 문서로 들어오는 인링크 수를 수치화한 것이다. 그러므로 웹 문서의 랭킹이 높다는 의미는 그 웹 문서는 다른 수많은 웹 문서에 의해 링크되고 있고 따라서 인기가 높고 더 유용하다고 말할 수 있다. 반면에 웹 문서의 랭킹이 낮으면 그 웹 문서는 다른 웹 문서에 의해 거의 링크되지 않으므로 인기가 없고 유용하지 않다는 의미이다.

한편 구글에서는 링크구조검색[2]이라는 기능을 제공하는데 구글 검색창의 옵션에 'link:'를 주게 되면 구글이 이미 색인한 결과에서 그 웹 문서를 링크하는 웹 문서의 수와 해당 URL을 알 수 있다. 예를 들면 구글의 검색창에 'link:http://www.egov.go.kr/'과 같이 대한민국 전자정부 홈페이지에 'link:'옵션을 주어 검색하면 대한민국 전자정부 홈페이지를 링크하는 2220개(2006년 1월 기준)의 웹 문서를 볼 수 있다. 대한민국 전자정부 홈페이지를 링크하는 웹 문서는 2220개나 되므로 대한민국 전자정부 홈페이지는 인기가 있고 유용하다고 말할 수 있다.

만약 한 전자메일 안에 링크가 여러 개가 있으면 중복되지 않는 각각의 링크에 해당하는 웹 문서를 링크하고 있는 웹 문서의 수를 합하여 해당 전자메일의 중요도로 사용한다. 여기서 한 전자메일 문서 D의 중요도 P는 식 (1)과 같다.

$$P=\sum_{l\in L} C_l \quad (1)$$

여기서 L은 문서 D안에 존재하는 링크의 집합이고 C_l은 링크 l이 가리키는 웹 문서를 링크하는(in-link) 웹 문서의 개수이다.

2) http://www.google.com/intl/en/help/features.html#link

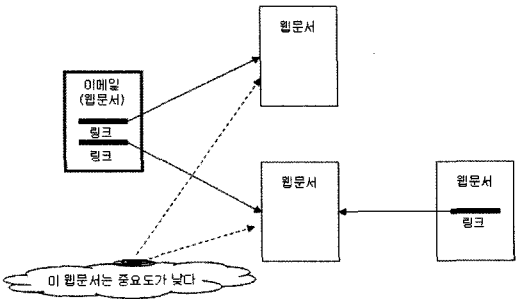


그림 3 스팸메일의 웹 링크 구조(중요도가 낮음)

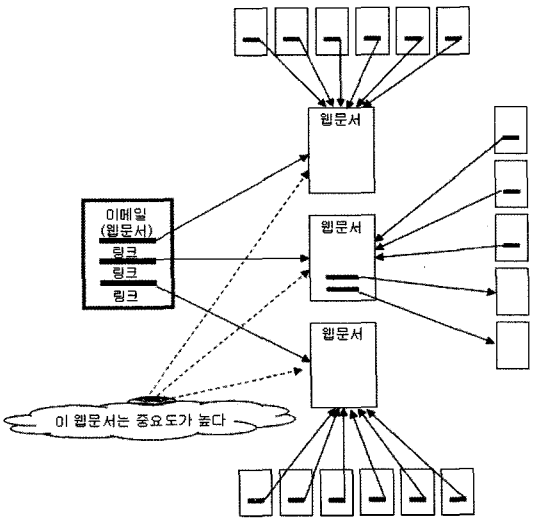


그림 4 정상메일의 웹 링크 구조(중요도가 높음)

스팸메일은 대부분 다른 웹 문서로의 링크를 가지고 있다. 그러나 그 웹 문서를 링크하는 웹 문서는 거의 없고 따라서 중요도가 낮다고 말할 수 있다(그림 3).

반면에 정상메일이 링크를 가지고 있을 때, 그 웹 문서들을 링크하는 웹 문서는 상대적으로 많고 중요도가 높다고 말할 수 있다(그림 4).

이 때 전자메일의 링크구조분석을 통해 다음과 같은 가설을 세울 수 있다.

<가설> 전자메일이 링크를 가지고 있을 때 그 링크의 웹 문서를 링크하는 웹 문서가 거의 없다면 스팸메일일 가능성이 높다. 반면에 전자메일이 링크를 가지고 있지 않거나(텍스트로만 이루어져 있거나), 링크를 가지고 있지만 그 웹 문서를 링크하는 웹 문서가 많다면 정상메일일 가능성이 높다.

3.2 링크구조분석 스팸메일 분류 알고리즘

그림 5에서는 링크구조분석을 통한 스팸메일 분류 과정을 나타내고 있다. 크게 학습 과정과 분류 과정으로 나눌 수 있는데, 학습 과정에서는 우선 학습 데이타로

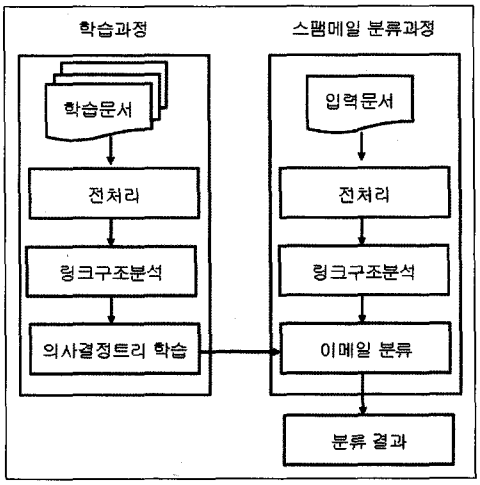


그림 5 링크구조분석 스팸메일 분류 과정

사용할 스팸메일과 정상메일에서 하이퍼링크를 추출하는 전처리 과정을 거친다. 전자메일에서 HTML 태그형태('<a href=URL>')인 링크뿐만 아니라, 텍스트 안에서 'http://URL' 형식의 링크도 추출한다. 추출한 링크를 구글 SOAP 검색 API[3]를 사용하여 링크에 해당하는 웹 문서를 링크하고 있는 웹 문서들의 수를 계산함으로써 그 웹 문서의 중요도를 구한다.

이 때 얻을 수 있는 하나의 전자메일 안의 링크의 개수와 전자메일의 중요도를 특징으로 하여 의사결정 트리를 학습시켜 모델을 생성한다.

학습하는 과정에서 의사결정 트리의 규칙 가지치기(pruning)의 신뢰도를 변화시킴에 따라 다양한 크기의 트리가 생성된다.

분류 과정은 스팸메일인지 아닌지 알 수 없는 새로운 입력에 해당하는 전자메일을 분류하는 과정이다. 학습 과정과 마찬가지로 우선 전자메일을 전처리하여 링크를 추출해 내고 중요도를 측정한다. 그 후 링크와 중요도를 학습 과정에서 생성한 의사결정 트리의 모델을 적용하여 스팸메일인지 아닌지를 분류하게 된다.

3.3 서버 주소 링크구조분석 방법

또한 전자메일에서 추출한 링크의 URL에서 서버 주소 부분을 취하여 서버 주소의 중요도를 측정할 수 있다. 예를 들면, 전자메일에서 'http://www.makdrim.com/md/shaver/shaver.php? pcode=inventad'와 같은 링크를 추출해 내어 이 링크에 해당하는 웹 문서가 다른 웹 문서에 의해 링크되는 수를 측정하는 것 외에 'http://www.makdrim.com/'와 같이 서버 주소를 취하여 이 서버 주소에 해당하는 웹 문서가 다른 웹 문서에 의해 링크되는 수를 측정할 수 있고 이것은 전자메일의 중요도를 측정하는 또 다른 특징이 될 수 있다.

3.4 링크구조분석과 내용기반 방법의 통합

전자메일의 링크와 링크의 서버 주소 각각의 중요도를 측정하고, 내용기반 방법을 서로 통합하면 스팸메일 분류율을 더욱 높일 수 있다.

그림 6은 3가지 방법을 통합한 시스템 구조도이다. 통합하기 위한 방법으로는 전문가들의 의견을 통합하기 위해 사용되는 대표적이고 간단한 방법인 다수결 원칙(majority voting)[7]을 사용하였다. 이 방법은 반수 이상의 전문가들이 같은 의견으로 동의하면 그 의견으로 집단의 전체 의사를 결정하는 방법이다.

따라서 전자메일의 링크 분석을 통한 분류 결과, 서버 주소 분석을 통한 분류 결과, SVM을 사용한 내용기반 방법의 분류 결과의 세 가지 분류 결과 중 두 가지 이상 같은 분류 결과를 최종 분류 결과로 채택한다.

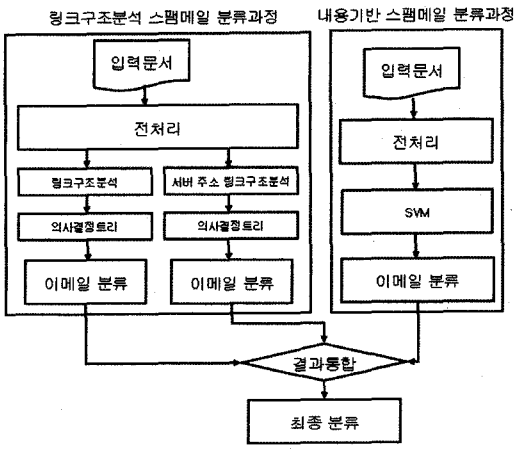


그림 6 링크구조분석과 내용기반 방법을 통합한 스팸메일 분류 시스템

4. 실 험

4.1 데이타 수집

어떤 전자메일이 스팸메일인지 정상메일인지 구분하는 기준은 개인마다 다르다. 한 광고 메일이 대부분의 사용자에게는 스팸메일일지라도 어떤 사용자에게는 필요한 광고 메일이 될 수 있기 때문이다. 따라서 스팸메일 분류에 있어서 궁극적으로는 개인화 기술을 적용하는 것이 바람직하나 이것은 본 논문의 연구범위를 벗어난다.

따라서 본 논문에서는 일반적인 스팸메일 분류 기준을 표 3과 같이 정의하였다.

3) Google SOAP Search API, http://code.google.com/apis/soapsearch/

표 3 일반적인 스팸메일 분류 기준

사용자 본인이 모르는 발신자로 부터 발신되었고 요청한 적이 없는 광고성 전자메일.

국외에서는 [8-10]과 같이 스팸메일을 지속적으로 수집하거나 실험용으로 스팸메일과 정상메일을 데이타베이스화한 스팸 아카이브(spam archive)를 제공한다. 그러나 본 논문에서는 국외의 스팸 아카이브를 사용할 수가 없었는데 국외의 스팸 아카이브는 전자메일의 텍스트만 추출하고 하이퍼링크 정보를 제거하였거나, 정상메일 제공자의 사생활 보호를 위하여 내용을 암호화하여 제공하였기 때문이다. 이것은 내용기반 스팸메일 분류 실험을 위해서는 문제가 없으나 전자메일 내의 하이퍼링크 정보를 이용해야만 하는 본 논문에서는 사용할 수 없었다. 또한 국내에는 스팸차단 연구를 위한 표준화된 스팸 아카이브는 전무하다.

따라서 실험에 사용하기 위해 전자메일 총 4000개를 별도로 수집하였는데 이는 스팸메일 2000개, 정상메일 2000개로 되어 있으며 25세부터 35세까지의 남성 8명, 여성 5명으로부터 2004년 9월부터 2005년 1월까지 수신된 국내 메일이다. 전자메일 제공자들은 제공한 전자메일을 스팸메일과 정상메일로 분류하는데 혼동을 겪곤 했는데, 가령 본인이 신청하여 받아 보고 있는 뉴스 레터가 너무 자주 오고 잘 읽지 않는다는 이유로 스팸메일로 분류하는 경우가 있었다. 그러나 이는 본인이 의도하여 신청했고 또 수신 거부 설정을 하면 되므로 스팸메일로 볼 수 없다. 따라서 전자메일 제공자들에게 표 3의 기준에 의해 스팸메일과 정상메일을 분류할 것을 요청하였다.

수집한 전자메일 중 스팸메일이 가지고 있는 평균 링크 수는 2.24개인 반면 정상메일의 평균 링크 수는 13.37개로 정상메일이 스팸메일에 비해 많은 링크 수를 가지고 있었다. 또 전자메일의 평균 중요도는 스팸메일이 5797.94, 정상메일이 22143.09로 3.1절에서 예측한대로 정상메일의 중요도가 스팸메일에 비해 매우 높았다.

4.2 성능 평가 방법

L을 정상메일, S를 스팸메일이라고 할 때 $n_{L\to L}$, $n_{S\to S}$는 각각 정상메일과 스팸메일을 올바르게 분류한 수이고, $n_{L\to S}$, $n_{S\to L}$는 각각 정상메일을 스팸메일로, 스팸메일을 정상메일로 잘 못 분류한 수를 나타낸다. 스팸메일 분류기의 precision, recall, accuracy, F-measure는 식 (2)와 같다.

$$\text{Precision}=\frac{n_{S\to S}}{n_{S\to S}+n_{L\to S}}$$

$$\text{Accuracy}=\frac{n_{S\to S}+n_{L\to L}}{n_{S\to S}+n_{L\to L}+n_{S\to L}+n_{L\to S}}$$

$$\text{Recall}=\frac{n_{S\to S}}{n_{S\to S}+n_{S\to L}}$$

$$\text{F-measure}=\frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \quad (2)$$

precision은 분류기에 의해 스팸메일이라고 분류된 전자메일 중 올바르게 분류된 전자메일의 비율이고, recall은 분류기에 의해 스팸메일이라고 분류된 전자메일이 전체 스팸메일 중에서 차지하는 비율을 나타낸 것이다. 또한 accuracy는 총 전자메일 중에서 분류기에 의해 스팸메일과 정상메일이 올바르게 분류된 비율이고, F-measure는 precision과 recall의 조화평균이다. 이 네 가지 척도는 정보 검색 분야에서 흔히 사용되며 스팸메일 분류기의 성능을 측정하는 기본적인 평가방법으로도 사용할 수 있다.

그러나 스팸메일을 분류할 때는 스팸메일을 정상메일로 잘못 분류 하는 false negative(S→L)보다 정상메일을 스팸메일로 잘못 분류하는 false positive(L→S)를 복구하는데 비용이 매우 많이 소요된다. 따라서 스팸메일 분류기는 false positive를 최소화 하도록 설계되어야 하며 false positive와 false negative에 가중치를 다르게 주어야 한다. 그런데 식 (2)의 성능 평가 척도는 이를 반영하지 못하므로 본 논문에서는 [2]에서 제안하였고 [11]과 [12]에서도 사용한 가중치를 준 accuracy (WAcc)와 에러(WErr=1-WAcc)를 함께 사용하였다.

WAcc와 WErr는 식 (3)과 같이 정의된다.

$$WAcc=\frac{\lambda\cdot n_{L\to L}+n_{S\to S}}{\lambda\cdot N_L+N_S},\ WErr=\frac{\lambda\cdot n_{L\to S}+n_{S\to L}}{\lambda\cdot N_L+N_S} \quad (3)$$

N_L은 총 정상메일 수이고, N_S는 총 스팸메일 수이다. λ는 가중치이고 1, 9, 999의 값을 가질 수 있다. λ=1일 때에는 스팸메일과 정상메일의 가중치가 같고, λ=9일 때에는 false positive가 false negative보다 9배 높은 패널티를 가지게 된다. λ=999일 때에는 false positive에 더욱 높은 패널티를 준 것으로 1개의 정상메일을 스팸메일로 잘못 분류한 것이 999개의 스팸메일이 차단되지 않고 스팸메일 분류기를 그냥 통과한 것과 같은 비용이 소요된다는 의미이다.

그런데 λ가 999의 높은 값을 갖게 되면 WAcc의 값은 매우 높아질 수 있고 따라서 잘못해석 될 수 있다 [2]. 이것을 피하기 위해 가중치가 적용된 WAcc와 WErr를 스팸메일 분류기가 사용되지 않는 기준선(baseline)과 비교해야 한다. [2]에서 제안하고 [12]에서 사용한 것과 같이 기준선은 스팸메일 분류기가 적용되지 않은 경우로, 정상메일은 분류기에 의해 전혀 차단되지

않고 스팸메일은 언제나 분류기를 통과하는 것을 의미한다. 이 기준선에서 가중치를 적용한 accuracy($WAcc^b$)와 에러($WErr^b$=1-$WAcc^b$)는 식 (4)와 같다.

$$WAcc^b = \frac{\lambda \cdot N_L}{\lambda \cdot N_L + N_S}, WErr^b = \frac{N_S}{\lambda \cdot N_L + N_S} \quad (4)$$

기준선과의 비교를 위해 [2]에서는 식 (5)와 같은 TCR(total cost ratio)을 제안하였다.

$$TCR = \frac{WErr^b}{WErr} = \frac{N_S}{\lambda \cdot n_{L \to S} + n_{S \to L}} \quad (5)$$

여기서 TCR이 1.0보다 크면 스팸메일 분류기를 사용하지 않았을 때보다 좋은 의미인데, 실제로 응용하기 위해 스팸메일 분류기는 TCR이 높도록 설계되어야 한다.

4.3 실험 방법

본 논문에서는 웹 문서의 중요도를 측정하기 위해 구글에서 제공하는 구글 SOAP 검색 API를 사용하였고 링크구조분석 알고리즘을 학습하는 데에는 의사결정 트리인 weka의 J48을 사용하였다. 내용기반 분류로 사용한 SVM은 mySVM의 java버전인 Stefan Rüping의 JMySVM을 사용하였다. J48과 JMySVM은 모두 독일 Dortmund대학 AI unit의 기계 학습 툴인 YALE[13] 안에 내장되어 있는 것을 사용하였다.

링크구조분석을 하기 위해 전자메일의 본문에서 링크에 대한 HTML 태그뿐만 아니라 텍스트에서 'http://' 형태의 텍스트도 링크로 간주하여 추출하였다. 추출한 링크는 구글 SOAP 검색 API를 사용하여 링크된 횟수를 측정하여 전자메일의 중요도를 계산하고 전자메일의 링크의 개수와 전자메일의 중요도를 특징으로 하여 의사결정 트리를 학습하고 분류를 수행하였다.

SVM의 커널 함수로는 다항식(polynomial)을 사용하였고 차수(degree)는 3으로 고정하여 실험하였다. 실험은 10-fold cross validation을 수행하였다.

4.4 실험 결과

그림 7은 링크구조분석을 통한 분류(LSA), 3.3절에서 설명한 링크의 서버 주소 부분만을 이용한 분류(LSA(HOST)), SVM을 사용한 내용기반 분류(CBC), 그리고 세 가지 방법을 통합한 시스템의 결과(통합)를 보여주고 있다.

그림 8은 학습 샘플의 크기를 400개씩 증가하여 4000개 까지 증가하는 동안의 recall, precision, accuracy, F-measure이다. 분류율은 학습샘플의 크기의 변화에 따라 크게 변화가 없었으며 통합, LSA, CBC의 순으로 높은 분류율을 보였다.

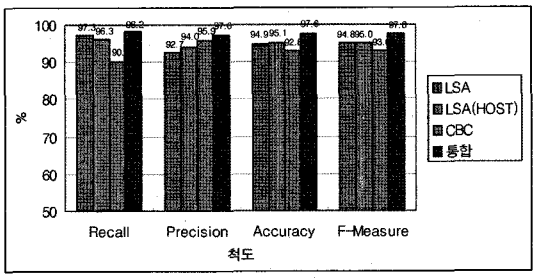


그림 7 링크구조분석을 통한 분류(LSA)와 링크의 서버 주소 부분만을 이용한 분류(LSA(HOST))와 SVM을 사용한 내용기반 분류(CBC), 그리고 세 가지 방법을 통합(통합)한 스팸메일 분류 결과

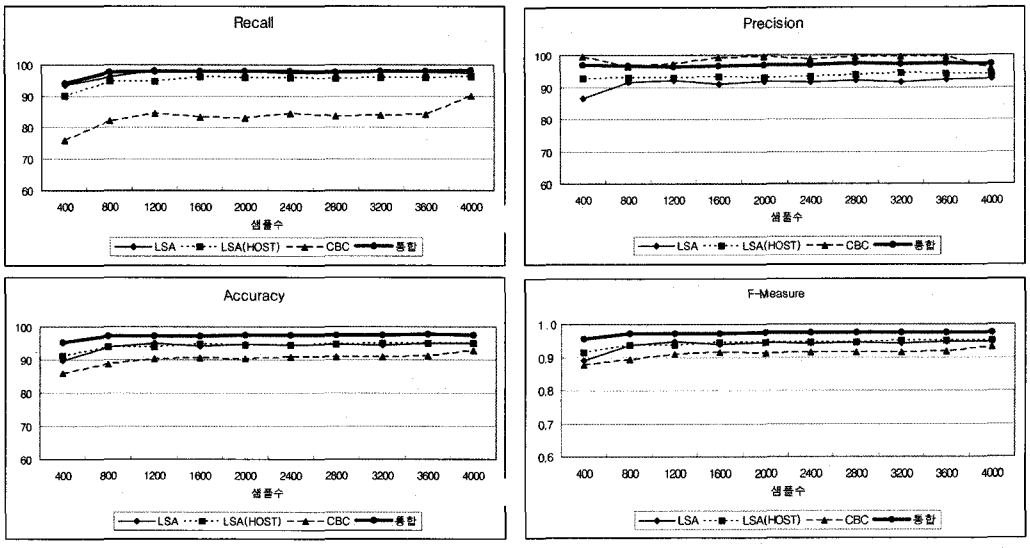


그림 8 샘플을 400씩 4,000까지 증가시켰을 때의 스팸메일 분류율의 변화

recall, accuracy, F-measure에서 전반적으로 LSA와 LSA(HOST)가 CBC보다 더 좋은 결과를 나타냈고 통합은 가장 좋은 결과를 나타냈다. 이는 제안한 링크구조분석 분류가 SVM을 이용한 기존의 내용기반 분류보다 효율적임을 뜻한다.

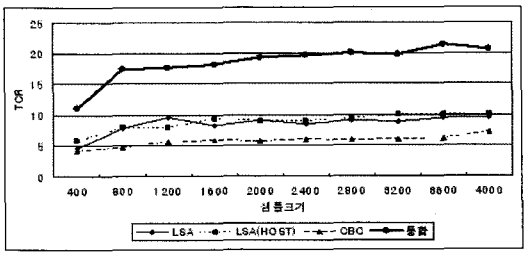
표 4에서는 λ의 값을 1, 9, 999로 변화를 주어 가중치를 둔 accuracy(WAcc)와 기준선에서의 accuracy($WAcc^b$), 그리고 TCR의 값을 나타내고 있다. recall과 precision은 패널티를 적용할 수 없는 척도이기 때문에 λ가 1일 경우만 사용하였다.

표 4를 보면 WAcc의 척도에서 λ가 1, 9, 999로 증가함에 따라 CBC는 점점 좋은 결과를 보여주지만, LSA와 통합은 조금씩 분류율이 떨어지는 것을 볼 수 있다. 이것은 내용기반 방법의 결과가 링크구조분석 알고리즘의 분류 결과나 통합한 시스템의 결과보다, false positive로 오분류한 결과가 false negative로 오분류한 결과보다 상대적으로 적은 비율임을 뜻한다.

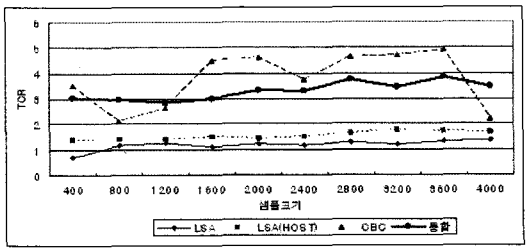
표 4의 TCR을 보면 λ가 1, 9일 때에는 TCR>1 이므로 스팸메일 분류기를 사용하지 않은 기준선보다 스팸메일 분류가 효율적이라는 결과를 보여주고 있다. 그러나 λ를 999로 하여 정상메일에 패널티를 많이 준 경우에는 TCR<1이고 스팸메일 분류기가 효율이 없는 것을 알 수 있다. 그림 9는 λ가 각각 1, 9, 999이고 학습 샘플의 크기를 400개씩 증가했을 때의 TCR을 나타낸다. λ가 1일 때에는 통합 시스템이 LSA보다 분류율이 높았고, LSA는 CBC보다 분류율이 높았으며 샘플 크기 간의 큰 변화는 없었다.

반면 λ가 9일 때에는 통합 시스템은 여전히 좋은 분류율을 유지하였으나 CBC가 샘플 크기에 따라 민감하게 변하며 통합 시스템과 유사한 분류율을 보였다. 그러나 λ가 999일 때에는 모든 방법의 TCR이 1보다 매우 낮아서 스팸메일 분류기로서의 효용성이 거의 없다.

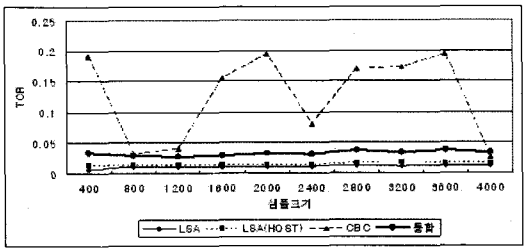
본 실험에 사용한 스팸메일 데이타는 본문의 내용에



λ=1



λ=9



λ=999

그림 9 LSA, LSA(HOST), CBC, 통합의 샘플 크기에 따른 *TCR*의 변화

텍스트가 거의 없고 이미지로만 이루어졌기 때문에 링크구조분석 알고리즘이 내용기반 방법의 분류율보다 높았다. 이는 주로 이미지로 이루어지고 텍스트가 거의 없는 스팸메일의 분류에 내용기반 분류보다 링크구조분석 분류가 효과적임을 의미한다.

표 4 λ가 1, 9, 999일 때의 recall, precision, *WAcc*, $WAcc^b$, *TCR*

	λ	recall	precision	*WAcc*	$WAcc^b$	*TCR*
LSA	1	97.25%	92.67%	94.75%	50.0%	9.524
LSA(HOST)	1	96.25%	93.96%	95.03%	50.0%	10.060
CBC	1	89.95%	95.88%	93.03%	50.0%	7.174
통합	1	98.15%	97.04%	97.58%	50.0%	20.661
LSA	9			92.75%	90.0%	1.379
LSA(HOST)	9			94.05%	90.0%	1.679
CBC	9			95.49%	90.0%	2.215
통합	9			97.12%	90.0%	3.466
LSA	999			92.26%	99.9%	0.013
LSA(HOST)	999			93.80%	99.9%	0.016
CBC	999			96.09%	99.9%	0.026
통합	999			97.00%	99.9%	0.033

표 5에서는 기존의 스팸메일 분류 방법과 본 논문에서 제안한 링크구조분석 스팸메일 분류의 장단점을 비교하였다. 표 5에서와 같이 제안한 링크구조분석 스팸메일 분류는 현재까지 알려진 스팸메일 분류 방법의 취약점을 극복하고 있다.

5. 결론 및 향후 연구

기존의 스팸메일 분류 방법이 전자메일의 내용을 분석하여 분류한 반면에 본 논문에서는 전자메일의 웹의 링크 구조 분석을 통하여 스팸메일을 분류하는 방법을 제안하였다. 이것은 전자메일의 링크와 중요도라는 새로운 특징을 사용한 스팸메일 분류 알고리즘을 제안한 것이고 SVM을 사용한 내용기반 분류 방법을 통합한 스팸메일 분류 시스템 또한 제안하였다. 최근의 스팸메일은 텍스트를 이미지로 표현한 양상을 많이 보이고 있는데, 이런 경우 내용기반 분류에 한계가 있다. 본 논문에서는 스팸메일 데이타를 텍스트가 적고 이미지가 대부분인 데이타를 수집하여 제안한 방법으로 실험하였고 이렇게 이미지가 대부분인 스팸메일에 대한 분류 시도는 본 논문에서 제안한 독창적 아이디어이다. 이렇게 전자메일의 내용분석이 아닌 웹의 구조를 분석하는 방법은 다양하고 빠르게 공격하는 스팸메일을 효과적으로 방어할 수 있다. 실험 결과, 본 논문에서 제안한 링크구조분석 알고리즘이 효율적임을 보였고 기존의 내용기반 분류와 통합한 결과는 더욱 효율적임을 보였다.

향후 연구로는 false positive를 더욱 최소화 하는 방법론의 연구와 링크구조분석 분류 결과를 데이타베이스에 저장하여 기존에 분석한 링크에 대해서는 데이타베이스의 결과를 참조하여 분류 시간을 줄이는 것을 기대할 수 있다. 또한 스팸메일뿐만 아니라 웹 게시판이나 블로그의 포스트에 광고 게시물을 올리는 스팸 게시물에 대한 피해도 급증하고 있는데, 스팸 게시물 또한 스팸메일과 같이 웹 로봇에 의해 무작위적이고 대량으로 작성되는 유사한 특징을 가진다. 또한 스팸 게시물에도 광고를 하려는 사이트로의 링크를 가지고 있는 경우가 대부분이고 본 논문에서 제안한 링크구조분석을 사용하면 스팸 게시물에 대한 분류도 가능할 것으로 기대된다.

참 고 문 헌

[1] 민도식, 송무희, 손기준, 이상조, "SVM 분류 알고리즘을 이용한 스팸메일 필터링", 한국정보과학회 2003년 춘계학술대회, 2003.

[2] Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., Spyropoulos, C., "An evaluation of naive bayesian anti-spam filtering," In Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning(ECML 2000), 2000.

[3] 서정우, 손태식, 서정택, 문종섭, "Support Vector Machine을 사용한 스팸메일 탐지 방안", 한국정보과학회 2003 추계학술대회, 2003.

[4] Drucker, H., Wu, D., "Support vector machines for spam categorization," IEEE Transactions on Neural Networks, VOL. 10, NO. 5, 1999.

[5] Boykin, O., Roychowdhury, V., "Personal email network: an effective anti-spam tool," Arxiv preprint cond-mat/0402143, 2004 - arxiv.org, 2004.

[6] Page, L., Brin, S., Motwani, R., Winograd, T., "The pagerank citation ranking: bringing order to the web," Technical Report, Stanford University, Stanford, CA, 1998.

[7] Vieira, C., Mather, P., "A comparative study of multiple classifier combination methods in remote sensing," In Proceedings of the IC-AI'2000, Vol. 1, pp.39-46, 2000.

[8] i-config: Internet Content Filtering Group, http://www.iit.demokritos.gr/skel/i-config/.

[9] SpamArchive.org, http://spamarchive.org/.

[10] The Apache SpamAssassin Project, http://spamassassin.apache.org/.

[11] Carreras, X., Marquez, L., "Boosting trees for anti-spam email filtering," In Proceedings of RANLP-2001, 4th International Conference on Recent Advances in Natural Language Processing, 2001.

[12] ZHANG, L., ZHU, J., YAO, T., "An evaluation of statistical spam filtering techniques," ACM Transactions on Asian Language Information Processing, Vol.3, No. 4, pp.243-269, 2004.

[13] YALE(Yet Another Learning Environment), http://rapid-i.com/.

표 5 기존 스팸메일 분류 방법과 링크구조분석 방법의 장단점 비교

전자메일 특성 / 분류방법	단어의 변조	스팸메일주소나 서버의 잦은 변경을 통한 공격	최신의 스팸메일	본문에 텍스트가 거의 없고 이미지로 이루어진 경우	단일의 스팸 차단 서버에서만 사용가능 하도록 제약
목록기반 분류	강건	취약	취약	강건	제약
규칙기반 분류	취약	강건	취약	취약	제약없음
내용기반 분류	보통	강건	취약	취약	제약없음
협업적 분류	강건	취약	보통	강건	제약
사회 연결망 분석	강건	보통	취약	강건	제약
링크구조분석	강건	강건	강건	강건	제약없음

이 신 영

1999년 숭실대학교 컴퓨터학부(학사). 2006년 숭실대학교 대학원 컴퓨터학과(석사) 2006년~현재 (주)시뮬레이션연구소 연구원. 관심분야는 인공신경망, 기계학습, 로보틱스

길 아 라

1987년 이화여자대학교 전산학과(이학사). 1990년 한국과학기술원(KAIST)(공학석사). 1997년 한국과학기술원(KAIST)(공학박사). 1993년~1996년 (주)한국 IBM 초청강사. 1995년~1997년 (주)새롬기술 선임연구원. 1997년~2002년 (주)새롬기술 기술자문. 2003년~2005년 Dialpad Comm. Inc.(미국) 사외이사. 1997년~현재 숭실대학교 컴퓨터학부 교수. 2006년~현재 (주)상권홀딩스 이사. 관심분야는 실시간 운영체제, 실시간 통신시스템, 멀티미디어 네트워크 시스템, 임베디드 운영체제, 센서네트워크 응용시스템

김 명 원

1972년 서울대학교 응용수학과(학사). 1981년 University of Massachusetts(Amherst) Computer Science(석사). 1986년 University of Texas(Austin) Computer Science(박사). 1975년~1978년 한국과학기술 연구소 연구원. 1982년~1985년 Institute for Computing Science & Computer Application(Univ. of Texas). 1975년~1987년 AT&T Bell Labs. Member of Technical Staff. 1987년~1994년 한국전자통신연구소 책임 연구원. 1991년~1993년 충남대학교 전자계산학과 겸임부교수. 2000년~2001년 미국 IBM T.J WATSON 연구소 방문 과학자. 1994년~현재 숭실대학교 컴퓨터 학부 교수. 2002년~2003년 숭실대학교 정보지원처장. 2004년~2006년 숭실대학교 정보과학대학원장. 1992년~1993년 한국신경회로망 연구회 회장. 1992년~1993년 한국정보과학회 뉴로컴퓨팅 연구회 회장. 1993년~1995년 한국정보과학회 뉴로컴퓨팅 연구회 위원장. 1993년~1995년 IEEE Neural Network Council 한국지부장. 1998년~2000년 한국인지학회 부회장. 1997년~2000년 한국뇌학회 부회장. 2001년~2002년 한국뇌학회 회장. 관심분야는 유연주론, 신경회로망, 퍼지시스템, 진화알고리즘, 패턴인식, 자동추론, 기계학습, 데이타마이닝, creativity engineering 등