Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 4, pp. 735~744

The Exploratory Analysis for Spam Mail Data Using Correspondence Analysis

Yang Kyu Shin[1)]

Abstract

The number of electronic mail(E-mail) has been increased dramatically as a result of expanding internet and information technology. Although there are many conveniences of E-mail in the bright side, some serious problems occur because of E-mail in its dark side. One of the problems is spam-mail which is unsolicited mail and also called bulk mail. This paper presents a set of patterns of spam-mail occurrences within a week using the correspondence analysis. The correspondence analysis is an exploratory multivariate technique that converts data into a particular type of graphical display in which the rows and columns are depicted as points. One of the meaningful patterns is a great increment of adult and phishing related spam-mails at weekends so any spam-mail filters should be designed to cope with this pattern.

Keywords : Correspondence analysis, E-mail, Exploratory data analysis, Spam-mail

1. 서론

컴퓨터를 기반으로 하는 통신 수단 중 가장 보편적인 방법인 전자우편은 최근 정보통신 기술의 급속한 발전에 힘입어 개인도 쉽게 메일 서버를 운영할 수 있게 되었고, 그에 따라 전자우편의 수도 기하급수적으로 증가하고 있다. 전자우편은 개인이나 조직이 필요로 하는 통신 수단으로서의 다양한 장점이 있지만, 스팸메일로 인한 물질적, 정신적 피해 또한 심각한 문제가 되고 있다.

스팸메일 차단을 위한 가장 기본적인 방법은 메일 수신자가 스팸메일에 대해 일일이 수신거부를 하고 스패머의 메일주소를 스팸메일 발신자 목록에 등록하여 다음부터는 등록된 스패머가 발신하는 메일을 자동으로 스팸처리 하도록 하는 것인데, 이는 수신자들이 수동으로 작업해야 하므로 번거롭고 또한 요즘과 같이 자동으로 스팸메일

1) 경상북도 경산시 유곡동 290번지 대구한의대학교 자산운용과학과 교수
E-mail: yks@dhu.ac.kr

을 발송하는 경우에는 거의 효과가 없으므로 적극적인 대응이 되지 못하고 있다. 좀 더 효과적인 방법은 메일 수신자가 스팸 차단용 도구를 사용하여 자신에게 전달된 메일을 검사하여 스팸메일을 자동으로 확인하고 삭제하도록 하는 것이다. 기업이나 공공 기관의 경우에는 이를 위하여 스팸메일 관리시스템 즉 스팸메일 필터링시스템을 도입하고 있다. 스팸메일을 차단하기 위해서는 먼저 메일서버로 수신되는 메일을 스팸메일과 정상메일로 분류하여야 하는데, 스패머들이 스팸을 어떻게 만들어 내는지를 예측하기는 쉬운 일이 아니지만 기존의 스팸메일의 패턴을 분석하여 어느 정도 유추할 수 있다.

스팸메일 탐지를 위한 메일의 분류 및 패턴 분석에 대한 연구는 통계학적 방법을 포함하여 다양한 방법으로 이루어지고 있다. Ruvini & Gabriel(2002), Mock(2001), Giorgetti & Sebastiani(2003) 그리고 Manco, Macciari, Ruffolo & Tagarelli(2003)은 일반적인 문서분류 기법을 이용한 스팸메일을 탐지하는 방법을 연구하였으며, Graham(2002)은 베이지안 분류 기법을 확장하여 일반 메일과 스팸메일을 분리하는 방법을 제시하였고, Robinson(2002, 2003)은 Graham 이 제안한 방법에서 문제점으로 남아있던 희소단어처리와 단어의 발생빈도 반영 등에 대한 문제점을 일부 해결한 방법을 제시하였다.

하지만 이와 같은 방법들 역시 스팸메일을 완전하게 탐지하여 차단할 수는 없다. 근본적 이유는 스팸메일에 대한 판단이 궁극적으로는 각 사람마다 달라지기 때문이다. 예를 들어, 세일 정보가 필요한 수신자에게는 특정 백화점의 세일 소식을 알리는 메일이 스팸이 아니기 때문이다. 따라서 스팸을 탐지하고 차단하는 어떤 시스템도 정상적은 메일을 스팸메일로 잘못 판정하는 false positive 오판율을 낮추어야 한다. 특히 다수의 사용자를 가진 기관이나 조직의 메일서버에 설치된 스팸메일 차단기의 false positive 오판율은 0%가 되어야 한다.

본 연구에서는 탐색적 자료 분석 기법인 대응분석을 이용하여 스팸메일의 요일별 패턴분석을 하고자 한다. 이는 앞에서 논의한 대로 다수의 사용자에게 대량의 메일이 들어오는 기관의 경우 스팸메일 차단기의 false positive 오판율을 0%로 낮추기 위해서는 요일별 패턴분석에 따른 스팸메일의 탐지가 결정적인 역할을 할 수 있기 때문이다. 따라서 본 연구에서는 대응분석을 이용하여 요일별 스팸메일 패턴에 대해 조사하였다. 대응분석에서 행 점수와 열 점수의 계산에 사용할 정규화 방법 중에서 두 변수간의 차이나 유사성을 확인하려는데 목적이 있으면 대칭적 방법을, 두 변수의 범주간의 차이나 유사성을 확인하려면 주성분방법을 선택하는 것이 적합한데, 여기서는 주성분방법을 이용하였다.

본 논문의 구성은 다음과 같다. 2절에서는 탐색적 자료분석에 대하여 간략히 살펴보고 3절에서는 대응분석에 대하여 기술하였다. 4절에서는 스팸메일 자료에 대하여 대응분석을 하여 요일별 스팸메일의 패턴 및 요일과 스팸메일 유형간의 관계에 대하여 분석하고, 그 결과를 이용하여 탐색적 방법을 통한 스팸메일의 특성 연구에 관하여 논하였다. 본 연구에서는 스팸메일 종류들 간의 패턴이 기관별로 거의 유사하다는 스팸메일 차단시스템 개발관련전문가들의 경험적 판단에 근거하여 대구한의대학교 웹메일시스템에서 6개월 동안 수신한 4,556,389건의 스팸메일 자료를 대상으로 분석을 하였다.

2. 탐색적 자료분석

n건의 스팸메일을 수신자에게 전달된 q 개의 요일별로 정리하고 이 결과를 다시 p개의 스팸메일 종류별로 분류하면 행의 수가 p이고 열의 수가 q인 이원 분할표

$$F = f_{ij},\ i = 1, \cdots,\ p : j = 1, \cdots,\ q \tag{2.1}$$

로 표현된다. 이때 f_{ij}는 스팸메일 종류 i 와 요일 j가 관측된 빈도로 $n = \sum_{j=1}^{q} \sum_{i=1}^{p} f_{ij}$이다.

이원 분할표에 대해 흔히 적용되는 통계적 방법은 카이제곱검정으로 행과 열의 독립성을 검정하거나 열 범주의 상대적 빈도에 대한 행 표본들의 균일성을 검정 할 수 있다. 식 (2.1)의 이원 분할표에 대한 카이제곱검정통계량은

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{(f_{ij} - f_{i+}f_{+j}/\ n)^2}{f_{i+}f_{+j}/\ n} \tag{2.2}$$

로 정의된다. 이때 $f_{i+} = \sum_{j=1}^{q} f_{ij},\ f_{+j} = \sum_{i=1}^{p} f_{ij}$ 이다.

그러나 카이제곱검정은 행과 열의 범주들 간의 결합양상을 보여주지는 못한다. 대응분석은 명목형인 두 변수의 범주들을 저차원 공간상의 점들로 동시에 나타내어 그들의 관계 및 결합양상을 파악하고자 하는데 목적이 있는 탐색적 자료 분석기법이다 (Greenacre and Hastie(1987)). 대응분석은 사회학, 생태학, 심리학, 기상학, 교육학, 지리학 그리고 경제학 등 여러 분야에 걸쳐 널리 응용되고 있다. 특히 최근 들어 범주형 자료 분석에 이용하여 응용성 및 실용성을 높이고 있다(Greenacre(1984)와 최용석(1993)).

3. 대응일치분석

행과 열의 결합 여부에 대한 가설검정을 하기위해서는 카이제곱분석을 행과 열의 결합 양상을 조사하기 위해서는 대응분석을 한다. 본 절에서는 대응분석에 대하여 살펴보기로 한다.(참고: 허명회(1999) 3장, 허명회, 양경숙(2002) 10장)

이원분할표를 나타내는 식 (2.1)에서 행 i 프로파일을 다음과 같이 나타내자.

$$a_i = \frac{(f_{i1}, \cdots, f_{iq})^t}{f_{i+}},\ i = 1, \cdots,\ p$$

$\sum_{j=1}^{q} a_{ij} = 1$이므로 p개의 행 프로파일들 $a_1, \cdots, a_p$는 $q-1$차원 S^q에 속하게 된다. 이때

$r_i = f_{j+}/n$, $i = 1, \cdots, p$로 정의하면 p개 행 프로파일들의 중심은

$$c = (f_{+1}. \cdots, f_{+q})^t / n = \sum_{i=1}^{p} r_i a_{ij} \quad , \quad I_q^t c = 1$$

로 표현된다. 그러므로 식 (2.2)의 카이제곱검정통계량은

$$\begin{aligned} \chi^2 &= n \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{f_{i+}}{n} \frac{(f_{ij}/f_{i+} - f_{+j}/n)^2}{f_{+j}/n} \\ &= n \sum_{i=1}^{p} r_i (a_i - c)^t Dc^{-1} (a_i - c) \end{aligned} \tag{3.1}$$

이다. 이때 $Dc = diag(c)$이다.

한편 $a_i - c$가 행 i프로파일과 행 중심과의 편차이므로 행 편차벡터를 다음과 같이 정의 할 수 있다.

$$B = \begin{bmatrix} b_1^t \\ \vdots \\ b_i^t \\ \vdots \\ b_p^t \end{bmatrix} : BI_q = O_p, \; r^t B = O_q^t$$

식 (3.1)을 이용하여 $q-1$차원 행 공간 L에서 임의의 두 벡터 b_i와 b_j사이의 제곱거리를 다음과 같이 정의한다.

$$d^2(b_i, b_j) = (b_i - b_j)^t Dc^{-1} (b_i - b_j)$$

$q-1$차원 행 공간 L에서 단위벡터 v_1에 행 편차벡터 $b_1, \cdots, b_p$들을 사영하기로 하자.

$b_i = \overrightarrow{OB}_i$에 대해 차원축소로 인한 손실을 최소로 한 최적화된 식은 다음과 같다.

$$\max \; v_1^t Dc^{-1} B^t DrBDc^{-1} v_1 ; \; v_1^t Dc^{-1} v_1 = 1, \; I_q^t v_1 = 0 \tag{3.2}$$

이때 $B = Dr^{-1}(F/n - rc^t)$이다.

식 (3.2)에 라그랑지 상수 λ_1을 이용한 편미분을 적용하여 $\widetilde{v}_1$를 $Dc^{-1/2}v_1$라 하면 고유체계

$$Dc^{-1/2}B^tDrBDc^{-1/2}\widetilde{v}_1 = \lambda_1\widetilde{v}_1 \tag{3.3}$$

가 유도된다. 식 (3.3)에 스펙트럼 분해를 하면 v_1에 내린 b_i의 정사영 좌표는 다음과 같다.

$$b_i^tD_c^{-1}v_1 (= b_i^tD_c^{-1/2}\widetilde{v_1})$$

다음으로 v_2를 v_1과 직교하는 행 공간에서의 단위벡터라 하고 v_2에 행 편차벡터 $b_1, \cdots, b_p$를 사영하기로 하자.

선형부공간 $L(v_1, v_2)$에 내린 $b_i = \overrightarrow{OB_i}$에 대해 차원 축소로 인한 손실을 최소로 한 최적화된 식은 다음과 같다.

$$\max \; v_2^tDc^{-1}B^tDrBDc^{-1}v_2;\; v_1^tDc^{-1}v_2 = 0,\; I_q^t v_2 = 0 \tag{3.4}$$

위의 식 (3.4)에 1축 차원 축소와 유사하게 라그랑지 승수 λ_2와 편미분을 적용하면 1축과 동일한 고유체계가 유도되고 스펙트럼분해를 통하여 v_2에 내린 b_i들의 정사영좌표, $b_i^tD_c^{-1}v_2$를 구한다. 같은 방법에 의하여 s축 차원 축소가 가능하다. 스팸메일 차고유체계 식 (3.2)와 (3.4)에 비정칙분해(Singular Value Decomposition)를 적용하여도 정사영좌표를 구할 수 있다. (Shin(1987))

열 범주는 행 공간에서 j번째 요소만 1이고 나머지 요소는 모두 0인

$$e_j = (0, \cdots, 1, \cdots, 0)^t,\; j = 1, \cdots, q \tag{3.5}$$

를 가정하여 식 (3.5)를 v_k에 사영시키면 그 좌표는 $e_j^tDc^{-1}v_k$ 이다.
그러므로 k축에서 행벡터와 열벡터는 각각 다음 좌표 값으로 표현된다.

$$b_1^tDc^{-1}v_k, \cdots, b_i^tDc^{-1}v_k, \cdots, b_p^tDc^{-1}v_k$$
$$e_1^tDc^{-1}v_k, \cdots, e_j^tDc^{-1}v_k, \cdots, e_q^tDc^{-1}v_k$$

위의 행 표본과 열 범주 표시는 0이 아닌 고유값 λ_k와 짝이 되는 고유벡터 v_k마다 가능하지만 효과적인 차원축소에는 처음 s개의 축만 활용한다. 차원수 s는 고유값의

감소패턴을 고려하여 결정하거나 근사도를 일정 수준(예를 들면 80%) 보다 크도록 정한다. 일반적으로는 평면공간에서의 분석이 효과적이므로 2차원공간이 주로 이용된다.

s차원 행렬도의 근사도는 총 카이제곱 중에서 어느 정도 설명되었는가로 정의되는데 총 카이제곱이 고유값의 총합과 비례하므로 $s(\leq \min(p-1, q-1) \equiv r_0)$개축에 의한 근사도는 $\sum_{j=1}^{s} \lambda_j / \sum_{j=1}^{r_0} \lambda_j$로 정의된다.

4. 스팸메일 자료 분석 결과

수신되는 스팸메일이 요일별로 상이한 범주에 속하는 특이한 양상을 보인다면 스팸메일 차단기 역시 해당 범주에 대한 확률을 조절하여 false positive 오판율을 낮춤과 동시에 스팸메일의 차단율을 높일 수 있다. 예를 들어, 토요일에 피싱(Phishing) 공격 스팸메일이 많이 들어온다면 스팸메일 차단기는 그와 관련된 단어들의 확률을 토요일에만 높여서 문제를 해결할 수 있다. 따라서 요일별로 수신되는 스팸메일의 종류가 다를 수 있다는 가정 하에서 대응분석을 이용하여 스팸메일 자료를 분석하기로 한다. 분석에는 SPSS/Data Reduction/Correspondence Analysis가 사용되었다.

4.1 분석 자료

분석 자료는 대구한의대학교 웹메일시스템에 2005년 1월 1일부터 6월 30일까지 수신된 4,556,389건의 스팸메일로 스팸메일 종류에 대한 설명은 (표4.1)과 같다.

(표 4.1) 스팸메일 종류 설명

스팸메일종류	설 명	수신건수
Format	헤더에 비정상적인 포맷이 들어있는 스팸메일	833,607
Robot	필터링 시스템에 탑재된 로봇이 차단한 스팸메일	533,840
Financial	금융, 카드, 대출관련 스팸메일	1,216,028
Product	제품홍보, 불법상품, 쇼핑몰, 관광, 레저관련 스팸메일	1,012,984
Spammer	스패머, 전문발송꾼, 메일수집관련 스팸메일	194,806
Education	교육, 자격증관련 스팸메일	165,585
Subject	제목으로 차단한 스팸메일	146,563
Phishing	피싱공격 스팸메일	134,247
Sender	스패머 주소로 차단(주로 head format)되는 스팸메일	59,538
Adult	성인, 포르노, 몰카, 누드관련 스팸메일	59,491
Etc.	사기, 도박, 게임, 웹 사이트 광고관련 스팸메일	199700

다음 (표 4.2)는 스팸메일 종류를 요일별로 분류한 자료이다.

(표 4.2) 스팸메일 자료

	일	월	화	수	목	금	토
Format	77,461	114,099	121,588	123,404	115,854	162,049	119,152
Robot	52,995	82,150	82,666	82,314	84,126	76,850	72,739
Financial	137,361	183,348	181,993	182,105	185,754	184,111	161,356
Product	71,778	147,708	167,566	171,847	163,318	176,559	114,208
Spammer	22,180	33,339	29,707	27,414	37,267	25,099	19,800
Education	18,723	28,215	23,328	23,933	22,965	27,209	21,212
Subject	15,618	27,083	24,089	20,554	20,298	18,425	20,496
Phishing	20,479	16,437	18,979	16,890	15,946	20,211	25,305
Sender	7,791	9,353	11,119	10,994	9,428	6,708	4,145
Adult	11,573	7,851	6,094	6,830	5,275	9,561	12,307
Etc.	23,755	30,240	33,608	27,822	33,439	26,176	24,660

(표 4.2)의 자료에 대하여 카이제곱검정통계량 값은 66,566.863으로 자유도 60의 카이제곱분포에서 p값은 0%에 근사한다. 그러므로 스팸메일 종류와 요일 간에 어떤 연관이 있음을 알 수 있다. 그러나 카이제곱검정은 행과 열의 결합양상을 보여주지는 못하므로 대응분석을 통하여 스팸메일 종류와 요일간의 결합양상 즉, 패턴을 분석해 보도록 하자. 요일이 다항표본, 스팸메일 종류가 범주가 되므로 3절에 의하면 (표4.2)에서 요일을 주좌표로 스팸메일 종류를 표준좌표로 계산하여야 차원 축소로 인한 손실을 최소화한 좌표 설정이 된다. SPSS를 이용하여 자료 분석을 하는 경우에는 모형설정의 정규화방법 단계에서 열주성분(column principal)을 선택하면 된다. (표 4.3)은 SPSS 대응분석 결과에 의한 비정칙값, 고유값, 점유비율 그리고 누적비율이다.

(표 4.3) 요약표

차원	비정칙값	고유값	점유비율	누적비율
1	0.091	0.008	0.562	0.562
2	0.069	0.005	0.330	0.892
3	0.024	0.001	0.039	0.931
4	0.022	0.001	0.034	0.965
5	0.020	0.000	0.026	0.991
6	0.011	0.000	0.009	1.000
전체		0.015	1.000	1.000

1축 기여율이 총 카이제곱 변동의 56%를, 2축이 33%를 차지함으로써 첫 두 축이 전체의 89%를 설명한다. 그러므로 축소차원수를 2로 하면 될 것으로 판단된다. (표 4.4)와 (표4,5)는 계산된 행과 열의 좌표 값이다.

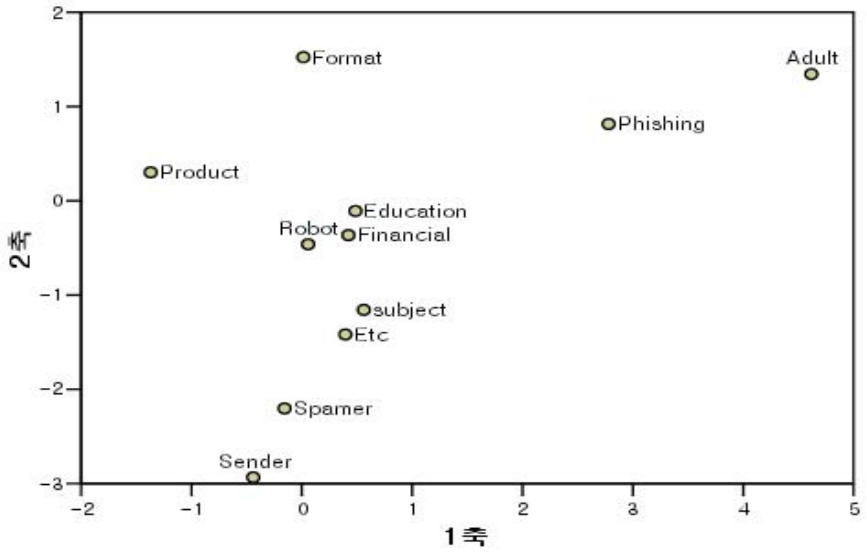
(표 4.4) 행 표준좌표

	Format	Robot	Financial	Product	spamer	Education	subject	Phishing	Sender	Adult	Etc
1축	0.014	0.057	0.422	-1.370	-0.156	0.485	0.561	2.780	-0.441	4.617	0.394
2축	1.523	-0.461	-0.364	0.301	-2.203	-0.109	-1.157	0.814	-2.933	1.344	-1.419

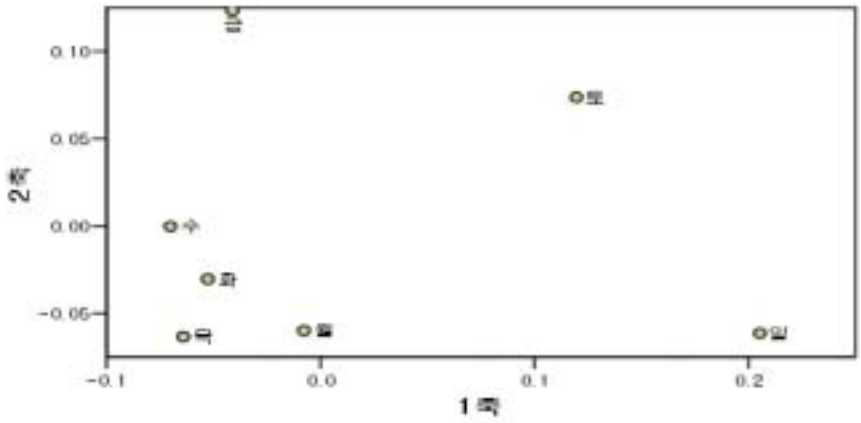
(표 4.5) 열 주좌표

	일	월	화	수	목	금	토
1축	0.025	-0.008	-0.053	-0.070	-0.064	-0.041	0.120
2축	-0.061	-0.060	-0.030	0.000	-0.063	0.123	0.074

(그림4.1)은 (표4.4)의 행 표준좌표를, (그림4.2)는 (표4.5)의 열 주좌표를 2차원 공간에 플롯한 것이다.



(그림 4.1) 스팸메일 종류의 표준좌표플롯



(그림 4.2) 요일의 주좌표 플롯

주좌표 플롯인 (그림4.2)에서 먼저 설명력이 높은 1축을 중심으로 살펴보면 왼쪽에 월요일, 화요일, 수요일, 목요일, 금요일이 밀집해 하나의 군을 이루고 있고, 오른쪽으로 토요일과 일요일이 각각 개별적으로 떨어져 위치하고 있다. 제 2축의 위에는 금요일과 토요일이 차례로 위치해 있고 아랫부분에 나머지 요일들이 모여 하나의 군을 이루고 있다고 할 수 있다. 이로부터 주말인 금요일, 토요일과 일요일에 수신되는 스팸메일 유형이 주중의 다른 요일들과 구별됨을 알 수 있다. (그림4.1)에 의하면 일요일에는 Adult, Phishing메일이 많고 토요일에는 Adult, Phishing메일, 그리고 헤더에 비정상적인 포맷이 들어있는 스팸메일이 많다고 할 수 있다. 금요일에는 헤더에 비정상적인 포맷이 들어있는 스팸메일이 다른 요일에 비해 특히 많음을 알 수 있다. 주중인 월요일, 화요일, 수요일 그리고 목요일에 수신되는 스팸메일의 종류는 유사한 패턴을 보인다고 할 수 있다.

5. 결론

본 연구는 스팸메일 차단시스템 개발 또는 적용 시 스팸메일의 차단율을 높이기 위한 방법의 하나로 대응분석을 이용한 탐색적 분석을 통하여 스팸메일 내에 존재하는 패턴을 찾는데 목적을 두었다. 분석결과에 의하면 요일별로 스팸메일 종류 간에 특성이 다름을 알 수 있다. 즉, 주말과 주중에 수신되는 스팸메일의 패턴이 다르고, 주중에는 대체적으로 모든 요일에 유사한 패턴을 보이지만 주말에는 요일마다 패턴이 다르게 나타난다. 현재 개발되고 있거나 사용 중인 스팸메일 차단시스템들은 스팸메일 유형이 그룹별 또는 개인별로 차이를 보이므로 차단율을 높이기 위하여 다양한 분류기법들을 사용할 뿐 아니라 그룹별, 개인별 관리를 하고 있으므로 스팸메일 차단시스템 설계 시 대응분석을 한 결과를 활용하면 좀 더 효율적인 스팸메일 관리를 할 수 있으리라 기대된다. 물론 본 연구에서 고려한 요일별 요인이외의 다른 요인들도 고려

해볼 수 있고 또 본 연구에서 이용된 자료는 대학교내의 스팸메일 자료로 다른 기관들에 적용할 시에는 대상 기관의 스팸메일 자료를 분석하여 특별한 스팸메일 패턴을 보이는지 확인해볼 필요가 있다.

참고문헌

1. 최용석 (1993). SAS 대응분석, 자유아카데미, 서울.
2. 허명회 (1999). 다변량 수량화, 자유아카데미, 서울.
3. 허명회, 양경숙 (2002). SPSS 다변량자료분석, SPSS아카데미, 서울.
4. Giorgetti, D. and Sebastiani, F. (2003). Automating Survey Coding by Multiclass Text Categorization Techniques, *Journal of American Society for Information Science and Technology,* 54, 1269-1277.
5. Graham. P. (2002). A Plan for Spam. http://www.paulgraham.com/spam.html
6. Greenacre, M. J. (1984). *Theory and Applications of Correspondence Analysis*. Academic Press, London.
7. Greenacre, M. and Hastie, T. (1987). The geometric interpretation of correspondence analysis, *Journal of the American Statistical Association*, 82, 437-447.
8. Manco, G., Macciari, E., Ruffolo, M. and Tagarelli, A. (2003). *Towards an Adaptive Mail Classifier*, Technical report, ISI-CNR.
9. Mock, K. (2001). *An Experimental Framework for Email Categorization and Management*, SIGIR '01.
10. New York Times, March 19, 1998.
11. Robinson, G. (2002). *Spam Detection.* http://radio.weblogs.com/0101454/stories/2002/0916/spamDetection.html
12. Robinson, G. (2003). Article in the Linux Journal march 2003 issue 107. http://www.linuxjournal.com/article.php?sid=6467
13. Ruvini, J. and Gabriel, J. (2002). *Do Users Tolerate Errors from their Assistant?,* Experiments with an E-mail Classifier, IUI'02.
14. Shin, Y. K. (1982). *The Singular Value Decomposition in Data Analytic Multivariate Analysis*, MSc Thesis, Korea University.

[2005년 7월 접수, 2005년 9월 채택]