퍼지관계곱을 이용한 내용기반 정크메일 분류 모델
(A Junk Mail Checking Model using Fuzzy Relational Products)

박 정 선 † 김 창 민 † 김 용 기 ††
(Jeong-Seon Park) (Chang-Min Kim) (Yong-Gi Kim)

요 약 인터넷의 발전을 기반으로 전자메일 서비스는 기존 우편 기능을 대체하여 현재의 대표적인 정보 전달 수단으로 자리잡고 있다. 전자메일 사용자의 확산에 따라 많은 기업들은 전자메일을 통한 개인별 카탈로그 보급 식의 광고에 투자를 하게 되었는데, 이는 개인별 취향을 고려한 광고가 가능하다는 잇점을 가진다. 그러나 전자메일 사용자들은 인터넷상에 개인 전자메일 주소가 노출됨에 의해서 많은 정크메일(junk mail)을 수신하게 되었는데, 정크메일이란 기업의 광고 선전물과 같이 수신을 원하지 않는 전자메일을 의미한다. 정크메일의 증가에 따라 정크메일을 분류하는 수단이 필요하게 되었는데, 현재까지는 사용자가 입력한 송신자의 전자메일 주소 또는 도메인 주소를 등록하여 차단하거나 제목에 특정 단어를 포함한 메일을 완전히 삭제하여 버리는 기술수준에 머무르고 있다. 본 논문에서는 퍼지관계곱을 기반으로 메일의 내용에 의미적으로 접근하여 정크메일을 추출하는 정크메일 분류 모델을 제안한다. 이는 퍼지관계곱 연산을 이용하여 미리 정의한 정크용어들과 사용자에게 수신되는 전자메일 내의 용어들 간 의미적 포함관계를 분석하고 그를 통해 전자메일의 정크도(degree of junk)를 추출한다. 각 전자메일별로 추출된 정크도는 사용자가 부여하는 정크 기준치(SVJ, Standard Value of Junk)를 기준으로 정크메일과 비정크메일로 분류한다. 제안된 기법은 사용자가 특정 개수의 동일한 전자메일에 대해 느끼는 정크도를 기준으로 분류한 정크메일 수를 비교하여 그 효용성을 증명하였다.

키워드 : 퍼지관계곱, 정크메일, 정크용어베이스, 정크도

Abstract E-mail service has been a general method for communication as internet is widely used instead of post mails. Many companies have invested in e-mail advertisement as e-mail service is spread. E-mail advertisement has an advantage that it can consider personal characters.

A lot of e-mail users have been given e-mails that they did not want to receive because their e-mail addresses were opened out to companies on internet. Therefore, they need junk mail checking systems and several e-mail service providers have supported junk mail filters. However, the junk mail filters can check the junk mail with constraint because they don't check the junk degree of mails by the contents of e-mail. This paper suggests a content-based junk mail checking model using fuzzy relational products. The process of the junk mail checking model using fuzzy relational products is as following: (1) analyzes semantic relation between junk words-base and e-mails, (2) checks the junk degree of the e-mail using the semantic relation, (3) checks the mails with SVJ(Standard Value of Junk) if those are junk mail or non-junk mail.

The efficiency of the proposed technique is proved by comparing the junk degree of the e-mail and the number of junk mails that was checked by e-mail users and checked by the proposed junk mail checking model.

Key words : fuzzy relational products, junk mail, junk words base, junk degree

† 비 회 원 : 경상대학교 컴퓨터학과 및 정보통신연구원
kong@ailab.gsnu.ac.kr
nuno@ailab.gsnu.ac.kr
†† 종신회원 : 경상대학교 컴퓨터학과 및 정보통신연구원 교수
ygkim@nongae.gsnu.ac.kr
논문접수 : 2002년 1월 23일
심사완료 : 2002년 8월 9일

1. 소 개

20세기 후반 인터넷의 발전을 기반으로 전자메일(e-mail, electronic mail)은 기존의 우편 메일(post mail)을 대신하여 대표적인 정보전달 수단으로 자리잡고

있다[1][2][3]. 전자메일은 기업간의 정보교환 뿐만 아니라 개인간의 정보교환 서비스를 제공함으로써 그 사용자가 급속히 확산되었는데, 이러한 현상에 따라 기업들은 전자메일을 통한 개인별 전자 카탈로그를 보급하는 형태의 광고에 많은 투자를 하게 되었다[2][3][4]. 전자메일 사용자들은 인터넷상의 다양한 서비스를 제공받고자 전자메일 주소를 여러 사이트에 등록하여 사용하는데, 이 때 기업간의 회원정보 공유 또는 사용자의 부주의나 무관심 등의 이유로 인해 사용자의 전자메일 주소가 인터넷상에 공개된다. 이로 인해 전자메일 사용자들은 특정 기업이나 상품의 광고 선전물과 같이 자신이 원하지 않고 자신에게 불필요한 전자메일을 대량 받게 되었는데, 이를 정크메일(junk mail)이라 한다.

전자메일을 통한 광고의 증가에 따라 최근 전자메일 사용자는 자신의 메일 관리함에 쌓이는 정크메일을 처리하기 위해 많은 시간과 노력을 낭비하게 되었고, 이에 따른 스트레스도 크게 증가하고 있는 추세이다. 전자메일 사용자가 꾸준히 증가한다는 점을 감안해 보았을 때 보다 많은 기업들이 전자메일을 통한 광고에 투자하게 될 것이고, 전자메일 사용자들은 현재 보다 더 많은 정크메일로 고통받게 될 것이다.

이러한 추세에 따라 최근 몇몇 전자메일 서비스 회사에서 정크메일 분류 기능을 제공하고 있는데, 현재까지는 그 기술이 메일 송신자의 전자메일 주소나 도메인 주소를 등록하여 차단하거나 제목이나 본문에 특정 단어를 포함한 메일을 완전히 삭제하여 버리는 수준에 머무르고 있다. 이러한 방법들은 구조가 단순해 시스템 설계가 간단하다는 장점을 가지나 사용자가 입력하는 내용에만 의존하므로 정크메일의 증가량에 따라 수신자의 부담도 커지는 등의 단점을 가진다. 따라서 이러한 현재 기술의 한계를 극복하는 전자메일의 내용에 기반한 정크메일 분류 기법에 관한 연구의 필요성이 강조되고 있다.

본 논문에서는 사용자의 입력에만 의존하는 기존의 정크메일 기법의 한계를 극복하기 위해 퍼지관계곱 연산을 이용한 내용기반 정크메일 분류 모델을 제안한다. 본 논문에서 제안하는 정크메일 분류 모델은 정크용어 베이스를 정의하고, 수신된 전자메일에 대해 정규화 및 퍼지화 단계를 거쳐 퍼지관계곱 연산을 적용하여 정크도(degree of junk)를 추출하며, 최종적으로 추출된 정크도를 기반으로 개인 사용자의 의견을 수렴한 정크기준치(SVJ, Standard Value of Junk)를 부여하는 순서로 정크메일과 비정크메일을 분류한다.

본 논문의 2장에서는 제안하는 모델의 연구배경에 대해 소개하고, 3장에서는 본 논문의 기반 이론인 퍼지관계곱에 대해 살펴보고, 4장에서는 퍼지관계곱을 이용한 전자메일의 정크도 추출에 관해 살펴본다. 그리고 5장에서는 제안하는 정크메일 분류 모델의 세부 구성과 설계 및 구현에 대해 살펴본다. 그리고 6장에서는 다양한 인터넷 사용자로부터 설문한 결과와 제안하는 기법의 결과를 비교 및 평가하고, 마지막으로 7장에서는 본 논문의 결론에 대해 살펴본다.

2. 연구배경

정크메일로 인한 전자메일 사용자의 스트레스 증가에 따라 전자메일 서비스 제공업체에서는 간단한 구조를 가지는 정크메일 분류 기법을 제공하고 있다[5][6][7][8]. 이는 크게 두 가지로 구분할 수 있는데, 첫 번째는 등록된 단어에 의한 분류기법이고, 두 번째는 등록된 주소에 의한 분류 기법이다.

등록된 단어에 의한 분류기법은 사용자가 등록한 특정 단어를 제목이나 본문에 포함하고 있는 전자메일에 대해 특정 방법으로 처리하는 기법이다. 이를 위해 사용자는 수신하기를 원하지 않는 단어와 분류된 전자메일에 대한 처리 방법을 설정하여야한다. 분류된 메일에 대한 처리 방법으로는 특정 폴더로 분류하거나 복사 또는 완전히 삭제하는 방법 등이 있다.

등록된 주소에 의한 분류는 세부적으로 2가지 모델로 구분할 수 있다. 첫째, 수신을 거부하고자 하는 송신자의 전자메일 주소나 도메인 주소를 등록하여 이를 기반으로 정크메일과 비정크메일을 분류하는 기법이 있다. 둘째, 수신하고자 하는 송신자의 전자매일 주소나 도메인 주소를 등록하여 등록된 주소 이외의 전자메일은 모두 삭제하거나 정크메일로 분류하는 모델이 있다. 이를 위해서도 사용자는 송신자의 주소를 등록하고, 분류되는 메일을 위한 처리방법을 설정하여야 한다.

이러한 기존의 정크메일 분류 방법들은 구조가 단순해 시스템 설계 및 구현이 간단하다는 장점을 가지나 수신자가 입력하는 내용에만 의존하므로 정크메일의 증가량에 따른 수신자의 부담이 커지게 되고, 전자메일의 내용에 기반하지 않으므로 해서 등록되지 않은 송신자나 제목에 비중이 높지 않은 정크메일은 분류가 불가능하며, 일반 메일이 차단될 우려가 높다는 단점을 가진다.

3. 퍼지관계곱(fuzzy relational products)

3.1 퍼지 관계곱(fuzzy relational products)

퍼지관계곱(fuzzy relational products)은 Bandler와 Kohout이 이진 관계곱(crisp relational products)을 확장하여 제안한 연산으로 두 퍼지관계 내 원소들간의 의

미적 포함관계를 나타낸다. 퍼지관계곱은 실세계에 내재하는 모호성(imprecision)을 표현하는 퍼지시스템[9][10][11]의 한 예로서 인지능력(cognitive ability), 결정능력(decision ability) 그리고 행동능력(action ability)과 같은 특징을 하나 이상 포함하고 있는 복잡한 시스템의 분석(analysis)과 종합(synthesis)을 위해 활용될 수 있다[12][13][14][15][16][17].

퍼지집합 A, B, C와 그들 간의 퍼지관계 $\widetilde{S}: B\times C$과 $\widetilde{S}: B\times C$가 주어지고 $a_i\in A$, $c_k\in C$라 할 때, $\widetilde{R}$과 $\widetilde{S}$의 퍼지관계곱 $(\widetilde{S}@\widetilde{S})_{ik}$는 퍼지집합 A의 원소 a_i와 퍼지집합 C의 원소 c_k의 의미상 포함관계를 나타내는 것으로서 수식 (1)(2)(3)과 같이 세 가지 퍼지관계곱 연산 ◁, ▷ 또는 □로서 표현될 수 있다[18][19][20]. 이때 $|B|$는 두 퍼지관계 R과 S에 대해 퍼지관계곱 연산을 적용시킬 때 퍼지집합 B의 원소 개수를 의미한다.

$$(R\triangleleft S)_{ik}=\frac{1}{|B|}\sum(R_{ij}\rightarrow S_{jk}) \tag{1}$$

$$(R\triangleright S)_{ik}=\frac{1}{|B|}\sum(R_{ij}\leftarrow S_{jk}) \tag{2}$$

$$(R\square S)_{ik}=\frac{1}{|B|}\sum(R_{ij}\leftrightarrow S_{jk}) \tag{3}$$

수식 (1)의 ◁ 연산자는 퍼지삼각서브논리곱(fuzzy triangle sub-product)이라고 하고 이는 a_i가 c_k에 포함되는 정도를 의미한다. 수식 (2)의 ▷ 연산자는 퍼지삼각수퍼논리곱(fuzzy triangle super product)이라고 하고 이는 a_i가 c_k를 포함하는 정도를 의미한다. 수식 (3)의 □ 연산자는 퍼지사각논리곱(fuzzy square product)이라고 하고 이는 a_i와 c_k가 유사한 정도를 의미한다[12][13][14][21].

본 논문에서는 전자메일의 정크도를 추출하기 위해 퍼지관계곱 연산 중 일반적으로 가장 널리 쓰이는 퍼지삼각서브논리곱 연산자를 적용하여 전자메일의 내용을 의미적으로 분석한다.

3.2 퍼지 조건연산자(fuzzy implication operator)

퍼지관계곱은 퍼지 조건연산자(fuzzy implication operator)를 이용하여 적절히 처리되는데 퍼지 조건연산자는 이진 조건연산과 달리 다양한 방법으로 구현 가능하며 현재 수 십여 가지가 제안되어 있다. 수식 (4) ~ 수식 (12)는 대표적인 퍼지 조건 연산자를 보여주고 있다[14][22][23].

$$a\rightarrow_1 b=\begin{cases}1 & iff\ a\neq 1\ or\ b=1\\ 0 & otherwise\end{cases} \tag{4}$$

$$a\rightarrow_2 b=\begin{cases}1 & iff\ a\leq b\\ 0 & otherwise\end{cases} \tag{5}$$

$$a\rightarrow_3 b=\begin{cases}1 & iff\ a\leq b\\ b & otherwise\end{cases} \tag{6}$$

$$a\rightarrow_4 b=\min\left(1,\frac{b}{a}\right) \tag{7}$$

$$a\rightarrow_{4'} b=\min\left(1,\frac{b}{a},\frac{1-a}{1-b}\right) \tag{8}$$

$$a\rightarrow_5 b=\min(1,1-a+b) \tag{9}$$

$$a\rightarrow_6 b=(1-a)\vee b \tag{10}$$

$$\begin{aligned}a\rightarrow_7 b &= (a\wedge b)\vee(1-a)\\ &= (a\rightarrow_6 b)\wedge a\end{aligned} \tag{11}$$

$$\begin{aligned}a\rightarrow_8 b &= ((1-a)\vee b)\wedge(a\vee(1-b)\vee(b\wedge(1-a)))\\ &= (a\rightarrow_7 b)\wedge a\\ &= (a\rightarrow_6 b)\wedge a\wedge a\end{aligned} \tag{12}$$

본 논문에서는 전자메일의 정크도 추출시 이용되는 퍼지관계곱 연산의 수행을 위해 다양한 퍼지조건연산자 중 일반적으로 가장 널리 쓰이는 수식 (8)에 보인 퍼지조건연산자 $\rightarrow_4$를 적용하였다[14].

4. 퍼지관계곱을 이용한 전자메일의 정크도 추출

본 논문에서 제안하는 정크메일 분류 모델은 정크용어베이스의 정의, 전자메일의 정규화 및 퍼지화, 정크도 산출, 정크기준치 부여 등의 세부 모듈을 가진다. 본 절에서는 제안하는 모델의 핵심 모듈이 되는 퍼지관계곱을 이용한 전자메일의 정크도 추출을 위한 모듈에 대해 다룬다.

퍼지관계곱을 이용한 전자메일의 내용기반 정크도 추출은 다음과 같은 순서로 이루어진다. 수식 (13)과 같이 수신된 전자메일 집합을 M, 수식 (14)와 같이 정크용어 집합을 T라 두었을 때, 전자메일 m_i와 정크용어집합 T의 빈도수를 이용한 퍼지관계는 수식 (15)와 같이 $\widetilde{S}$로 표현된다. 그리고 수식 (16)에 나타난 정크용어 퍼지집합 $\widetilde{J}$는 정크용어집합 T내의 정크용어 t_i와 그들이 가지는 정크성 fv_i로 구성되는데, 수식 (16)으로부터 정크용어 t_i가 가지는 정크성 j_i를 수식 (17)과 같이 추출하고, 수식 (18)과 같이 퍼지관계 $\widetilde{R}$을 정크용어 집합 T와 정크용어 퍼지집합 $\widetilde{J}$의 정크성 j_i에 의한 퍼지관계로 표현한다. 두 퍼지관계 $\widetilde{S}$와 $\widetilde{R}$을 퍼지관계곱 연산

중 수식 (19)와 같이 퍼지삼각서브논리곱연산 적용하고 이를 처리하기 위해 기존의 퍼지조건연산자 중 수식 (8)에 보인 $\rightarrow_4$ 연산자를 적용하며, 이를 이용하여 전자메일 m_i에 대한 정크용어 t_i의 의미적 포함거리 r_j를 추출한다[24][25].

$$M = \{m_1, m_2, \cdots, m_n\} \tag{13}$$

$$T = \{t_1, t_2, \cdots, t_m\} \tag{14}$$

$$\tilde{S}_{m_i \rightarrow T} = [d_1, d_2 \quad \cdots \quad d_m] \tag{15}$$

$$\tilde{J} = \{fv_1/t_1, fv_2/t_2, \ldots\ldots, fv_m/t_m\} \tag{16}$$

$$j_i = \mu_{\tilde{J}}(t_i),\ t_i \in T \tag{17}$$

$$\tilde{R}_{T \rightarrow \tilde{J}} = \begin{bmatrix} j_1 \\ j_2 \\ \vdots \\ j_m \end{bmatrix} \tag{18}$$

$$\begin{aligned} \tilde{J} \rightarrow m_i &= {}_{m_i}\tilde{S} \triangleleft \tilde{R}_{\tilde{J}} \\ &= [d_1, d_2 \cdots d_m] \triangleleft \begin{bmatrix} j_1 \\ j_2 \\ \vdots \\ j_m \end{bmatrix} \\ &= r_i \end{aligned} \tag{19}$$

수식 (19)와 같이 퍼지관계곱 연산을 이용하여 추출한 결과 r_i는 전자메일 m_i와 정크용어 간 의미적 포함거리만을 추출하고 각 정크용어가 가지는 정크성의 참여비중을 고려하지는 못한다. 따라서, 본 논문에서는 정크용어가 가지는 정크성의 비중을 결과에 반영하기 위하여 퍼지관계곱 연산 처리를 위한 퍼지조건연산자를 새로이 제안한다.

수식 (20)은 본 논문에서 제안하는 정크도 추출을 위한 퍼지조건연산자를 보인다. 이는 정크용어가 가지는 정크성의 비중을 고려하기 위하여 기존의 퍼지조건연산자 $\rightarrow_4$에 정크용어의 정크성 j_i를 곱하는 특성을 가진다. 수식 (21)은 실제 본 연구에 사용된 퍼지조건연산자 $\rightarrow_{\mu}$를 보인다.

$$a \rightarrow_j b = a \times (a \rightarrow b) \tag{20}$$

$$\begin{aligned} a \rightarrow_{\mu} b &= a \times (a \rightarrow_4 b) \\ &= a \times \min\left(1, \frac{b}{a}\right) \end{aligned} \tag{21}$$

본 논문에서는 전자메일의 정크도 추출을 위해 새롭게 제안한 퍼지조건연산자를 적용하여 각 정크용어들의 전자메일 m_i에 대한 정크도를 구하고, 수식 (22)에서 보는 것과 같이 그들의 합을 구하며 정의된 정크용어집합의 원소 개수 m을 이용하여 평균을 구함으로써 메일 m_i에 대한 전체 정크도 r_{m_i}를 추출한다. 이와 같이 산출된 결과의 표준화를 위하여 수식 (23)과 같이 r_{m_i}에 상수 c를 곱하고, 1과 min연산하여 최종적으로 전자메일 m_i에 대한 정크도 f_i를 추출한다[24][25].

$$r_{m_i} = \frac{\sum_{i=1}^{m} r_i}{m}$$

(단, m은 정크용어집합 T의 원소 개수임) (22)

$$f_i = \min(1, r_{m_i} \times c) \tag{23}$$

5. 퍼지관계곱을 이용한 내용기반 정크메일 분류 모델

본 논문에서 제안하는 정크메일 분류 모델은 사용자의 입력에만 의존하는 기존의 정크메일 기법의 한계를 극복하기 위한 모형으로, 퍼지관계곱을 이용하여 수신된 전자메일의 내용을 기반으로 정크메일과 비정크메일을 분류하는 특성을 가진다[24][25]. 퍼지관계곱을 기반으로 전자메일의 정크도를 추출하고, 추출된 정크도를 기반으로 개인사용자의 의견을 수렴한 정크기준치(SVJ, Standard Value of Junk)를 부여하여 정크메일과 비정크메일로 분류하는 단계를 거쳐 수행된다.

제안하는 모델은 전처리 작업 단계와 실시간 작업 단계로 분리하여 설계되었다. 먼저 전처리 단계에서는 정크용어베이스를 구축하고, 정크기준치(SVJ)를 부여하는 작업을 수행하며, 실시간 처리 단계에서는 수신된 전자메일에 대해 정규화 및 퍼지화 작업, 그리고 퍼지관계곱을 이용한 전자메일의 정크도 추출, SVJ의 적용, 전자메일의 분류 작업이 순서대로 이루어진다. 그림 1은 본 논문에서 제안하는 퍼지관계곱을 이용한 내용기반 정크메일 분류 모델의 구성도를 보이고 있다.

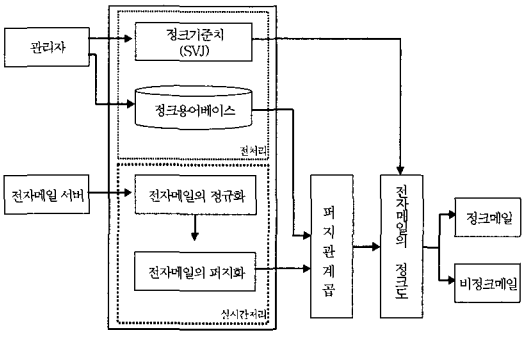

그림 1 퍼지관계곱을 이용한 정크메일 분류 모델 구성도

표 1 정크용어베이스

번호	정크용어	정크도	번호	정크용어	정크도	번호	정크용어	정크도
1	가격	1	26	무통장	0.9	51	주소	0.2
2	가입	0.9	27	문의	0.9	52	증정	0.7
3	감사합니다	0.9	28	부가세	0.8	53	초대	1
4	개최	0.6	29	사은	1	54	최저가	0.8
5	경품	0.8	30	상담	0.9	55	추천	0.8
6	고객	1	31	상품	1	56	추첨	0.3
7	공지	0.6	32	샘플	0.9	57	쿠폰	0.8
8	공짜	0.6	33	서비스	1	58	퀴즈	0.2
9	관리자	0.8	34	세일	1	59	테마	0.4
10	광고	0.8	35	소식	0.7	60	판매	1
11	구독	0.9	36	쇼핑	0.9	61	팩스	0.6
12	구매	0.8	37	수신거부	0.9	62	편집자	0.4
13	구입	0.8	38	신용카드	0.5	63	할부	1
14	귀하	0.6	39	안내	0.9	64	할인	0.8
15	기념	0.6	40	운영자	0.7	65	해지	0.7
16	기회	0.6	41	웹마스터	1	66	행사	1
17	뉴스	0.8	42	응모	0.9	67	회원	1
18	담당자	0.6	43	이벤트	1	68	copyright	1
19	당첨	0.7	44	이용	0.8	69	fax	0.6
20	대표	0.9	45	인터뷰	0.5	70	webmaster	1
21	등록	0.6	46	인하	0.7			
22	멤버	1	47	적립	1			
23	모집	0.8	48	정보	0.4			
24	무료	0.6	49	제품	0.9			
25	무이자	1	50	(주)	1			

5.1 정크용어베이스

정크용어베이스는 정크메일 분류 모델의 기반 자료로서 정의된 정크용어들과 그들이 가지는 정크성으로 구성된다. 이는 정크메일 분류 모델의 전처리 단계에서 수행하여야할 작업으로, 미리 수신된 전자메일을 기반으로 관리자가 다양한 방법을 적용하여 구축한다.

정크용어베이스 내의 정크용어는 전자메일의 정크도를 파악하는데 있어 핵심이 되는 자료로서 '정크하다'는 특수한 성질을 포함하는 단어들로서 구성되며, 그 크기가 방대해져 정크용어베이스의 특수성을 잃지 않는 크기로 구축한다[26]. 정크하다는 성질은 매우 추관적이기 때문에 사용자의 생각을 설문조사를 통해 통계적으로 정립하는 것이 가장 일반적이다. 본 논문에서는 각 정크용어가 가지는 정크성을 정의하기 위해 일반적인 정크적 지식을 바탕으로 현재 수신된 다양한 광고성 전자메일로부터 메일 내용에 포함되어 있는 용어들의 빈도수[18][19][26]를 추출하여 정크용어집합을 정의하고, 정의된 정크용어들에 대해 사용자가 생각하는 정크성을 0과 1 사이의 퍼지 값으로 부여하도록 인터넷 설문조사하여 0.1이상의 정크성을 가지는 정크용어들을 추출한다. 추출된 정크용어들을 기반으로 제안하는 정크메일 분류 모델에서의 정크용어베이스를 구축하고, 구축된 정크용어베이스는 정크메일 분류시 사용자마다의 특성을 고려하도록 하기 위해 관리자가 수정할 수 있도록 설계한다. 표 1은 제안하는 시스템을 위해 구축한 정크용어베이스 내의 정크용어와 각 정크용어가 가지는 정크성을 보이고 있다.

5.2 전자메일의 정규화

전자메일의 정규화란 IETF(Internet Engineering Task Force)에서 발표한 RFC(Request For Comments)문서들[27]에 근거하여 전자메일 서버에 수신된 순수메일(raw mail)이 포함하고 있는 수많은 헤더 필드와 본문을 "송신자", "제목", "내용" 등으로 구성요소를 분류하고 전자메일 내에 포함된 메일 내용과 관계없는 HTML 태그 등을 제거하여 정규메일(regulated mail)로 변환시키는 작업을 의미한다[28].

본 논문에서 순수메일은 한글 텍스트 메일을 기반으로 하며, 전자메일 또는 메일은 정규화 과정을 거친 정규메일을 의미한다. 전자메일의 정규화 과정을 수행하기 위해 본 논문에서는 수신된 전자메일을 하나의 텍스트 파일로 저장하고 각 파일 단위로 처리하였으며, 메일 내용을 단어 단위로 용어데이타베이스(words database)에

저장하였다. 표 2는 예제 메일 m_1에 나타난 정크용어와 그 정크용어가 가지는 정크도를 보이고 있다.

표 2 예제 메일 m_1에 나타난 정크용어와 정크도

정크용어번호	정크용어	정크도
1	가격	1
6	고객	1
12	구매	0.8
24	무이자	1
27	사은	1
29	상품	1
35	쇼핑	0.9
36	수신거부	0.9
46	적립	1
48	정보	0.4
49	제품	0.9
52	증정	0.7
57	추첨	0.3
58	쿠폰	0.8
64	할인	0.8
67	회원	1
68	copyright	1

5.3 내용기반 메일 퍼지화

수신된 전자메일이 정크메일인지를 판별하기 위해서는 전자메일의 정크도를 추출하는 작업이 이루어져야한다. 전자메일의 정크도를 추출하기 위해서는 먼저 메일 내용의 의미를 이해하여야 한다. 이는 전자메일 내에 포함되어 있는 특정 용어의 빈도수를 이용하는 방법으로 해결할 수 있는데, 용어의 빈도수를 이용하는 방법은 정보검색모델에서 문서의 의미를 파악하는 방법 중 가장 대표적인 방법이다. 용어의 빈도수가 해당 문서의 의미를 모두 반영한다고는 할 수 없지만, 현 기술수준에서는 가장 적절한 방법이다[18][19][23][26][29].

여러 가지 정보검색 모델 중 Bandler와 Kohout이 제안한 BK-퍼지정보검색모델에서는 특정 문서의 의미를 파악하기 위해서 문서 내에 등장하는 용어의 빈도수로서 문서의 상대적 관련성을 추출하고 이를 퍼지화한다[12]. 그리고 BK-퍼지정보검색모델을 개선시킨, 개선된 BK-퍼지정보검색모델(A-FIRM, Advanced Bandler-Kohout Fuzzy Information Retrieval Model)에서는 용어의 빈도수를 퍼지화하기 위한 멤버쉽 함수를 고안하고, 이를 이용하여 문서에 대한 용어의 소속도를 산출하였다[19]. 수식 (24)는 A-FIRM에서 이용되는 용어 빈도 퍼지화 멤버쉽 함수를 보이고 있고, 그림 2는 용어 빈도 퍼지화 멤버쉽 함수 그래프를 보이고 있다. 수식 (24)와 그림 2에 나타난 x는 용어의 빈도수를 의미하고, m은 빈도수에 대한 임계값을 의미한다.

$$\mu_{T_r}(x)=\begin{cases}\frac{1}{2}x, & 0\le x\le 1\\ \frac{0.5}{m-1}(x-1)+0.5, & 1<x\le m\\ 1, & m<x\end{cases} \quad (24)$$

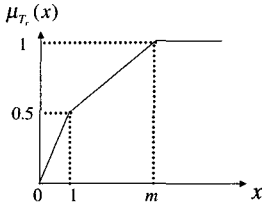

그림 2 A-FIRM의 용어 빈도 퍼지화 멤버쉽 함수 그래프

본 논문에서는 A-FIRM에서 제안한 용어의 빈도수 퍼지화 멤버쉽 함수를 적용하여 수신된 전자메일 내에 존재하는 정크용어의 빈도수를 퍼지화하였다. 빈도수 퍼지화를 위한 함수의 빈도수 임계값 m은 3으로 부여하여 처리하였는데 이는 한 전자메일 내에 한 단어가 빈도수 3 이상의 수를 가진다면 모두 1의 의미를 가진다는 뜻이다. A-FIRM의 용어 빈도 퍼지화 멤버쉽 함수에 따르면 $m=3$일 때, 빈도수 x에 대한 퍼지화 멤버쉽 함수 값 $\mu_{T_r}(x)$은 각각 $x=1$일 때 0.5, $x=2$일 때 0.75, $x\ge 3$일 때 1의 퍼지 값을 가진다. 빈도수 임계값 m은 일반적으로 전자메일의 내용이 일반 정보검색을 위한 웹문서에 비해 많지 않다는 특성 등을 고려하여

표 3 예제 메일 m_1에 나타난 정크용어의 빈도수와 퍼지화

정크용어번호	정크용어	빈도수	빈도수 퍼지화
1	가격	3	1
6	고객	2	0.75
12	구매	1	0.5
24	무이자	6	1
27	사은	1	0.5
29	상품	1	0.5
35	쇼핑	1	0.5
36	수신거부	1	0.5
46	적립	7	1
48	정보	1	0.5
49	제품	1	0.5
52	증정	1	0.5
57	추첨	1	0.5
58	쿠폰	2	0.75
64	할인	9	1
67	회원	2	0.75
68	copyright	1	0.5

최적의 결과를 산출하도록 하는 시스템 튜닝을 통해 얻은 기준이다. 표 3은 수신된 예제 메일 m_1에 나타난 정크용어의 빈도수와 빈도수를 퍼지화한 결과를 보인다.

5.4 전자메일의 정크도 추출과정

본 논문에서는 전자메일의 정크도를 추출하기 위해 2단계의 과정을 거친다. 첫 번째 단계에서는 수신된 전자메일 m_i와 정크용어 집합 T의 빈도수에 의한 퍼지관계 $\widetilde{S}$를 구하고, 정크용어 집합 T와 정크용어 퍼지집합 J의 정크성에 의한 퍼지관계 R을 구하여 두 퍼지관계 $\widetilde{S}$와 $\widetilde{R}$을 퍼지관계곱 연산의 퍼지삼각서브논리곱 $(_{m_i}\widetilde{S} \triangleleft \widetilde{S_j})$연산을 적용한다. 이 때, 퍼지관계곱 연산을 처리하기 위한 퍼지조건연산자는 본 논문에서 새롭게 제안한 $\rightarrow_{\mu}$ 연산자를 적용하였으며, 이는 기존의 퍼지조건연산자에 의해 추출되는 메일 m_i에 대한 정크용어 t_i의 의미적 포함정도에 각 정크용어가 가지는 정크성을 고려하도록 선언된 퍼지조건연산자이다. 두 번째 단계에서는 퍼지삼각서브논리곱연산을 적용하여 얻은 각 정크용어의 전자메일 m_i에 대한 정크도를 통합하여 평균계산하며 결과의 표준화를 위하여 결과 r_{m_i}에 상수 c를 곱하고 1과 min연산하여 최종결과 전자메일 m_i의 정크도 f_i를 추출한다[24][25]. 본 논문에서 결과의 표준화를 위한 상수 c에는 10을 부여하여 처리하였는데, 이는 시스템 튜닝에 의해 최적의 결과를 보이는 수치를 설정한 것이다.

표 4는 수신된 예제 전자메일 m_1에 대해 퍼지관계곱을 이용하여 정크도를 추출하는 전 과정을 보인다. 전자메일에 대한 정크도를 추출하기 위해 각 정크용어들이 전자메일 m_1에 대해 가지는 정크도를 추출하는데, 전자메일 m_1에 나타난 정크용어들 중 "구매"를 기준으로 정크도 추출 과정을 살펴보겠다. 먼저 정크용어의 정크성에 의한 퍼지관계 $\widetilde{R}$을 적용한 결과로 "구매"의 정크성 0.8을 추출하고, 전자메일 m_1 내에서 "구매"가 빈도수 1의 값을 가지므로 전자메일 m_1과 정크용어 "구매"의 빈도수에 의한 퍼지관계 $\widetilde{S}$를 적용하여 0.5의 퍼지값을 추출하였다. 두 퍼지관계 $\widetilde{S}$와 $\widetilde{R}$을 퍼지삼각서브논리곱 연산자를 적용하여 퍼지관계곱 연산을 수행하며, 퍼지관계곱을 위한 퍼지조건연산자로 $\rightarrow_{\mu}$를 적용하여 메일 m_1에 대해 정크용어 "구매"의 정크도 0.8을 추출하였다. 표 4에서 보는 것과 같이 각 정크용어들이 메일 m_1에 대해 가지는 정크도를 추출하고 정크용어집합내의 원소 수를 이용하여 평균을 구하며, 결과의 표준화를 위해 상수를 곱하고 1과 min 연산하여 최종결과 전자메일 m_1의 정크도 f_i=1을 구하였다.

표 4 예제 메일 m_1에 나타난 정크용어와 빈도수

정크용어 번호	정크용어	정크성	빈도수	빈도수 퍼지화	$\rightarrow_{\mu}$적용
1	가격	1	3	1	1
6	고객	1	2	0.75	1
12	구매	0.8	1	0.5	0.8
24	무이자	1	6	1	1
27	사은	1	1	0.5	1
29	상품	1	1	0.5	1
35	쇼핑	0.9	1	0.5	0.9
36	수신거부	0.9	1	0.5	0.9
46	적립	1	7	1	1
48	정보	0.4	1	0.5	0.32
49	제품	0.9	1	0.5	0.9
52	증정	0.7	1	0.5	0.7
57	추첨	0.3	1	0.5	0.18
58	쿠폰	0.8	2	0.75	0.8
64	할인	0.8	9	1	0.64
67	회원	1	2	0.75	1
68	copyright	1	1	0.5	1
			$r_{m_1} = \frac{\sum_{i=1}^{70} r_i}{70}$		0.202
			$f_i = \min(1, r_{m_1} \times c)$		1

5.5 정크기준치(SVJ, Standard Value of Junk)

정크기준치는 퍼지관계곱의 적용에 따라 추출된 전자메일의 정크도를 기반으로 정크메일과 비정크메일로 분류시 적용되는 기준 값을 의미한다. 본 논문에서 제안하는 정크메일 분류 모델에서는 정크기준치를 부여하여 분류할 정크메일의 범위를 넓히거나 좁혀서 결과를 조절할 수 있게 한다. 즉, 퍼지관계곱을 적용해 추출한 정크도에 사용자의 의견을 수렴한 정크기준치를 부여하여 정크메일과 비정크메일로 분류한다. 이는 사용자로부터 0과 1사이의 값을 입력받음으로써 수행하는데, 예를 들어 관리자가 정크기준치를 0.4로 입력한다면 0.4 이상의 정크도를 가지는 전자메일은 정크메일로 분류된다[25].

6. 비교 및 평가

본 논문에서 제안하는 퍼지관계곱을 이용한 정크메일 분류 모델은 동일한 전자메일에 대해 사용자가 느끼는 정크도와 사용자의 의견을 수렴한 정크기준치(SVJ)의 부여에 따른 정크메일 분류 개수에 대해 비교 및 평가하였다. 본 절에서는 제안하는 모델의 비교 및 평가를

위한 평가 기준 및 결과 비교에 대해 살펴보겠다.

6.1 평가 기준

제안하는 모델의 평가를 위해 정크메일 내의 정크성을 가지는 용어들 중 70개의 용어에 대해 설문을 통하여 다양한 분야에 종사하는 수십 명의 전자메일 사용자들에게서 의견을 수렴하고 그 결과를 평균하여 정크용어베이스를 구축하였다. 그리고 정크기준치를 SVJ=0.3, SVJ=0.4, SVJ=0.6으로 부여하여 사용자의 의견 수렴 시 어떤 기준에서도 유사한 결과가 유도됨을 보였다.

6.2 사용자가 느끼는 전자메일에 대한 정크도

제안하는 시스템을 평가하기 위하여 전자메일 100개를 제공하고 다양한 분야에 종사하는 100명의 사용자를 대상으로 수신된 메일이 각 사용자에게 얼마나 필요 없는지를 기준으로 0과 1 사이의 값으로 부여하도록 설문하였으며, 설문 결과에 대해 최대 값과 최소 값을 제외한 나머지 결과를 평균하여 계산하였다.

6.3 결과 비교

표 5는 수신된 전자메일 20개에 대해 사용자가 정의한 정크도와 제안하는 기법을 통하여 추출한 정크도를 비교하여 보인다.

표 5 전자메일 20개에 대한 정크도

	정크도	
	메일사용자	제안하는 기법
m_1	0	0
m_2	0.4	0.4
m_3	0.9	1
m_4	1	1
m_5	0.3	0
m_6	0.5	0.5
m_7	0.9	0.9
m_8	0.4	0.5
m_9	0.9	1
m_{10}	0.9	1
m_{11}	0.8	1
m_{12}	0.6	0.6
m_{13}	0.5	0.2
m_{14}	0.7	0.9
m_{15}	0.8	0.7
m_{16}	0.7	0.3
m_{17}	0.7	1
m_{18}	0.7	1
m_{19}	0.7	0.2
m_{20}	0.8	0.8

표 6, 표 7, 표 8은 표 5에 나타난 20개의 전자메일에 대해 추출된 정크도를 기반으로 각각 다른 정크기준치(SVJ)를 부여하였을 경우에 정크메일로 분류되는 메일 예제 번호와 개수를 보인다.

표 6 전자메일 20개에 SVJ를 0.3으로 부여한 결과

구분 / SVJ	정크메일번호	
	메일사용자	제안하는 기법
0.3	m_2, m_3, m_4, m_5, m_6, m_7, m_8, m_9, m_{10}, m_{11}, m_{12}, m_{13}, m_{14}, m_{15}, m_{16}, m_{17}, m_{18}, m_{19}, m_{20}	m_2, m_3, m_4, m_5, m_6, m_7, m_8, m_9, m_{10}, m_{11}, m_{12}, m_{14}, m_{15}, m_{16}, m_{17}, m_{18}, m_{20}
정크메일 수	19	17

표 7 전자메일 20개에 SVJ를 0.4로 부여한 결과

구분 / SVJ	정크메일번호	
	메일사용자	제안하는 기법
0.4	m_2, m_3, m_4, m_6, m_7, m_8, m_9, m_{10}, m_{11}, m_{12}, m_{13}, m_{14}, m_{15}, m_{16}, m_{17}, m_{18}, m_{19}, m_{20}	m_2, m_3, m_4, m_6, m_7, m_8, m_9, m_{10}, m_{11}, m_{12}, m_{14}, m_{15}, m_{16}, m_{17}, m_{18}, m_{20}
정크메일 수	18	15

표 8 전자메일 20개에 SVJ를 0.6로 부여한 결과

구분 / SVJ	정크메일번호	
	메일사용자	제안하는 기법
0.6	m_3, m_4, m_7, m_9, m_{10}, m_{11}, m_{12}, m_{14}, m_{15}, m_{16}, m_{17}, m_{18}, m_{19}, m_{20}	m_3, m_4, m_7, m_9, m_{10}, m_{11}, m_{12}, m_{14}, m_{15}, m_{17}, m_{18}, m_{20}
정크메일 수	14	12

표 6, 표 7, 표 8에서 보았을 때 제안하는 퍼지관계곱을 이용한 정크메일 분류 모델은 전자메일 사용자와 유사한 개수의 전자메일을 정크메일로 분류하며, 제안하는 시스템이 분류한 정크메일들은 모두 전자메일 사용자가 분류한 정크메일 리스트에 포함되어 있음을 알 수 있다.

7. 결론 및 향후과제

본 논문에서 제안하는 퍼지관계곱을 이용한 정크메일 분류 모델은 수신자의 입력에 의존하는 기존의 정크메일 분류기법의 한계를 극복하고 전자메일의 내용에 기반한 정크메일 분류 시스템의 기초 모델을 제시하였다. 본 모델은 클라이언트 기반 정크메일 분류보다는 서버 기반 정크메일 분류에 적합하다. 제안하는 모델은 동일한 전자메일들에 대해 사용자가 느끼는 정크도를 설문하여 비교 평가하였는데, 제안하는 모델의 결과와 전자메일 사용자들이 느끼는 정크도에 동일하거나 매우 근접한 결과로 추출하였다. 또한, 추출된 정크도를 기반으로 전자메일 사용자의 의견을 수렴하여 부여한 정크기준치를 반영하여 분류하였는데, 일정 양의 전자메일에

대해 전자메일 사용자와 유사한 개수로 유사한 전자메일을 정크메일로 분류하였다.

본 연구에 이어 향후에 이루어져야 할 과제는 다음과 같다. 첫 번째로 전자메일 서버로의 전자메일 수신에 관한 연구, 두 번째는 이미지 기반 텍스트 인식에 관한 연구가 이루어져야 하고, 세 번째는 사용자의 특성을 고려한 정크메일 분류 시스템에 관한 연구, 네 번째는 보다 많은 설문조사 결과를 기반으로 시스템의 신뢰성을 향상시키는 연구가 이루어져야 할 것이다.

참 고 문 헌

[1] Technology News, January 2000.
[2] 야마우치 요시유키, 니시다 도오루, 이메일 마케팅, 삼각형프레스, 2001.
[3] 정재윤, 이메일 마케팅.com, 비비컴, 2001.
[4] 이재규, 전자상거래원론, 법영사, 2000.
[5] Daum, http://www.daum.net.
[6] Yahoo, http://www.yahoo.co.kr.
[7] Hotmail, http://www.hotmail.com.
[8] Zero Junk Mail, http://www.zerojunkmail.com.
[9] 전인홍, 이광로, 퍼지이론과 응용, 교학사, 1992.
[10] 이광형, 오길록, 퍼지이론 및 응용, 홍릉출판사, 1997.
[11] 김도현, 권기호 공역, 핵심 퍼지시스템 이론 및 응용서, 에드텍, 1994.
[12] Kohout, L. J., Keravnou E. and Bandler W., Automatic Documentary Information Retrieval by means of Fuzzy Relational Products, In Gaines, B. R., Zadeh L. A. and Zimmermann, H. J., editors Fuzzy Sets in Decision Analysis, pages 308-404, North-Holland, Amsterdam, 1984.
[13] Kohout, L. K., Bandler, W., Fuzzy Relational Products as a Tool for Analysis and Synthesis of the Behaviour of Complex Natural and Artificial Systems, in: Wang S. K. and Chang P. P. eds., Fuzzy Sets: Theory and Application to Policy Analysis and Information Systems, Plenum Press, New York, 341-367, 1980.
[14] Bandler, W., and Kohout, J., Semantics of Implication operators and fuzzy relational products, Intl. Journal of Man-Machine Studies, 1980.
[15] Bandler, W., and Kohout L. J., Fuzzy Power Sets and Fuzzy Implication Operator, Fuzzy Set and System 4, 13-30, 1980.
[16] Kim, Yong-Gi and Kohout, L. J., Use of Fuzzy Relational Products and Algorithms for generating Control strategies in resolution based Automated Reasoning, Proceedings of the fourth International Fuzzy System Association (IFSA) world congress, (Brussels, Belgium), p109-p112, July 7-12, 1991.
[17] Ying Zou, Elicitation of the Groups and Group Cognitive Structures: An Application of Ternary Fuzzy Relational Products, The Florida State University, Feb. 5, 1997.
[18] 김창민, 김용기, 개선된 BK-퍼지정보검색모델(A-FIRM)과 BK-퍼지정보검색모델(BK-FIRM)의 성능평가, 한국 퍼지 및 지능시스템학회 추계학술발표논문집, 8(2), 1998.
[19] 김창민, A-FIRM: 개선된 BK-퍼지정보검색모델, 전자계산학과, 경상대학교, 1999.
[20] 이영일, 퍼지기법을 이용한 자율수중운동체의 휴리스틱 항행 탐색, 경상대학교, 2001.
[21] Kohout, L. J., and Harris, M., Computer Representation of Fuzzy and Crisp Relations by Means of Threaded Trees Using Foresets and Aftersets, Journal of Fuzzy Logic and Intelligent Systems, 3(1), 1993.
[22] Kim, Yong-Gi and Kohout, L. J., March 1-3, Comparison of Fuzzy Implication Operators by means of Weighting Strategy in on Applied Computing (SAC'92), Kansas City, 1992.
[23] Keravnou, E., June-July, System for Experimental Verification of Deviance of Fuzzy Connectives in Information Retrieval Application, Second World Conference on Mathematics at the Service of Man. Topic 7, Measuring "Deviance in Non-Classical Logics and Modelling, Las Palmas (Canary Islands), 1982.
[24] 박정선, 김창민, 김용기, 퍼지관계곱을 이용한 전자메일의 정크도 추출, 한국퍼지 및 지능 시스템 학회 춘계학술대회 학술발표논문집, 11(1) : 224-227, 2001.
[25] 박정선, 김창민, 김용기, 퍼지관계곱을 이용한 정크메일 분류 시스템, 한국퍼지 및 지능 시스템 학회 추계학술대회 학술발표논문집, 11(2), 2001.
[26] William B. Frakes, Ricardo Baeza-Yates, Information Retrieval : Data Structures & Algorithms, PRENTICE HALL, 1992.
[27] IETF(Internet Engineering Task Force), http://www.ietf.org/
[28] Dave Wood, Mark Stone, Programming Internet Email, Oreilly, 1999.
[29] Santon, G., and M. McGill, Introduction to Modern Information Retrieval, New York: McGraw-Hill, 1983.

김 용 기

1978년 서울대학교 공과대학(공학사). 1987년 University of Montana(전산학석사). 1992년 Florida State University(전산학박사). 1982년 ~ 1984년 KIST 시스템공학연구소 연구원. 1992년 ~ 현재 경상대학교 컴퓨터과학과 교수. 관심분야 인공지능, 지식기반시스템, 자율무인잠수정, 지능항해시스템, 퍼지정보검색시스템

김 창 민

1997년 경상대학교 컴퓨터과학과(이학사). 1999년 경상대학교 컴퓨터과학과(공학석사). 1999년 ~ 현재 경상대학교 컴퓨터과학과 박사과정. 관심분야는 퍼지정보검색, 퍼지문서분류기법, 지능제어아키텍처, 자율무인잠수정아키텍처, 충돌회피

박 정 선

1998년 경상대학교 컴퓨터과학과(이학사). 2002년 경상대학교 컴퓨터과학과(공학석사). 2002년~현재 연암공업대학 시간강사. 관심분야는 인공지능, 지식기반시스템, 퍼지정보검색기법, 지능항해시스템, 퍼지정크메일분류기법