퍼지관계곱을 이용한 정크메일 분류 시스템

A Junkmail Checking System Using Fuzzy Relational Products

박정선, 김창민, 김용기
경상대학교 컴퓨터과학과
Jeong-Seon Park, Chang-Min Kim, Yong-Gi Kim
Department of Computer Science, Gyeongsang National University
kong@ailab.gsnu.ac.kr, nuno@ailab.gsnu.ac.kr, ygkim@nongae.gsnu.ac.kr

요 약

20세기 후반 인터넷의 발전을 기반으로 전자메일은 현재의 대표적인 개인간 정보전달 수단으로 자리잡게 되었다. 그러나 전자메일 사용자들은 인터넷상에 개인 전자메일 주소가 노출되므로 해서 많은 정크메일(junkmail)을 수신하게되었는데, 정크메일이란 기업의 광고 선전물과 같이 수신을 원하지 않는 전자메일을 의미한다. 이러한 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 분류하는 시스템을 제안한다. 이는 퍼지관계곱 연산을 이용하여 미리 정의한 정크용어들과 사용자에게 수신되는 전자메일 내의 용어들간 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 기분으로 정크메일과 비정크메일로 분류한다. 제안된 기법은 사용자가 특정 개수의 동일한 전자메일에 대해 느끼는 정크도를 기준으로 분류한 정크메일 수를 비교하여 그 효용성을 증명하였다.

Key Words : 전자메일, 정크도, 정크메일, 퍼지관계곱

1. 소개

20세기 후반 크게 발전한 인터넷을 기반으로 전자메일(electronic mail)은 현재의 정보전달 수단 중 기업간의 정보교환 뿐만 아니라 개인간의 정보교환 기능을 제공함으로써 그 사용자가 급속도로 확산되었다[1]. 그로 인해 많은 기업들은 전자메일을 통한 개인별 전자 카탈로그를 보급하는 형태의 광고에 많은 투자를 하게 되었고 개인의 전자메일 주소의 확보에 크게 관심을 두었다. 많은 개인 전자메일 사용자들은 인터넷상의 다양한 서비스를 제공받고자 전자메일 주소를 여러 사이트에 등록하여 사용하는데, 이 때 사용자의 부주의나 무관심에 의해 사용자의 전자메일 주소가 인터넷상에서 공개되는 경우가 허다하게 발생한다. 이로 인해 전자메일 사용자들은 특정 기업이나 상품의 광고 선전물과 같이 자신이 원하지 않고 자신에게 불필요한 전자메일을 대량 받게 되었는데, 이를 정크메일(junkmail)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 증가하고 있는 추세이다. 꾸준히 증가하고 있는 전자메일 사용자 수를 감안해 본다면 더 많은 기업들이 전자메일을 통한 광고에 투자하게 될 것이고, 그에 따라 전자메일 사용자들은 현재 보다 더 많은 정크메일로 고통받게 될 것이다.

이러한 추세에 따라 최근 몇몇 전자메일 서비스 회사에서 정크메일 분류 기능을 제공하고 있는데, 현재까지는 메일 송신자의 전자메일 주소나 도메인 주소를 등록하여 차단하거나 제목에 특정 단어를 포함한 메일을 완전히 삭제하여 버리는 기술 수준에 머무르고 있다. 이러한 방법들은 구조가 단순해 시스템 설계 및 구현이 간단하다는 장점을 가지고 있으나 수신자가 입력에만 의존하므로 정크메일의 증가량에 따라 수신자의 부담이 비례하여 커지게 되는 단점을 가진다. 이러한 현재 기술 수준의 한계를 극복하기 위해서 수신된 전자메일을 그 내용에 기반하여 의미적으로

접근하여 정크메일과 비정크메일을 분류해주는 기술에 관한 연구의 필요성이 크게 대두되었다.

본 논문에서는 퍼지관계곱 연산을 이용하여 미리 정의한 정크용어들과 전자메일의 내용에 기반한 의미적 포함관계를 분석하여 수신된 전자메일의 정크도를 추출하고 정크메일 분류시스템에 대한 사용자의 정크 기준치를 부여받아 정크메일과 비정크메일을 분류하는 시스템을 제안한다.

2. 퍼지관계곱

Bandler와 Kohout는 이진 관계곱 연산을 확장하여 퍼지관계곱(fuzzy relational product) 연산을 제안하였다. 이는 퍼지집합 A, B, C와 그들 간의 퍼지관계 $\tilde{R}:A\times B$과 $\tilde{S}:B\times C$가 주어지고 $a_i\in A$, $c_k\in C$라 할 때, $\tilde{R}$과 $\tilde{S}$의 퍼지관계곱 $(\tilde{R}@\tilde{S})_{ik}$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 $\triangleleft$, $\triangleright$ 또는 $\square$로서 표현될 수 있다[2][3][4][5].

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

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

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

수식 (1)의 $\triangleleft$ 연산자는 퍼지삼각서브논리곱(fuzzy triangle sub-product)이라고 하고 이는 a_i가 c_k에 포함되는 정도를 의미한다. 수식 (2)의 $\triangleright$ 연산자는 퍼지삼각수퍼논리곱(fuzzy triangle super product)이라고 하고 이는 a_i가 c_k를 포함하는 정도를 의미한다. 수식 (3)의 $\square$ 연산자는 퍼지사각논리곱이라고 하고 이는 a_i와 c_k가 유사한 정도를 의미한다[2][3][4][5].

퍼지관계곱은 퍼지조건연산자(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 (10)은 대표적인 퍼지조건연산자를 보여주고 있다[5][6][7].

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

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

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

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

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

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

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

3. 퍼지관계곱을 이용한 정크메일 분류 시스템

본 논문에서 제안하는 정크메일 분류 시스템은 퍼지관계곱을 기반으로 전자메일의 내용에 의미적으로 접근하여 전자메일의 정크도를 추출하고, 추출된 정크도에 개인사용자의 의견을 수렴하여 정크메일과 비정크메일로 분류한다. 이는 정크용어베이스의 구축, 수신된 전자메일에 대한 정규화, 퍼지화, 정크도 추출, 사용자의 정크기준치 부여, 정크메일 분류의 순서로 이루어지며 그림 1은 그 구성도를 보이고 있다.

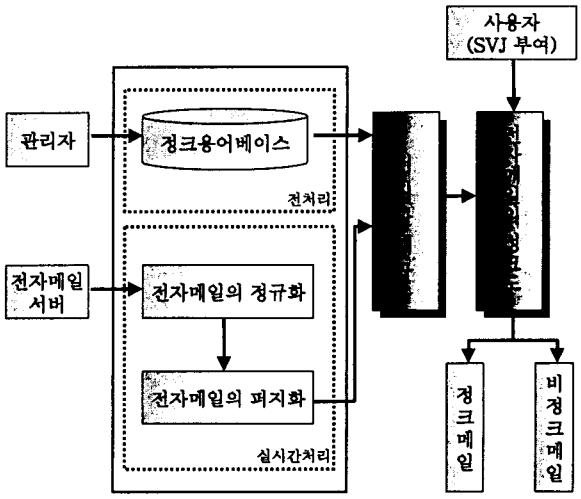

그림 1. 퍼지관계곱을 이용한 정크메일 분류 시스템 구성도

(1) 정크용어베이스

본 논문에서 제안하는 기법을 위해 일반적인 지식을 바탕으로 현재 수신된 다양한 전자메일 내에 포함되어 있는 용어들 중 설문을 통해 정크성을 가지는 용어들을 추출하여 정크용어집합을 정의하고, 각 정크용어들에 대해 정크성을 [0,1] 의 퍼지 값으로 부여하여 정크용어베이스를 구축한다.

(2) 전자메일의 정규화

수신된 순수 메일(raw mail)을 "송신자", "제목", "내용" 등으로 구성요소를 분류하고, 전자메일 내에 포함된 메일 내용과 관계없는 HTML 태그 등을 제거하여 정규메일(regulated mail)로 변환시킨다. 본 논문에서는 순수메일을 정규메일로 변환하는 작업에 대해서는 다루지 않고 전자메일이 이미 정규화 되었

다고 가정하며, 이후로 언급되는 '전자메일'은 정규메일을 의미한다.

(3) 내용기반 전자메일 퍼지화

수신된 전자메일의 정크도를 추출하기 위해서는 메일 내용의 의미를 이해하여야 하는데, 이는 정보검색 모델에서 문서의 의미를 파악하는 방법 중 문서 내에 포함되어 있는 특정 용어의 빈도수를 이용하는 방법을 적용하여 해결할 수 있다. 용어의 빈도수가 해당 문서의 의미를 모두 반영한다고는 할 수 없지만, 현 기술수준에서는 가장 적절한 방법이다[8].

정보검색 모델 중 Bandler와 Kohout가 제안한 BK-퍼지정보검색모델(Bandler -Kohout Fuzzy Information Retrieval Model)에서는 특정 문서의 의미를 파악하기 위해서 문서 내에 등장하는 용어의 빈도수로서 문서의 상대적 관련성을 추출하고 이를 퍼지화하였다. 또한 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[9]. 그림 2는 A-FIRM에서 이용되는 용어 빈도 퍼지화 멤버쉽 함수와 그래프를 보여주고 있다.

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

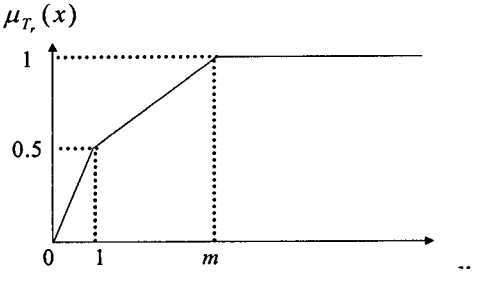

그림 2. A-FIRM의 용어 빈도 퍼지화 멤버쉽 함수

(4) 퍼지관계곱을 이용한 전자메일의 정크도 추출

본 논문에서 제안하는 전자메일의 정크도 추출 기법은 다음과 같다[10]. 수식 11과 같이 수신된 전자메일 집합을 M, 수식 12와 같이 정크용어집합을 T라 두었을 때, 전자메일 m_i와 정크용어 집합 T의 빈도수를 이용한 퍼지관계는 수식 13과 같이 $\tilde{S}$로 표현된다. 그리고 수식 14와 같이 표현되는 퍼지집합 $\tilde{J}$로부터 수식 15와 같이 정크용어 t_i가 가지는 정크성을 추출하고, 수식 16과 같이 정크용어 t_i와 정크성의 퍼지관계 $\tilde{R}$로 표현한다. 최종적으로 수식 17과 같이 두 퍼지관계 $\tilde{S}$와 $\tilde{R}$을 퍼지삼각서브논리곱연산 적용하며, 이를 처리하기 위해 퍼지조건연산자 4번을 적용하여 전자메일 m_i에 대한 정크용어 t_i의 의미적 포함관계 r_i를 추출한다.

$$M=\{m_1,m_2,\cdots,m_n\} \tag{11}$$

$$T=\{t_1,t_2,\cdots,t_m\} \tag{12}$$

$$\tilde{S}_{m_i\to T}=[d_1,d_2\ \cdots\ d_m] \tag{13}$$

$$\tilde{J}=\{fv_1/t_1,fv_2/t_2,\ \ldots\ldots,fv_m/t_m\} \tag{14}$$

$$j_i=\mu_{\tilde{J}}(t_i),\ t_i\in T \tag{15}$$

$$\tilde{R}_{T\to\tilde{J}}=\begin{bmatrix}j_1\\ j_2\\ \vdots\\ j_m\end{bmatrix} \tag{16}$$

$$\begin{aligned}\tilde{J}\to m_i &= {}_{m_i}\tilde{S}\triangleleft\tilde{R}_{\tilde{J}}\\ &=[d_1,d_2\cdots d_m]\triangleleft\begin{bmatrix}j_1\\ j_2\\ \vdots\\ j_m\end{bmatrix}\\ &=r_i\end{aligned} \tag{17}$$

수식 17에 나타난 퍼지관계곱 연산을 이용한 결과 r_i는 메일 m_i에 대한 정크용어퍼지집합 $\tilde{J}$의 의미적 포함정도를 의미한다. 그러나 위와 같은 방법은 메일 m_i와 정크성 간 의미적 포함거리만을 추출하므로 각 정크용어가 정크성에 기여한 참여비중을 고려하지는 못한다. 따라서, 본 논문에서는 정크용어의 비중을 결과에 반영하기 위하여 퍼지관계곱 연산에 이용되는 퍼지조건연산자를 새로이 제안한다.

수식 18은 본 논문에서 제안하는 정크도 추출을 위한 퍼지조건연산자로 정크용어의 비중을 고려하기 위하여 기존의 퍼지조건연산자에 용어의 정크도 j_i를 곱한다. 수식 19는 실제 본 연구에서 사용된 퍼지조건연산자이다. 본 연구에서는 결과의 표준화를 위하여, 수식 20과 같이 수식 17의 r_i에 상수 c를 곱하여 최종결과 f_i를 추출한다.

$$a\overset{j}{\to}b\ =\ a\times(a\to b),\quad a\neq 0 \tag{18}$$

$$\begin{aligned}a\overset{j_4}{\to}b\ &=\ a\times(a\overset{4}{\to}b)\\ &=\ a\times\min\left(1,\frac{b}{a}\right)\quad a\neq 0\end{aligned} \tag{19}$$

$$f_i\ =\ \min(1,r_i\times c) \tag{20}$$

(5) 정크메일 분류 기준치 적용

본 논문에서 제안하는 정크메일 분류 시스템에서는 개인 사용자마다의 의견을 수렴하기 위해 사용자가 정크기준치(SVJ, Standard Value of Junk)를 입력하며 이를 기준으로 수신된 전자메일에 대해 최종적으로 정크메일 또는 비정크메일로 분류하도록 설계하였다.

4. 비교 및 평가

본 논문에서 제안하는 퍼지관계곱을 이용한 정크메일 분류 시스템은 동일한 전자메일들에 대해 사용자가 느끼는 정크도와 정크메일 분류 개수를 기준으로 비교·평가한다. 이를 위하여 70개의 정크용어로 정크용어베이스를 구축하였으며, 다양한 분야에 종사하는 50명의 사용자를 대상으로 설문하여 적용하였다.

표 1은 전자메일에 대해 사용자가 정의한 정크도와 제안하는 기법을 통하여 추출한 정크도를 보이며, 표2는 20개의 전자메일에 대해 추출된 정크도를 기반으로 각각 다른 정크기준치를 부여하였을 때의 정크메일 분류 정도를 수치로 표현하였다.

	정크도	
	메일사용자	제안하는 기법
메일 #1	0	0
메일 #2	0.4	0.4
메일 #3	0.9	1
메일 #4	1	1
메일 #5	0.3	0

표 1. 전자메일에 대한 정크도

구분 / 정크기준치	정크메일 수	
	메일사용자	제안하는 기법
0.3	19	17
0.4	18	15
0.6	14	12

표 2. 전자메일 20개에 대한 정크메일 수

5. 결론 및 향후과제

본 논문에서 제안하는 퍼지관계곱을 이용한 정크메일 분류 시스템은 동일한 전자메일에 대해 사용자가 느끼는 정크도에 매우 근접한 정크도를 추출하였으며, 추출된 정크기준치를 기반으로 분류하였을 시 전자메일 사용자와 유사하게 정크메일과 비정크메일을 분류하였다. 그리고 수신자의 정보 입력에 의존하는 기존의 정크메일 분류기법의 한계를 극복하고, 전자메일의 내용에 기반한 정크메일 분류 기법의 기초모델을 제시하였다.

본 연구에 이어 향후에 이루어져야 할 과제는 다음과 같다. 첫 번째로 입력되는 순수메일의 정규화에 관한 연구가 이루어져야 하고, 두 번째는 사용자의 특성을 고려한 정크메일 분류 시스템에 관한 연구, 세 번째는 보다 많은 설문조사 결과를 기반으로 시스템의 신뢰성을 향상시키는 연구가 이루어져야 할 것이다.

6. 참 고 문 헌

1. Technology News, January 2000.
2. 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, vol. 3, no.1, 1993
3. 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
4. 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
5. Bandler, W., and Kohout, J. "Semantics of Implication operators and fuzzy relational products," Intl. Journal of Man-Machine Studies, 1980
6. Kim, Yong-Gi and Kohout, L. J., "Comparison of Fuzzy Implication Operators by means of Weighting Strategy in on Applied Computing (SAC'92)," Kansas City, March 1-3, 1992
7. Keravnou, E., "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), June-July, 1982.
8. 김창민, 김용기, "개선된 BK-퍼지정보검색모델(A-FIRM)과 BK-퍼지정보검색모델(BK-FIRM)의 성능평가", 한국 퍼지 및 지능시스템학회 추계학술발표논문집, 제8권 제2호,1998.
9. 김창민, A-FIRM: 개선된 BK-퍼지정보검색모델, 전자계산학과, 경상대학교, 1999.
10. 박정선, 김창민, 김용기, "퍼지관계곱을 이용한 전자메일의 정크도 추출", 한국 퍼지 및 지능시스템 학회 춘계학술발표논문집, 제11권 제1호, 2001.