

퍼지관계곱을 이용한 전자메일의 정크도 추출

Extracting the Degree of Junk from E-mail 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

전자메일은 20세기 후반 인터넷의 발전으로 현재의 정보전달 수단 중 대표적인 개인간 인터넷 통신 수단으로 자리잡게 되었다. 그러나 전자메일 사용자들은 전자메일 주소가 노출되므로 해서 많은 정크메일(junk mail) 즉, 자신이 원하지 않는 전자메일을 수신하게되었다. 이로 인해 일반 전자메일과 정크메일을 분류하기 위한 수단이 필요하게 되었는데, 현재까지는 사용자가 입력한 송신자의 전자메일 주소 또는 도메인 주소를 등록하여 차단하거나 제목에 특정 단어를 포함한 메일을 완전히 삭제하여 버리는 수준에 머무르고 있다. 본 논문에서는 의미적 접근 기반 정크메일 분류 기법의 기초 모델을 제안한다. 퍼지관계곱을 이용한 전자메일의 정크도 추출은 퍼지관계곱 연산을 이용하여 미리 정의한 정크용어들과 사용자의 수신함에 있는 전자메일 내의 용어들 간 의미적 포함관계를 분석하고 그를 통해 전자메일의 정크도(degree of junk)를 추출하는 연구를 제안한다. 제안된 기법을 통해 추출한 정크도는 동일한 전자메일들에 대해 사용자가 느끼는 정크도와 비교하여 효용성을 증명하였다.

Keyword : 정크도, 정크메일, 퍼지관계곱

I. 서론

인터넷을 기반으로 전자메일(electronic mail)은 현재의 정보전달 수단 중 대표적인 개인간 인터넷 통신 수단으로 자리잡게 되었다[1]. 많은 전자메일 사용자들은 인터넷상의 다양한 서비스를 제공받고자 전자메일 주소를 여러 사이트에 등록하여 사용하는데, 이 때 사용자의 부주의나 무관심에 의해 사용자의 전자메일 주소가 인터넷상에서 공개되는 경우가 허다하다. 이로 인해 전자메일 사용자들은 특정 기업이나 상품의 광고 선전물과 같이 자신이 원하지 않는 불필요한 전자메일을 받게 되었는데, 이를

정크메일(junk mail)이라 한다.

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 그 기술이 메일 송신자의 전자메일 주소나 도메인 주소를 등록하여 차단하거나 제목에 특정 단어를 포함한 메일을 완전히 삭제하여 버리는 수준에 머무르고 있다. 이러한 방법들은 구조가 단순해 시스템 설계가 매우 간단하다는 장점을 가지고 있으나 수신자가 입력하는 내용에만 의존하므로 정크메일의 증가량에 따른 수신자의 부담이 커지게 된다. 이러한 현재 기술 수준의 한계를 극복하기 위해서 수신된 전자메일을 그 내용에 기반하여 의미적으로 접근하여 일반 전자메일과 정크메일을 분류해주는 기술에 관한 연구의 필요성이 크게 대두되었다.

본 논문에서는 퍼지관계곱 연산을 이용하여 미리 정의한 정크용어들과 전자메일의 내용에 기반한 의미적 포함관계를 분석하여 수신된 전자메일의 정크도를 추출하는 연구를 제안한다.

본 논문의 구성을 살펴보면 본론의 제 1절에서는 본 연구의 기반이 되는 퍼지관계곱 이론에 대해 살펴보고 제 2절에서는 본 연구에 앞서 이루어져야 할 전처리 작업에 대해 살펴본다. 제 3절에서는 전자메일의 내용기반 퍼지화에 대해 다루고 제 4절에서는 본 논문에서 제안하는 퍼지관계곱을 이용한 전자메일의 정크도 추출에 대해 살펴본다. 제 5절에서는 제안하는 기법을 수신된 전자메일에 대해 인간이 느끼는 정크도와 비교하여 그 성능을 평가하였고, 마지막으로 결론에 대해 살펴본다.

II. 본론

1. 퍼지관계곱

Bandler와 Kohout는 이진 관계곱 연산을 확장하여 퍼지관계곱(fuzzy relational product) 연산을 제안하였다. 집합 A, B, C 와 퍼지관계 $\tilde{R}: A \times B$ 와 $\tilde{S}: B \times C$ 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 A 의 원소 a_i 와 C 의 원소 c_k 의 의미상 포함관계를 나타내는 것으로서 수식 1, 2, 3과 같이 세 가지 퍼지관계곱 연산 $\triangleleft, \triangleright$ 또는 \square 로 표현될 수 있다.

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

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

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

수식 1의 \triangleleft 연산자는 퍼지삼각서브논리곱(fuzzy triangle sub-product)이라고 하고 수식 1은 a_i 가 c_k 에 포함되는 정도를 의미한다. 수식 2의 \triangleright 연산자는 퍼지삼각수퍼논리곱(fuzzy triangle super product)이라고 하고 수식 2는 a_i 가 c_k 를 포함하는 정도를 의미한다. 수식 3의 \square 연산자는 퍼지사각논리곱이라고 하고 수식 3은 a_i 와 c_k 가 유사한 정도를 의미한다[2][3][4][5].

퍼지관계곱은 퍼지 조건 연산자(fuzzy implication operator)를 이용하여 적절히 처리되는데 퍼지 조건 연산자는 이진 조건 연산과 달리 다양한 방법으로 구현 가능하며 현재 수십여 가지가 제안되어 있다. 수식 4 ~ 수식 7은 대표적인 퍼지 조건 연산자를 보여주고 있다[5][6][7].

$$a \rightarrow_1 b = \begin{cases} 1 & \text{iff } a \neq 1 \text{ or } b = 1 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$a \rightarrow_2 b = \begin{cases} 1 & \text{iff } a \leq b \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

$$a \rightarrow_4 b = \min\left(1, \frac{b}{a}\right) \quad (6)$$

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

2. 정크용어베이스의 구축 및 전처리 단계

본 논문에서 제안하는 기법은 두 가지의 전처리 작업을 요구한다. 첫 번째는 일반적인 지식을 바탕으로 현재 수신된 다양한 전자메일 내에 포함되어 있는 용어들 중 정크성을 가지는 용어들을 추출하여 정크용어집합을 정의하고, 각 정크용어들에 대해 정크성을 $[0, 1]$ 의 퍼지 값으로 부여하여 정크용어베이스를 구축하는 작업이다.

두 번째는 수신된 순수 메일(raw mail)을 “송신자”, “제목”, “내용” 등으로 구성요소를 분류하여 정규메일(regulated mail)로 변환시키는 작업이다. 본 논문에서는 순수메일을 정규메일로 변환하는 작업에 대해서는 다루지 않고 정크메일이 이미 정규화 되었다고 가정하며, 이후로 언급되는 ‘정크메일’ 또는 ‘메일’은 정규메일을 의미한다.

3. 내용기반 메일 퍼지화

수신된 전자메일의 정크도를 추출하기 위해

서는 먼저 메일 내용의 의미를 이해하여야 하는데, 이는 정보검색모델에서 문서의 의미를 파악하는 방법을 이용하여 해결할 수 있다. 문서의 의미를 파악하는 방법 중 대표적인 것은 문서 내에 포함되어 있는 특정 용어의 빈도수를 이용하는 방법이다. 용어의 빈도수가 해당 문서의 의미를 모두 반영한다고는 할 수 없지만, 현 기술수준에서는 가장 적절한 방법이다 [8].

여러 가지 정보검색 모델 중 Bandler와 Kohout가 제안한 BK-퍼지정보검색모델에서는 특정 문서의 의미를 파악하기 위해서 문서 내에 등장하는 용어의 빈도수로서 문서의 상대적 관련성을 추출하고 이를 퍼지화할 수 있게 했다. 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 1은 개선된 BK-퍼지정보검색모델에서 이용되는 용어 빈도 퍼지화 멤버쉽 함수와 그래프를 보여주고 있다.

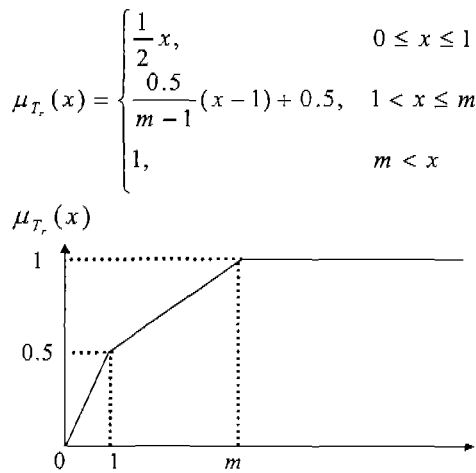


그림 1. A-FIRM의 용어 빈도 퍼지화 멤버쉽 함수

4. 퍼지관계급을 이용한 전자메일의 정크도

본 논문에서 제안하는 전자메일의 정크도 추출 기법은 다음과 같다. 수식 8과 같이 수신된 전자메일 집합을 M , 수식 9와 같이 정크용어 집합을 T 라 두었을 때, 전자메일 m_i 와 정크용어 집합 T 의 빈도수를 이용한 퍼지관계는 수식 10과 같이 \tilde{S} 로 표현된다. 그리고 수식 11과 같

이 퍼지집합 \tilde{J} 로부터 정크용어 t_i 가 가지는 정크성을 추출하고, 수식 12와 같이 정크용어 t_i 와 정크성의 퍼지관계 \tilde{S} 로 표현한다. 수식 13과 같이 두 퍼지관계 \tilde{S} 와 \tilde{R} 을 퍼지삼각서브논리곱연산에 적용하여 전자메일 m_i 에 대한 정크용어 t_i 의 의미적 포함관계 r_i 를 추출한다.

$$M = \{m_1, m_2, \dots, m_n\} \quad (8)$$

$$T = \{t_1, t_2, \dots, t_m\} \quad (9)$$

$$\tilde{S}_{m_i \rightarrow T} = [d_1, d_2 \dots d_m] \quad (10)$$

$$j_i = \mu_{T_j}(t_i), t_i \in T \quad (11)$$

$$\tilde{R}_{T \rightarrow J} = \begin{bmatrix} j_1 \\ j_2 \\ \vdots \\ j_m \end{bmatrix} \quad (12)$$

$$\begin{aligned} \tilde{J} \rightarrow m_i &= m_i \tilde{S} \triangleleft \tilde{R}_J \\ &= [d_1, d_2 \dots d_m] \triangleleft \begin{bmatrix} j_1 \\ j_2 \\ \vdots \\ j_m \end{bmatrix} \\ &= r_i \end{aligned} \quad (13)$$

수식 13과 같이 퍼지관계급 연산을 이용한 결과 r_i 는 메일 m_i 에 대한 정크용어퍼지집합 \tilde{J} 의 의미적 포함정도를 의미한다. 그러나, 위와 같은 방법은 메일 m_i 과 정크성 간 의미적 포함거리만을 추출하므로 각 정크용어가 정크성에 기여한 참여비중을 고려하지는 못한다. 따라서, 본 논문에서는 정크용어의 비중을 결과에 반영하기 위하여 퍼지관계급 연산에 이용되는 퍼지조건연산자를 새로이 제안한다.

수식 14는 본 논문에서 제안하는 정크도 추출을 위한 퍼지조건연산자이다. 정크용어의 비중을 고려하기 위하여 기존의 조건연산자에 용어의 정크도 j_i 를 곱한다. 수식 15는 실제 본 연구에서 사용된 퍼지조건연산자이다. 본 연구에서는 결과의 표준화를 위하여, 수식 16과 같이 수식 13의 r_i 에 상수 c 를 곱하여 최종결과 f_i 를 추출한다.

$$a \rightarrow b = a \times (a \rightarrow b) \quad (14)$$

$$\begin{aligned} a \rightarrow b &= a \times (a \overset{j_i}{\rightarrow} b) \\ &= a \times \min\left(1, \frac{b}{a}\right) \end{aligned} \quad (15)$$

$$f_i = r_i \times c \quad (16)$$

5. 평가

본 논문에서는 퍼지관계음을 이용한 전자메일의 정크도 추출 기법을 평가하기 위하여 70개의 정크용어, 다양한 분야에 종사하는 20명의 사용자를 대상으로 하여 적용한다. 제안된 기법은 채택된 전자메일에 대한 사용자의 반응과의 유사도로써 비교·평가 한다. 표 1은 전자메일에 대해 사용자가 정의한 정크도와 제안하는 기법을 통하여 추출한 정크도를 비교하여 보여준다.

	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
메일 #6	0.5	0.5
메일 #7	0.9	0.9
메일 #8	0.4	0.5
메일 #9	0.9	1
메일 #10	0.9	1

표 1. 전자메일에 대한 정크도

III. 결론

본 논문에서 제안하는 퍼지관계음을 이용한 전자메일의 정크도 추출 기법은 동일한 전자메일에 대해 사용자가 느끼는 정크도에 매우 근접한 결과를 추출하였다. 본 연구는 수신자의 정보 입력에 의존하는 기존의 정크메일 분류기법의 한계를 극복하고, 전자메일의 내용에 기반한 정크메일을 분류하는 방법의 기초모델을 제시한다.

본 연구에 이어 향후에 이루어져야 할 과제는 다음과 같다. 첫 번째로 입력되는 순수메일의 정규화에 관한 연구가 이루어져야 하고, 두 번째는 본 연구를 기반으로 하는 퍼지관계음을 이용한 내용 기반 의미적 접근 정크메일 분류모델의 설계 및 구현에 관한 연구가 이루어져야 한다.

IV. 참고 문헌

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.