제21회 한국정보처리학회 춘계학술발표대회 논문집 제11권 제1호 (2004. 5)

문서요약 및 동적 분류체계를 사용한 E-mail 분류의 재구성

안찬민*, 박선*, 김태순*, 최범기**, 이주홍*
*인하대학교 컴퓨터정보공학과
**㈜쿼크
e-mail : {ahnch1, sunpark, kts429, bgchoi}@datamining.inha.ac.kr,
juhong@inha.ac.kr

Reconstruction of E-mail Category Using Dynamic Category Hierarchy and Document Summarization

Chan-Min Ahn*, Sun Park*, Tae-Soon Kim*, Bum-ghi Choi**, Ju-Hong Lee*
*Dept. of Computer Science, Inha University
** Quark

요 약

현재의 이메일 분류는 규칙기반, 베이시안, SVM 등을 이용하여 스팸메일을 필터링 하는 이원분류가 주로 연구되어지고 있고, 이외에도 다원분류에 대한 연구로는 클러스터링을 이용한 방법이 있다. 그러나 클러스터링에 의한 방법은 단순히 유사도에 의해 메일을 묶는 수준에 그치고 있다. 본 논문에서는 자동 문서요약 방법과 동적분류체계 방법을 결합하여 새로운 이메일 자동 다원분류 방법을 제안했다. 본 논문에서 제안한 방법은 이메일을 자동으로 분류하며 분류한 결과를 검색할 때 사용자의 요구사항을 만족하지 못하면 재분류하여 분류 및 검색의 정확성을 높였다.

1. 서론

이메일은 일반 사용자 뿐 아니라 전자상거래, 광고, 사업등에 사용되고 있다. 기업 및 일반사용자가 받는 메일의 하루량은 수십에서 수천통에 이른다. 이러한 수신 메일량의 증가함에 따라 사용자는 메일을 좀더 효율적으로 관리할 수 있는 방법을 요구하고 있다.

현재 이러한 문제를 해결하기 위해 필터(filters)나 메시지를 분류할 수 있는 많은 도구들이 개발되었으나 대부분의 도구들은 사용자가 직접 필터링 규칙이나 메시지가 분류될 수 있도록 색인어 목록을 작성해야 한다. 이렇게 사용자가 직접 설정해야 하는 도구들은 내용이 중복되거나 색인어를 많이 포함하는 대량의 메일인 경우 부적절하다. 또한 시간이 지나 사용자의 변화되는 요구사항에 맞추어 재분류하거나 재필터링 할 수 없는 단점이 있다[1].

이를 위해 본 논문에서는 자동 문서요약 방법의 PCA(Principal Component Analysis)과 SVD(SingularValueDecomposition)를 기반으로 이메일을 디렉토리로 자동 분류하는 방법과 동적분류체계 방법을 이용하여 디렉토리 분류(directory categorization) 결과를 재분류 할 수 있도록 하였다. 본 논문에서는 자동 디렉토리 분류방법을 지원하여 사용자의 수동개입을 줄였다. 또한 이메일을 재분류할 수 있도록 하여 사용자의 주관에 맞추어 메일을 분류 할 수 있다.

2. 관련 연구

지금까지 이메일 분류는 대부분 스팸메일을 찾는 이원분류가 주로 연구되었다. 스팸 분류를 위해 사용된 방법으로는 규칙기반 분류(rule-based classifiers), 베이시안 분류(bayesian classifiers), SVM(Support-Vector Machines)등이 있다. Cohen[2]은 일치하는 이메일을 분류하기 하기 위해 텍스트 마이닝 기법을 이용한 두개의 규칙기반 시스템을 기술하였다. 두 방

법은 전처리 단계에 단순 불리안이나 빈도기반의 벡터모델을 사용하는가에 주로 차이가 있다. Androutsopoulos[3]와 Sakkis[4]은 안티스팸 필터링을 하기 위해 베이시안 분류자를 이용하였다. 그들의 접근방법은 규칙기반 분류자를 사용하는 것에 비해 좀더 좋은 정확성을 보였다. Drucker[5]는 SVM을 이용하여 이메일의 스팸과 스팸이 아닌 것을 분류하는데 Ripper, Rocchio, boosting decision trees 등의 세 알고리즘과 비교하여 SVM 이 이메일 분류시 가장 좋은 성능임을 보였다. Kunlun[6]는 스팸을 분류하기 위해 활성 학습 정책을 이용하는 SVM 기반의 새로운 방법을 제안하였다. Woitaszek[7]는 simple SVM을 이용하여 상업적 e-mail을 구분하였다.

다원분류 연구로는 비관리자 분류(unsupervised classification) 기법으로 수신된 메시지 집합으로부터 메시지 폴더를 자동으로 구성하여 이메일을 분류한다. Mock[8]는 tf-idf 값으로 역색인방법에 기반한 이메일 자동분류시스템을 제안하였다. Manco 와 Masciari[1]는 이메일 메시지를 관리 및 유지하기 위하여 데이터마이닝 알고리즘에 기반을 두고 수신 메시지를 분류하였다. 수신메시지로부터 정보를 추출하여 유사도를 구하고, 군집 기술(clustering techniques)로 각 메시지를 구룹으로 묶고(grouping), 폴더를 생성하여 이메일을 분류한다. 이 방법은 여러 단계의 전처리와 다양한 추출 정보에 의하여 유사도를 얻기 때문에 계산이 복잡 하여 분류하는데 시간이 오래 걸린다.

3. 문서 요약

문서요약에 대한 기존연구는 크게 통계적 방법, 문맥 구조기반 방법, 지식기반 방법이 있다. 통계적 방법은 단어의 출현 빈도, 제목, 문장의 길이, 중요한 단어나 구 등을 특질로 각 문장 및 문단의 중요도를 계산하여 요약하는 방법이다[10]. 문맥 구조기반 방법은 문장들 사이의 문맥관계를 계산하여 문서를 요약한다[11]. 지식기반 방법은 요약을 원하는 문서에 관련된 지식들을 이용하여 요약하는 방법이다[12].

본 논문에서 이용하는 문서요약 방법은 통계적 방법인 PCA(Principal Component Analysis)와 SVD(Singular Value Decomposition)를 이용하는 문서 요약 방법을 사용한다[13].

4. 동적 분류체계 방법

동적분류체계 방법에서 사용되는 퍼지 이론은 다음과 같다[8].

(정의 1) 퍼지 함의 연산자 (Fuzzy Implication Operator) 는 크리스프 함의 연산자 (Crisp Implication Operator) 를 확장하여 퍼지에 적용한 것으로서, 크리스프 함의 연산자는 $\{0,1\}\times\{0,1\} \rightarrow \{0,1\}$ 로 정의되는데 반해, 퍼지 함의 연산자는 $[0,1]\times[0,1] \rightarrow [0,1]$ 로서 단위 구간의 다치 논리로 확장된 것이다. 퍼지 함의 연산자의 종류는 무수히 많으며 대표적인 Kleene-Diense 퍼지함의 연산자의 예는 다음과 같다[9].

$$a \rightarrow b = (1 - a) \vee b = \max(1 - a, b), \quad a = 0 \sim 1,\ b = 0 \sim 1 \qquad (1)$$

본 논문에서는 위의 식(1)의 Kleen-Diense 퍼지함의 연산자를 사용한다. 퍼지 함의 연산자를 식(2)의 퍼지관계곱을 적용하여 분류들 간의 퍼지함의관계, $C_i \rightarrow C_j$ 를 유도할 수 있다.

$$\pi_{m,\beta}(C_i \subseteq C_j) = \left(R^T \triangleleft_\beta R\right)_{ij} = \frac{1}{|C_{i\beta}|}\sum_{K_k \in C_{i\beta}}\left(R^T_{ik} \rightarrow R_{kj}\right) \qquad (2)$$

여기서, K_k 는 k번째 검색어이고, C_i , C_j 는 i번째와 j 번째 분류이며, $C_{i,\beta}$는 C_i 의 β-제약, $\{x|\mu_{Cj}(x) \geq \beta\}$이고 $|C_{i,\beta}|$ 는 $C_{i,\beta}$ 의 원소의 갯수이다. R는 $m \times n$ 행렬로서 R_{ij} 는 $\mu_{Cj}(K_i)$,즉, $K_i \in C_j$ 인 정도이다. R^T 는 행렬 R 의 전치 행렬로서 $R_{ij} = R^T_{ji}$ 이다.

5. 자동 이메일 분류

5.1 문서 요약에 의한 자동 이메일 디렉토리 분류

본 절에서는 3 장의 문서요약 방법을 이용하여 디렉토리 분류를 한다. 분류방법은 이메일을 제목에 가중치를 부여하여 메일의 내용과 묶어 하나의 이메일 문서로 만든다. 이 메일문서로부터 명사들을 추출하여 표 1 과 같은 메일문서와 명사 행렬을 만든다.

표 1. 메일문서-명사 행렬의 예

구분	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	…
E1	1	0	1	1	0	0	1	0	1	0	…
E2	0	0	3	0	0	1	1	0	2	0	…
E3	0	1	0	0	0	4	0	1	0	0	…
E4	0	1	0	0	3	0	0	1	0	4	…
E5	0	3	1	0	0	5	0	0	3	0	…
E6	1	2	2	1	0	1	0	1	0	1	…
…	…	…	…	…	…	…	…	…	…	…	…

N:명사, E:메일문서

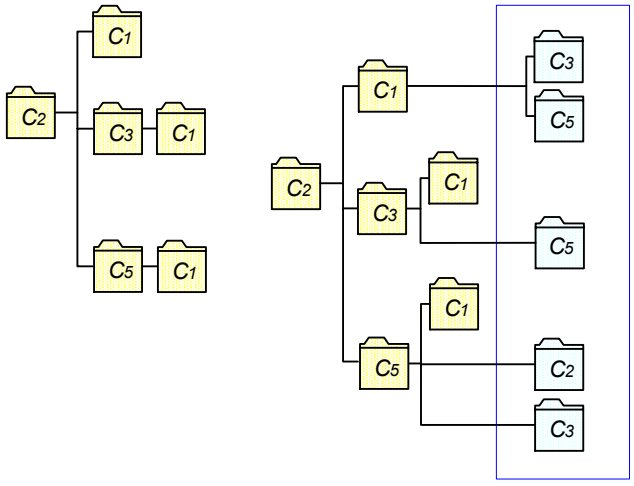
위의 행렬에서 PCA 를 이용하여 주제어를 추출하고 SVD 를 이용하여 중요 메일문서를 추출한다. 다음으로 주제어를 이름으로 하는 디렉토리를 생성하고, 주제어와의 중요도에 따라 메일문서를 각 디렉토리로 분류한다. 그러나, 이메일의 제목이 아무런 의미도 갖지 못함은 물론 이메일의 의도 조차도 내포하지 못한다면 PCA 와 SVD 을 사용한 방법은 불필요하거나 메일을 잘못 분류할 수 있다. 또한 메일의 내용이 제목과 유사한 내용이라도 중요한 의미를 담고 있는 특질을 포함하고 있지 않다면 중요한 문장이 될 수 없으며, 반대로 제목과 유사성이 없는 내용이라도 중요한 의미를 포함한 특질등이 나타나는 내용이라면 중요하

게 고려해야 한다.

본 논문에서는 이러한 문제를 해결하기 위해 자동 문서요약 방법을 이용하여 디렉토리 별로 분류한 결과를 동적분류체계 방법을 이용하여 동적으로 재분류할 수 있게 하였다.

5.2 동적분류체계 방법에 의한 이메일 재구성

본 논문에서는 이메일을 동적분류체계로 구성하기 위해 색인어와 분류주제 간의 관계를 규정해야 한다. 그러나 색인어와 분류주제 간의 관계를 직접 결정할 수 는 없으므로 색인어와 메일 간의 관계 및 메일과 분류주제 간의 관계에 의해서 결정 한다. 이러한 관계는 5.1 절의 자동 문서 요약에 의한 디렉토리 분류로 유도할 수 있다. 여기서, 메일문서를 색인어로 구성된 퍼지 집합으로 간주 할 수 있고, 마찬가지 분류주제를 분류된 메시지들의 색인어들로 구성된 퍼지 집합으로 간주할 수 있다. 메시지가 속한 두 분류주제 간의 관계는 생성된 두 분류주제의 퍼지 집합의 함의 정도를 계산하여 결정할 수 있다. 두 퍼지 집합의 함의 정도는 퍼지 함의 연산자를 이용하여 한 퍼지 집합이 다른 퍼지 집합에 포함되는 정도를 계산하여 구할 수 있고, 이를 이용하여 서로 다른 두 분류주제의 유사관계를 동적으로 생성할 수 있다.



(a) $\alpha = 0.94$ (b) $\alpha = 0.76$

그림 1. 동적분류체계방법에 의한 이메일의 재분류

퍼지 함의 연산자는 각 응용의 필요성에 맞게 제시되어야 하는데 본 논문에서는 식(1)의 퍼지 함의 연산자를 사용한다. 퍼지 함의 연산자를 식(2)의 퍼지 관계곱을 적용하여 분류주제들 간의 퍼지 함의 관계, $C_i \rightarrow C_j$ 를 유도할 수 있다. 이렇게 유도된 $C_i \rightarrow C_j$ 는 $C_i \subseteq C_j$ 의 포함 정도를 나타낸다. 다음에는 분류간의 포함정도를 α-cut 하여 크리스프 값으로 바꾸면 최종 결과로서 각 분류간의 관계를 얻을 수 있다. 여기서 α 값을 조정하여 분류주제와 분류주제의 포함관계를 그림 1 과 같이 동적으로 축소하던지 확장할 수 있다.

6. 실험 결과

본 논문에서는 Visual C++6.0 과 Visual Basic 을 사용하여 프로토타입을 구현하였다. 이 절에서는 본 논문에서 설명한 방법을 이용하여 재분류할때의 정확률을 측정한 결과를 보여준다.

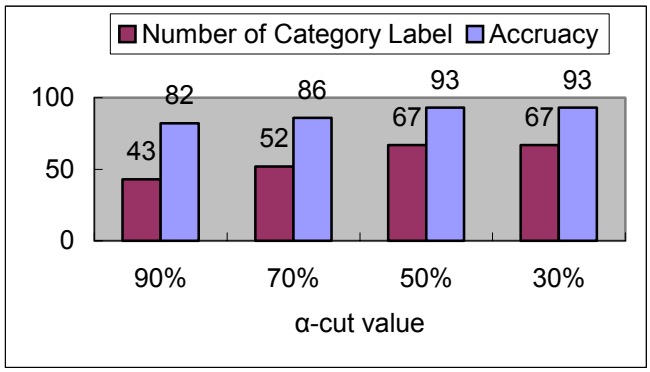


그림 2. 자동 문서 요약과 동적분류체계 방법을 이용하여 재분류 했을 때의 평균 정확률

그림 2 는 재분류시 α 값 변화에 따른 분류정확률을 평가 한다. 여기서, α 값이 작아질수록 분류주제는 많아지면서 분류정확률은 높아지는 것을 알 수 있다. 그리고, α 값이 50% 이하일 경우에는 더 이상 분류주제 및 정확률이 변하지 않는 것을 알 수 있다. 그리고, 이메일 재분류시 정확률을 82%에서 93%로 향상시킨 것을 알 수 있다.

7. 결론

이 논문에서 우리는 이메일 메시지를 자동으로 분류할 수 있는 방법을 설명 및 구현하였다. 본논문에서 제시한 방법은 자동 문서 요약 방법을 이용하여 수신받는 이메일을 디렉토리 분류하였으며, 이메일의 분류결과가 사용자의 주관적에 맞지 않으면 동적분류계층 방법을 이용하여 재분류 할 수 있도록 하였다.

본 논문에서 제안한 방법은 다음과 같은 장점을 가진다. 첫째 제시된 방법은 메일의 분류주제가 자동 생성됨으로 사용자의 간섭 필요 없다. 둘째, 동적분류체계 방법을 이용하여 사용자가 필요하면 언재든지 재분류 할 수 있다. 셋째, 대량의 메일을 효율적으로 관리할 수 있도록 색인검색과 디렉토리 검색방법을 지원한다. 넷째, 학습이 필요 없이 메일을 빠르게 재분류함으로서 유동적인 이메일 환경에 적합하다.

참고문헌

[1] Manco G., Masciari E., A Framework for Adaptive Mail Classification. In Proceedings of the 14th IEEE International Conference on Tools with Artificial Intelligence. 2002.

[2] Cohen W.W.. Learning Rules that classify E-

mail. In Proc. AAAI Spring Symposium in Information Access, 1999.
[3] Androutsopoulos I. et al. An Evaluation of Naïve Bayesian Anti-Spam Filtering. In Proc. Workshop on Machine Learning in the New Information Age, 2000.
[4] Sakkis G. et al. Stacking classifiers for anti-spam filtering of e-mail. In Proc. 6th Conf. On Empirical Methods in Natural Language Processing, 2001.
[5] Drucker H., Wu D., and Vapnik V. N., Support Vector Machines for Spam Categorization. IEEE Transactions on Neural network, 10(5), 1999.
[6] Kun-Lun L., Kai Li, Hou-Kuan H. , Sheng-Feng T., Active Learning with Simplified SVMS for SPAM Categorization. In Proc. First Conf. On Machine Learning and Cybernetics, Beijing, 4-5, November, 2002.
[7] Woitaszek M., Shaaban M.. Identifying Junk Electronic Mail in Microsoft Outlook with a Support Vector Machine. In Proc. 2003 Symposium. On Application and the Internet. 2003.
[8] Baeza-Yates R. and Ribero-Neto B. Modern Information Retrieval. Addison Wesley, 1999.
[9] Ogawa Y., Morita T., and Kobayashi K.. A fuzzy document retrieval system using the keyword connection matrix and a learning method. Fuzzy Sets and System, pp. 163-179, 1991.
[10] Kupiec J., Pedersen J., Chen F., " A Trainable Document Summarizer", Proc. 18th ACM-SIGIR Conf., 1995.
[11] Barzilay R., Elhadad M., " Using Lexical chains for Text Summarization," proc. Association for Computational Linguistics, pp.10-17, 1997.
[12] Eduard H., Chin Y. L., " Automated Text Summarization in SUMMARIST", Proc. Association for Computational Linguistics, pp. 18-24, 1997.
[13] 이창범, 김민수, 백장선, 박혁로, " 주성분 분석과 비정치 분해를 이용한 문서 요약", 정보처리학회 논문지 B 제 10-B 권 제 7 호, 2003.