2003 정보 및 제어 학술회의 (2003.11.21-22)

반자동 방식을 이용한 이메일 추천 시스템

An E-Mail Recommendation System using Semi-Automatic Method

정 옥 란*, 조 동 섭**

* 이화여자대학교 컴퓨터공학과(전화:(02)3277-2309, 팩스:(02)3277-2306, E-mail : orchung@ewha.ac.kr)
** 이화여자대학교 컴퓨터공학과(전화:(033)3277-2309, 팩스:(02)3277-2306, E-mail : dscho@ewha.ac.kr)

Abstract : Most recommendation systems recommend the products or other information satisfying preferences of users on the basis of the users' previous profile information and other information related to product searches and purchase of users visiting web sites. This study aims to apply these application categories to e-mail more necessary to users. The E-Mail System has the strong personality so that there will be some problems even if e-mails are automatically classified by category through the learning on the basis of the personal rules. In consideration with this aspect, we need the semi-automatic system enabling both automatic classification and recommendation method to enhance the satisfaction of users. Accordingly, this paper uses two approaches as the solution against the misclassification that the users consider as the accuracy of classification itself using the dynamic threshold in Bayesian Learning Algorithm and the second one is the methodological approach using the recommendation agent enabling the users to make the final decision.

Keywords : personal rules, automatic classification, accuracy, dynamic threshold, Bayesian Learning Algorithm,

Ⅰ. 서론

급속도로 발전하는 인터넷으로 인한 정보 과부하와 이메일의 급증은 이젠 모든 네티즌들이 겪는 불편함이 아닐 수 없다. 개인화 기법을 이용한 기존의 많은 추천시스템이나 텍스트 분류는 대부분 웹 문서나 상업적 목적의 상품 추천에 집중되어 있다. 추천 시스템의 적용분야는 Usenet news[1], web pages[2,3], 비디오[4], 영화, 음악[5], 책등 다양하지만, 아직까지 이메일을 위한 추천시스템은 연구가 시도되지 않았다. 그러나 이런 추천시스템을 이메일에 적용하려면 메일의 특수성이 먼저 고려되어야 할 것이다.

대다수 웹 정보는 대중을 위한 많은 정보들이 산재되어 있으나, 이메일의 정보는 어느 정도 개인적인 정보로 이루어져 있다고 할 수 있다. 개인이 속해 있는 단체, 가입되어 있는 회원사, 구매한 적이 있거나 관련 있는 상품 회사 사회적 관련이 있는 사람들, 메일 주소의 오용으로 원하지 않은 메일 문서들로 이루어져 있다. 이런 메일을 관리하기 위해서는 기존의 텍스트 분류을 적용하여 자동 분류되는 메일 관리 시스템이 있으나, 사용자의 여러 가지 상황에 따른 유연하게 관리하기 위해서는 사용자의 의견을 직접적으로 반영할 수 있고, 또한 기존의 메일 자료에 의해 관련 카테고리를 보여줄 수 있는 추천 시스템이 적합할 것이다. 이메일 시스템은 개인적인 성향이 강하게 개입되는 점이 있기 때문에 학습을 통해 개인적 룰을 형성하여, 그 룰로 카테고리별 자동 분류를 하더라도 약간의 문제점이 있을 것이다. 이런 점에서 사용자의 만족도를 높이기 위해서는 자동 분류 방식보다는 추천방식과 부분적 자동 분류가 모두 가능한 반자동 방식이 필요할 것이다. 사용자가 가장 우려하는 오분류에 대한 해결책으로 두가지 접근 방식을 이용한다. 첫째는 동적 임계치를 이용해서 분류 자체의 정확도를 향상 시키는 알고리즘적 접근 방식과 자동 분류가 아닌 사용자가 최종 판단을 할 수 있는 추천 방식을 이용하는 방법론적 접근 방식이다.

이 논문은 2003년도 학술진흥재단의 지원에 의하여 연구되었음. (KRF-2003-41-D00460)

Ⅱ. 일반적인 텍스트 분류

메일을 추천하기 위한 전처리 작업으로 텍스트 분류가 가장 기본이 될 것이다. 이것을 기반으로 메일을 분류하게 되는데, 메일 분류의 의미는 정해진 해당 카테고리에 각각의 메일에 할당하는 것이다. 메일의 수가 증가할수록 각각의 메일을 효과적으로 검색 및 색인화하고, 내용 요약과 같은 작업을 수행할 때 많은 시간 소비와 어려운 작업을 하여야 한다. 이를 해결하기 위해 각 메일들을 카테고리별로 귀속시키는 작업을 수행하여, 휴리스틱을 이용하는 방법 대신 컴퓨터를 이용하는 자동화된 기계 학습 기술이 이용된다.

1. 기계 학습 알고리즘

메일 문서를 분류하는 과정에서 룰을 형성하고, 카테고리에 맞게 분류할 때 학습 알고리즘이 이용된다.

문서 자동 분류를 위한 기계 학습법에는 대표적으로 나이브 베이지안 기법(Naive Bayesian Method), 개체 기반 학습 기법인 k-NN(k-Nearest Neighbor), 단어 출현 빈도수를 이요한 TFIDF(Term Frequency Inverse Document Frequency)들이 있는데, 본 논문에서는 가장 많이 사용되고 있는 학습 알고리즘인 나이브 베이지안 기법을 응용하여 이용하였다. 이 학습 기법은 베이스 정리 (Bayes theorem)에 기초한 확률 모델을 이용한다. 이 방법에서는 분류하고자 하는 문서 d에 대한 벡터 모델을 입력하여, 분류 가능한 카테고리들 가운데 이 문서를 관찰할 수 있는 가능성이 가장 높은 클래스를 찾아 그 클래스에 분류한다.

2. 특징 추출

문서 분류 시 중요한 처리 과정이 특징 추출이라 할 수 있는데, 기존의 분류 시스템들도 이 부분에 중점을 두고 연구 진행 되어 왔다. 문서 전처리 과정을 통하여 학습에 이용될 중요한 속성들을 추출하는 과정에서 신뢰성을 향상시키기 위해서는 해당 문서의 공통적인 특징을 가려내어 이를 기준으로 각 속성마다 가중치를 차별적으로 두어 더욱 정확한 중요 속성을 추출하는 방법들이 이용되고 있다. 즉 학습 데이터들의 특징을 고려하여 구분된 카테고리별로 다시 한번 중요도를 정의하는 특징 추출 가중치 설정 기법이다. 이를 위하여 각 학습 자원들의 특징을 고려하여 구분된 카테고리들을 대상으로 일련의 구별 작업을 두어 이를 기반으로 한 속성 추출 작업을 수행할 필요가 있다. 이러한 가중치 설정 작업은 해당 키워드가 속해있는 카테고리의 정보를 고려하여 이루어지며 이로써 카테고리를 대표하는 키워드에게 더욱 높은 가중치가 설정된다. 이러한 특징 추출에 대한 기계학습은 서로 다른 몇 개의 카테고리가 존재하는 경우, 각각의 카테고리 별 키워드에 가중치를 주는 것이다[6][7]. 특징을 선택하는 방법은 크게 네 가지로 나누어 생각해 볼 수 있다. 간략하게 정리해 보면 분류방법에 비의존적인 방법, 분류방법에 의존적인 방법, 개별적인 특징을 취하는 방법, 부분 집합을 취하여 이용하는 방법이다. 본 연구는 분류방법에 비의존적인 방법을 택하여 수행하였으며, 즉 계획에 의존하지 않는 접근 (Scheme-independent approach) 방식이다. 하부 특징 선택은 선처리 작업으로 수행되며 추론 알고리즘을 작동하는데 있어서 선택된 하부 특징 집합의 영향을 고려하지 않는다. 각각의 특징들을 측정하는 작업은 특징 선택을 위해 문서를 학습하는 과정 동안에 이루어진다. 일반적으로 가장 많이 사용되는 특징 선택 방법으로는 문서 빈도수 임계치(Document Frequency threshold), 정보 획득(Information Gain), x^2 통계치(CHI) 등이 있으며, 본 연구에서는 단어 빈도수 임계치를 이용하는 방법이 사용되었다.

3. 추천 시스템의 전처리 과정

이메일 사용자에게 각 메일 문서에 맞는 카테고리를 추천하기 위해서는 먼저 내부적으로 미리 정해진 카테고리에 분류하는 전처리 과정이 있어야 할 것이다. 이 과정에서 '추천을 얼마나 적합하게 하느냐'의 정확도의 문제를 해결할 수 있는 키를 가지고 있다고 할 수 있다. 즉 카테고리별 분류가 정확하게 분류되어야 만이 추천도 잘 할 수 있는 것이다. 문서 분류에 대한 연구는 여러 분야에서 활발하게 진행되어 왔다. 현재까지 이메일 분류 시스템에 대한 연구는 MIT대학에서 만든 자동 분류하는 Maxims[8]이 대표적이다. 일반적으로 기계 학습법에 사용하여 분류 작업을 자동으로 수행할 수 있는 자율적인 소프트웨어를 분류 에이전트라 한다. 이와 같은 분류 에이전트의 대표적인 한 예로 카네기 멜론 대학의 Personal Webwatcher[9]가 있다. 이 분류 에이전트는 웹 브라우저를 통해 사용자의 행동을 모니터링하여 사용자의 관심 영역을 학습한 뒤, 브라우징하는 웹 문서내의 링크들에 대해 사용자의 관심영역에 속하는 것들과 그렇지 않은 것들을 분류하여 관심 있는 링크들만을 제안 해 주는 시스템이다. 또한 엔더슨 컨설팅 연구실에서 개발된 InfoFinder[9] 역시 사용자의 관심 프로파일을 바탕으로 온라인 문서들에 대한 분류 작업을 통해 사용자가 관심을 가질 문서들을 찾아주는 에이전트 시스템이다. 이외에도 엔터테인먼트 선별 에이전트인 Ringo, 뉴스 기사 분류 에이전트인 NewT[10]등이 모두 문서 분류 기법을 이용한 대표적인 분류 에이전트 시스템이다. 또한 협업필터링(collaborative filtering)를 이용한 성공적인 추천 시스템으로는 Tapestry, GroupLens, PHOKS등이 있다.

III. 이메일 추천 시스템

1. 반자동 방식을 이용한 이메일 분류

본 연구에서 이메일 분류를 반자동 방식으로 한다는 것은 기존 자동 분류 방식이 이메일 시스템에 적용하기에는 약간의 문제점이 있기 때문이다. 이메일 시스템은 개인적 성향이 강하기 때문에 학습을 통해 형성된 개인적 룰로 카테고리별 자동 분류를 하더라도 사용자를 만족시키기는 어려울 것이다. 그래서 필요없는 메일이나 스팸으로 간주되는 메일은 분류에서 미리 삭제하는 자동 기능과 그 외의 메일은 사용자가 열었을 때 관련 카테고리를 우선순위로 추천해주는 방식을 제안하는 것이다. 해당 카테고리를 추천 받은 사용자는 우선순위에 따라 하나이상의 카테고리에 저장하거나, 또 시간차에 의해 해당 카테고리가 변동이 될 경우 유연하게 메일을 관리 할 수 있으므로 잘못된 메일 분류를 방지할 수 있을 것이다. 물론 메일의 양이 너무 많거나 추천에 대한 신뢰도가 만족될 때는 자동 분류를 할 수 있도록 사용자 인터페이스에 체크 박스를 설정하였다. 그 체크박스에 체크를 하게 되면 생성된 룰을 바탕으로 자동 분류가 가능하다.

2. 이메일 추천 시스템의 구조

본 논문에서 제안한 이메일 추천 시스템은 크게 두 가지 특징을 가지고 있다. 첫째로는 모듈별로 구성되어 효율적인 특징 추출 및 룰 형성, 카테고리별 분류

를 할 수 있게 하였으며, 두번째는 동적 임계치를 이용한 베이지안 학습 알고리즘을 적용하여 분류의 정확도를 높혔다는 것이다. 다음 그림 1은 본 시스템의 전체적인 구성도로서, 사용자 정보를 기반으로 베이지안 학습에 의한 필터링, 분류, 추천 과정을 보여주고 있다.

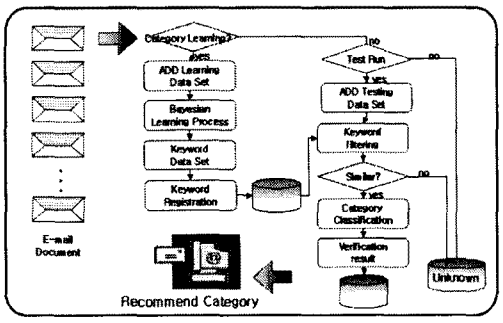

그림 1. 이메일 추천 시스템의 구조.

전체적인 시스템 구조가 모듈별 설계로 되어 있으며, 그 모듈들은 공유된 파일들에 의해서 서로 전달된다. 각 모듈별 역할은 다음과 같다.

①The Web Mail Interface Module: 새로운 메일이 도착하면, 먼저 사용자의 메일 처리 과정을 관찰하여 학습한다. 특징 추출 및 규칙(rule)형성에 도움을 주는 모듈이며, 또한 사용자가 개인에 맞는 카테고리 설정을 할 수 있는 과정이다.
②The Category Rule Generation Module: 메일 처리 과정에서 특징을 추출하여 응용된 베이지안 알고리즘을 적용하여 개인에 맞는 룰(rule)을 형성한다.
③The Mail Classification & Recommendation Module: 생성된 룰을 기반으로 새로운 메시지가 도착하면 카테고리별 분류를 한 다음 사용자에게 해당 카테고리를 우선순위별로 추천해준다. 또한 불필요한 메일이나 스팸메일은 자동 삭제한다.

다음 두 번째 특징으로 기존의 고정된 임계치를 동적으로 개선하여 필터링의 정확도를 향상시켰다. 그 식을 살펴보면, 식(1)의 C는 전체 카테고리 집합이고 식(2)의 D는 메일 문서로 정의하자. 식(3)에서는 문서D에 대한 각 카테고리c에 대한 조건부 확률을 구해서, 분류 문서에 대하여 가장 높은 확률값 식(4)을 가지는 카테고리에 분류하게 되는 것이다. 기존의 베이지안 알고리즘 식(5)의 T값을 고정해서 사용을 하였는데, 본 연구에서는 메일 문서의 학습 상황에 맞게 에이전트가 T값을 동정으로 정할 수 있게 개선하였다. 이로 인한 정확도의 향상은 뒤의 실험 결과에서 비교 분석하였다.

$$Category\ Set\ C = \{c_0, c_1, c_2, c_3, ..., c_k\},\ C_0 = unknown\ category \quad (1)$$

$$Document\ Set\ D = \{d_1, d_2, d_3, ..., d_i\} \quad (2)$$

$$\Re(d_i) = \{p(d_i \mid c_1), p(d_i \mid c_2), p(d_i \mid c_3), ..., P(d_i \mid c_k)\} \quad (3)$$

$$P'_{max}(d_i) = \max\{p(d_i \mid C_t)\} \quad , t = 1, ..., k \quad (4)$$

$$Cbest(d_i) = \begin{cases} \{c_j \mid P(d_i \mid c_j) = P'_{max}(d_i),\ if\ P'_{max}(d_i) \geq T \\ \qquad\qquad whereT = 1 - \dfrac{P'_{max}(d_i)}{\sum_{t=1}^{k} P(d_i \mid C_t)} \\ c_0 \quad , \quad otherwise \end{cases} \quad (5)$$

3. 시스템 구현 및 실험

이메일 추천 시스템은 언제 어디서나 로그인이 가능하고 시스템에 제한이 없으며, 또한 별도의 메일 클라이언트 프로그램이 필요 없는 장점을 가지고 있는 웹 메일 기반으로 구현하였다. 구현환경으로는 Windows 2000 Professional, 데이터 베이스 컨트롤을 위해 MS SQL 2000, 룰 형성 및 알고리즘 실행을 위해 MS visual C++ 6.0과 ASP, ASP 콤포넌트를 이용하였다. 다음 그림 2는 실제 구현된 사용자 인터페이스 화면를 보여준다.

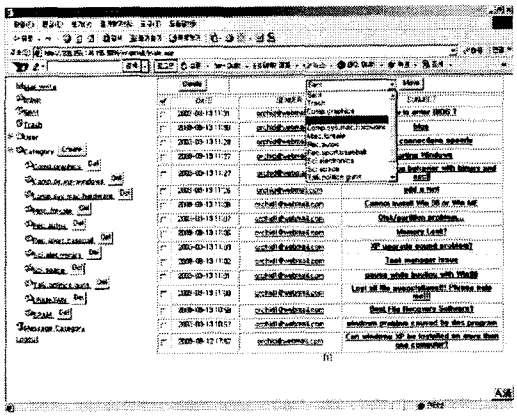

그림 2. 사용자 인터페이스

사용자 인터페이스는 사용자 관찰과정에 이용되며, 실제 카테고리 생성 및 저장을 할 수 있다. 본인이 자주 쓰는 카테고리를 생성하고 필요 없는 카테고리를 삭제할 수도 있으며, 학습하는 과정에서 특정 추출하여 메일 분류를 내부적으로 하여, 사용자에게 추천 카테고리를 제공하는 것이다. 메일 사용자는 새로운 메일 문서를 열 때 다음 그림 3과 같이 추천 카테고리를 제공 받게 되는 것이다. 본 연구의 성능 평가는 '사용자에게 얼마나 정확한 카테고리를 추천하느냐'인데, 이는 먼저 메일 내용을 해당 카테고리에 맞게 분류하는지를 체크하면 될 것이다.

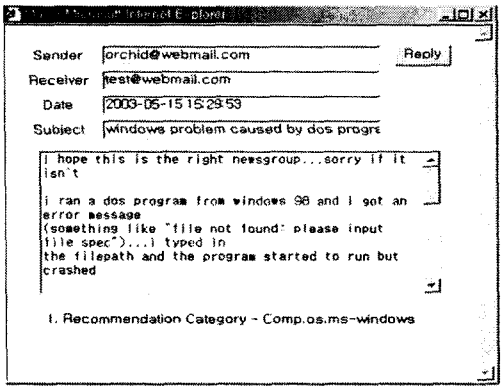

그림 3. 메일 추천 카테고리

실험방법으로는 먼저 카테고리를 생성하고, 각각 카테고리별 실험데이터를 학습 시킨 후 룰을 형성한다. 다음으로 카테고리별 자료를 수집하여 룰에 의해서 정확한 분류를 하는지를 체크하는 것이다. 이 실험을 위해 10가지 카테고리를 미리 설정하고, 룰을 위해 샘플 데이터와 성능 평가를 위한 데이터를 수집하여 실행하였다. 본 실험은 시스템의 기능 중 FileCheck기능을 이용하여 정확률을 체크하였다. 실제 실험은 많은 양의 데이터를 실험해야 하기 때문에 카테고리별 만통 정도의 메일을 하나의 데이터 포맷으로 만들어 테스트하였다. 이런식으로 각 카테고리를 테스트 했을때 정확률은 그림4와 같다. 그림4에서 correct data(1)은 기존의 베이지안 알고리즘을 적용하였을때 정확률이고, correct data(2)는 동적 임계치를 이용한 베이지안 알고리즘을 적용한 결과이다.

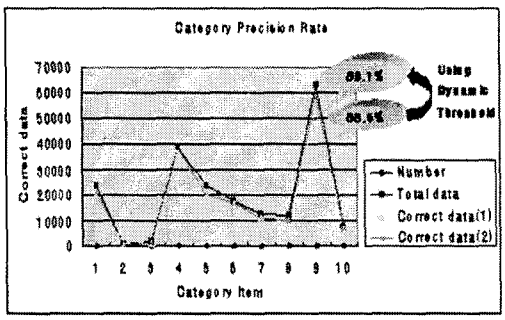

그림 4. 메일 추천 카테고리 정확률

평균 정확률은 88.6%로 측정되었으며, 동적 임계치를 적용한 결과는 89.1%를 보여주었다. 기존의 알고리즘을 사용하였을 때 보다 0.5%의 향상을 보여주었다.

Ⅳ. 결론

본 연구에서는 이메일 사용자에게 도움이 될 수 있는 추천 시스템을 설계 및 구현하였다. 현재 이메일을 통해 많은 양의 정보들이 오가고 있고, 사용자들은 본인에 맞는 맞춤 이메일 인터페이스를 요구하게 될 것이다. 메일 문서를 텍스트 분류에 적용하여 좀 더 정확도 높은 카테고리를 추천받게 된다면 사용자의 메일 관리가 훨씬 편리해 질 것이다. 또한 문서 분류에 가장 문제가 되고, 사용자가 우려하는 오분류에 대한 해결책으로 두가지 방법으로 해결하고자 하였다. 첫째는 동적 임계치를 이용해서 분류 자체의 정확도를 향상시키는 알고리즘 접근 방법과 둘째는 자동 분류가 아닌 사용자가 최종 판단을 할 수 있는 추천 방식을 이용하는 것이다. 갈수록 늘어나는 메일을 관리할 때 우리가 제안한 이메일 추천 시스템은 매우 유용하게 활용될 수 있을 것이다. 향후 연구 방향으로는 카테고리를 사용자가 직접 설정하는 방법으로 되어 있는 데 자동 카테고리 설정과 추천 할 수 있는 방법을 동시에 가능한 시스템으로 확장 시켜 나갈 것이다.

참고문헌

[1] Konstan, J.A., Miller, B.N., Maltz, D., Herlocker, J.L., Gordon, L.R. and Riedel, J.GroupLens: Applying Collaborative Filtering to Usenet News. *CACM, 40(3).* 77-87.

[2] Balabanovic, M. and Shoham, Y. Fab: Content-Based, Collaborative Recommendation.*CACM,40(3).* 66-72.

[3] Hill, W. and Terveen, L., Using Frequency-of-mention in Public Conversations for Social Filtering. *CSCW'96,* 106-112.

[4] Hill, W., Stead, L., Rosenstein, M. and Furnas, G., Recommending and Evaluating Choices in a Virtual Community of Use. *CHI'95,* 194-201.

[5] Shardanand, U. and Maes, P., Social Information Filtering: Algorithms for Automating "Word of Mouth". *CHI'95,* 210-217.

[6] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, "Modern Information Retrieval," Addison-wesley, 1999.

[7] Willian W.Cohen, "Learning Rules that Classify E-Mail", AAAI Spring symposium on Machine Learnning in Information Access, pp18-25,1996.

[8] P.Maes, "Agents That Reduce Work and Information Overload", Communications of the ACM, Vol.37, No.7, pp.30-40, 1994.

[9] Haejung Bak, Yeongdaek Park, Sukhwan Yun, "Web Agent using user's favorite", Korea Information Processing Society Review, September1999.

[10]Jeffrey M.Bradshaw, "Software agent", AAAI Press/ The MIT Press, pp. 151-161.