이메일 분류를 위한 추천 에이전트 시스템

정옥란° 조동섭

이화여자대학교 컴퓨터학과
{orchung°, dscho}@ewha.ac.kr

A Recommendation Agent System for E-Mail Classification

Ok-Ran Jeong° Dong-Sub Cho
Dept. of Computer Science and Engineering , Ewha Womans University

요 약

급속도로 발전하는 인터넷의 발달로 인한 정보의 과부하와 이메일의 급증은 이젠 모든 네티즌들이 겪는 불편함이 아닐 수 없다. 본 논문에서는 이런 이메일 관리를 사용자가 효율적으로 할 수 있도록 추천 에이전트(Recommendation Agent)를 제안하고자 한다. 추천 에이전트 시스템에서는 이메일의 자동 분류에서 가장 핵심인 정확도(Accuracy)를 개선시키기 위해 최종 결정을 사용자가 하는 방식으로 접근하였으며, 또한 여기에 이용되는 학습 및 분류 알고리즘을 동적 임계치을 적용한 베이지안 학습 알고리즘을 이용하여 알고리즘적 방법도 병행하였다. 새로운 메일이 도착했을 때 최적의 분류를 할 수 있도록 메일 카테고리를 추천하는 시스템이다. 또한 사용자 편의를 위하여 필요없는 메일이나 스팸으로 간주되는 메일은 자동 삭제하는 기능을 추가하였다.

1. 서 론

웹을 통한 이메일의 사용량은 네트워크의 발달과 함께 기하급수적으로 많아지고 있으며, 인터넷 사용자들이 가장 애용하는 프로그램이다. 이를 위한 웹기반 이메일은 사용자들의 다양한 기능적 요구에 부응해야 하며, 프로그래머와 메일 관리자들은 더 많은 기능을 추가하기 위해 노력해야 한다. 빠르고 편리하다는 이유로 일반 사용자들도 쏟아져 나오는 메일을 관리하는데 매일 일정량의 시간을 소비해야 하는 현실이다.

본 논문에서 제안한 추천 에이전트 (Recommendation Agent)는 일정기간 사용자의 메일 처리 과정을 관찰하여 사용자에게 맞는 룰을 형성하고, 형성된 룰을 바탕으로 새로운 메시지가 도착하면 적합한 카테고리를 추천하는 것이다. 즉 최종 카테고리 분류 결정은 메일 사용자가 하는 것이다. 문서 자동 분류에 대한 연구가 많이 진행되고 있지만, 개인적 성향이 강한 이메일 시스템에서는 자동 분류 보다는 추천 받는 방법이 적합할 것이다. 또한 전처리 과정으로 학습과정과 룰 형성이 필요한데, 본 연구에서는 베이지안 알고리즘을 개선하여 이용하였다.

2. 메일 분류를 위한 전처리 작업

2.1 문서 필터링

본 연구의 메일 분류를 위해서 중요한 전처리 작업은 메일 문서 필터링이라 할 수 있다. 여기서 필터링이란 메일의 내용을 분석하여 그 내용에 따라 카테고리별 폴더에 저장하는 것이다. 이를 수행하기 전에는 일정량 이상의 학습이 필요하고, 학습과 분류에 어느 정도 시간이 걸린다. 본 연구에 이용되는 학습 알고리즘은 응용된 베이지안 알고리즘을 사용하였다. 형태적인 면에서 단어의 의 종류, 단어의 위치, 단어의 빈도수 등의 특징(feature)으로 표현될 수 있다. 또한 이렇게 단어를 표현하는 방법에 있어서도 단어구성면에서 단일어인지, 구문적 어구(syntactic pharase)인지, 시소러스(thesaurus) 인지에 따라 다르게 표현될 수 있는 요소들이 대단히 많다.[1,2] 메일 내용을 분류한다 함은 미리 정의되어 있는 여러 카테고리에 각각의 메일들을 할당하는 것이다. 하지만 메일의 수가 증가할수록 각각의 메일을 효과적으로 검색(retrieval) 및 색인화(indexing)하고, 내용 요약(summarization)과 같은 작업을 수행할 때 많은 어려움을 겪게 된다. 이를 해결하기 위해 각 메일들을 카테고리별로 귀속시키는 작업을 수행하며, 휴리스틱(heuristic)을 이용하는 방법 대신 컴퓨터를 이용하는 자동화된 기계학습 기술이 이용되었다. 대표적인 분류기법으로는 최근접 이웃분류(nearest neighbor classification), 베이지안 알고리즘(Bayesian Algorithm), r 결정트리(decision tree), 신경망(neural networks), 그리고 지지벡터기계(support vector machine)들이 있다[1,2]. 이러한 분류 알고리즘은 문서의 특징을 선택하는 여러 방법과 함께, 최근 활발한 연구가 많이 적용되고 있다. 여기서 본 연구에 가장 적합한 베이지안 알고리즘을 이용하여 학습시킨후 개인적 룰을 형성하였다. 이 룰을 기반으로 새로운 메시지에 대해 추천 에이전트(Recommendation Agent)를 이용하여 메일을 필터링 하고자 했다.

이 논문은 2003년도 두뇌한국21(BK21)사업에 의하여 지원되었음.

2.2 문서 특징 추출

문서 분류시 기본적으로 미리 잘 정의되어야 할 부분이 특징 추출(Feature Extraction)이다. 문서 전처리 과정을 통해서 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 방법이 이용되고 있다. 이러한 속성 추출방법을 특징추출이라고 한다. 즉, 특징 추출은 학습 자원의 중요 속성들을 자원이 구분된 카테고리별로 다시 한번 중요도를 정의하는 특징 추출 가중치 설정 기법이다. 이를 위해서 각 학습 자원들의 특징을 고려하여 구분된 클래스들을 대상으로 일련의 구별 작업을 두어 이를 기반으로 한 속성 추출 작업을 수행할 필요가 있다. 이러한 가중치 설정 작업은 해당 키워드가 속해 있는 클래스의 정보를 고려하여 이루어지며 이로써 클래스, 즉 각 카테고리를 대표하는 키워드에게 더욱 높은 가중치가 설정된다. 이러한 특징추출에 대한 기계학습 방법은 서로 다른 몇 개의 카테고리가 존재하는 경우, 각각의 카테고리별 키워드에 가중치를 주는 것이다[3]. 본 연구에서는 사용자 인터페이스 모듈(The Web Mail Interface Module)에서 얻은 특징추출은 카테고리 룰 형성 모듈(The Category Rule Generation Module)과 카테고리 분류 모듈(The Web Mail Classification Module)에서 이용된다.

2.3 Dynamic Threshold를 이용한 베이지안 알고리즘

메일 필터링(Mail Filtering)을 하는 과정에서 카테고리별 룰(Rule)을 형성하고, 내용을 분류할 때 학습 알고리즘(learning algorithm)이 필요하다. 본 논문에서는 가장 많이 사용되고 있는 학습 알고리즘인 베이지안 알고리즘을 이용하였다. 이 학습방법은 모든 문서에서 특정단어의 출현으로 구별되는 이진 속성 벡터(vector of binary attributes)로 표현된 모델로 문서를 정형화 하는데, 모델은 다형성 베이누이 사건 모델(multi-variate Bernoulli event model)을 기초로 하여 각 클래스의 문서 마다 다르게 모델을 만들게 된다. 여기서 사용되는 가설은 문서들의 모든 속성은 주여진 전체 클래스의 다른 문서의 전후 관계에 대해서 독립적이다. 여기서 모델화 작업으로 만들어진 문서의 모델을 사용하여 각각의 클래스에 대한 문서의 확률 값을 구하고, 구해진 확률 값 중 가장 높은 확률 값을 가진 클래스에 문서를 분류하게 되는 것이다[4]. 또한 본 논문에서는 기존의 고정된 임계치(threshold)를 동적으로 개선하여 필터링의 적합도를 향상시켰다. 응용된 알고리즘은 그림 1과 같다.

$$Category\ \ Set\ C = \{c_0, c_1, c_2, c_3, \ldots, c_k\},\ C_0 = unknown\ \ category \quad [1]$$

$$Document\ \ Set\ D = \{d_1, d_2, d_3, \ldots, d_i\} \quad [2]$$

$$\Re(d_i) = \{p(d_i|c_1), p(d_i|c_2), p(d_i|c_3), \ldots, P(d_i|c_k)\} \quad [3]$$

$$P_{max}(d_i) = \max\{p(d_i|C_t)\},\quad t = 1, \ldots, k \quad [4]$$

$$C_{best}(d_i) = \begin{cases} \{c_t \mid P(d_i|c_t) = P_{max}(d_i),\ if\ P_{max}(d_i) \geq T \\ \qquad whereT = 1 - \dfrac{P_{max}(d_i)}{\sum_{t=1}^{k} P(d_i|C_t)} \\ c_0, \qquad otherwise \end{cases} \quad [5]$$

그림 1 Dynamic Threshold를 이용한 베이지안 알고리즘

3. 추천 에이전트 시스템

3.1 모듈별 시스템 설계

이메일 분류(E-Mail Classification)를 위한 추천 에이전트 시스템은 크게 세가지 모듈로 구성되어 있으며, 각각의 모듈들을 차례로 자세하게 살펴보겠다. 먼저 대략적인 모듈별 역할은 다음과 같다.

- 사용자 인터페이스 모듈 (The Web Mal Interface Module) : 새로운 메시지가 도착하면 먼저 사용자의 메시지 처리과정을 관찰하여, 학습한다. 이 모듈은 특징 추출 및 규칙(Rule) 형성에 도움을 준다. 실제 본 시스템에서는 사용자가 받은 메일을 설정된 카테고리 중 어느 카테고리에 분류하는지를 관찰 및 학습 데이터로 이용하는 것이다.
- 카테고리 룰 형성 모듈 (The Category Rule Generation Module) : 메시지 처리과정에서 학습 후 추출된 특징을 베이지안 알고리즘을 이용하여 개인적 룰(Personal Rule)을 형성한다.
- 카테고리 분류 모듈 (The Web Mail Classification Module) : 형성된 개인적 룰을 기반으로 새로운 메시지가 도착하면 최적의 카테고리를 추천한다. 또한 불필요한 메일은 삭제한다.

3.2 시스템의 전체적인 구조

본 시스템의 목적은 사용자가 메일을 처리하는데 도움을 주는 에이전트 개발에 있다고 할 수 있으며, 또한 각 사용자의 메일 관리가 편리하도록 인터페이스 환경을 제공하는데 있다. 전체적인 시스템의 흐름은 그림 2와 같다.

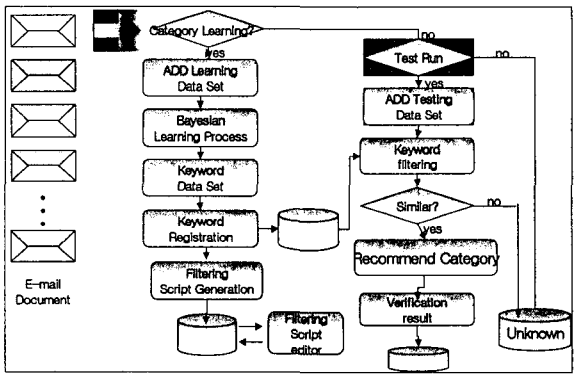

그림 2 추천 에이전트 시스템 흐름도

3.3 시스템의 구현

본 시스템은 언제 어디서나 로그인이 가능하고 시스템에 제한이 없는 장점을 가지고 있는 웹 메일(Web Mail)을 기반으로 하였으며, VC++6.0, MS SQL 2000 Server, ASP, ASP 콤포넌트로 구현하였다.

실제 구현된 User Interface는 다음 그림 3과 같다. 일반적으로 사용하는 메일 쓰기, 읽기, 휴지통등의 기능은 별 차이가 없으나 본 메일 시스템만이 가지고

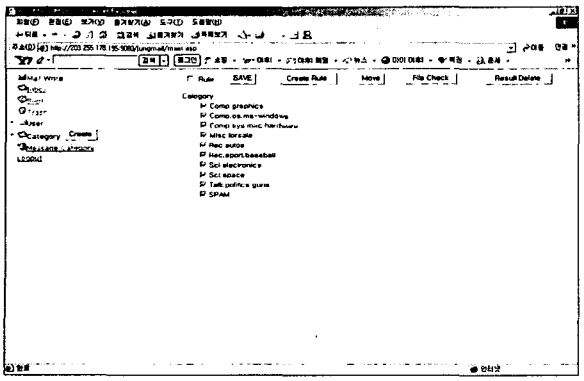

그림 3 사용자 인터페이스

있는 화면상 주요 기능은 다음과 같다.
- Rule Check Box : 체크박스의 기능은 체크가 안 되어있을 때는 추천 카테고리가 받은 메일 하단에 나타난다. 메일을 받는 동시에 자동 분류될 수 있는 기능은 여기 체크 박스에 체크를 하면 된다.
- Save Button: 첫 번째 Rule 체크박스나 하단의 메일 폴더 트리에서 체크 등 수정작업을 한 후 현재 체크된 상태를 저장할 때 사용하며, 저장을 해야 체크한 상황이 효력을 발휘하게 된다.
- Create Rule Button: 모든 폴더에 있는 룰을 초기화하고, 그 후 하단의 폴더 트리에서 체크를 한 폴더 안에 메일의 내용을 기준으로 새로운 룰을 생성한다.

카테고리 설정 및 학습과정은, 먼저 룰 형성을 위해서 사용자가 카테고리를 미리 설정하고 받은 메일들을 추천 카테고리 도움을 받아 설정된 카테고리에 저장, 불필요한 메일이나 스팸으로 간주되는 메일을 삭제한다. 이 과정을 일정 시간동안 학습을 한 다음 개인에 맞는 카테고리 룰(Rule)을 형성하게 되는 것이다. 그 룰이 형성된 후로 받는 메일은 추천 에이전트 시스템 관리가 가능해 지는 것이다. 다음 그림 4와 같이 메일을 여는 순간 적합한 카테고리를 우선순위로 추천하여 주는 것이다.

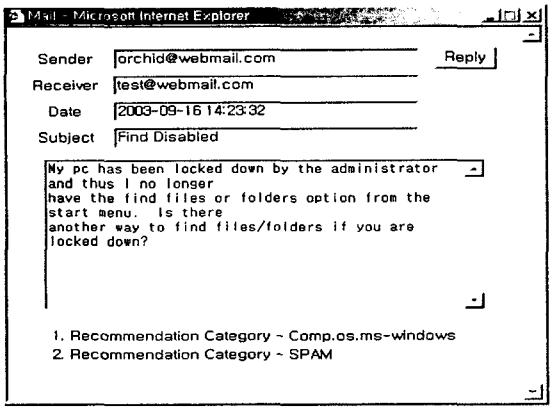

그림 4 추천 에이전트 실행 화면

3.4 실험 및 결과

본 시스템의 성능 평가를 위해 카테고리를 미리 10가지로 정하고 시뮬레이션 하였다. 이 카테고리 폴더는 웹에 많이 나와있는 주제를 감안하여 선정하였다.

그림3의 File Check를 이용하여 실험을 한 결과, 88.6%의 정확률을 보여 주었다. 여기서 카테고리 정확률(Category Precision Rate)이 뜻하는 것은 받은 메일이 해당 카테고리 폴더에 적합하게 분류되었는지를 나타낸다. 학습데이터를 많이 가지게 될수록 학습기간이 길어질수록 정확률은 더 높아질 것이다. 기존의 베이지안 알고리즘을 응용해서 실험을 했을 경우 앞에서 언급했듯이 88.6%의 정확률을 나타냈으며, 동적 임계치(Dynamic Threshold)를 적용한 개선된 베이지안 알고리즘을 이용했을 경우는 89.1%의 정확률을 나타냈다. 실험 결과를 세부적으로 정리한 내용은 다음 표1과 같다.

표 1 카테고리 정확률 결과

Num	Category Item	Total data	Correct data(1)	Precision1 (%)	Correct data(2)	Precision2 (%)
1	Comp.graphics	23579	20278	86	21457	91
2	comp.os.ms-window	573	527	92	470	82
3	comp.sys.mac.hardw	1578	1404	89	1215	77
4	Misc.forsale	38434	36128	94	38050	99
5	rec.autos	23712	19918	84	20155	85
6	Rec.sport.baseball	17124	15925	93	16610	97
7	sci.electronics	12354	11242	91	10624	86
8	Hea.comp.lea.g.c	11694	9706	83	9940	85
9	Hea.politics	62915	53478	85	59140	94
10	Spam	7260	6461	89	6897	95
	total	199223	176512		184558	
	Average			88.6		89.1

4. 결론

개인에 맞는 룰을 형성하여, 효율적인 이메일 분류를 위해 카테고리 선정을 위한 추천 에이전트를 활용하는 것이다. 즉, 추천 카테고리를 선정 받아 개인의 메일 분류를 편리하게 할 수 있는 것이다. 또한 스팸 메일도 스팸카테고리에 분류되어 자동 삭제하므로 스팸처리에도 효과를 볼 수 있다. 본 논문에서는 이러한 기능을 가지고 있는 추천 에이전트 시스템(Recommendation Agent)를 설계 및 구현하였다. 현재 이메일을 통해 많은 양의 정보들이 오가고 있고, 사용자들은 또한 이 중에서 개인 각자에게 맞는 맞춤 이메일 인터페이스를 요구하게 될 것이다. 이러한 현안을 해결하고자 한 것이며, 현실적으로도 매우 유용하게 사용될 것이다. 향후 연구로는 미리 사용자가 카테고리를 설정하는 방법을 자동 카테고리 설정방법으로 확장할 것이다.

[참 고 문 헌]

[1] Dunja Mladenic, Marko Grobelnik, Feature selection for classification based on text hierarchy , *Proc. of the Workshop on Learning from Text and the Web*, Pittsburgh, USA, 1998

[2] Yiming Yang, Jan O. Pedersen, A Comparative Study on Feature Selection in Text Categorization , *Proc. of ICML97*, pp. 412-420, 1997

[3] 이상섭, 오재준, 박영택,"웹에이전트의 핵심 기술" *http://member.tripod.lycos.co.kr/ironjohn/agent/agent2.html*

[4] McCallum, A. Nigam, K. 1998 A Comparison of Event Models for Naïve Bayes Text Classification In AAAI 98 Workshop on Learning for Text Categorization., *http://www.cs.cmu.edu/~mccallum 1998*