

이메일 관리를 위한 룰 필터링 컴포넌트 기반 능동형 추천 에이전트 시스템

A Dynamic Recommendation Agent System for E-Mail Management based on Rule Filtering Component

정 옥 란*, 조 동 섭**
(Ok-Ran Jeong*, Dong-Sub Cho**)

Abstract – As e-mail is becoming increasingly important in every day life activity, mail users spend more and more time organizing and classifying the e-mails they receive into folder. Many existing recommendation systems or text classification are mostly focused on recommending the products for the commercial purposes or web documents. So this study aims to apply these application to e-mail more necessary to users. This paper suggests a dynamic recommendation agent system based on Rule Filtering Component recommending the relevant category to enable users directly to manage the optimum classification when a new e-mail is received as the effective method for E-Mail Management. Moreover we try to improve the accuracy as eliminating the limits of misclassification that can be key in classifying e-mails by category. While the existing Bayesian Learning Algorithm mostly uses the fixed threshold, we prove to improve the satisfaction of users as increasing the accuracy by changing the fixed threshold to the dynamic threshold. We designed main modules by rule filtering component for enhanced scalability and reusability of our system.

Key Words :recommendation agent, rule filtering component, dynamic threshold

1. 서 론

대부분 기존의 추천 시스템들은 웹 사이트에 방문하는 기존 사용자들의 사전 프로파일 정보, 상품 검색 및 구입 관련 정보를 기반으로 사용자의 기호에 맞는 상품이나 기타 정보를 추천하고 있다. 이런 추천 시스템의 방법으로 협업 필터링 방식을 사용하면 다른 사용자들의 피드백 정보를 이용하여 동적으로 연관 링크를 제공할 수 있는 장점이 있다[1]. 추천시스템의 적용 분야는 Usenet news[2], web pages, 비디오, 영화, 음악, 책 등 다양하지만, 아직까지 이메일을 위한 추천시스템은 연구가 시도되지 않았다. 그러나 이런 추천시스템을 위한 많은 정보들이 산재되어 있으나, 이메일의 정보는 어느 정도 개인적 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 않는 메일 문서들로 이루어져 있다. 이런 메일을 관리하기 위해서는 기존의 텍스트 분류에 적용하여 자동 분류되는 메일 관리 시스템이 있으나, 사용자의 여러 가지 상황에 따라 능동적으로 관리하기 위해서는 사용자의 의견을 직접적으로 반영할 수 있고, 또한 기존의 메일 자료에 의해 관련 카테고리를 보여줄 수 있는 능동형 추천 에이전트 시스템이 적합할 것이다.

이메일 시스템은 개인적인 성향이 강하게 개입되는 점이 있기 때문에 학습을 통해 개인적 룰로 카테고리별 자동 분류

를 하더라도 약간의 문제점이 따를 것이다. 이런 점에서 사용자의 만족도를 높이기 위해서는 자동 분류 방식보다는 추천 방식과 부분적 자동 분류를 모두 가능하게 할 수 있는 반자동 방식이 적합 할 것이다. 본 연구에서는 사용자가 가장 우려하는 오분류에 대한 해결책으로 두 가지 접근방식을 이용한다. 첫째는 동적 임계치를 이용하여 분류 자체의 정확도(accuracy)를 높이는 알고리즘적 접근방식과 자동 분류가 아닌 사용자가 최종 판단을 할 수 있는 추천 에이전트를 이용하는 방법론적 접근 방식이다. 또한 이 접근 방식의 주요 모듈 부분을 룰 필터링 컴포넌트기반으로 시스템을 설계하여 시스템의 확장성과 재사용성을 높이고자 하였다.

2. 기반 연구

2.1 기계 학습 알고리즘

메일 문서를 분류하는 과정에서 룰을 형성하고, 카테고리에 맞게 분류할 때 학습 알고리즘이 이용된다. 문서 자동 분류를 위한 기계 학습법에는 대표적인 나이브 베이지안 기법, 개체 기반 학습 기법인 k-NN(k-Nearest Neighbor), 단어의 출현 빈도수를 이용한 TFIDF(Term Frequency Inverse Document Frequency)들이 있는데, 본 논문에서는 가장 많이 사용되고 있는 학습알고리즘인 나이브 베이지안 기법을 이용하였다. 이 학습 기법은 베이즈 정리(Bayes theorem)에 기초한 확률 모델을 이용한다. 이 방법에서는 분류하고자 하는 문서에 대한 벡터 모델을 입력하여, 분류 가능한 카테고리들 가운데 이 문서를 관찰할 수 있는 가능성이 가장 높은 클래스를 찾아 그 클래스에 분류하는 방식이다. 여기서 이용되는

저자 소개

* 正會員 : 梨花女子大學校 컴퓨터學科 博士課程

** 正會員 : 梨花女子大學校 컴퓨터學科 教授 · 工博

가설은 문서들의 모든 속성이 주어진 카테고리 안에서 다른 문서의 전후 관계에 대해서 서로 독립적이라는 나이브 베이지안 가정을 적용한다. 본 연구에서는 적용된 베이지안 알고리즘의 임계치를 동적으로 변환하여 분류의 정확도를 높혔다.

2.2. 컴포넌트 특성

컴포넌트는 인터페이스를 통하여 서비스를 제공하고 요구할 수 있는 것으로 독립적이고, 작은 단위로 나누어질 수 있는 시스템의 요소이며, 시스템의 필요에 의하여 다른 요소로 대체 가능한 것이다. 그리고 미리 개발된 응용 코드의 일부분이다[3]. 제시된 컴포넌트의 정의를 기초로 하여 컴포넌트의 특성을 나타낼 수 있다. 특성으로는 식별 가능성, 추적 가능성, 교체 가능성, 인터페이스에 의한 접근 가능성, 인터페이스가 제공하는 서비스 고정성, 문서로 정확히 기록되는 서비스, 물리적 구현의 은닉, 독립성, 언어 및 개발 툴에 독립적인 재사용성, 플랫폼과 무관한 재사용성, 동적 재사용성을 제시할 수 있다. 또한 컴포넌트는 일반적으로 논리적인 관점과 물리적인 관점의 컴포넌트로 나눌 수 있다[3]. 컴포넌트의 특성을 기반으로 하여 논리적인 컴포넌트와 물리적인 컴포넌트의 개념을 제시하면 다음과 같다. 논리적인 컴포넌트는 비즈니스 컴포넌트를 의미하며, 비즈니스 영역에서 실세계의 개념을 모델링한 것을 나타낸다. 반면에 물리적인 컴포넌트는 비즈니스 컴포넌트를 독립적인 소프트웨어로 나누어서 공학적인 관점으로 구축하는 것을 의미한다. 따라서 본 연구에서는 논리적인 컴포넌트 관점으로 접근하여, 메일 관리에 있어서 가장 주요하게 이용되는 필터링 부분을 컴포넌트로 재 구현 하여 확장성과 재사용성을 높이고자 하였다.

3. 제안된 시스템의 접근 방식과 특징

3.1 이메일 분류 방식

본 연구에서 이메일 관리를 추천 방식으로 한다는 것은 기존 자동 분류 방식이 이메일에 적용하기에 약간의 문제점이 있기 때문이다. 이메일 시스템은 개인적 성향이 강하게 개입되기 때문에 학습을 통해 형성된 개인적 룰로 카테고리별 자동 분류를 하더라도 사용자를 만족시키기는 한계점이 있다. 그래서 형성된 룰을 바탕으로 먼저 메일을 카테고리별 분류를 한 다음에, 사용자에게 그 결과를 바탕으로 관련 카테고리를 우선순위에 따라 추천해주는 추천 에이전트 방식을 이용하는 것이다. 해당 카테고리를 추천 받은 사용자는 우선순위에 따라 하나이상의 카테고리에 저장하거나, 또 시간차에 의해 해당 카테고리가 변동이 될 경우 적절하게 메일을 관리할 수 있으므로 잘못된 메일 분류를 방지할 수 있을 것이다. 물론 메일의 양이 너무 많거나 추천에 대한 신뢰도가 만족될 때에는 자동분류를 할 수 있도록 사용자 인터페이스에 체크 박스를 설정하였다.

3.2 추천 에이전트의 특징

본 연구에서 제안하는 추천 에이전트의 특징은 크게 두 가지로 볼 수 있다. 첫째로 동적 임계치(dynamic threshold)를 이용한 베이지안 학습 알고리즘을 적용하여 시스템 관리의 핵심이 될 수 있는 분류의 정확도(accuracy)를 개선하였다. 둘째로는 특징 추출 및 룰 형성, 카테고리별 분류등 주요 기능 부분을 룰 필터링 컴포넌트(Rule filtering component)로

작성하여 본 시스템의 확장성과 재사용성을 높혔다. 다음 그림1은 시스템의 전체적인 흐름도로서, 사용자 정보를 기반으로 베이지안 학습에 의한 필터링, 분류, 추천과정을 보여주고 있다.

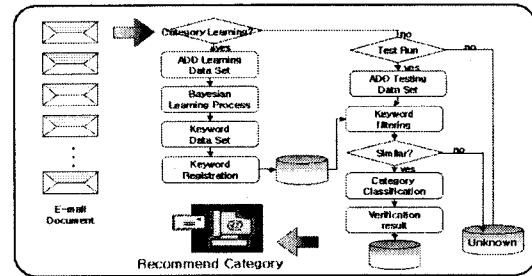


그림 1. 시스템의 전체적인 흐름도

3.3 룰 필터링 컴포넌트

이메일 관리를 위한 추천 에이전트 시스템은 다음과 같은 기능을 가진다.

첫째, 새로운 메일이 도착하면, 먼저 사용자의 메일 처리 과정을 관찰하여 학습한다. 특징 추출 및 룰(rule) 형성에 도움을 주는 모듈이며, 또한 사용자가 개인에 맞는 카테고리를 미리 설정할 수 있는 과정이다.

둘째, 메일 처리 과정에서 특징을 추출하여 응용된 베이지안 알고리즘을 적용하여 개인에 맞는 룰을 형성한다.

셋째, 생성된 룰을 기반으로 새로운 메일이 도착하면 카테고리별 분류를 한 다음 사용자에게 해당 카테고리를 우선순위별로 추천 해준다.

이런 주요 기능을 COM+로 작성하여 다른 응용 시스템과의 확장성을 고려하였으며 분산 환경에서도 편리하게 재사용될 수 있도록 설계하였다. 주요 인터페이스와 메소드는 다음 그림2와 같다.

```

IRuleFilter
{
    HRESULT SetDBOpen(BSTR bstrDBConnect,BSTR
bstrDBID,BSTR bstrDBPW);
    HRESULT MergeMail(BSTR bstrID,BSTR
bstrRuleName,BSTR bstrMailData);
    HRESULT CheckFolder(BSTR bstrID, BSTR
bstrMailData,[out,retval] BSTR *pRet);
    HRESULT CheckFolders(BSTR bstrID, BSTR
bstrMailData,[out,retval] BSTR *pRet);
}

SetDBOpen(DBOPEN, DBID, DBPW)
MergeMail(ID, RuleName, MailData);
CheckFolder(ID, MailData);
CheckFolders(ID, MailData);
  
```

그림 2. 룰 필터링 컴포넌트

3.4 동적 임계치를 이용한 베이지안 알고리즘

제안된 시스템의 두 번째 특징으로 기존의 고정된 임계치를 동적으로 개선하여 필터링의 정확도를 향상시켰다.

본 연구에서는 문서 분류를 위한 대표적인 교사 학습 알고리즘인 베이지안 학습 기법을 통하여 이메일 문서 분류가 이루어진다. C 를 [1]과 같이 전체 카테고리 집합이라고 하고, C_0 는 분류가

불가능 할 때, 다음과 같이 정의하자.

$$\text{Category Set } C = \{c_1, c_2, \dots, c_k\}, C_0 = \text{unknown category} [1]$$

이메일 문서들의 전체 집합을 D 로 한다면,

$$E-\text{Mail Document Set } D = \{d_1, d_2, \dots, d_n\} [2]$$

나이브 베이지안 분류법[2]에서 한문서 d_i 의 각 클래스 c_j 에 대한 조건부 확률을 식[3]과 같이 구해준다.

$$\mathcal{R}(d_i) = \{p(d_i | c_1), p(d_i | c_2), p(d_i | c_3), \dots, P(d_i | c_k)\} [3]$$

대부분 시스템에서는 분류 대상 문서에 대해서 식[4]와 같이 가장 높은 확률값을 가지는 클래스로 분류하게 된다.

$$P_{\max}(d_i) = P_{\max}\{(d_i) | (c_t)\}, t = 1, \dots, k [4]$$

그러나 본 연구에서는 기존의 베이지안 알고리즘을 이용되었던 고점임계치 T 를 식 [5]와 같이 동적임계치 T' 로 변환하였다.

$$C_{best}(d_i) = \begin{cases} \{c_j | P(d_i | c_j) = P_{\max}(d_i), \text{ if } P_{\max}(d_i) \geq T \\ \quad \text{where } T = 1 - \frac{P_{\max}(d_i)}{\sum_{i=1}^k P(d_i | C_i)} \} & [5] \\ c_0, \quad \text{otherwise} \end{cases}$$

4. 구현 및 실험 결과 분석

4.1 시스템의 구현

제안된 시스템은 별도의 메일 클라이언트 프로그램이 필요없는 웹 메일 서버를 기반으로 하였으며, 구현환경으로는 Windows 2000 Professional, 데이터베이스 컨트롤을 위해 MS-SQL 2000 Server, 를 형성 및 알고리즘 실행을 위해 MS visual C++ 6.0, 를 필터링 컴포넌트를 위해 COM+, 기타 기능을 위해 ASP, ASP 컴포넌트를 이용하였다. 다음 그림 3은 구현된 사용자 인터페이스이다.

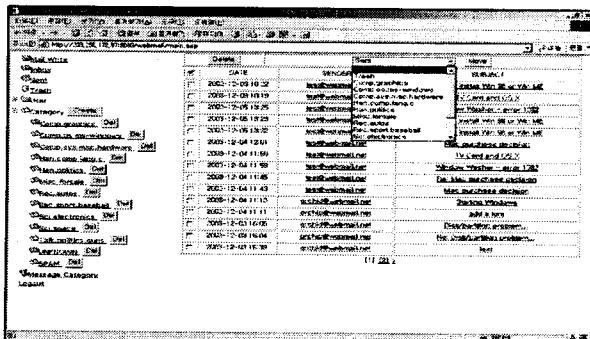


그림 3. 사용자 인터페이스

그림 3의 사용자 인터페이스는 사용자 관찰 과정에 이용되며, 실제 사용자가 카테고리를 생성 및 저장을 할 수 있다. 사용자가 자주 쓰는 카테고리를 생성하고, 필요없는 카테고리를 삭제할 수도 있다. 학습하는 과정에서 특정 추출하여 형성된 룰을 기반으로 메일 분류를 내부적으로 먼저 실행하여, 사용자에게 다음 그림 4와 같이 해당 카테고리를 확률값과 함께 추천하여 주는 것이다. 사용자는 추천된 카테고리를 참고로 메일을 해당 카테고리에 저장하게 되는 것이다.

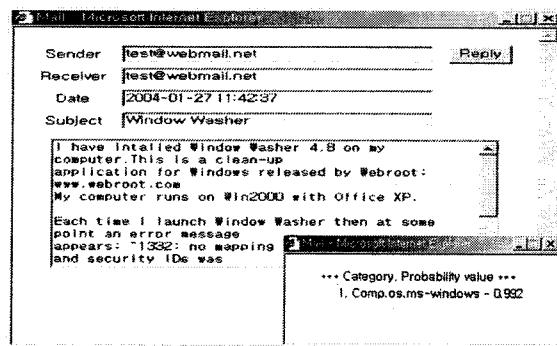


그림 4. 추천 카테고리

4.2 실험 결과 분석

정보 분류 시스템의 적합성은 일반적으로 재현율(recall ratio)과 정확률(precision ratio)이라는 척도에 의하여 측정된다. 재현율은 정보 분류 시스템에 들어있는 적합한 문헌 중에서 분류 시스템에 의하여 분류된 적합한 문헌의 비율이고, 정확률은 분류된 전체 문헌 중에서 적합한 문헌의 비율이다[4]. 본 연구의 성능 평가는 '사용자에게 얼마나 정확한 카테고리를 추천하느냐'인데, 이는 정확률과 재현율을 먼저 체크한 후, 메일 내용을 해당 카테고리에 맞게 분류하였는지를 실험하게 된다. 이 실험을 위해 12개의 카테고리를 미리 설정하고, 룰을 위해 샘플 데이터와 성능 평가를 위해 데이터를 수집하여 실행하였다. 각 카테고리당 1000통으로 학습 시킨 후, 받은 메일 문서 221,374통의 문서를 기준의 알고리즘으로 실행한 경우 정확도는 88.6%(196,137통)였고, 동적 임계치를 이용한 정확도는 89.5%(204,795통)로 측정되어 0.9% 향상을 보여주었다.

5. 결 론

본 논문에서는 이메일 사용자에게 도움이 될 수 있는 추천 이전트 시스템을 설계 및 구현하였다. 현재 이메일을 통해 많은 양의 정보들이 오가고 있고, 사용자들은 메일 관리에 편리한 맞춤 이메일 인터페이스를 요구하게 될 것이다. 자동 분류가 아닌 사용자가 최종 판단을 할 수 있는 추천 방식을 이용하였으며, 가장 핵심이 되는 필터링 부분을 컴포넌트로 구현하여 확장성과 재사용성을 높이고자 하였다. 향후 연구 방향으로는 카테고리를 사용자가 직접 설정하는 방법으로 현재 구현 되어있는데, 자동 카테고리 설정과 추천할 수 있는 방법이 동시에 가능한 에이전트로 확장시켜 나갈 것이다.

참 고 문 헌

- [1] M. Pazzani, D. Billsus , Learning and Revising User Profiles :The Identification of Interesting Web sites, Machin Learning 27, Kluwer Academic Puglishers, pp.313-331, 1997.
- [2] 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.
- [3] Chris Frye, "Undeerstanding Components," Andersen Consulting Knowledge Xchange, 1998.
- [4] 강승식, "한국어 형태소 분석과 정보 검색", 홍릉과학 출판사, 2002.