2004년도 대한전기학회 하계학술대회 논문집 2004.7.14-16

룰 필터링 컴포넌트 기반 이메일 추천 에이전트 시스템

정 옥란, 조 동섭
이화여자대학교 컴퓨터공학과

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

Ok-Ran Jeong, Dong-Sub Cho
Computer Science & Engineering Department , Ewha Womans University

Abstract - 본 연구에서는 갈수록 늘어나는 이메일 문서의 관리를 효율적으로 하기 위한 방법으로 새로운 메일이 도착했을 때 해당 카테고리를 추천받아 사용자가 직접 최적의 분류를 할 수 있는 이메일 추천 에이전트 시스템을 제안한다. 이메일 문서들의 카테고리별 분류 및 해당 폴더 저장에서 핵심이 될 수 있는 정확한 분류를 위해 동적 임계치를 이용한 베이지안 학습 알고리즘을 적용하였으며, 또한 주요 모듈 부분을 확장성과 재사용성을 위해 컴포넌트화 하였다.

1. 서 론

정보 기술의 발달로 사용자가 접할 수 있는 정보의 양은 기하급수적으로 늘어나고 있다. 이런 흐름에 따라 정보과잉속에서 사용자가 원하는 정보를 제공해주는 추천 시스템들이 빠르게 확산되고 있다. 그러나 대부분 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 적용분야는 다양하지만 아직까지 이메일 관리를 위한 연구는 시도되지 않았다. 따라서 본 연구에서는 기존의 텍스트 분류를 적용하여 자동 분류되는 메일 관리 시스템보다는 사용자의 여러 가지 상황에 따라 능동적으로 관리하며 사용자의 의견을 직접 반영할 수 있는 이메일 추천 에이전트 시스템을 제안하고자 한다.

이메일 시스템 관리시 학습을 통해 개인적 룰로 카테고리별 자동분류를 하더라도 개인적인 성향이 강한 이메일 사용자의 만족도를 높이기 위해서는 자동 분류 방식보다는 추천 방식을 겸할 수 있는 반자동 방식이 적합할 것이다. 따라서 우리는 분류의 핵심이 될 수 있는 오분류에 대한 해결책으로 두가지 접근방식을 제시한다. 첫째는 동적 임계치를 이용하여 분류 자체의 정확도(accuracy)를 높이는 알고리즘적 접근 방식과 자동 분류가 아닌 사용자가 최종 판단을 할 수 있는 추천 에이전트를 이용하는 방법론적 접근 방식이다. 이 접근 방식의 주요 모듈 부분을 룰 필터링 컴포넌트 기반으로 하여 시스템의 확장성과 재사용성을 높이고자 하였다.

2. 관련 연구

2.1 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 있으며, 메일 문서 분류에 적합한 학습알고리즘인 나이브 베이지안 기법을 이용하였다. 이 학습 기법은 베이즈 정리(Bayes theorem)에 기초한 확률 모델을 이용하였다. 이 방법에서는 분류하고자 하는 문서에 대한 벡터 모델을 입력하여, 분류 가능한 카테고리들 가운데 해당문서를 관찰할 수 있는 가능성이 가장 높은 클래스를 찾아 그 클래스에 분류하는 방식이다. 여기서 이용되는 가설은 문서들의 모든 속성이 주어진 카테고리 안에서 다른 문서의 전후 관계에 대해서 서로 독립적이라는 나이브 베이지안 가정을 적용한다. 본 연구에서는 적용된 베이지안 알고리즘은 기존의 고정된 임계값을 동적으로 변환하여 메일 문서 분류의 정확도를 높혔다.

2.2 분류 에이전트

현재까지 이메일 자동 분류 에이전트 시스템에 대한 연구는 MIT대학에서 만든 Maxims[2]이외에 많이 이루어지지 않았지만 문서 분류에 대한 연구는 여러 분야에서 활발하게 진행되어 왔다. 일반적으로 기계 학습법을 이용하여 분류 작업을 자동으로 수행할 수 있는 자율적인 소프트웨어를 분류 에이전트라 한다. 이와 같은 분류 에이전트의 대표적인 한 예로 카네기 멜론 대학의 Personal Webwatcher[3]가 있다. 이 분류 에이전트는 웹 브라우저를 통해 사용자의 행동을 모니터링하여 사용자의 관심 영역을 학습한 뒤, 브라우징하는 웹 문서내의 링크들에 대해 사용자 관심영역에 속하는 것들과 그렇지 않은 것들을 분류하여 관심 있는 링크들만을 제안 해 주는 시스템이다. 또한, 앤더슨 컨설팅 연구실에서 개발된 InfoFinder[3] 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[4]등이 모두 문서 분류기법을 이용한 대표적인 분류 에이전트 시스템이다. 또한 협업 필터링(collaborative filtering)를 이용한 성공적인 추천 시스템으로는 Tapestry, GroupLens, PHOKS[5]등이 있다. 이런 기존의 에이전트시스템과 추천시스템을 기반으로 본연구를 설계하였다.

2.3 컴포넌트의 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[6]. 제시된 컴포넌트의 정의를 기초로 하여 컴포넌트의 특성을 나타낼 수 있다. 특성으로는 식별가능성, 추적가능성, 교체가능성, 인터페이스에 의한 접근 가능성, 인터페이스가 제공하는 서비스 고정성, 문서로 정확히 기록되는 서비스, 물리적 구현의 은닉, 독립성, 언어 및 개발 툴에 독립적인 재사용성, 동적 재사용성을 제시할 수 있다. 또한 컴포넌트는 일반적으로 논리적인 관점과 물리적인 관점의 컴포넌트로 나눌 수 있다[6]. 컴포넌트의 특성을 기반으로 하여 논리적인

컴포넌토와 물리적인 컴포넌트의 개념을 제시하면 다음과 같다. 논리적인 컴포넌트는 비즈니스 컴포넌트를 의미하며, 비즈니스 영역에서 실세계의 개념을 모델링 한 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 재구현하여 확장성과 재사용성을 높이고자 하였다.

3. 이메일 추천 에이전트 시스템

3.1 이메일 분류 방식

기존의 자동 문서 분류 방법을 이메일에 적용하기에는 약간의 문제점이 있다. 이메일 문서는 개인적 성향이 강하게 개입되기 때문에 학습을 통해 형성된 개인적 룰로 카테고리별 자동 분류를 하더라도 사용자를 만족시키기는 한계점이 있다. 따라서 형성된 룰을 바탕으로 먼저 메일을 카테고리별 분류를 한 다음에, 사용자에게 그 결과를 바탕으로 관련 카테고리를 우선순위에 따라 추천해 주는 추천 에이전트 방식을 제안하는 것이다. 해당 카테고리를 추천 받은 사용자는 우선순위에 따라 하나 이상의 카테고리에 저장하거나, 또 시간차에 의해 해당 카테고리가 변동이 될 경우 적절하게 메일을 관리 할 수 있으므로 적합하지 않은 메일 분류를 방지할 수 있을 것이다. 물론 메일의 양이 너무 많거나 추천에 대한 신뢰도가 만족될 때에는 자동분류를 할 수 있도록 사용자 인터페이스에 체크 박스를 이용할 수 있도록 설계되었다.

3.2 제안된 시스템의 특징

제안된 시스템의 큰 특징은 크게 두가지로 요약할 수 있다. 첫째, 이메일 문서에서의 특징 추출 및 룰 형성, 카테고리별 분류등 주요 기능 부분을 룰 필터링 컴포넌트(Rule Filtering Component)로 작성하여 본 시스템의 확장성과 재사용성을 높였다. 둘째로는 동적 임계치를 이용한 베이지안 학습 알고리즘을 적용하여 시스템 관리의 핵심이 될 수 있는 분류의 정확도(Accuracy)를 개선하였다. 시스템의 전체적인 흐름은 다음 그림 1과 같다.

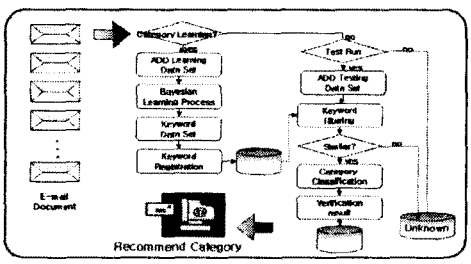

그림 1. 시스템의 전체적인 흐름도

그림 1은 전체적인 시스템의 흐름도로서, 사용자 정보를 기반으로 베이지안 학습에 의한 필터링 · 분류 · 추천 과정을 보여주고 있다.

이메일 관리를 위한 추천 에이전트 시스템은 다음과 같은 기능을 가지고 있는 세가지 모듈로 구성되어 있다.
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 관찰 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 해 준다.

이런 주요 기능을 COM+로 작성하여 다른 응용시스템과의 확장성을 고려하였으며 분산환경에서도 편리하게 재사용될 수 있도록 설계하였다. 주요 인터페이스와 메소드는 다음 그림 2와 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);
```

그림2. 룰 필터링 컴포넌트

제안된 시스템의 두 번째 특징으로 기존의 고정된 임계치를 동적으로 개선하여 필터링의 정확도를 향상시켰다. 본 연구에서는 문서분류를 위한 대표적인 교사학습 알고리즘인 베이지안 학습 기법을 통하여 이메일 문서 분류가 이루어진다. C를 [1]과 같이 전체 카테고리 집합이라고 하고, C_0는 분류가 불가능 할 경우 이다. 이메일 문서들의 전체 집합을 D로 한다면 [2]와 같이 정의할 수 있다.

$$Category\ Set\ C = \{c_1, c_2, \cdots, c_i\},$$
$$C_0 = Unknown\ Category \quad [1]$$
$$E-Mail\ DocumentSet\ D = \{d_1, d_2, \cdots, d_n\} \quad [2]$$

나이브 베이지안 분류법에서 한문서 d_i의 각 클래스 c_j에 대한 조건부 확률을 식[3]과 같이 구해준다.

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

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

$$P'_{\max}(d_i) = P_{\max}\{(d_i) \mid (c_t)\},\ t = 1, \cdots, k \quad [4]$$

그러나 본 연구에서는 기존의 베이지안 알고리즘을 이용되었던 고정임계치 T를 식 [5]에 의해서 동적임계치 T'로 변환하였다. 본 시스템의 성능 평가를 위해 동적임계치 T'를 적용했을때 향상된 정확도 결과를 보여주었다.

$$C_{bes}(d) = \begin{cases} \{c \mid P(d|c) = P'_{max}(d),\ if\ P_{max}(d) \geq T' \\ \qquad\qquad where T = 1 - \dfrac{P_{max}(d)}{\sum_{i=1}^{k} P(d|C)} \\ c_0, \quad otherwise \end{cases} \quad [5]$$

4. 시스템 구현 및 결과 분석

4.1 시스템 구현

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

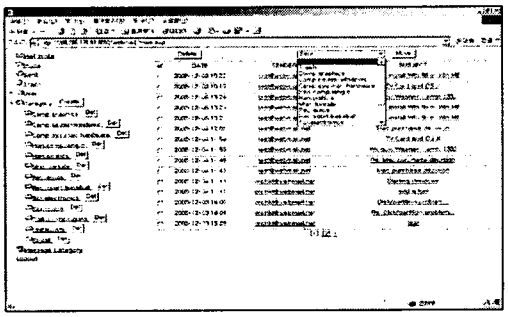

그림 3. 사용자 인터페이스

그림 3의 사용자 인터페이스는 사용자 관찰 과정에 이용되며, 실제 사용자가 카테고리 생성 및 저장을 할 수 있다. 사용자가 자주 쓰는 카테고리를 생성하고, 필요없는 카테고리를 삭제할 수도 있다. 학습하는 과정에서 특징 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.

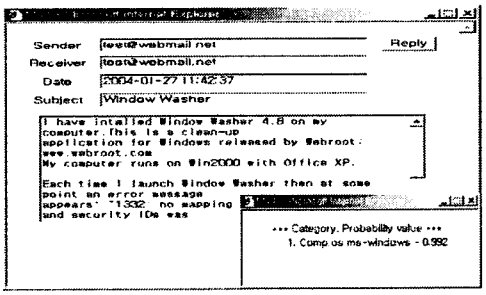

그림 4. 추천 카테고리

4.2 실험 결과 분석

정보 분류 시스템의 적합성은 일반적으로 재현율(recall ratio)와 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[7]. 본 연구의 성능 평가는 '사용자에게 얼마나 정확한 카테고리를 추천하느냐'인데, 이는 정확률과 재현률을 먼저 체크한 후, 메일 내용을 해당 카테고리에 맞게 분류하였는지를 실험하게 된다. 이 실험을 위해 12개의 카테고리를 미리 설정하고, 룰을 위해 샘플 데이터와 성능평가를 위해 데이터를 수집하여 실행하였다. 실제 실험은 많은 양의 데이터를 테스트해야 하기 때문에 카테고리별 만통 정도의 메일을 하나의 데이터 포맷으로 만들어 테스트 하였다. 이 실험의 데이터 포맷은 날짜, 보낸사람, 제목, 메일 문서 총 라인수, 빈라인, 실제 메일 내용, 4개의 빈라인으로 구성하였다. 이런 데이터 포맷을 이용하여 정확도를 측정하였다. 실험 결과 다음 그림 5과 같은 결과를 보여주었다. 각 카테고리당 1000통으로 학습 시킨 후, 받은 메일 문서 221,374통의 문서를 기존의 알고리즘으로 실행한 경우 정확도는 88.6%(196,137통)였고, 동적 임계치를 이용한 정확도는 89.5%(204,795통)로 측정되어 0.9% 향상을 보여주었다.

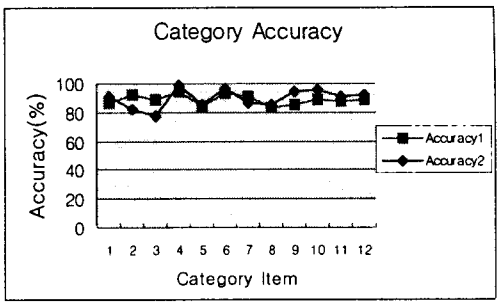

그림 5. 카테고리별 정확률

좀 더 많은 학습 데이터와 좀 더 긴시간 학습기간을 갖게 될 경우 정확도는 더 높아질 것이다.

5. 결론

본 논문에서는 이메일 사용자에게 도움이 될 수 있는 추천 에이전트 시스템을 설계 및 구현하였다. 현재 이메일을 통해 많은 양의 정보들이 오가고 있고, 사용자들은 메일 관리를 위해 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 구현되어 있는 방식을 자동 카테고리 설정과 동시에 추천할 수 있는 방법이 가능한 에이전트로 확장시켜 나갈 것이다.

이 논문은 2004년도 두뇌한국21(BK21) 사업에 의하여 지원되었음.

[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] P.Maes, "Agents that Reduce Work and Information Overload", Communications of the ACM, Vol.137, No.7, pp.30-40,1994.
[3] 백혜정, 박영택, 윤석환, "사용자 관심도를 이용한 웹 에이전트", 정보처리학회지, 1999.9.
[4] Jeffrey M.Bradshaw, "Software agent", AAAI Press/ The MIT Press, pp.151-161.
[5] Terveen, L., Hill, W., Amento, B., MacDonald, D. and Creter, J., PHOAKS: A System for Sharing Recommendations, CACM,40(3), 59-62
[6]Chris Frye, "Undeerstanding Components," Andersen Consulting Knowledge Xchange, 1998.
[7]강승식, "한국어 형태소 분석과 정보 검색", 홍릉과학 출판사, 2002.