온톨로지와 Semantic Enrichment를 이용한 스팸 메일 필터링 시스템

김현준°, 김흥남, 정재은, 조근식
인하대학교 컴퓨터공학부
{dannis°, nami4596, j2jung}@eslab.inha.ac.kr gsjo@inha.ac.kr

Spam Mail Filtering System using Ontology and Semantic Enrichment

Hyun-Jun Kim°, Heung-Nam, Kim, Jason, J. Jung, Geun-Sik Jo
School of Computer Science & Engineering, Inha University

요 약

최근 인터넷의 급속한 성장과 더불어 전자메일(E-Mail)은 의사교환의 필수적인 매체로 사용 되어지고 있다. 그러나 편리하고 비용이 들지 않는 장점을 이용해 엄청난 양의 스팸 메일이 매일같이 쏟아져 오고, 이를 해결하기 위한 다양한 연구들이 제시되어져 왔다. 특히, 문서 분류에 널리 쓰이는 베이지안 분류자(Bayesian classifier)가 가장 널리 이용되어지고 있는데, 정확도와 재현율에서 비교적 우수한 성능을 보이고 있다. 그러나 몇 가지 문제점을 갖고 있는데, 첫째, 사전에 사용자에 의해 스팸, 논스팸 메일에 대한 충분한 학습이 선행되어야 하는 점, 둘째, 필터링을 위한 연산시간이 소요되는 점, 셋째, 필터링의 대상이 되는 메일 본문의 내용이 적을 경우 정확한 필터링이 어렵다는 점 등의 문제점이 있다. 본 논문에서는 마지막 문제점으로 지적된 메일 본문의 내용이 적을 경우 즉, 연산을 위한 특징적인 단어들의 부족으로 정확한 분류가 불가능한 경우의 해결방안으로 온톨로지와 Semantic Enrichment 기법을 이용한 스팸 메일 필터링 시스템을 제안한다. 실험 결과, 제안하는 시스템이 베이지안 분류자를 이용한 분류 시스템보다 정확도에서 4.1%, 재현율에서 10.5%, 그리고 F-measure에서 7.64%의 성능향상을 보였다.

1. 서 론

최근 인터넷의 성장과 더불어 전자우편(E-Mail)은 현재 통신, 정보, 의사 교환의 필수적인 매체로 사용 되어지고 있다. 그러나 송, 수신에 있어 편리하고 비용이 들지 않는 장점을 이용하여 많은 개인이나 업체들은 자신들의 상업적 광고를 무차별적으로 발송하고 있으며, 그 양은 매년 증가하고 있는 추세이다 [1]. 따라서 메일 서비스 업체들은 저장장치의 용량부족 등의 문제를, 일반 사용자들은 쏟아져 들어오는 상업성 광고 및 불법, 음란광고로 인해 자신의 계정부족 및 스팸 메일을 지우는데 시간을 투자하는 등의 불편을 겪고 있다. 최근 초고속 인터넷 서비스 업체에 의한 조사 결과 전체 메일의 약 84.4%가 스팸 메일이었다 [2].

이러한 문제의 해결을 위한 방법의 한가지로 전자문서 분류에 많이 이용되는 베이지안 분류자(Bayesian Classifier)는 정확도와 재현율에서 비교적 정확한 성능을 얻을 수 있는 대표적인 방법이다. 하지만 분류를 위해서는 사전에 사용자에 의한 학습이 필요하며, 분류 시에 시간이 소요되는 문제, 그리고 내용이 짧은 메일의 경우 분류를 위한 데이터가 부족하여 결국 분류의 정확도가 저하되는 문제를 갖고 있다. 본 논문은 마지막 제시된 문제를 해결하기 위해 온톨로지(Ontology)와 데이터베이스 분야에서 적용되어져 왔던 Semantic Enrichment 기법을 제안하고 사용한다. 즉, 적은 수의 단어만을 갖는 분류대상 메일에 대하여, 보다 내용의 의미를 정확히 파악 가능하도록 관련 단어들을 추가하여 분류를 실행함으로서 정확도와 재현율을 향상시키는 방법을 제안하고 실험으로 성능을 증명하였다.

2. 스팸 메일 필터링 성능 향상을 위한 방법론들

2.1 전처리 단계(Pre-processing)

베이지안 분류자를 이용하여 메일을 분류하기 이전에 처리의 효율성을 위해 필요한 단계이며, 보통 다음과 같은 전처리 방법들을 사용하게 된다 [3].

- 텍스트 어휘 분석 (Text Lexical Analysis) : 기본적으로 문서를 단어들로 변환하는 과정이며, 숫자, 하이픈, 문장부호, 대소문자 등의 구별을 함으로서 문서에서 의도하는 의미들로 순수하게 인식할 수 있도록 구별하는 방법이다.
- 불용어 제거 (Elimination of Stopwords) : 문서에서 빈도수가 너무 높은 단어들은 오히려 변별력을 저하시키는 원인이 될 수 있다. 보통 특정 문서집합에서 80% 이상의 문서에서 공통으로 출현한 단어는 검색이나 분류를 위해 의미 없는 것으로 간주될 수 있는데 이를 불용어(Stopword)라 한다. 일반적으로 전치사, 관사, 접속사 등은 불용어로 간주할 수 있다.
- 스테밍 (Stemming) : 스템(Stem)이란 단어에서 접두사와 접미사를 제거하고 남는 부분이다. 보통 사용자에 의한 질의되어진 단어와 검색시의 관련 문서들에 포함된 단어가 실제 내용은 일치하지만, 구문적 변형으로 인해 검색 결과의 성능을 저하시키는 원인이 될 수 있는데, 이러한 문제는 각 단어들을 스템으로 대체하면 해결 할 수 있다.
- 시소러스 (Thesaurus) : 시소러스란 용어의 사용법과 용어들 사이의 관계에 대한 정보를 제공하는 어휘 도구를 말한다. 즉, 정보 검색을 위한 키(색인)와 단어간의 관계, 즉 동의어, 하위어(그 색인에 속하는 용어), 관련어 등의 관계를 나타낸 색인표를 통하여 검색시 질의에 포함된 용어의 의미를 확대하기 위해 주로 사용된다.

2.2 베이지안 분류자 (Bayesian Classifier)

베이지안 분류자는 주어진 데이터에 대해 이미 학습된 정보를 바탕으로 분류를 하는 확률적인 방법론으로서, 일반적으로 정보검색, 문서 분류에 많이 사용되어져 왔으며 식 (1)과 같다.

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)} \quad (1)$$

여기서, P(H)는 H의 사전확률(Priori Probability)이고, P(H|D)는 D가 주어졌을 때 H의 사후확률(Posterior Probability) 이다 [4]. 주어진 데이터 D는 분류의 대상이 되는 메일이며, P(H|D)는 사용자에 의해 스팸, 논스팸 메일에 대해 학습한 내용을 바탕으로 얻게 된 확률을 의미한다.

2.3 온톨로지 (Ontology)

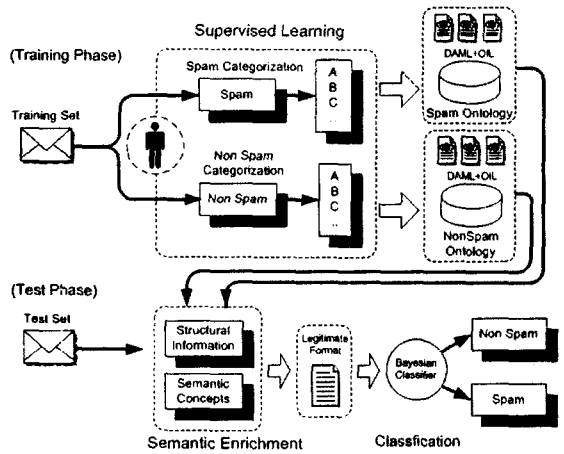
온톨로지는 웹 기반의 지식 처리나 응용 프로그램 사이의 지식 공유, 재사용들을 가능하게 하는 아주 중요한 요소로서, 특정 도메인 내의 지식들을 개념화하고 이를 명세화함으로서 어플리케이션 간의 정보 공유와 재사용을 도울 수 있도록 하는 것이다. 문서 분류나 본 논문의 스팸 메일 필터링 시스템의 경우, 컴퓨터가 본문의 내용을 이해하고 연산을 수행할 수 있도록 도와주는 필수적인 구성요소중의 하나라 할 수 있다.

2.4 Semantic Enrichment

Semantic Enrichment란 서로 다른 이종의 데이터베이스를 호환성 있게 사용하기 위하여 연구되어진 방법이다 [5]. 그러나 최근 시멘틱 웹의 등장과 함께 정보검색 분야에서도 적용되어 사용되고 있는 기법이며, 본 논문의 시스템에서는 텍스트 기반의 메일이 온톨로지를 기반으로 한 분류에 적합하도록 그 형태를 변환시키고 의미적으로 관련 있는 단어들을 추가시키는 역할을 하게 된다.

3. 온톨로지와 Semantic Enrichment를 사용한 스팸 메일 필터링 성능 향상 기법

제안하는 시스템은 [그림1]과 같으며, 학습데이터를 통한 학습과 온톨로지 구축 단계, 그리고 분류를 위해 대상 메일을 적합한 형태(Legitimate Format)로 변환시켜 분류하는 과정의 두 단계로 이루어진다.



[그림1] 시스템 구조도

3.1. 학습단계 (Training Phase)

보통 문서 분류, 스팸 메일 필터링을 위해서 사전에 학습이 필요하다. 본 논문에서는 학습단계를 위해 기본적으로 베이지안 분류자를 사용하였으며, 기존에 단순히 스팸/논스팸에 대한 학습이 이루어졌던 것과는 다르게 사용자에 의해 스팸/논스팸의 여부뿐만 아니라, 각각의 하위 항목까지도 지정해 주도록 하였다. 즉, 학습되어지는 메일은 보다 세부적인 분류까지 이루어지며, 곧바로 DAML+OIL 형식을 통해 [그림2]와 같이 온톨로지를 구성하게 된다.

```
<daml:Class rdf:about="#Email_add">
    <rdfs:subClassOf rdf:resource="#Spam"/>
</daml:Class>
<daml:Class rdf:about="#Adult">
    <rdfs:subClassOf rdf:resource="#Spam"/>
</daml:Class>
        .....
<daml:ObjectProperty rdf:about="#Email_add">
    <rdfs:domain rdf:resource="#Spam"/>
    <rdfs:range rdf:resource="#Spam"/>
</daml:Class>
        .....
```

[그림2] DAML+OIL을 이용한 Spam 온톨로지의 예

3.2. 분류단계 (Test Phase)

학습을 통해 스팸/논스팸 온톨로지가 구성되어진 후, 새로운 메일이 도착하였을 때 시스템은 두 단계로 구성된 Semantic Enrichment과정을 통해 베이지안 분류자에 의한 분류가 잘 이루어질 수 있도록 적합한 형태의 문서를 만들게 된다 [6].

- **Structural Information** : 우선 텍스트 기반의 메일을 RDF로 변환시키는 역할을 하게 되며, 이렇게 변환된 문서가 기존에 DAML+OIL로 구성된 온톨로지의 어떤 클래스에 해당하는지를 판단하게 된다. 즉 스팸/논스팸 온톨로지의 각 클래스와의 유사도 판별을 통해 해당 클래스로 추정되는 후보 클래스 $C_{i,spam}$ 과 $C_{i,NonSpam}$ 를 추출하게 된다.
- **Semantical Concepts** : 온톨로지 내의 개념들을 문서 내의 특정 단어들과 매칭시키고 관계를 설정하는 단계이며, 위의 단계를 통해 후보로 선정된 클래스들의 리소스를 이용하여 대상 메일에 포함된 관련된 단어들을 추가(Enrichment)시킨다. 이렇게 함으로써 문서가 지니는 단어들에 대한 의미적으로 관련 있는 단어들이 풍부해 지게 되며, 이는 곧 분류에 적합한 형태(Legitimate Format)의 문서가 되는 것이다. 예를 들어, [표1]은 스팸/논스팸 두 가지의 온톨로지가 사용자에 의해 구성이 되어 있을 때의 각 온톨로지가 포함하는 단어(Terms)와 각 단어들에 대한 학습 빈도수(Frequency)를 나타내고 있다. 임의의 메일 D_i = {d1, d2}, i=spam에 대하여, T_{spam}의 경우, 해당 단어들의 학습된 횟수에 의하여 T_{spam}=6, $T_{NonSpam}$=8의 결과를 얻게 되며, 이는 곧 베이지안 분류자에 의해 논스팸으로 분류되는 원인이 되게 된다.

[표1] 스팸/논스팸만을 학습한 결과

	T_{spam}				$T_{NonSpam}$			
Terms	d1	d2	d_3	d_4	d_1	d_2	d_3	d_4
Freq.	4	2	1	3	5	3	2	2

그러나 [표2]의 경우와 같이, 사용자에 의한 학습단계에서 스팸/논스팸에 대한 학습뿐만이 아닌, 각 분류의 세부적 클래스까지 학습이 된 결과로 온톨로지가 구성된 경우, 우선 시스템은 분류 대상 문서 D_i 의 속성들 d_i를 대상으로 스팸/논스팸의 후보 클래스를 선정하게 된다. T_{spam}={C1}, $T_{NonSpam}$={C2 , C3 , C4}로 선정이 되어진 후, 본문의 내용을 추정이 가능한 형태의 문서를 만들기 위해 시스템은 Enrichment 과정을 수행하게 된다 [그림3]. 즉, 각 후보 클래스 C_i의 속성들을 D_i에 추가함으로

서 베이지안에 의한 결과에 신뢰도를 높일 수 있는 방식이다.

[표2] 스팸/논스팸 각각의 클래스까지 학습한 결과

Category / Terms	T_{spam}				$T_{NonSpam}$			
	C_1	C_2	C_3	C_4	C_1	C_2	C_3	C_4
d_1	3	0	0	1	1	2	1	1
d_2	1	0	1	0	0	1	1	1
d_3	0	1	0	0	0	1	0	1
d_4	2	0	1	0	0	1	1	0
Total	6	1	2	1	1	5	3	3

Spam : $C_1 = \{d_1, d_2, d_4\} = \{3,1,2\} = \underline{6}$

NonSpam : $C_2 = \{d_1, d_2, d_3, d_4\} = \{2,1,1,1\} = \underline{5}$
$C_3 = \{d_1, d_2, d_4\} = \{1,1,1\} = 3$
$C_4 = \{d_1, d_2, d_3\} = \{1,1,1\} = 3$

※ d_i 는 후보클래스를 통해 Enrichment 된 속성들

[그림3] 후보 클래스를 통한 Enrichment과정

[그림3]과 같이 스팸/논스팸 최종 후보 클래스는 각각 Spam의 경우 C_1=6, NonSpam의 경우 C_2=5가 되며, 베이지안 분류자를 통해 D_i 은 스팸으로 분류된다. 결국, 본문에 대한 의미를 각각의 온톨로지를 사용함으로써 컴퓨터가 인식하게 되며, 단순 베이지안 분류시 보다 정확한 결과를 얻을 수 있게 된다.

4. 실험 및 결과

4.1 실험환경

본 논문의 실험을 위해서 IIS 5.0, Micorsoft Active Server Page 와 MS-SQL Server을 사용해서 구현하였으며, 실험환경은 펜티엄4 2.4GHz, 256MB RAM의 시스템이었다.

트레이닝 및 테스트에 사용된 데이터들은 수집된 실제 영문 메일이며, 데이터들의 구성은 [표3]과 같다. 테스트에 쓰인 모든 데이터는 트레이닝 데이터보다 나중에 온 메일이다.

[표3] 데이터셀의 구성

	Training Data	Test Data
Spam	329	148
NonSpam	247	53
Total	576	201

4.2 실험 평가 기준

제안하는 시스템의 분류 성능을 평가하기 위하여 정확도(Precision), 재현율 (Recall)과 F1-measure 측정식을 이용하였으며 각각의 정의는 다음과 같다.

$$\text{Precision} = \frac{\text{스팸으로 분류된 실제 스팸메일 수}}{\text{스팸으로 분류된 메일 수}} \quad (2)$$

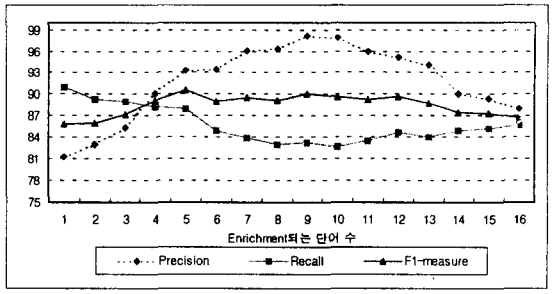
$$\text{Recall} = \frac{\text{스팸으로 분류된 실제 스팸메일 수}}{\text{전체 스팸메일 수}} \quad (3)$$

$$\text{F1-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad (4)$$

여기서, 식 (4)의 F1-measure의 값은 클수록 분류 성능이 우수함을 의미한다.

4.3 실험결과 및 분석

실험은 Enrichment 과정에서 추가되는 단어 개수에 대한 최적의 임계값을 얻기 위한 실험[그림4]와 기존의 베이지안 분류자에 의한 시스템과의 성능비교[표4] 두 가지로 행하여 졌다.



[그림4] 최적의 Enrichment를 위한 임계값 변화에 의한 실험

위의 그림과 같이 Enrichment를 위한 최적의 임계값은 단어 수가 9일때, 즉 추가되는 의미상 유사 단어가 5개까지 일때 가장 좋은 성능을 보였으며, 이 임계값을 사용하여 실험한 결과와 기존의 베이지안 분류자만을 이용한 시스템과의 분류 성능 결과는 [표4]와 같다.

[표4] 베이지안 분류자와 제안하는 시스템의 성능 비교

		Precision	Recall	F1-measure
1	Bayesian Classifier	89.21%	77.5%	82.94%
2	Semantic Enrichment	93.33%	88%	90.58%
Improved Performance		(2) 4.1%	(2) 10.5%	(2) 7.64%

정확도와 재현율, 그리고 F1-measure에 의한 측정치 모두가 베이지안 분류자를 이용한 시스템보다 각각 4.1%, 10.5%, 7.64% 향상된 결과를 얻을 수 있었는데, 이는 시스템으로 하여금 본문의 내용을 이해할 수 있도록 함으로써 분류의 성능을 향상시킬 수 있음을 의미한다.

5. 결론 및 향후 연구

본 논문은 기존에 데이터베이스 분야에 사용되어진 Semantic Enrichment 기법을 온톨로지와 함께 스팸 메일 필터링 시스템에 적용하고 실험을 통해 성능을 평가하였다. 분류가 모호한 대상 메일에 대해서는 컴퓨터로 하여금 본문의 내용을 이해할 수 있게 함으로서 기존의 베이지안 분류자를 이용한 시스템보다 향상된 결과를 얻을 수 있었다. 그러나 시스템의 성능에 큰 영향을 줄 수 있는 온톨로지 구축이 사람에 의해 이루어지기 때문에 구축당시 상당한 노력을 필요로 하는 점, 환경과 시간의 변화에 따른 시스템 갱신이 어려운 점, 그리고 분류에 시간이 소요되는 점 등은 향후 연구를 통해 해결되어질 것이다.

[참 고 문 헌]

[1] 한국전산원, "국가정보화백서(National Informatization White Paper)," pp 23, 2002.

[2] 한국통신, www.kt.co.kr, 2003.

[3] Ricardo, B.-Y. and Berthier, R.-N., Modern Information Retrieval, pp.27, Addison-Wesley, 1999.

[4] Mitchell, T. M., Machine Learning, Chapter 6: Bayesian Learning, McGraw-Hill, 1997.

[5] Hohenstein, U., Plesser, V., "Semantic Enrichment: A First Step to Provide Database Interoperability," In Proc of the Workshop Föderierte Datenbanken, Magdeburg, 1996.

[6] Saias, J., Quaresma, P., "Semantic enrichment of a web legal information retrieval system," In Proc. of the JURIX2002, volume 89 of Frontiers in AI and Applications, pp 11-20, London, UK, 2002.