확률 분포와 추론에 의한 이메일 분류 및 정리 방법

고남현[O], 김지윤[*], 최만규[**]
[O]한국방송통신대학교 컴퓨터과학과,
[*]한양대학교 컴퓨터소프트웨어학과
[**]인천대학교 동북아물류대학원
e-mail: gnh1201@gmail.com[O], kjycslab@gmail.com[*], zzigirang@naver.com[**]

Classification and Allocation method of e-mail using possibility distribution and prediction

Nam-Hyeon Go[O], Ji-Yun Kim[*], Man-Kyu Choi[**]
[O]Dept. of Computer Science, Korea Open National University
[*]Dept. of Computer Software, Hanyang University
[**]Graduate school of Logistics, Incheon National University

● 요 약 ●

본 논문에서는 디리클레 분포와 베이즈 추론 모델을 활용하여 전자우편을 분류하고 정리하는 방법을 제안한다. 과거 원치 않는 광고성 이메일인 스팸 탐지에서 시작한 전자우편 분류는 지속적인 송수신 량의 증가와 내용의 다양화로 인해 광고성과 정보성의 판단 기준이 모호해진 상태이다. 스팸 탐지와 같은 이분법적 분류 방식이 아닌 내용의 주제 별로 자동 분류할 수 있는 방법이 필요하다. 본 논문에서 다루는 제안 기법은 전자우편의 내용에서 다뤄질 수 있는 주제의 종류를 예측하기 위한 방법을 제공한다. 발신하거나 수신된 전자우편이 속한 주제를 자동으로 정할 수 있다. 본 제안 기법의 활용을 통해 전자우편의 분류만이 아닌 업무 및 시장 동향 분석과 정보보안 분야에서는 악성코드 분류에 사용될 수 있을 것으로 기대된다.

키워드: 인공지능(Artificial intelligence), 전자우편(E-mail), 정보보안(Information security), 디리클레 분포(Dirichlet distribution), 베이즈 이론(Bayes theory)

I. 서론

본 논문에서는 디리클레 분포와 베이즈 추론 모델을 활용한 전자우편 분류 방법을 제안한다. 내용의 다양성을 나타내는 전자우편 및 문서를 알고리즘을 통해 분류할 수 있는지를 확인한다. ‘

II. 관련 연구

1.1 잠재 디리클레 할당

잠재 디리클레 할당(LDA)은 문서의 주제를 찾는 발생적 모델이다. 문서에서 등장할 수 있는 주제들이 디리클레 분포를 따른다는 가정을 기반으로 한다. 찾고자하는 주제의 수를 지정하고 각 주제에서 나올 수 있는 단어를 추출하는 과정이 이뤄진다.[1][2]

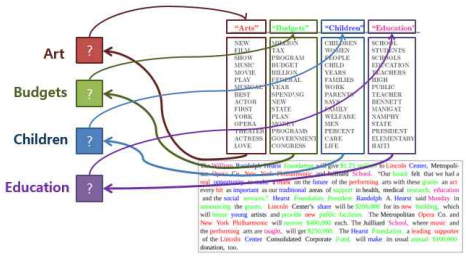

Fig. 1. Process of Latent dirichlet allocation

1.2 베이즈 추론 모델

베이즈 추론 모델은 기존의 정보를 토대로 얻은 사전 확률을 활용하여 사후 확률을 개선한다. 개선된 사후 확률로 새로운 정보의 예측을 진행한다.[3]

Fig. 2. Increase accuracy by Bayes theorem

III. 제안 기법

본 기법에서는 주어진 전자우편에서 발생 가능한 주제의 종류를 예측하고 분류한다. 정보 수집과 전처리부터 알고리즘 적용에 이르는 과정이 포함된다.

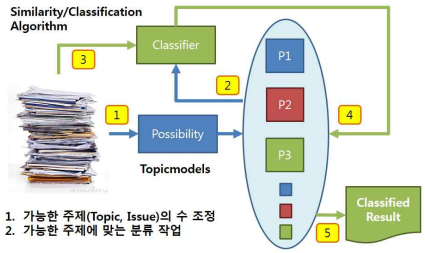

Fig. 3. All process map of proposal technique

3.1 정보의 수집과 전처리

정보를 수집하고자 하는 파일의 형식에 맞는 처리기로 본문을 추출한다. 자연어에 위배되는 문장(예, Tags)은 제거한다. 언어에 맞게 정의된 말뭉치를 활용하여 형태소 분석과 불용어 제거를 진행한다.

3.2 단어 출현의 수치화(행렬화)

추출된 단어는 단순히 저장만이 아닌 알고리즘에 의해 계산 가능한 형태로 변환되어야한다. 이러한 목적을 달성하는 방법으로는 문서와 단어 사이를 관계를 행렬(DT-Matrix)로 표현하는 방법이 있다.

3.4 데이터 세트의 확장

문서의 분류에 활용할 수 있는 자료 집합은 두 가지로 나뉜다. 사전 학습 자료와 예측의 정확성을 시험해보기 위한 시험 자료가 있다. 두 자료의 지속적 비교는 지속적인 학습량의 증가로 이어지게 된다.

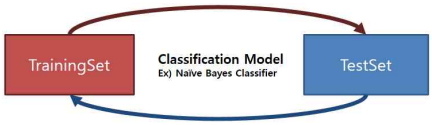

Fig. 4. Training structure of testing and training set

3.5 예측 알고리즘의 적용

전처리를 통해 알고리즘에 의한 학습과 연산이 가능한 형태로 변환된 자료를 활용하는 과정이다. 해당 과정을 통해 분류되지 않았던 수많은 전자메일이 다루는 의미를 파악한다.

IV. 실험 결과

본 실험은 추출하고자 하는 주제의 지정 개수를 10가지로 지정하여 진행되었다. 주제별로 등장하는 단어와 전체 분석 대상 문서에서 한 주제가 차지하는 비중의 정도를 확인하였다.

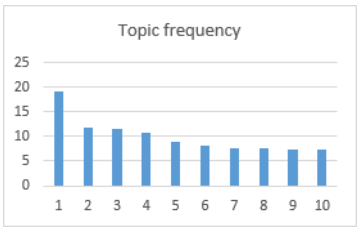

Fig. 5. Graph of frequency by topics

주제별 주요 내용은 아래와 같이 확인되었다. 첫 번째 표는 추출된 주제 중 3가지에 대한 상위 키워드 결과를 나타내고, 두 번째 표는 실험 환경이다.

Table 1. Example of recognized key words

Topic	Key words
5	브라우저;애플리케이션;로그인 상태;회원
4	관심 기업;인크루트;대비 공채;소식지
9	변경;업데이트;바이러스;이메일;신용등급

Table 2. Experiment Environments

Item	Value
Number of samples	64,828 samples
Size of samples	3,517,583,360 bytes
Character encoding	UTF-8, CP949(KR)

V. 결론

본 실험은 전자메일에서 발생할 수 있는 주제의 수를 예측하고 주제에 맞게 분류할 수 있음을 확인하였다. 이는 광고성과 정보성의 기준이 모호한 전자메일의 효과적인 분류에 효용이 있다. 따라서 본 결과는 다양한 문서 분류에 활용될 수 있다.

References

[1] D. Blei, "Probabilistic Topic Models. Communications of the ACM", Vol. 55, No. 4, 2012.

[2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, Volume 3, pp. 993-1022, Jan 2003.

[3] KP. Murphy, "Naive Bayes classifier", Oct 24, 2006 (https://www.cs.ubc.ca/~murphyk/Teaching/CS340-Fall06/reading/NB.pdf)