이력서의 Classification을 위한 Feature Selection 방안

이만유*, 조형석*, 이유진**, 홍지원**, 김상욱**
*한양대학교 공과대학 컴퓨터전공
**한양대학교 컴퓨터 소프트웨어 학과
e-mail : {e_manu, ognboss1}@naver.com,
{yujinlee, nowiz, wook}@hanyang.ac.kr

An Approach to Feature Selection for Classification of Resume

Manyu Lee*, Hyungsuk Cho*, Yu-jin Lee**, Jiwon Hong**, Sang-Wook Kim**
*Dept. of Computer Science & Engineering, Hanyang University
**Dept. of Computer and Software, Hanyang University

요 약

사람이 수많은 지원자의 이력서들을 모두 꼼꼼히 읽는 데에는 엄청난 시간과 노동이 필요하다. 만약 컴퓨터가 이력서를 알맞은 직군으로 분류해 줄 수 있다면 이러한 어려움을 해소할 수 있다. 이를 위해 본 논문에서는 알맞은 직군으로 분류하기 위한 이력서를 학습할 때에 feature를 어떤 방법으로 선택할 수 있는지 그리고 feature의 개수는 몇 개가 적절한지에 대해 알아본다.

1. 서론

이력서는 기업과 구직자에게 있어 중요한 문서이다. 하지만 기업이 수많은 지원자의 이력서를 일일이 읽고 판단하는 것은 어려운 일이다. 만약 컴퓨터가 수많은 이력서들을 직군별로 분류해 줄 수 있다면 편리할 것이다. 구직자 또한 자신의 이력서와 가장 잘 맞는 직군을 알 수 있다면 직장을 선택하는데 많은 도움이 될 것이다. 이를 위해 본 논문에서는 이력서를 알맞은 직군으로 분류하기 위한 학습에 이용 될 수 있는 feature에 대해 다룬다.

2. 본론

본 논문에서는 이력서의 여러 요소들을 무시하고 단순히 출현되는 단어를 이용해 이력서를 학습한다. 이력서에는 학교, 직장, 자기소개 등 여러 요소가 존재하지만 양식이 모두 다르고 항목이 제 각각인 이력서들을 일반화하여 학습하는 것은 어려움이 있기 때문이다. 또한 본 논문에서는 한 직군을 문서로 취급하여 단어들에 대한 IDF를 측정한다. 이는 각각의 이력서를 한 문서로 취급할 경우 직군이 아닌 이력서를 분류 할 수 있는 단어에 가중치가 부여되기 때문에 같은 직군의 이력서들을 한 문서로 취급하여 여러 직군 속에서 한 직군을 특정 지을 수 있는 단어에 가중치를 부여하기 위함이다.

2.1 기존의 방안

TF(Term Frequency) 방안은 이력서의 모든 단어를 feature로 이용하고, 각 이력서에 대한 단어의 빈도수를 value로 이용한다. 같은 직군에 속한 이력서들은 비슷한 단어들이 많이 사용될 수 있으므로 특정단어의 빈도수가 많은 이력서들은 같은 직군에 포함된다고 볼 수 있다 [1].

TF-IDF 방안은 기존 TF 방안의 value에 역문서 빈도 IDF(Inverse Document Frequency)를 곱해준다. a, is, the, of, for 등 어느 이력서에서나 무분별하게 자주 등장하는 단어의 가중치를 낮춰주기 위해 IDF를 이용한다. IDF는 식 (1)과 같다 [2].

$$IDF(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|} \quad (1)$$

$|D|$: 문서집합 D의 크기
$|\{d \in D : t \in d\}|$: 단어 t가 포함된 문서의 수

2.2 제안하는 방안

제안하는 방안 에서는 모든 단어를 feature로 이용하는 것이 아니라, 직군별 단어의 중요도를 산출하여 해당 직군을 특징지을 수 있는 단어들을 feature로 이용한다. 생년월일, 출신학교, 지역이름 등 직군과 관련이 없는 단어들은 학습에 좋지 않은 영향을 끼칠 가능성이 크기 때문이다. 본 논문에서 support와 중요도의 개념은 식 (2)와 같다.

$support$: 특정 단어가 존재하는 이력서의 개수

$$\text{중요도} : \frac{\text{본 직군에서 단어에 대한 } support}{\text{모든 직군에서 단어에 대한 } support} \quad (2)$$

단어의 중요도는 본 직군의 support에 비례하고, 타 직군의 support에 반비례하여 본 직군의 속한 이력서에 등장하는 빈도가 높을수록, 타 직군의 이력서에 등장하는 빈도가 낮을수록 중요도가 증가한다. 또한 각 단어에 대한 value는 단어의 빈도수가 아닌 단순히 단어의 유무에 따른 boolean값으로 이용한다. 이는 단어의 중요도를 빈도수와

무관한 단어의 유무만으로 산출하였기 때문이다.

3. 실험

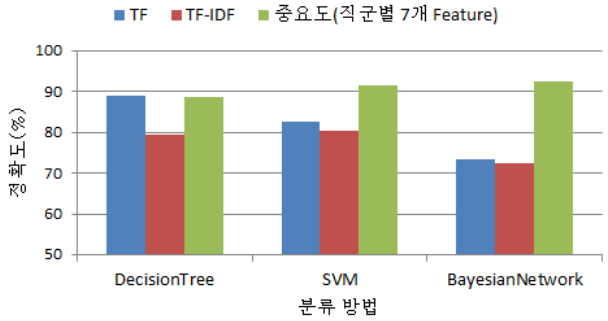
이 장에서는 2장의 방안을 이용하여 이력서를 학습한다. 제시한 방안 중 어떤 feature selection 방안이 가장 높은 정확도를 보이며 중요도를 통해 직군 별로 feature를 선택 할 때 적당한 개수는 몇 개인지 확인한다.

3.1 실험 환경

구인구직 사이트인 Open Resume와 LinkedIn에 실제 등록된 30만개의 이력서를 기반으로 실험을 진행하였다. 총 10개의 직군과 각 직군별 100개의 이력서로 총 1000개의 이력서를 이용하였고 직군이 라벨링 되지 않은 이력서들에 직접 직군을 라벨링 하였다. 오픈소스 WEKA를 이용하여 정확도를 측정하였고 이때 메모리는 4Gbyte를 할당하였다 [3]. 직군은 account, CEO, consultant, lawyer, doctor, network, product, sales, web, software engineer로 분류된다. 또한 분류 방법으로는 decision tree, SVM, bayesian network를 이용하였다 [4].

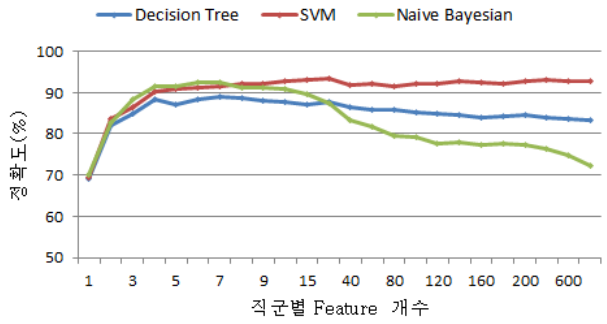
3.2 실험 결과 및 해석

그림 1은 TF, TF-IDF, 중요도 방안으로 얻은 feature를 이용하여 decision tree, SVM, bayesian network 세 가지의 분류기를 학습한 결과에 대한 정확도를 막대그래프로 나타낸 그림이다. 그림 1을 통해 중요도를 이용한 bayesian network의 학습이 정확도가 가장 높은 것을 알 수 있다. 반면 TF-IDF를 이용한 bayesian network 학습은 전체 실험 중 가장 낮은 정확도를 보인다. 이를 통해 단어를 통해 이력서를 학습할 때에 효과적인 feature를 사용하는 것이 얼마나 중요한지를 알 수 있다.



(그림 1) 기존방안과 제안하는 방안의 학습 결과

그림 2는 중요도 방안의 직군 당 feature 개수에 따른 정확도의 변화를 보여준다. SVM의 경우 feature 의 개수 변화와 상관없이 일정한 수준의 정확도를 유지하는 반면 decision tree와 bayesian network는 feature의 개수가 증가할수록 정확도가 떨어진다 [4]. 특히 bayesian network는 decision tree보다 정확도가 급격히 줄어드는데, 이는 decision tree가 자체적으로 효과적인 feature에 우선순위를 두는 특성 때문으로 보인다 [5]. 또한 decision tree와 bayesian network에서 직군별 중요도가



(그림 2) feature 개수변화에 따른 정확도

높은 7-8개의 feature를 이용한 학습의 정확도가 가장 높다. 이를 통해 이력서에 등장하는 모든 단어들 중 70개에서 80개 사이의 단어가 직군과 연관성이 있는 단어라는 것을 알 수 있다.

4 결론

본 논문에서는 이력서를 학습하여 직군을 분류 할 때 단어를 이용한 feature selection 방안에 대해 다뤘다. 이력서에 적용하기 어려운 기존 방안을 보완하여 중요도의 개념을 이용한 효과적인 feature selection 방안을 제안하였고 실험을 통해 기존의 방안 보다 정확도가 높은 것을 확인했다. 특히 decision tree와 bayesian network의 경우 중요도가 높은 순으로 단어들을 7-8개를 추출하여 학습할 때 가장 정확하게 분류한다. 본 논문에서는 직군을 주제로 실험을 진행하였지만 향후에는 연봉, 직무 등 기업과 구직자에게 필요한 여러 방면으로 활용 될 수 있다.

감사의 글

본 연구는 (1)2015년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업 (No. 2015R1A5A7037751), (2)2014년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구 사업(NRF-2014R1A2A1A10054151) 및 (3) 미래창조과학부 및 정보통신기술진흥센터의 서울어코드활성화지원사업 (IITP-2016-R0613-16-1149)의 결과물임을 밝힙니다.

참고문헌

[1] Juan Ramos, "Using TF-IDF to determine word relevance in document queries," ICML, 2003.

[2] TF-IDF, https://en.wikipedia.org/w/index.php?title=Tf%E2%80%93idf&oldid=738436457, 2016

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten; The WEKA Data Mining Software: An Update; SIGKDD Explorations, 2009.

[4] Micheline Kamber “Data Mining: Concepts and Techniques,” Morgan Kaufmann Pub, 2011.

[5] Harris, Earl. "Information Gain Versus Gain Ratio: A Study of Split Method Biases," ISAIM, 2002.