Journal of the Korea Academia-Industrial cooperation Society
Vol. 19, No. 11 pp. 730-737, 2018

https://doi.org/10.5762/KAIS.2018.19.11.730
ISSN 1975-4701 / eISSN 2288-4688

머신러닝을 이용한 의료 및 광고 블로그 분류

이기성[1], 이종찬[2*]
[1]호원대학교 컴퓨터게임·학부, [2]군산대학교 컴퓨터정보공학과

A Classification of Medical and Advertising Blogs Using Machine Learning

Gi-Sung Lee[1], Jong-Chan Lee[2*]
[1]Division of Computer & Game, Howon University
[2]Department of Computer Information Engineering, Kunsan National University

요 약 행복한 삶의 질을 목적으로 하는 의료소비자가 증가하면서 웹에 분산되어 있는 블로그의 의료 정보를 바탕으로 신뢰성 있는 의료 시설을 선택하고 고품질의 의료 서비스를 받음으로서, 시간과 비용을 절약할 수 있는 O2O 의료 마케팅 시장이 활성화 되고 있다. 인터넷, 모바일, SNS 등에서 증가하는 비정형 텍스트 데이터는 전문 의료 지식 이외에 작성자의 관심, 선호, 예상 등을 직간접적으로 반영하고 있기 때문에 의료정보의 신뢰성을 담보하기 어렵다. 본 연구에서는 빅데이터 및 MLP를 사용하여 의료정보 블로그를 분류 (의료블로그, 광고블로그)함으로서 사용자에게 보다 고품질의 의료정보 서비스를 제공하는 블로그 판단 시스템을 제안한다. 제안된 빅데이터 및 머신러닝 기술을 통해 인터넷상에 존재하는 국내의 다수 의료정보 블로그를 종합, 분석한 후 질환별 개인 맞춤형 건강정보 추천 시스템을 개발한다. 이를 통하여 사용자는 자신의 건강문제를 지속적으로 점검하고 가장 적절한 조치를 취함으로서 자신의 건강 상태를 유지하는 것이 가능할 것으로 기대된다.

Abstract With the increasing number of health consumers aiming for a happy quality of life, the O2O medical marketing market is activated by choosing reliable health care facilities and receiving high quality medical services based on the medical information distributed on web's blog. Because unstructured text data used on the Internet, mobile, and social networks directly or indirectly reflects authors' interests, preferences, and expectations in addition to their expertise, it is difficult to guarantee credibility of medical information. In this study, we propose a blog reading system that provides users with a higher quality medical information service by classifying medical information blogs (medical blog, ad blog) using bigdata and MLP processing. We collect and analyze many domestic medical information blogs on the Internet based on the proposed big data and machine learning technology, and develop a personalized health information recommendation system for each disease. It is expected that the user will be able to maintain his / her health condition by continuously checking his / her health problems and taking the most appropriate measures.

Keywords : Medical blog, Big data, Machine learning, Blog reading system, Medical information blogs

1. 서 론

현대 사회에서 행복한 삶의 질을 목적으로 하는 의료소비가 증가하면서 언제 어디서나 맞춤형 건강관리 서비스 제공하는 웰니스(Wellness) 산업에 대한 관심이 지속적으로 확대되고 있다. 이에 따라 ICT 융·복합 기술을 적용하여 개인에게 최적화한 건강관리 방법을 보다 쉽게, 상대적으로 적은 비용을 들여 최상의 결과를 제공하

이 논문은 2018년 호원대학교 교내연구비의 지원에 의하여 연구되었음.
*Corresponding Author: Jong-Chan Lee(Kunsan National Univ.)
Tel: +82-63-469-4863 email: chan2000@kunsan.ac.kr
Received September 29, 2018 Revised October 29, 2018
Accepted November 2, 2018 Published November 30, 2018

기 위하여 의료 관련 블로그를 종합, 분석하여 개인 맞춤형 의료 정보를 제공하려는 시도가 점차적으로 증가하고 있다[1-5].

웰니스 산업의 확대에 따라 사용자는 유사질환 커뮤니티를 활용하여 유사질환와의 정보를 공유하고 전문의를 통한 Q&A 서비스로 개인 주도 건강관리를 실현할 수 있다. 하지만 스마트 기기의 발달로 수많은 사람들이 인터넷 블로그에서 주요 의료 정보를 얻고 있지만 동종업체 간 치열한 홍보 경쟁으로 조회 수를 올리기 위한 홍보 기사가 범람하고 있다. 홍보성 블로그는 핵심에서 벗어난 유명인사 도용, 애매한 문장의 마무리, 제목과 내용의 불일치, 특정 물건 또는 방법의 선전 등의 특징을 갖는다[5-8]. 의료 블로그의 가장 중요한 기능인 의료 기사의 줄거리가 제공되지 않았을 뿐만 아니라, 무슨 내용이 실려 있는 지 독자의 주의를 끌게 하는 광고성 기사가 빈번히 생산 되고 있다. 따라서 사용자가 관심 있는 의료정보를 획득하기 위하여 반복해서 검색을 해야 하는 번거로움이 있고 의료정보 블로그의 게시자의 신뢰성이 떨어지는 문제점이 있다. 또한 의료 블로그 상의 비정형 텍스트 데이터는 사실에 입각한 지식 이외에 의약품의 광고나 의료 기기의 홍보 등을 직간접적으로 반영하고 있기 때문에 의료정보의 신뢰성을 담보하기 어렵다 [9-12]. 더구나 언어가 갖고 있는 다양한 특성들로 인하여 그 의미를 분석, 도출하는데 따른 어려움이 발생한다. 구체적으로 개발 시의 문제점을 기술하면 아래와 같다.

- 수집된 텍스트 데이터를 저장하기 위한 데이터베이스의 분류 체계 미흡으로 인하여 대용량, 비정형, 실시간 특성을 갖는 웹 데이터를 수용하기 어렵다.
- 의료, 음식, 운동, 문화, 상품 정보 등 다양한 유사 의료 데이터가 생성되지만 이를 일괄 또는 실시간으로 처리하기 어렵다.
- 기존 의료정보 제공서비스는 각 서비스 사마다 데이터를 개별관리 하여 실제 평가의 척도가 제한이며 객관적으로 판단할 정보가 부족하다.
- 의료 정보를 빙자하여 유사 광고를 게재하는 블로그가 증가하고 있다. 그 내용이 점차적으로 고묘해져서 일반 검색 방법으로는 광고블로그를 배제시키는 것이 어렵다.

본 연구에서는 인터넷상에 존재하는 국내의 의료 정보 블로그를 수집하고 빅데이터화 하여, 의료 및 광고성 기사의 특징을 추출하고, 이를 바탕으로 의료 블로그와 홍보 블로그를 분류하기 위한 블로그 판독 시스템을 제안한다. 이는 개인 맞춤형 의료 정보 제공 서비스로서, 기존 서비스와 차별화하여 블로그 판단 알고리즘을 통하여 게시 자를 분류 (전문 의료인, 일반인, 광고성)함으로서 사용자에게 보다 정확한 고품질의 의료정보 서비스의 제공이 가능하다.

2. 블로그 판독 시스템 구조

2.1 시스템 구성도

빅데이터 시스템에서, 의료 관련 블로그의 데이터를 전용 웹 크롤러를 통하여 수집하고 자료를 분석 가능한 형태로 저장하며, 처리 및 분석을 수행한다. 이 분석된 결과로부터 사용자 맞춤 형 의료 관련 개인 블로그 및 의료기관 블로그 정보를 동시에 제공함으로 양질의 서비스를 사용자에게 제공한다. 본 연구에서는 위 목적에 부합하는 빅데이터 분산처리 시스템을 구축한다. Fig. 1은 하둡(Hadoop) 기반 빅데이터 분산처리 시스템의 하드웨어 구성도이다[13].

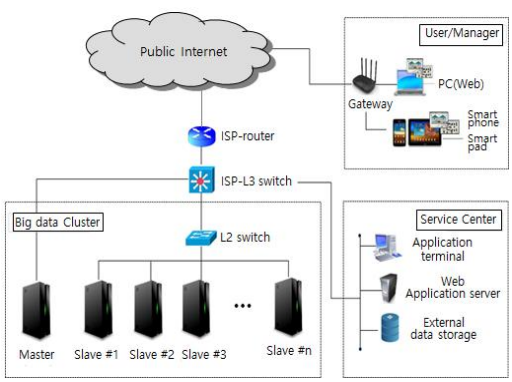

Fig. 1. Structure of big data system

이 시스템은 4대의 서버를 사용하여 그 중 1대 서버를 마스터 서버, 나머지 3대의 서버를 슬레이브 서버로 설정하여 분산처리 클러스터를 구축하였다. 빅데이터 처리 시스템의 고가용성을 위해 마스터 서버의 고장 시에도 슬레이브 #1 서버가 마스터 서버의 역할을 수행하며, 각각의 슬레이브 서버가 동시에 2대가 고장 나더라도 나머지 슬레이브 서버가 그 역할을 대신하도록 시스템을 구축하였다. 각 슬레이브 서버는 L2 스위치로 연결하고,

클러스터 간 접속을 위하여 마스터 서버는 L3 스위치에 배치하고, 관제 및 서비스를 위하여 필요한 서버 등은 L3 스위치에 배치하였다.

2.2 블로그 판독 기능

블로그 판독 시스템을 구축하기 위하여 Table 1과 같이 3단계의 처리 기능을 도입한다. 크롤링 된 데이터를 추가적으로 분류, 분석하여 최적화한 정보로 가공하고 이를 통하여 광고 블로그 및 의료정보 블로그를 분류할 수 있다.

1, 2단계의 처리 기능과 분석 방법의 전체적인 흐름은 아래와 같다.

- 자료 수집 단계에서는 개인 블로그 및 병원 블로그의 데이터를 전용 웹 크롤러를 통하여 수집하고 자료를 분석 가능한 형태로 저장한다.
- 저장된 블로그의 텍스트 데이터를 자연어 처리 및 형태소 분석을 실시하여 데이터 전처리를 수행한다. 블로그 판단에 부적절한 영향을 주는 부분을 제거하는 과정으로 한글, 숫자, 감정 분류를 위한 이모티콘을 제외한 한자, 영어를 제거하는 필터링 후에, 형태소 분석기를 통해 용언, 체언, 부사, 형식 형태소로 나누어 전처리 과정을 수행한다.
- 비정형화된 텍스트 데이터를 텍스트 마이닝 과정을 거쳐 구조화된 MongoDB에 입력 가능하도록 정형화한다.
- 선택된 자질들은 오름차순으로 의료 정보 사전(Word List)을 구성한다. 의료 정보 사전은 벡터 수치화를 할 때 그 단어의 고유번호를 구하기 위해서 필요하다. 형태소 분석기에 의해 추출된 의료 정보를 벡터 정보에서 구성할 수 있도록 인덱스화하여 구성한다. 의료 정보를 구성할 때는 중복된 정보을 제거하고 오름차순으로 정렬된 파일 형태로 생성된다.
- 학습처리는 학습에 기준이 되는 학습 데이터를 이용하여 분류기에서 사용될 학습모델을 생성하는 것이다. 학습처리의 처리기준에서 가장 중요한 것은 허위 학습 데이터와 허위 학습 데이터를 구분하여 학습시킨다.
- 블로그 분류 처리는 학습처리에서 생성된 범주 별 모델을 해당 범주의 분류기의 모델로 사용하여 분류를 처리하는 것이다. 학습처리에서 선택과 동일한 자질을 선택하여 벡터데이터를 생성하고 이렇게 생성된 데이터를 입력하여 분류의 결과를 판단한다.

2.2.1 크롤링을 이용한 블로그 수집 및 저장

다양한 의료 블로그 수집을 위해 비구조적 데이터 확보가 필요하다. 본 연구에서는 데이터의 최신 상태 유지를 위해 웹 크롤링을 수행한다. 크롤링 자동화 기능을 구축하고 링크 체크나 HTML 코드 검증과 같은 웹 사이트의 자동 유지 관리 작업 기능뿐만 아니라 웹 페이지의 특정 형태의 정보를 수집도 수행한다.

의료관련 블로그를 수집하기 위하여 2011년 1월부터 2018년 8월 사이의 블로그 15,000개를 수집하였다. 수집된 텍스트 데이터는 블로그 판독 시스템의 MongoDB에 존재하고, 분석에 필요한 속성들을 JSON 형태로 저

Table 1. Collection/processing/analysis of a blog

Steps	Functions	Implementation methods
Level 1: Data collection/ processing	Selection of blogs for data collection	- Individual and group websites - Website of middle and large hospital - Website of the pharmaceutical company
	Collection of data	- Data collection through web crawler
	Storage of data	- MongoDB-based information storage
	Processing of data	- Natural language processing - Morphological analysis - Extraction of keywords using KoNLy, Twitter, etc.
Level 2: Data analysis	Conversion of data	- Structured data conversion of unstructured data - Conversion of unstructured data into word-document matrix
	Primary filtering	- Blog filtering by applying bayesian method
	Secondary filtering	- Blog filtering using MLP
Level 3: Service provision	Provision of service	- Providing extracted medical information - Providing extracted medical facility information

장하였다. Fig. 2와 같이 모듈별로 파일을 생성하였고, 저장된 속성은 JSON의 키 값인 모듈(Module)과 제목(subject), 본문(description) 항목으로 구성된다.

```
{
"General" : [
   {
    "subject" : "Diabetes and Symptoms"
    "description" : "There are two types of diabetes. Type 1 diabetes usually appears in childhood and is sometimes referred to as childhood diabetes. It is a disease that insulin is artificially administered because insulin secretion does not occur in the body.\r\n\r\n"
   }
]}
```

Fig. 2. Storage structure for blog data

본 연구에서는 비정형 데이터의 저장에 특화된 MongoDB를 사용하여 수집된 데이터를 정제하고 저장한다. 블로그 데이터 처리과정을 통하여 MongoDB에 저장된 데이터 구조는 Fig. 3과 같다. 의료관련 블로그의 정보를 데이터화하고 블로그 판독 시스템을 이용하여 광고 블로그와 의료정보 블로그를 분류한다. Fig. 3에서 의료 정보 데이터는 id 항목을 통해서 날짜별로 관리되고 심근경색(Myocardial Infarction)이라는 하위 항목에 질환 특성으로 분류한다. 실제 데이터 분석과정에서는 이와 같이 MongoDB에 저장된 심근경색 데이터를 텍스트 마이닝을 통하여 분석한다.

```
{
"id" : ab76c1
"disease name" : "myocardial infarction (MI)"
"Related site" : "www.samsunghospital.com"
"MI_data" : [
    {
        "date" : 2018/08/11
        "activities" : [
            {
              "symptom": chest pains
              "cure" : vascular treatment
              "taking": incretin preparation
              “food”: garlic
              “exercise”: swimming
            },
            {
              "symptom": difficulty in breathing
              "cure": drug treatment
              "taking": aspirin
              “food”: onion
              “exercise”: walking
            }
        ]
    },
    {
        "date" : 2018/08/12
        "activities" : [
```

Fig. 3. Structure of data in MongoDB

2.2.2 크롤링을 이용한 블로그 수집 및 저장

MongDB에서 JSON으로 속성들을 저장하고 모듈(Module), 제목(Subject), 본문(Description) 정보를 대상으로 모듈 별 15개의 파일로 나누어서 저장한다. 문자열의 처리는 우선, 이슈 등록자가 리포트를 등록함에 있어 띄어쓰기를 대체로 고려하지 않는다고 가정하고 자동 띄어쓰기를 적용하여 수행한다. 그 후 문자열은 HTML tag, 특수문자, 숫자의 경우 블로그 분류와 무관하다고 판단하여 삭제하였다. 한글의 경우에는 KoNLPy의 Komoran 형태소 분석기를 사용하여 모든 형태소를 소문자로 변환한 후 구두점과 공백, 숫자를 제거한다. 또한 문장의 표현에 역할이 미비한 관사, 전치사, 조사, 접속사 등의 기능어 또한 제거함으로서 각 단어로 자연어 처리되었다. 그리고 어근을 추출하거나 접사를 제거하는 작업인 어간추출(Stemming)을 통하여 동일한 의미의 단어들을 하나의 단어로 처리하여 텍스트 처리의 효율성을 높였다.

2.2.3 블로그 특징 추출

광고 블로그, 의료정보 블로그를 판독하기 위하여, 블로그 판독 시스템에서는 텍스트 마이닝 기법을 적용하여 탐지확률을 높였다. Table 2에 블로그 판독 시스템에서 사용하는 파라미터를 보인다. 광고 블로그의 특징을 추출하기 위한 파라미터는 아래와 같다.

- 기사의 제목과 본문의 자수 비교 : 홍보 블로그는 일반 의료 블로그에 비해 대부분 적은 글자로 이루어진다. 따라서 본문의 자수를 비교하여 광고 블로그를 구분하는 특징으로 사용한다.
- 특정 병원, 특정 약품, 건상식품, 건강 음료 등의 사용빈도 확인 : 광고성 블로그의 특징은 특정 병원의 홍보, 제약 회사의 특정 약품의 홍보, 건강식품 회사의 식품, 음료 등의 홍보 등이 주를 이루므로 이 또한 주요 특징 중 하나이다.
- 주제와 본문의 불일치 : 홍보성 블로그의 경우 주제문과 본문의 일치 정도가 낮은 경우가 대부분으로 이를 수치화하여 주요 특징으로 사용한다.
- 선정 단어 단순 감탄어 사용 : 정확한 의미 전달을 방해하는 말줄임표의 등장여부, 문자의 말미에 특정 품사 (느낌표, 물음표 등) 사용 또한 중요한 특징 중 하나이다.
- 긍정, 부정 패턴 과다 사용 : 특정한 내용을 홍보하

Table 2. Parameters for determining blog type

Classification	Parameter	Criteria	Analysis method
Detection of ad blogs	Number of words in the body of the blog	Similarity	TFIDF, Bayesian, MLP
	Frequency of use of certain medicines, health foods and beverages	Similarity	
	Use of sensational words and simple exclamations	Similarity	
	Correspondence between the subject and the text	Similarity	
	Positive, negative pattern overuse	Positive/ negative	
Detection of medical blogs	Detection of use of medical terminology by disease	Similarity	
	Detection of disease-related symptoms	Similarity	

기 위하여 극단적인 긍정 또는 부정문을 사용하는 경우도 특징 중 하나이다.

의료정보 블로그 탐지를 위한 특징 추출 파라미터는 아래와 같다.

- 질병 별 의료 전문 용어 탐지: 텍스트 내용에 특정 질병이 탐지 되거나, 진료가능 질병으로 분류된 블로그를 탐지한다.
- 질병 별 관련 증상 탐지: 질병에 나타나는 관련 증상을 기반으로 의료정보 블로그 여부를 탐지한다.

2.2.4 의료 및 광고 블로그 분류

(1) TF-IDF를 이용한 블로그 분류

블로그 수집 및 자료 분석은 단어-질병 행렬로 변환된 의료 블로그를 바탕으로 텍스트 데이터를 의미 있는 정보로 저장하기 위하여, 단어별 빈도가 아닌 있는 TF-IDF (Term Frequency - Inverse Document Frequency)를 이용하여 각 단어별 가중치를 산출한다.

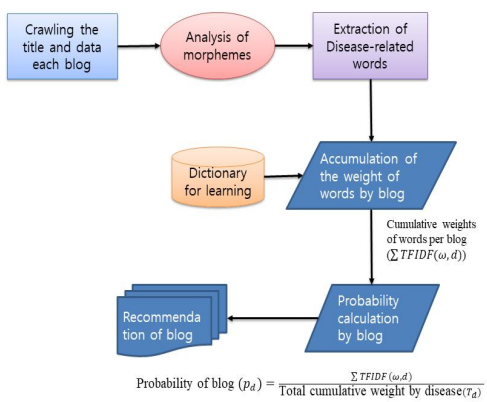

Fig. 4. Data processing structure

가중치가 부여된 단어들을 바탕으로 명사로 분리하여 빈도분석 및 비율분석을 통해 의료 정보 구성요인을 파악하기 위한 변수를 도출한다. Fig. 4와 같이 MLP(Multi Layer Perceptron)을 적용하여 요소 분석을 수행한다. 의료 정보의 구성요인을 도출하고 이를 바탕으로 질병별 판별요인을 파악하여 질병 유형에 따른 판별 요인을 결정한다.

Fig. 5와 같이 2차, 3차 카테고리 세분화 후 검색영역 확장 및 정밀 검색 기능을 구축한다. TF-IDF를 사용하여 구와 절 수준에서의 분석을 수행한다. 유사, 유관 블로그 검색을 위한 연관 검색어 맵을 구축하였다. 연관 검색어의 중요도를 판단하기 위하여 TF-IDF 방식을 이용하면 하나의 문서 중에서 가장 가중치 값이 큰 단어가 그 문서에 키워드로 채택한다.

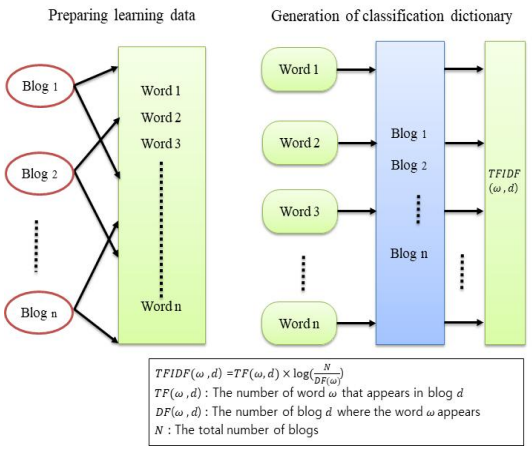

Fig. 5. Structure of TF-IDF

(2) 베이지안 필터를 적용한 1차 블로그 분류

베이지안 필터는 나이브 베이즈 분류(Naive bayes classifier) 알고리즘을 사용한다. 문장을 분류할 때 텍스트 내부에서의 단어 출현 비율을 조사한다. 이를 기반으

로 해당 텍스트 데이터를 적합한 블로그(광고 또는 의료 블로그)로 분류한다. (식 1)에서 $p(\omega_k|x)$는 여러 개의 의료 분야(ω_k) 중에 어떤 분야에 속할 확률이 가장 큰 가를 나타내는 정보이다.

$$p(w_k|x) = \frac{p(x|w_k)p(w_k)}{p(x)} \quad (1)$$

이때 $p(x|\omega_k)$는 각 의료 분야로 분류될 확률을 나타낸다. 전체 문서에서 해당 의료 분야의 문서가 나올 확률이다. $p(x|\omega_k)$에서 입력 텍스트 x는 의료 관련 단어들의 집합이다. 따라서 텍스트 데이터를 단어들로 분리한다. 입력 텍스트 x를 각 단어(n)의 집합이라고 할 때, $p(x|w_k)$는 (식 2)와 같다.

$$P(x|w_k) = p(x_1|w_k)p(x_2|w_k)p(x_3|w_k)\cdots p(x_n|w_k) \quad (2)$$

$P(x_n|w_k)$의 확률은 단어가 해당 블로그에 속할 확률이다. 따라서 어떤 의료 블로그에 해당 단어가 출현할 확률은 (식 3)과 같다. 여기서 μ_{kn}은 출현율, ν_n은 단어의 출현횟수, T_k는 의료 블로그 전체 단어 수를 의미한다.

$$\mu_{kn} = \frac{\nu_n}{T_k} \quad (3)$$

(3) MLP를 적용한 2차 블로그 분류

베이지안 필터로부터 필터링된 블로그 데이터를 MLP(Multi Layer Perceptron)의 입력 값으로 사용한다. MLP는 입력 층과 출력 층 사이에 다수의 은닉 층을 추가한 신경망이다. MLP는 문서를 그대로 입력 값으로 사용할 수 없으므로 데이터를 숫자로 표현할 수 있는 벡터로 변환해야 한다. 또한 텍스트 데이터는 이미지 데이터와 길이가 다르므로 고정 길이의 벡터로 변환해야 한다.

- 질병 정보 형태소 분석 파일: 질병관련 증상, 진료과목 등의 정보와 질병 분류를 학습시켜 모델을 만들고, 해당 모델을 사용해 새로운 블로그를 분류하도록 하였다. 베이지안 필터를 적용해 블로그를 분류한 결과를 추가 학습데이터로 사용하여 학습데이터를 증가시켰다.
- 의료 정보의 벡터 변환: 텍스트 데이터를 벡터 데이터로의 변환을 위하여, 단어 하나하나에 ID를 부여하고, 그러한 ID의 출현 빈도와 정렬 순서를 기반으로 벡터를 만드는 방법을 사용한다. 여기서는 단어의 정렬 순서는 무시하고, 출현빈도만 사용하였다. 텍스트에 한 단어가 있는지를 수치로 나타내는 BoW (Bag of Words)를 적용하여 입력 값의 수치화 작업을 수행한다. 우선 입력 텍스트의 형태소 분석을 수행하고 각 단어에 ID를 부여한다. 그리고 추가적으로 단어의 출현 횟수를 구한다. 이를 통하여 단어의 출현 횟수를 기반으로 문장을 표현할 수 있다.
- 의료 정보 텍스트 분류 과정: MLP의 텍스트 분류 과정은 다음과 같다.

- 텍스트 데이터에서 불필요한 품사를 제거한다.
- 사전을 기반으로 단어를 숫자로 변환한다.
- 파일 내부의 단어 출현 비율을 계산한다.
- 데이터를 학습시킨다.
- 시험 데이터를 입력하여 성공률을 확인한다.

MLP에서 사용된 파라미터는 Table 3과 같다. 입력 계층의 크기(dense)는 BoW를 적용하여 512로 결정하였다. 벡터로 변환된 단어들은 순차적으로 은닉 층(Hidden layer)에 입력되어 학습된다. 은닉 층의 활성화함수로는 ReLU, 출력 층(Output layer)의 활성화함수로는 소프트맥스(Softmax)를 사용한다. 한 번의 학습량을 조정하는 학습률은 0.001로 설정한다. 손실률 계산을 위하여 CEE(Cross Entropy Error)를 적용하였고, 학습방법은 오류역전파법(Backpropagation)을 사용한다.

Table 3. Parameter in the MLP

Parameter	Set Value
Input size	512
Hidden layer	5
Activation(Hidden layer)	ReLU
Activation(Output layer)	Softmax
Dropout Rate	0.5
Learning Rate	0.001
Loss rate	CEE
Learning method	Backpropagation

3. 성능 평가

성능 평가를 Table 4와 같이 질환별 의료 키워드 사전을 구축하였다[13]. 의료관련 사이트로부터 크롤링으로 수집된 정보를 문장 단위로 분리하고 각 문장을 형태

소 분석 후 토큰 배열로 변환한다. 변환된 토큰을 의료 정보 사전에 저장한다. 이 의료 사전은 TF-IDF에서 중요도 척도로서, 베이지안 필터에서 필터링 조건으로, MLP에서는 입력 층의 입력 값과 학습 파라미터로서 사용된다.

Table 4. Medical dictionary by disease

Medical Dictionary	Disease related medical word
Diabetes	blood sugar, insulin, hyperglycemia
High blood pressure	hypertension, diastole, atherosclerosis
Obesity	Overweight, body/abdominal obesity
Hyperlipidemia	dyslipidemia, cholesterol, High bleeding
Atherosclerosis	heart disease, cerebrovascular disease
Angina pectoris	heart disease, ischemia, cardiomyopathy
Stroke	paralysis, ischemia, ischemic attack
Lung disease	pneumonia, pulmonary tuberculosis, asthma
Liver disease	sarcoidosis, fatty liver, liver cancer
Gastric disease	gastritis, gastric ulcer, gastric cancer
Joint disease	joints, arthritis, degenerative, rheumatic
Osteoporosis	skeletal system, fracture, bone density
thyroid	thyroid cancer, thyroiditis
headache	migraine, tension headache, temporal arteritis
Rhinitis	sinusitis, sinusitis, sore throat
Urinary disease	urinary incontinence, prostate, prostatitis
Heart disease	cardiovascular, cardiovascular, arrhythmia
Eye diseases	conjunctivitis, cataract, glaucoma
Skin disease	dermatology, atopy, atopic dermatitis
Depression	depressive disorder, sleep depression
Stress	adaptive disorder, anxiety disorder
Gastric cancer	gastric cancer, lymph node, gastroscopy
Colon cancer	colon, colon polyps, colitis
Liver cancer	cirrhosis, liver cirrhosis, jaundice
Lung cancer	malignant tumor, tumor, chemotherapy

Table 5는 제안된 시스템의 평가 항목과 성능 목표치이다. TF-IDF, 베이지안 필터, MLP를 거쳐서 MongoDB에 저장 블로그들에 대하여 평가를 수행하고, 광고 블로그와 의료 블로그의 소속 정보를 평가한다.

Table 5. Test items and performance objectives

Test item	performance objectives
Detection of ad blogs	More than 80%
Detection of medical blogs	More than 90%
Accuracy of Big Data Classification by Disease	More than 90%

MongoDB에 저장된 블로그의 평가를 통하여 광고 블로그로 탐지된 블로그 중에서 실제 광고 블로그가 80% 이상 포함되었음을 확인하였다. 또한 의료정보 블로그로 탐지된 블로그 중에서 실제 의료정보 블로그가 90% 이상 포함됨을 확인하였고, 진환별로 키워드 분류가 된 정보와 실제 블로그의 주체가 일치 하는 수가 90% 이상 포함됨을 확인하였다.

4. 결 론

건강은 현대인의 주요 관심 사항이지만 인터넷 상에는 다양한 의료 유사 정보가 존재하므로 신뢰하기 어렵고, 더불어 특정 사용자에게 적합한 맞춤형 의료 정보의 제공은 불가능에 가깝다. 본 연구에서는 인터넷상에 존재하는 국내의 수많은 의료 정보 블로그를 빅데이터 기술을 통해 수집, 분석하여 광고 블로그를 탐지하여 제거하고, 질환별 분류를 가능케 하는 블로그 판독 시스템을 제안하였다. 제안된 시스템을 통하여 유사질환자와의 정보공유 및 전문의를 통한 상담 서비스로 개인 주도 건강관리를 실현시킬 수 있는 개인 맞춤형 건강정보 추천을 가능하게 하였다. 추후 의료 정보 사전 및 광고 사전 등을 정밀하게 구축하고, 머신러닝으로서 CNN(Convolution Neural Network)과 RNN(Recurrent Neural Network)을 적용한 연구가 필요하다.

References

[1] Y. Y Ou, P. Y Shih, Y. H. Chin, T. W. Kuan, "Framework of ubiquitous healthcare system based on cloud computing for elderly living," *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, 2014.
DOI: https://doi.org/10.1109/APSIPA.2013.6694298

[2] J. Y. Lee, K. D. Jung, "Proposed Architecture for U-Healthcare Systems," *Advanced Culture Technology*, Vol. 4, No. 2, pp. 43-46, 2016.

DOI: https://doi.org/10.17703/IJACT.2016.4.2.43

[3] Y. E. Gelogo and H. K. Kim, "Integration of Mobile Computing to Ubiquitous Healthcare," *Software Engineering and Its Applications*, Vol. 9, No. 9, pp. 295-302, 2015.
DOI: https://doi.org/10.14257/ijseia.2015.9.9.26

[4] M. Rostami, S. Ayat, I. Attarzadeh, and F. Saghari, "Proposing a Method to Classify Texts Using Data Mining," *Advances In Computer Research*, Vol. 6, No. 4, pp. 125-137, 2015.

[5] O. H. Shin, "Demystifying Big Data: Anatomy of Big Data Developmental Process," *Telecommunication Policy (ELSEVIER)*, 2015.
DOI: https://doi.org/10.1016/j.telpol.2015.03.007

[6] P. Mohata and S. Dhande, "Web Data Mining Techniques and Implementation for Handling Big Data," *Computer Science and Mobile Computing*, Vol. 4, No. 4, pp. 330-334, 2015.

[7] E. Ferrara and P. Meo, Giacomo Fiumara, Robert Baumgartner, "Web Data Extraction, Applications and Techniques: A Survey," *Knowledge-Based Systems (ELSEVIER)*, pp. 301-323, 2014.
DOI: https://doi.org/10.1016/j.knosys.2014.07.007

[8] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M. Arif Bijaksana, "Relevance Feature Discovery for Text Mining," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 6, No. 1, 2015.
DOI: https://doi.org/10.1109/TKDE.2014.2373357

[9] B. Dayley, “Node.js, MongoDB, and AngularJS Web Development”, *Addison-Wesley*, 2014.

[10] A. S. Oh, "A Study on Design of Health Device for U-Health System," *Bio-Science and Bio-Technology*, Vol. 7, No. 2, pp. 79-86, 2015.

[11] F. N. Afrati and J. D. Ullman, "Optimizing Multiway Joins in a Map-Reduce Environment," *IEEE Transactions on Knowledge and Data Engineering*, Vol.23, No.9, pp.1282-1298, 2011.
DOI: https://doi.org/10.1109/TKDE.2011.47

[12] B. Singh and S. Kumar, "An Effective Information Retrieval with Keyword Optimization," *Computer Technology and Applications*, Vol. 5, No. 1, pp. 174-177, 2014.

[13] J. C. Lee and K. J. Park, "Design of Food Management System Using NFC Tag," *Journal of korean society of computer and information*, Vol. 23, No. 5, pp. 25-29, 2018.

이 기 성(Gi-Sung Lee) [종신회원]

- 1996년 2월 : 숭실대학교 컴퓨터과학과 (공학석사)
- 2001년 2월 : 숭실대학교 컴퓨터과학과 (공학박사)
- 2001년 3월 ~ 현재 : 호원대학교 컴퓨터게임학과 교수

<관심분야>
모바일 멀티미디어, 네트워크 보안, 머신러닝

이 종 찬(Jong-Chan Lee) [정회원]

- 1996년 8월 : 숭실대학교 컴퓨터과학과 (공학석사)
- 2000년 8월 : 숭실대학교 컴퓨터과학과 (공학박사)
- 2000년 10월 ~ 2005년 2월 : 한국전자통신연구원 선임연구원
- 2005년 3월 ~ 현재 : 군산대학교 컴퓨터정보공학과 교수

<관심분야>
머신러닝, 빅데이터, 블록체인