

이미지와 텍스트 정보의 카테고리 분류에 의한 SNS 팔로잉 추천 방법

(Recommendation Method of SNS Following to Category Classification of Image and Text Information)

홍택은*, 신주현**

(Taek Eun Hong, Ju Hyun Shin)

요약

다양한 스마트 디바이스의 발전에 따라 거리, 공간의 제약 없이 실시간으로 의사소통, 정보공유 등이 가능한 SNS(Social Network Service)를 즐기는 사용자(User)가 증가하고 있다. 의사소통, 관계 형성에 중점을 두었던 SNS 사용자들이 정보공유의 기능으로 SNS를 활용하는 추세이다. 본 논문에서는 사용자의 SNS 게시글을 이용하여 카테고리를 추출하고 정보제공자(Information provider)를 팔로잉 추천해주는 방법을 기술한다. 게시글의 텍스트에서 단어를 분류하고 빈도수를 측정하며, 머신러닝 기법 중 하나인 CNN(Convolutional Neural Network)을 바탕으로 구축한 Inception-v3 모델을 이용하여 이미지를 단어로 분류한다. 텍스트와 이미지에서 분류한 단어를 DMOZ 기준으로 카테고리 분류하여 정보제공자 DB를 구축한다. 정보제공자 DB의 카테고리화 게시글에서 분류한 사용자의 카테고리를 비교한다. 카테고리가 일치할 경우 카테고리에 분류되어 있는 정보제공자들을 대상으로 유사도를 측정하여 가장 비슷한 정보제공자의 계정을 추천해주는 방법에 대해 제안한다.

■ 중심어 : 소셜네트워크 ; 사용자 정보 ; 추천시스템 ; 오픈 디렉토리 프로젝트

Abstract

According to many smart devices are development, SNS(Social Network Service) users are getting higher that is possible for real-time communicating, information sharing without limitations in distance and space. Nowadays, SNS users that based on communication and relationships, are getting uses SNS for information sharing. In this paper, we used the SNS posts for users to extract the category and information provider, how to following of recommend method. Particularly, this paper focuses on classifying the words in the text of the posts and measures the frequency using Inception-v3 model, which is one of the machine learning technique - CNN(Convolutional Neural Network) we classified image word. By classifying the category of a word in a text and image, that based on DMOZ to build the information provider DB. Comparing user categories classified in categories and posts from information provider DB. If the category is matched by measuring the degree of similarity to the information providers is classified in the category, we suggest that how to recommend method of the most similar information providers account.

■ keywords : Sosial Network Service; User Information ; Recommendation System ; Open Directory Project

I. 서론

SNS(Social Network Service)는 거리, 공간의 제약 없이 사용자 간 실시간으로 의사소통, 관계 형성, 정보 공유를 할 수 있는 사이버 공간이다. 전 세계적으로 SNS(Social Network Service)의 인기가 높아지면서 사용자가 증가하고 있으며, 사회적인 관심의 대상으로 떠오르고 있다[1]. SNS를 통해서 다양한 사용자들과 관계를 형성하고 유지하는 데 중점을 두었던 사용자들이 사이버 공간에서 이루어지는 관계 형성과 유지에 싫증과

지루함을 느끼면서 SNS를 정보공유의 기능으로 활용하기 시작했다[2]. 관계 형성에 중점을 두고 있는 SNS에는 페이스북, 트위터 등이 있으며, 최근 대세를 이루고 있는 정보공유 중심의 SNS에는 인스타그램, 텀블러, 링크드인 등이 있다. SNS를 이용한 다양한 연구들이 진행되고 있는데, 특히 게시글을 통해 사용자의 특성이나 관심사 등을 판별하여 사용자에게 맞춤형 서비스를 제공하기 위한 연구가 활발하게 이루어지고 있다[3]. SNS 사용자들은 자신의 계정을 통해 관심을 갖고 있는 다양한 정보를 게시하며 팔로잉을 통해 자신과 같은 관심사를 갖고 있는 다른 사용

* 준회원, 조선대학교 소프트웨어융합공학과

** 정회원, 조선대학교 제어계측로봇공학과

이 논문은 2015학년도 조선대학교 학술연구비의 지원을 받아 연구되었음.

접수일자 : 2016년 09월 05일

수정일자 : 2016년 09월 24일

게재확정일 : 2016년 09월 26일

교신저자 : 신주현 e-mail : jhshinkr@chosun.ac.kr

자들과 정보를 공유한다[4]. 이처럼 팔로잉을 통해 관심사가 비슷한 사용자의 게시글에서 정보를 습득할 수 있다. 대부분의 SNS에서 이미 팔로잉 추천을 제공하고 있지만 사용자 사이의 관계를 이용한 추천을 제공하고 있으므로 정보 습득을 목적으로 하는 사용자에게는 적합하지 않다. 따라서 본 논문에서는 SNS 게시글의 이미지와 텍스트를 이용하여 사용자의 관심사에 맞는 정보기반의 팔로잉 추천 방법을 제안한다. 정보제공자의 게시글에서 텍스트와 이미지를 추출하여 ODP(Open Directory Project)인 DMOZ의 카테고리를 기준으로 관심사 데이터베이스를 구축한 후 사용자 게시글을 분석하여 데이터베이스 내에서 사용자와 같은 관심사를 갖는 정보제공자를 팔로잉 추천 한다. 본 논문의 구성은 다음과 같다. 2장에서는 SNS를 이용한 추천 시스템에 관련된 기존 연구에 관해서 설명한다. 3장에서는 SNS를 이용하여 사용자와 같은 관심사를 갖는 카테고리의 정보제공자를 추천해주는 방법에 대해 서술하며, 4장에서는 본 논문에서 제안한 방법에 대한 성능평가와 유사도 측정을 통해 추천한 카테고리의 정보제공자들 중에서 가장 유사한 정보제공자를 추천하는 방법에 대해 서술한다. 5장에서 결론 및 향후 연구를 기술하며 마무리한다.

II. 관련 연구

SNS를 이용한 사용자 맞춤형 서비스는 사용자의 게시글에서 수집할 수 있는 자료를 토대로 사용자의 개별적인 특성을 분석하여 이루어지고 있으며, 개인의 관심사와 특성을 이용하여 사용자 선호도에 알맞은 콘텐츠를 추천하기 위한 추천 기법에 대한 연구가 활발히 진행되고 있다[5-8]. 장현웅의 연구에서는 폭발적으로 증가하고 있는 SNS의 데이터 중 이미지 데이터를 분류하기 위한 기반의 연구를 하였다. 대용량의 이미지 데이터에 CNN(Convolutional Neural Network) 기술을 이용하여 라벨을 학습시킨 후 이미지 기반 SNS인 인스타그램의 이미지로부터 라벨

정보를 추출하였으며, 그 결과 70.44%의 정확도를 얻었다[5]. 홍명덕의 연구에서는 SNS의 사용자 게시글을 이용하여 사용자의 최근 관심사를 추출하고 사용자 선호도에 알맞은 뉴스를 제공하기 위한 콘텐츠 기반의 추천 기법에 관한 연구를 하였다. 사용자의 정보와 최근 게시글을 이용하여 사용자 프로파일을 생성하고 뉴스 카테고리를 이용하여 뉴스 방송 원고를 분석한 후 사용자 프로파일과 뉴스 프로파일의 유사도 측정을 통해 가장 유사도가 높은 뉴스를 사용자에게 제공했다[6]. 유소엽의 연구에서는 SNS 사용자의 온라인 사회적 관계와 게시글을 이용하여 관심사를 추천해 주는 연구를 했다. 사회적 관계와 사용자 게시글의 카테고리를 이용하여 사용자 선호도를 추출하고 팔로워와 공유하는 선호도를 추출한 후 사용자와 팔로워가 공유하는 선호도를 비교하여 사용자와 팔로워 사이의 관심사를 추천하였다[7]. 전련화의 연구에서는 소셜 네트워크 추천에 적합한 다양한 요소들을 추출하고 심층적으로 분석함으로써 관심사를 명시적으로 적용할 수 있는 사용자 맞춤형 서비스를 제공하는 방법을 제안하였으며, SNS적인 요소와 사용자의 평가이력 요소를 고려함으로써 개개인의 특징을 활용했다[8]. 기존의 연구들은 SNS 게시글에 존재하는 이미지 또는 텍스트를 각각 사용하였기 때문에 사용자가 올린 게시글이 어떤 의도인지 정확하게 파악하기 어렵다는 단점이 존재한다. 따라서 본 논문에서는 기존 연구의 문제점을 해결하기 위해 SNS 게시글의 텍스트와 이미지를 이용하여 사용자의 관심사를 카테고리를 기준으로 분류하고, 추천하는 방법을 제안한다.

III. 팔로잉 추천 방법

1. 시스템 구성도

SNS 중에서 FaceBook을 대상으로 실험을 진행하였으며, Java 기반 Facebook API인 Facebook4J를 이용하여 125개의 사

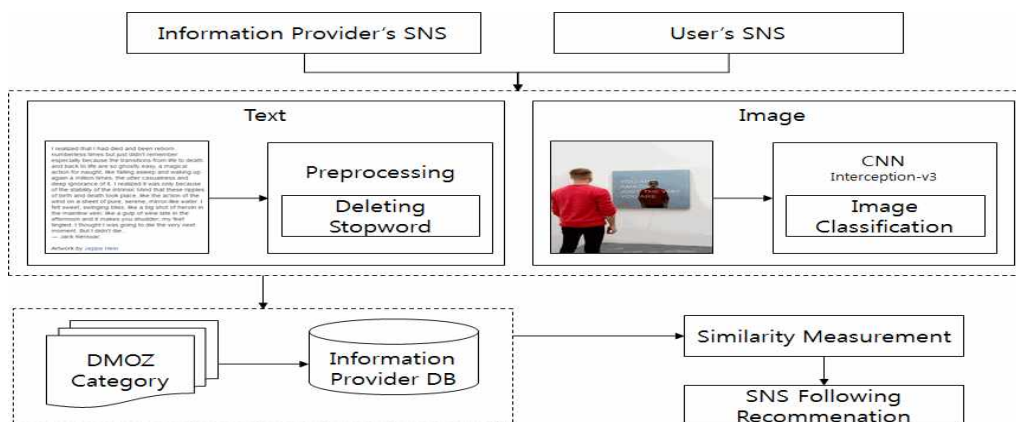


그림 1. SNS 팔로잉 추천 시스템 구성도

용자 계정에서 3,125개의 게시글을 수집하였다. DMOZ 카테고리의 대부분을 기준으로 각 10명의 정보제공자를 선정하였으며, 정보제공자의 계정에서 25개의 게시글을 수집하였다. 텍스트와 이미지를 분석하여 DMOZ와 비교하여 카테고리를 분류하고 카테고리별로 정보제공자 DB에 저장한다. 같은 방법으로 사용자의 게시글을 분석하여 카테고리를 분류하고 정보제공자 DB와 비교하여 같은 카테고리에 해당하는 정보제공자 계정을 팔로잉 추천한다. 그림 1은 본 논문에서 제안하는 팔로잉 추천 방법에 대한 시스템 구성도이다.

2. DMOZ 카테고리 재정의

DMOZ는 ODP로 진행되는 프로젝트이며, 전 세계의 다양한 사용자의 참여로 구축되는 디렉토리 형식의 카테고리이다. 그림 2는 DMOZ 내용의 일부이다.

```
Sports/Baseball/Major_League
Sports/Baseball/Major_League/All-Star_Game
Sports/Baseball/Major_League/Audio
Sports/Baseball/Major_League/Awards
Sports/Baseball/Major_League/Awards/Cy_Young
Sports/Baseball/Major_League/Awards/Gold_Glove
Sports/Baseball/Major_League/Awards/MVP
Sports/Baseball/Major_League/Awards/Rookie_of_the_Year
Sports/Baseball/Major_League/Chats_and_Forums
Sports/Baseball/Major_League/Directories
Sports/Baseball/Major_League/Fan_Pages
Sports/Baseball/Major_League/News_and_Media
Sports/Baseball/Major_League/News_and_Media/Organizations
Sports/Baseball/Major_League/Rules
Sports/Baseball/Major_League/Scores_and_Schedules
Sports/Baseball/Major_League/Spring_Training
Sports/Baseball/Major_League/Stadiums
Sports/Baseball/Major_League/Statistics
Sports/Baseball/Major_League/Strikes
Sports/Baseball/Major_League/Teams
Sports/Baseball/Major_League/Teams/Arizona_Diamondbacks
Sports/Baseball/Major_League/Teams/Arizona_Diamondbacks/Minor_League_Affiliates
Sports/Baseball/Major_League/Teams/Arizona_Diamondbacks/News_and_Media
Sports/Baseball/Major_League/Teams/Arizona_Diamondbacks/Players
Sports/Baseball/Major_League/Teams/Atlanta_Braves
Sports/Baseball/Major_League/Teams/Atlanta_Braves/Minor_League_Affiliates
Sports/Baseball/Major_League/Teams/Atlanta_Braves/News_and_Media
Sports/Baseball/Major_League/Teams/Atlanta_Braves/Players
Sports/Baseball/Major_League/Teams/Baltimore_Orioles
Sports/Baseball/Major_League/Teams/Baltimore_Orioles/Minor_League_Affiliates
Sports/Baseball/Major_League/Teams/Baltimore_Orioles/News_and_Media
Sports/Baseball/Major_League/Teams/Baltimore_Orioles/Players
Sports/Baseball/Major_League/Teams/Boston_Braves
Sports/Baseball/Major_League/Teams/Boston_Red_Sox
```

그림 2. DMOZ 카테고리의 내용

DMOZ는 Arts, Business, Computers, Games, Health, Home, News, Recreation, Reference, Regional, Science, Shopping, Society, Sports, Kids&Teens Directory, World의 총 16개의 카테고리로 구성되어 있으나 Reference, Regional, Kids&Teens Directory, World는 중복되는 소분류 카테고리가 너무 많으므로 정확한 카테고리 분류를 위해 제외하였다. 카테고리를 분류하기 위해 최상위 카테고리라 최하위 카테고리를 이용하여 DMOZ를 재정의하였으며, 표 1은 재정의한 DMOZ이다.

3. 텍스트를 이용한 카테고리 분류

수집된 텍스트는 전처리과정을 통해 불용어를 제거하여 단어 단위로 분할한 후 빈도수를 측정하여 상위 5개 단어를 후보 단어로 선정한다. 각 후보 단어를 재정의한 DMOZ의 내용과 비교하여 일치할 경우 후보 카테고리로 선정하고 후보 카테고리의 빈

표 1. 재정의한 DMOZ 카테고리

Root category	Contents
Arts	arts animation anime characters clubfandom cliques cosplay
Business	business accounting associations australia research canada india
Computers	computers algorithms animated compression conferences past
Games	games abstract contigo morabaraba pentegonia quarto renju tilez
Health	health addictions food games articles internet organizations alcohol
Home	home roommates states organizations appliances blenders coffee
News	news alternative directories newspapers radio columnists directories
Recreation	recreation antiques appliances irons china directories art tractors
Science	science agriculture animals birds poultry chickens ducks quail
Shopping	shopping paperweights signs neon reproductions soda cocacola pepsi
Society	society activism mattel microsoft nestlé nike cacerolazo consumer
Sports	sports clubs races schools teams airsoft organizations

도수를 측정하여 가장 높은 빈도수를 보이는 단어를 텍스트의 카테고리로 분류한다. 표 2는 텍스트의 카테고리 분류과정을 나타낸다.

표 2. 정보제공자 sns의 텍스트 분류 결과

Original text	I realized that I had died and been reborn numberless times but just didn't remember especially because the transitions from life to death and back to
Deleting Stopword	realized died reborn numberless times just didn't remember especially transitions life death back life ghostly easy magical action naught like falling asleep
Candidate word	(love, 12) (hope, 9) (like, 7) (one, 7) (people, 7)
Candidate Category	love(Society, Sports, Arts) hope(Society, Sports, Arts) like(-) one(Arts) people(News, Games, Computers, Arts , Sports, Society, Shopping, Science)
Category	(Arts , 4) (Society, 3) (Sports, 3) (News, 1) (Games, 1) (Computer, 1) (Shopping, 1) (Science, 1)

표 2에서 후보 단어로 love, hope, like, one, people이 분류되었

고 DMOZ와 비교한 결과 love의 후보 카테고리 Society, Sports, Arts가 분류되었으며, hope의 후보 카테고리는 Society, Sports, Arts가 선정되었다. like의 경우 DMOZ와 후보 단어 간 일치하는 단어가 존재하지 않았기 때문에 '-'으로 표기하였다. 후보 카테고리의 빈도수를 측정한 결과 Arts가 가장 높으므로 카테고리는 Arts로 분류된다.




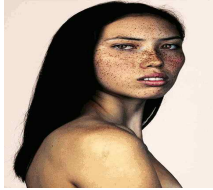



4. 이미지를 이용한 카테고리 분류

이미지를 분류하기 위한 다양한 방법이 있으나 본 논문에서는 딥러닝 기법중 하나인 CNN을 이용한다. 딥러닝은 신경망 네트워크를 이용하여 많은 수의 계층을 만들어 학습하는 기계학습 분야이다. 신경망 네트워크는 1950년대부터 인간의 뇌와 비슷한 형태의 학습 모델로 연구가 지속되었으나 낮은 컴퓨팅 성능 문제로 인해 진전되지 못하다가 컴퓨팅 성능 향상, 빅데이터의 중요성이 대두되면서 주목받기 시작했다[9]. Inception-v3 모델을 이용하여 이미지를 분류하였다. Inception-v3 모델은 구글에서 ImageNet Large Visual Recognition Challenge2012 데이터를 이용해서 만든 State-Of-Art 이미지 인식 트레이닝 모델이며, Top-5-Error Rate가 3.49%로 성능이 뛰어나다[10, 11]. 기계학습과 딥러닝을 위해 구글에서 만든 오픈소스 라이브러리인 TensorFlow를 사용하여 CNN 기법을 적용한다. TensorFlow는 데이터 플로우 그래프 방식을 이용하여 수학 계산과 데이터 흐름을 노드와 엣지를 사용하는 방향 그래프로 표현한다[12]. VMware Workstation 10.0.7을 이용하여 RAM 2GB, Hard Disk 20GB의 가상머신 환경을 구축하고 OS는 Ubuntu 64-bit를 이용하여 실험 환경을 구축하였다. 표 3은 TensorFlow의 Inception-v3 모델을 이용하여 이미지를 분석한 결과이다. 표 3을 통해 이미지에서 분류한 단어를 DMOZ와 비교하여 후보 카테고리를 분류하는 과정을 확인할 수 있다. 분류한 단어에 해당하는 후보 카테고리가 없을 경우 '-'로 표기하고 가장 빈도수가 높은 후보 카테고리를 이미지의 카테고리로 분류했으며, art가 카테고리로 분류되는 것을 알 수 있다. 카테고리를 분류하는 과정에서 후보 카테고리가 빈약하게 추출되는 결과가 보이는데 이는 Inception-v3 모델이 ImageNet Large Visual Recognition Challenge2012을 통해 학습되었기 때문으로 추측된다.

5. 정보제공자 DB 구축

3.2절과 3.3절에서 분류한 텍스트와 이미지의 카테고리를 통해 정보제공자 DB를 구축한다. 정보제공자에게 아이디를 부여하며, 텍스트의 카테고리라와 이미지의 카테고리를 태깅한다. 정보제공자 DB의 카테고리는 총 16개이고 Arts, Business, Computers, Games, Health, Home, News, Recreation, Science, Shopping, Soc

표 3. 정보제공자 sns의 이미지 분류 결과

No	Image	word classification	Candidate Category
1		jersey	-
		sweatshirt	-
		moving van	-
		-	-
		cash machine	-
2		tub	-
		bathtub	-
		mask	art
		shower curtain	-
		book jacket	-
3		quilt	-
		throne	-
		prayer rug	-
		pillow	-
		jigsaw puzzle	-
4		brassiere	-
		bikini	-
		book jacket	-
		bath towel	-
		swimming trunks	-
∴	∴	∴	∴
23		revolver	-
		cloak	-
		mask	art
		academic gown	-
		wig	-
24		book jacket	-
		comic book	-
		pickelhaube	-
		bassoon	shopping
		mask	art
25		mask	art
		lipstick	-
		book jacket	-
		ski mask	-
		bow tie	-

ety, Sports로 구성되어 있다. 텍스트의 카테고리라와 이미지의 카

태고리, 대분류 카테고리가 일치할 경우 DB에 저장한다. 표 4는 정보제공자 DB를 보여준다.

표 4. 정보제공자 DB

Root category	Contents
Arts	Arts_info_1(Arts, Arts) Arts_info_2(Arts, Arts) Art_info_3(Arts, Arts) ...
Games	Games_info_1(Games, Games) Game_info_2(Games, Games) ...
⋮	⋮
Sports	Sports_info_1(Sports, Sports) Sports_info_2(Sports, Arts) ...

표 4에서 Arts 카테고리의 경우 Arts_info_num으로 id를 부여하였으며, num에는 index가 들어가게 된다. id의 뒤에는 텍스트와 이미지의 카테고리가 순서대로 태깅되는 것을 알 수 있다. 수집한 120개의 정보제공자 계정에서 DB에 저장된 정보제공자 계정의 수는 109개이며, 11개는 텍스트와 이미지 카테고리가 일치하지 않아서 분류하지 못한 경우에 해당한다.

6. 정보제공자 DB를 이용한 팔로잉 추천

3.2절과 3.3절에서 기술한 방법과 동일한 방법으로 사용자 게시글의 텍스트와 이미지에서 카테고리를 분류한다. 텍스트에서 불용어를 제거하고 빈도수를 기반으로 후보 단어를 추출한 후 재정의한 DMOZ와 비교하여 후보 카테고리를 추출하고 빈도수를 측정하여 가장 빈도수가 높은 카테고리를 텍스트의 카테고리로 분류한다. 이미지는 TensorFlow의 Inception-v3 모델을 이용하여 단어를 분류하고 DMOZ와 비교하여 후보 카테고리를 추출한 후 빈도수가 가장 높은 후보 카테고리가 이미지의 카테고리가 된다. 표 5와 표 6은 사용자의 게시글에서 텍스트와 이미지에서 카테고리를 분류하는 과정을 나타낸다.

표 5. 사용자 sns의 텍스트 분류 결과

Original text	uovastrapazzate_Un'amica nel verde ???#abstract #archidaily #archilovers #architecture
Deleting Stopword	Uovastrapazzate un amica nel verde abstract archidaily archilovers architecture
Candidate word	(beautiful, 22) (design, 18) (style, 18) (art, 12) (love, 12)
Candidate Category	beautiful(Arts) design(Science) style(Computers) art(Arts) love(Society, Sports, Arts)
Category	(Arts, 3) (Science, 1) (Computers, 1) (Society, 1) (Sports, 1)

표 6. 사용자 sns의 이미지 분류 결과

No	Image	word classification	Candidate Category
1		obelisk	-
		bell cote	-
		mosque	-
		stupa	-
		beacon	-
2		book jacket	-
		comic book	-
		mask	-
		packet	-
3		pillow	-
		wig	-
		mask	Arts
		book jacket	-
⋮	⋮	lipstick	-
		hair spray	-
22		hair spray	-
		hay	society
		repressed	-
		balloon	-
		sandar	-
23		lakeside	-
		bannister	-
		planetarium	-
		grand piano	Arts
		stage	-
		ping-pong ball	sports

표 5를 통해 사용자 텍스트에서 카테고리가 Arts로 분류되는 것을 알 수 있으며, 표 6을 통해서 사용자 이미지에서 Arts가 카테고리로 분류되는 것을 알 수 있다. 사용자에게 id와 이미지, 텍스트 태깅을 부여하면 Users_1(Arts, Arts)가 되며, 사용자 또한 Arts 카테고리로 분류할 수 있다. 정보제공자 DB에서 Arts 카테고리에 해당하는 Arts_info_1, Arts_info_2 등의 정보제공자를 팔로잉 추천해 줄 수 있다.

IV. 실험 결과 및 고찰

본 절에서는 유사도를 이용하여 카테고리 내의 정보제공자들

중 사용자와 가장 유사한 정보를 가지고 있는 정보제공자를 추천하는 방법에 대해 기술한다. 이미지와 텍스트에서 분류한 단어들을 이용하여 코사인 유사도와 자카드 유사도를 측정하여 사용자에게 적합한 정보제공자를 추천한다[13, 14]. 식 1은 코사인 유사도 측정 수식이며, 식 2는 자카드 유사도 측정 수식이다.

$$S_c(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (1)$$

$$S_j(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2)$$

식 1은 내적 공간의 두 벡터 사이의 각도를 측정하여 유사한 정도를 구하는 방법이며, 사용자의 단어 분류 집합과 정보제공자의 단어 분류 집합이 벡터 A, B이다. 두 벡터의 내적을 각 벡터 크기의 곱으로 나누어 코사인 값을 도출한다. 식 2는 두 데이터 집합의 교집합의 크기를 합집합의 크기로 나눈 것이며, A는 사용자 단어 분류 집합이고 B는 정보제공자 단어 분류 집합이다. 식 1과 식 2의 결과 값은 0에서 1 사이의 값을 갖고 1에 가까울수록 A와 B가 유사하며, 0에 가까울수록 A와 B가 유사하지 않다. 식 2의 경우 단순 매칭 계수를 이용하기 때문에 수치가 다소 낮게 측정될 수 있으므로 식 1과 식 2의 평균을 측정하였으며 다음의 식 3과 같다.

$$S_{mean} = \frac{S_c(A, B) \times S_j(A, B)}{2} \quad (3)$$

본 논문에서 제안한 텍스트와 이미지를 함께 사용하여 사용자에게 정보제공자를 팔로잉 추천하는 방법이 텍스트와 이미지를 각각 사용하는 방법보다 정확하지 확인하기 위해 3.5절에서 팔로잉 추천한 Arts 카테고리의 정보제공자와 사용자를 대상으로 식 1과 식 2, 식 3을 이용한 결과는 그림 3, 4, 5와 같다.

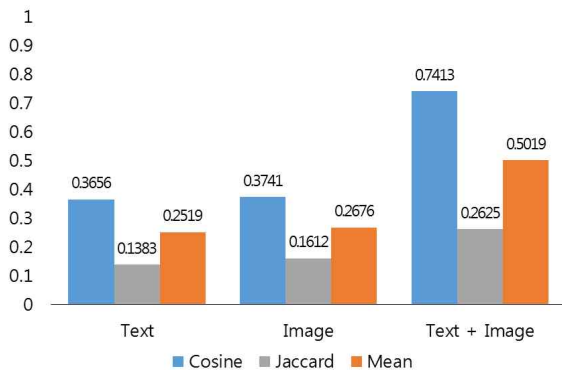


그림 3. Users와 Art_info_1의 유사도 결과

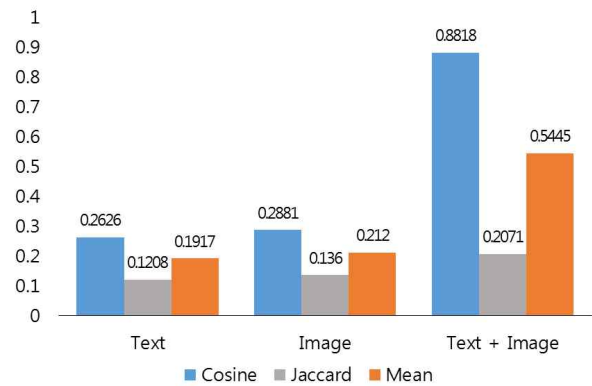


그림 4. Users와 Art_info_2의 유사도 결과

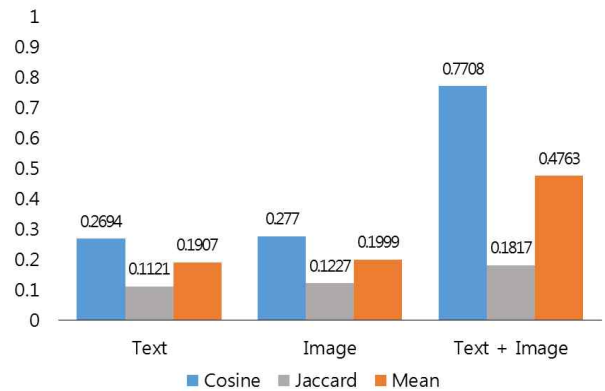


그림 5. Users와 Art_info_3의 유사도 결과

그림 3, 4, 5는 Users와 Art_info_1, 2, 3의 게시글에서 분류한 단어와 식1, 식2, 식3을 이용하여 유사도를 측정한 결과이다. 전반적으로 코사인 유사도가 자카드 유사도의 값보다 높게 도출되는 현상이 보인다. 자카드 유사도의 경우 단순 매칭 기법을 사용하기 때문에 수치가 다소 낮게 측정된 것으로 보인다. 또한, 이미지의 경우 Inception-v3 모델에서 추출하는 한정적인 단어를 이용하여 유사도 값을 도출했기 때문에 텍스트의 평균값보다 이미지의 평균값이 더 높게 측정되는 것으로 보인다. 실험 결과 텍스트와 이미지를 각각 사용한 경우의 평균값보다 텍스트와 이미지를 함께 사용했을 때 평균값이 더 높게 측정되는 것을 알 수 있다. 사용자에게 정보제공자 팔로잉 추천을 할 때 텍스트와 이미지를 따로 사용하는 것보다 텍스트와 이미지를 함께 사용하는 것이 더 유사한 관심사를 가진 정보제공자를 추천해 줄 수 있다는 것을 알 수 있다. 또한, 평균값을 이용하여 사용자와 가장 유사한 정보제공자를 추천해 줄 수 있다. 텍스트와 이미지를 함께 사용했을 때 Arts_info_2인 정보제공자의 평균이 0.5445로 가장 높으므로 사용자와 가장 유사하다고 할 수 있으며, 사용자에게 정보제공자 팔로잉 추천하기에 가장 적합하다.

V. 결론

본 논문에서는 SNS 게시글의 이미지와 텍스트를 이용하여 정보기반의 팔로잉 추천 방법에 대해 제안하였다. SNS 정보제공자의 게시글에서 텍스트와 이미지를 추출하여 재정의한 DMOZ를 이용하여 정보제공자 DB를 구축하였다. 기존의 DMOZ는 디렉토리 형식의 16개의 대분류 카테고리 구성되어 있으나 최상위 카테고리 및 최하위 카테고리를 이용하여 12개의 카테고리로 재정의 하였다. 텍스트에서 불용어를 제거하여 단어를 분류한 후 재정의한 DMOZ를 이용하여 카테고리를 분류했다. 이미지는 TensorFlow 라이브러리를 이용하여 CNN 기법을 모델링 한 Inception-v3를 이용하여 분류한 후 재정의한 DMOZ를 이용하여 카테고리를 분류했다. 분류한 텍스트와 이미지를 태깅하여 정보제공자 DB에 저장한 후 정보제공자의 게시글을 카테고리 분류한 방법과 동일하게 사용자의 게시글을 카테고리 분류한다. 사용자 카테고리 및 일치하는 정보제공자 DB의 카테고리에서 정보제공자들의 게시글과 사용자의 게시글을 유사도 분석하여 가장 높은 값이 추출되는 정보제공자의 계정을 추천해준다. 향후 연구로 다양하고 복잡한 정보를 제공하기 위한 복합 대분류 카테고리를 정의하는 방법과 DMOZ 카테고리 및 SNS에 적합한 이미지 분류를 위한 머신러닝 모델 구축에 관한 연구를 진행할 계획이다.

References

- [1] Y.W. No, D.Y. Kim, J.E. Han, M.S. Yook, J.T. Lim, K.B. Bok, et al., "Hot Topic Prediction Scheme Considering User Influences in Social Networks", *Journal of the Korea Contents Association*, Vol. 15, No. 8, pp. 24-36, 2015.
- [2] K.J. Cha, E.M. Lee, "An Empirical Study of Discontinuous Use Intention on SNS : From a Perspective of Society Comparison Theory", *The Journal of Society for e-Business Studies*, Vol. 20, No. 3, pp. 59-77, 2015.
- [3] I.K. Ha, "Analysis of Research Trends on Social Network Service: focusing on the Studies of Twitter", *Journal of the Korea Contents Association*, Vol. 14, No. 9, pp. 567-581, 2014.
- [4] S.H. Hur, K.S. Choi, "A Study on characteristics and types of tweet in twitter", *Hanminjok Emunhak*, Vol. 61, pp. 455-494, 2012.
- [5] H.W. Jang, S.S. Cho, "Automatic Tagging for Social Images using Convolution Neural Networks", *Journal of Korea Information Science Society*, Vol. 43, No. 1, pp. 47-53, 2016.
- [6] M.D. Hong, K.J. Oh, M.H. Ga, G.S. Jo, "Content-based Recommendation Based on Social Network for Personalized News Services", *Journal of Intelligent Information Systems*, Vol. 19, No. 3, pp. 27-71, 2013.
- [7] S.Y. Yoo, O.R. Jeong, "Social Category based Recommendation Method", *Journal of Korean Society for Internet Information*, Vol. 15, No. 5, pp. 73-825, 2014.
- [8] L.H. Tian, Y.J. Kim, B.H. Kim, M.S. Lee, "Personalized Information Recommendation Technique based on Social Networks", *Journal of Korea Information Science Society : Computing Practices and Letters*, Vol. 19, No. 12, pp. 668-672, 2013.
- [9] S.M. Ahn, "Deep Learning Architectures and Applications", *Journal of Intelligent and Information Systems*, No. 22, Vol. 2, pp. 127-142, 2016.
- [10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, "Rethinking the Inception Architecture for Computer Vision", *Arxiv*, 2015.
- [11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., "ImageNet Large Scale Visual Recognition Challenge", *International Journal of Computer Vision*, Vol. 115, Issue. 3, pp. 211-252, 2015.
- [12] M. Abadi, A. Agarwal, P. Barham, et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", *Arxiv*, 2016.
- [13] M.S. Hong, O.J. Lee, W.J. Lee, J.D. Lee, "Meta-data Configuration and Wellness Feature Analysis Technique for Wellness Content Recommendation", *Journal of the Korea Society of Computer and Information*, Vol. 19, No. 8, pp. 83-93, 2014.
- [14] K.M. Kim, D.Y. Kim, J.H. Lee, "Measuring Similarity Between Movies Based on Polarity of Tweets", *Journal of Korean Institute of Intelligent Systems*, Vol. 24, No. 3, pp. 292-297, 2014.

저 자 소 개

**홍택은(준회원)**

2015년 조선대학교 컴퓨터공학부 공
학사

2015년~현재 조선대학교 소프트웨어
융합공학과 석사과정

<주관심분야 : 소셜 네트워크, 오피니
마이닝, 감성정보 분석>

**신주현(정회원)**

1986년~2011년 (주)청진정보 팀장, (주)
투루텍 기술이사

2007년 조선대학교 전자계산학과 이
학박사

2011년~현재 조선대학교 제어계측로
봇공학과 산학협력중점교수

<주관심분야 : 멀티미디어 데이터베
이스, 빅 데이터 처리, 텍스트마이닝, 감성정보 처리
등>