「정보시스템연구」 제26권 제4호
한국정보시스템학회
2017년 12월, pp. 285~301

http://dx.doi.org/10.5859/KAIS.2017.26.4.285

온라인 커뮤니티에서의 건강 관련 콘텐츠 분류 모형 개발*

김태윤** · 김유신*** · 최상현**** · 김도훈***** · 장유진******

〈목 차〉

Ⅰ. 서 론

1.1 연구 배경

웹 2.0과 소셜미디어 같은 협력 플랫폼 기술이 발달하면서 대중의 지혜를 활용하는 집단지성 서비스가 다양한 분야에서 제공되고 있다(Joo and Normatov, 2012). 이러한 대표적인 플랫폼으로서 WikiAnswer, Yahoo! Answer, Naver Knowledge-in 등과 같은 온라인 커뮤니티의 Q&A 사이트들이 등장하게 되었고 국내 사이트들의 점유율을 비교한 결과 네이버 지식인은 다음 지식, 네이트 지식에 비해 전체 80%가 넘는 점유율을 보임으로서 국내 1위의 입지를 갖고 있다(권순찬, 2009).

네이버 지식인의 경우 교육, 게임, 생활, 경제, 쇼핑, 건강 등 다양한 영역으로 세분화되었으며 2015년 11월 9일 기준 하루 11,532건의 질문, 12,427건의 답변이 새로이 달리는 것을 확인할 수 있었다(네이버 지식인). 하지만 위키

* “본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음” (IITP-2017-2013-0-00881)
** 충북대학교 경영정보학과, kim-taeyoun@hanmail.net (주저자)
*** 서울시립대학교 빅데이터분석학과, yoosin25@uos.ac.kr
**** 충북대학교 경영정보학과, chois@cbnu.ac.kr (교신저자)
***** 충북대학교 의과대학, mwille@naver.com
****** 인제대학교 의과대학, yjchang0110@gmail.com

피디아, 네이버 지식인과 같이 집단지성을 활용한 Q&A 서비스의 정보가 일반인에 의한 정보인 만큼 그 신뢰성에 대한 논란이 끊이지 않고 있다(Joo and Normatov, 2012).

Adams(2010)은 웹 2.0 시대에 무궁무진하게 발생하는 온라인 건강 관련 정보들의 신뢰성에 대한 의문을 제기하고 문헌연구와 몇 가지 원칙들을 사용하여 비전문가들로부터 생성되고 있는 품질 낮은 정보들을 평가하고 이에 대한 자각 및 고취에 관한 연구를 진행하였다. 이와 흡사하게 Savolainen(2011)은 인터넷 정보의 신뢰성 및 품질을 측정하기 위해 콘텐츠의 유용성, 정확성, 구체성 등의 속성 분류 기준을 세웠고 저자에 대해서도 평판도, 전문성, 정직성으로 나누어 인터넷 정보에 대한 평가를 진행하였다. 이렇듯 웹 2.0 시대와 더불어 새로운 지식 구축의 메커니즘으로 부상하고 있는 집단지성이 새로운 지식의 구축의 대안이자 혁신적인 모델로 평가받고 있지만 네이버 지식IN이나 다음 아고라 같은 온라인커뮤니티가 지식 정보의 오류에 부닥칠 수밖에 없는 현실이다(정종철, 2014).

특히 인터넷 검색을 통해 제공되는 건강 관련 정보는 이용자들이 향후 질병 치료, 예방과 같은 건강 행동을 정할 때 매우 중요한 정보가 될 수 있기 때문에 정보원의 신뢰성이나 정보의 정확성은 매우 중요한 이슈가 될 수 있다(조수영, 2011). 게다가 인터넷을 통한 건강정보의 검색은 간편하다는 장점이 있으나 전문가에 의해 집필되는 의학서적과는 달리, 누구라도 정보를 올릴 수 있기 때문에 일반 이용자들에게 검증되지 않은 지식을 제공하여 혼란과 부작용을 일으킬 수 있다(Lee and Moon, 1997).

본 논문에서는 국내 집단지성의 가장 대표적인 사이트인 네이버 지식인을 대상으로 지식 소비자들에게 얼마나 유용하고 신뢰성 있는 답변을 제공하는지 실태에 대해 실증적으로 계측하여 문제점을 고찰하는데 목적이 있다. Gao(2015)는 웨이보에서 빈번하게 발생하고 있는 근거 없는 건강 관련 정보들의 확산으로 인해 해당 환자나 정보를 믿고 의사결정을 하는 사람들에게 심각한 결과를 초래한다고 주장하면서 중국인들이 인지하고 있는 신뢰 요인들을 분석한 결과를 제시하였다.

이에 우리는 네이버 지식인에서 생성되고 있는 건강 관련 정보를 수집하여 Content 분석을 실시하여 질문과 답변의 유형을 확인하였고 특정 질병에 관해 어떤 유형의 질문과 답변이 생성이 되는지 확인 고찰하였다. 또한 전문가 및 일반 지식인들로부터 생성된 답변의 유효성을 파악하기 위해 답변의 의학적 근거 유무를 판별하였으며 의학적 근거가 있다고 판단한 정보에 대해서는 의학적 수준을 점수화하여 데이터를 분류하였다. 질문 답변의 유형 분류 및 의학적 수준에 대한 평가는 실제 병원에 종사하고 있는 해당 분야의 전문 의료진의 도움을 받아 연구를 진행하였다.

답변유형의 경우 치료, 관리, 진단 등의 양질의 정보가 있었던 반면, 지식인 사용자들을 자칫 호도할 가능성이 있는 홍보 및 광고 답변이 있었기에 이를 자동으로 분류 할 수 있는 필터링 방법론을 제안하고자 한다.

이를 통해 네이버 지식인의 건강 관련 지식의 유효성을 객관적으로 분석하였고 환자들을 호도할 가능성이 있는 건강 정보의 위험요인에 대해 자동 필터링을 할 수 있는 시스템을 제안

함으로서 웹 2.0 시대의 집단지성의 문제점을 보완하고 관리하는데 기여 할 수 있을 것이라 기대하는 바이다.

Ⅱ. 이론적 배경

2.1 집단지성의 발현

웹 2.0 확산으로 인터넷 및 모바일 이용자가 적극적으로 참여하여 정보와 지식을 생산・공유・소비하고 있다. 웹 2.0 환경 하에서 이용자가 직접 지식 정보의 생산에 참여하고 활발한 공유・전파 활동을 수행하면서 지식 정보의 생산과 유통 비용도 획기적으로 감소하며 그 결과 양질의 지식 정보를 저비용으로 획득하는 것이 가능해진다(김종길, 2010).

이와 더불어 집단지성의 개념이 1994년 Levy에 의해 본격적으로 탐구되었는데 네이버 지식in에서 하나의 질문에 대해서 여러 네티즌들이 답변을 하는 것, 다음 카페에서 네티즌들이 서로 노하우를 공유하는 것 모두 집단지성으로 불린다. 집단지성의 '집단'은 전문가로 공인 받지 않았지만 일상에서 체험한 지식들을 공유하며 지식의 공동생산에 나서는 일반인들을 의미하는 것으로 받아들이는 경향이 나타나게 되었다(최항섭, 2009).

이를 통해 집단지성의 개념이 웹2.0의 발달로 온라인상의 다양한 형태의 커뮤니티를 형성하게 되었고 이를 통해 지식의 생성・공유・소비 활동이 더욱 활발하게 진행되어 온 것을 알 수 있다.

이러한 온라인상의 집단지성을 통한 지식이 반드시 현명한 판단을 내리는데 도움을 줄지는 아직 의문이다. 집단지성이 발현되기 위한 조건중의 하나가 누구나 참여가 가능해야 하며, 이를 통해 다양한 구성원들이 어려움 없이 참여할 수 있게 되어 다양한 의견들이 제시되어야 한다고 말하고 있다(김태원, 2013).

그러나 집단의 어리석음에 대하여 집단적으론 지혜롭지만 개인적으론 멍청한 개미와 달리 인간은 개인적으로 지혜롭지만 집단적으로는 어리석다며 집단사고의 위험성을 강조하였다(김태원, 2013).

인터넷 집단지성으로 대표되고 있는 네이버 지식인 및 다음 지식의 경우 일반 인터넷 이용자 즉 네티즌들이 직접 질문하고 그에 대답하는 구조로 되어 있다. 이런 구조 하에서 지식검색 프로그램 상에 올라 온 답변들 혹은 지식정보들이 사실이나 진실만을 담고 있는지 확신할 수 없는 문제가 발생하게 된다. 별도의 책임있는 단위를 통해 담보되지 못함으로 인해 많은 지식 정보들이 잘못된 채로 올라오고 그대로 방치될 수 있는 가능성이 제기 되는 것이다(이종철, 2014).

지식검색 사이트상의 지식 정보가 오류를 갖게 될 가능성은 결국 '집단지성'을 통한 지식정보의 구축이 갖는 한계와도 직결된다(이종철, 2011).

2.2 인터넷 정보의 신뢰성 및 품질 측정

본 연구에서는 집단지성의 대표적인 예인 네이버 지식인을 대상으로 건강 관련 정보의 콘텐츠 분석을 통해 유효성을 측정하고 유효하지 않은 글의 특징을 추출하여 자동 필터링 할 수

있는 방법론을 제시하고자 한다.

오래 전부터 Q&A사이트 정보의 신뢰성 및 품질을 측정하는 연구가 국내외에서 활발히 진행되었다. Shachaf(2010)는 Q&A사이트의 답변 신뢰성을 분석하기 위해 Askville, WikiAnswers, Wikipedia, Yahoo! Answers 으로부터 1522개의 랜덤 샘플데이터를 수집하여 정확성, 완전성, 검증가능성에 대한 분석을 진행하여 사이트별로 비교 연구를 진행하였다. 그 결과, Wikipedia가 정확성과 검증가능성 면에서 다른 사이트들 보다 우수하였고 Yahoo! Answers의 경우 가장 낮은 수준의 검증 가능성을 보였다고 결과를 제시하였다. 이를 통해 사이트의 대중성과 정보의 품질은 서로 상관관계가 없다는 것을 증명하는 연구를 진행하였다.

또 다른 연구는 Gao et al.(2015)는 중국의 Weibo의 건강관련 정보들을 대상으로 잘못된 의학정보들의 영향으로 인해 환자들의 치료를 위한 의사결정 시 심각한 결과를 초래할 수 있는 가능성이 있다고 주장하며 중국인들이 해당 사이트의 어떠한 요인으로 인해 건강정보에 신뢰성을 갖게 되는지를 연구하였다. 이에 Weibo 사이트의 신뢰성 요인을 출처, 메시지, 정보의 구분 등으로 나누고 메시지의 경우는 주장의 타입(객관성, 주관성) 과 그의 정도를 (상, 중, 하)로 나누어 분석을 진행하였고 정보의 구분에서는 저자, 긍·부정 댓글, 인용수 등으로 나누어 건강 정보의 신뢰성에 미치는 영향요인을 분석하였다.

그 결과 출처가 명확할수록, 재인용수가 많을수록 신뢰도가 늘었고 부정적 댓글이 많을수록 신뢰도가 떨어짐을 증명해냈다.

이렇듯 지식 검색 사이트가 주는 편리함은 있지만 잘못된 건강정보의 제공은 질병악화, 유병기간 연장, 치료비용의 증가뿐만 아니라 사람의 생명과도 직결되기 때문에 국가차원의 신뢰할 수 있는 양질의 정보를 제공하는 사이트의 운영과 건강정보를 제공하는 민간사이트의 질 관리를 위한 방안이 필요하다(송태민, 2006).

최근 SNS의 발달로 인해 트위터나 페이스북에서도 서로간의 건강 정보 메시지를 주고 받는데 부적절한 정보가 비일비재하게 발생하는 것을 우려한 존스홉킨스 의료진들이 트윗 메시지를 읽고 저자와 내용에 따라 카테고리를 분류하고 어떤 유형의 글이 주로 나타나는지를 질적 연구 통해 밝혀냈다.

Lee et al.(2014)는 700개의 건강관련 트윗 메시지를 저자의 직업별로 먼저 분류를 하였고 직업별로 내용의 카테고리를 분류한 결과 기관 및 기업의 경우 자신의 상품과 서비스 판촉 행위를 위한 메시지를 전달하였고 환자 대변인은 건강에 대한 개인의 경험을 공유하고자 트윗 메시지를 이용하였다. 트위터는 검증 가능한 내용, 뉴스, 광고 등의 내용이 주를 이루고 있음을 콘텐츠 분석을 통하여 밝혀내었다.

이러한 선행 연구를 바탕으로 본 논문에서는 네이버 지식인의 건강관련 질문과 답변의 콘텐츠 분석을 통해 유형을 정의 및 분류 하였으며 의학적 수준이 어느 정도인지를 파악하기 위해 충북대학병원 의료진과 공동연구를 진행하여 평가하였다.

2.3 콘텐츠 분류 알고리즘

질문&답변 유형을 분류하여 기초통계를 확인한 결과 상당수의 사익을 목적으로 한 의학

적 근거가 없는 광고 글이 많이 포함되어 있음을 확인하였다. 이에 우리는 정보성 답변과 광고성 답변을 자동 분류할 수 있는 방법론을 제시하려고 한다. 이러한 문서 분류에 관련 연구는 오래 전부터 활발히 진행 되어 왔다.

김현준 등(2004)은 상업성 광고 및 불법, 음란광고로 인해 기업 및 개인이 쏟는 시간과 비용의 문제가 사회문제로 부각되자 이를 방지하고자 스팸 메일 필터링 시스템을 개발하여 보다 정확한 분류 정확도를 높이고자 하였다. 분류 알고리즘 중에 하나인 나이브베이지안을 활용하여 본문의 텍스트, HTML Tag, HTML Link, HTML subject, 4가지 종합한 ALL 등 5가지 경우의 수를 갖고 nonspam과 spam 데이터 분류실험을 진행하였다. 그 결과 각 속성을 개별 독립변수들로 활용한 나이브베이지안 보다 4가지를 종합하여 가중치를 부여한 나이브베이지안의 성능이 분류정확도, 재현율 F-Measure 측면에서 높게 나오는 결과를 보였다.

이태원, 홍태호(2015)는 고객리뷰 감성분류를 위해 대표적인 기계학습기법인 SVM을 적용하고, SVM의 입력변수 선정과정에 품사태깅 방식과 용어추출기법을 다르게 조합하고 사용하여 긍정적/부정적 문서를 분류하였다. 이때 추출된 용어는 문서빈도, TF-IDF, 정보획득량, 카이제곱 통계량으로부터 나왔으며 각 상위 20개에 해당하는 최적의 용어를 선정한 후 SVM을 이용하여 고객 감성 분류를 시도하였다. SVM(Support Vector Machine), NB(Naive Bayesian)등을 포함하는 기계학습 기반의 분류 알고리즘은 트레이닝 데이터와 그것을 분류하기 위한 특징 추출이 필요하다. Pang et al.(2002)는 다양한 크기의 용어, 빈도수, 단어의 존재여부 등을 변수로 사용하여 실험을 진행하여 2000개의 영화 감성 관련 용어를 활용한 SVM 분류 실험에서 81.4%의 분류 성능을 보인바가 있다. Sohn et al.(2009)는 스팸 필터링을 위해 스팸메시지의 길이, 단어의 빈도수, 이모티콘, n-grams 그리고 특별한 단어들의 빈도수를 갖고 실험을 진행하였다.

2.3.1 SVM을 활용한 유효 콘텐츠 분류

Vapnik(1995) 는 이진 분류 문제 해결을 위해 기계학습기법을 활용한 학습이론에 기반하여 알고리즘을 개발하였다. SVM(Support Vector Machine)비선형문제를 선형문제로 해결하기 위해 데이터를 고차원적인 특정 공간에 서로 다른 클래스로 분리하여 최적분리경계면(Maximum margin hyperplane)을 찾고 이와 가장 가까운 점인(Support vector)와의 최대거리를 확보하여 클래스를 분류하는 방법이다. 이태원, 홍태호(2015)는 SVM이 분류와 회귀문제를 해결할 수 있으며, 다른 분류기법들과 비교하여 우수한 성능을 보인다고 하였으며 SVM에서 사용하는 커널함수와 파라미터의 설정값에 따라 분류모형의 성능이 달라지기 때문에 학습용 데이터를 통해 최적의 파라미터 값을 도출하고 검증용 데이터를 예측하였다.

이와 같이 본 논문에서는 네이버 폐암 관련 답변의 유형을 분석하여 텍스트 특징을 추출하고 이를 변수로 활용한 분류실험을 진행하였다. 해당 실험에서는 분류알고리즘의 성능비교보다는 비정형데이터에서 도출된 변수에 초점을 두었기 때문에, 기계학습에서 대체로 가장 높은 성능을 보이는 SVM만을 활용하였다. 기존에

선행되었던 웨이보, 위키피디아의 의료 Q&A에 대한 신뢰성 및 품질에 대한 연구는 주로 지식 소비자로부터 설문을 통한 양적연구였으며 저자, 댓글수, 인용수 등을 활용한 기초분석의 통계치를 활용하여 신뢰성 여부를 판단하였다. 하지만 본 연구에서는 비정형 콘텐츠로부터 의료진의 도메인 지식을 활용하여 특징을 추출하고 콘텐츠 유형을 분류하여 콘텐츠 자체에 대한 질적 분석과 분류알고리즘을 활용한 유해 콘텐츠를 분류할 수 있는 방법론을 제시함으로써 기존 연구들과 차별화하였다. 또한 의료 콘텐츠의 의학적 수준을 측정함으로써 질적 연구의 신뢰성을 높였으며 집단지성의 플랫폼의 품질 수준 향상을 제고하기 위한 기초를 마련하였다.

Ⅲ. 연구 모형 및 방법론

3.1 데이터

선행연구에 따르면 의학적 근거 없는 건강 정보로 인해 심각한 결과를 초래할 수 있다고 우려를 표명 하였다. 따라서 우리는 네이버 지식인의 건강 콘텐츠 유효성 분류 검증을 위해 폐암을 주제로 선정하였다.

질문	답변	저자	날짜
폐암4기 폐렴/폐혈증/저혈압	폐암은 우리나라에서 4번째로 흔히 발생하는데,암사망의	서라벌 대명리	2014-04-12
이거폐암 의심?	질문자의 연령과 기술한 증상을 고려했을때 걱정하는 폐암	duderak	2013-06-17
이거폐암 의심?	폐도 이제 지쳤다는 겁니다.담배30년 동안핀사람이 폐암	junghr1234	2013-06-17
폐암인지 가리켜주세요..	안녕하세요. 하이닥-네이버 지식IN 내과 상담의 진성림입	bestdr1	2012-09-10
폐암4기 환자 대처	병세를 호전시킬 수 있고 신체적 여건(기초체력과 면역체	레스톤프로폴	2015-03-06
20살남자 폐암초기증상인가요	1. 열은 없는데 숨을쉬면 목이 간지러워서 기침을 많이하	비공개	2012-06-21
폐암초기증상질문	어떤 검사를 하는냐는 병원의 진료과정에 맞기셔야죠 보통	shinjiya	2012-08-02
폐암초기증상질문	폐암 증상으로는 기침이랑 각혈 등이 있고X-RAY찍어보셔	wishsub	2012-08-10
폐암에 대한 질문	역학과 고려수지침의 자석요법을이용하면 담배를 금단현	자석으로 담배	2014-12-08

<그림 1> 네이버 지식인 데이터 수집

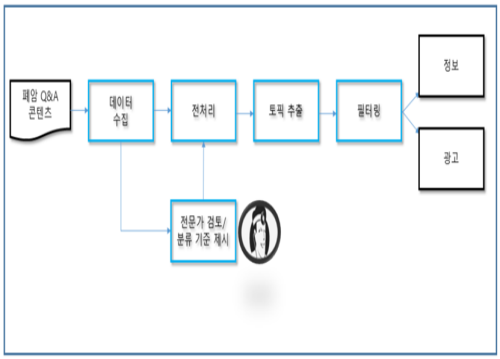
데이터는 2012-1-5부터 2015-6-9의 기간 동안 질문, 답변, 저자, 날짜를 <그림 1>과 같이 수집하였다.

수집한 데이터는 총 질문 784개, 답변 1542개이며 특정 저자로부터 편향된 답변이 중복적으로 나오지 않는지 파악하기 위하여 저자와 글을 올린 날짜까지 수집을 하게 되었다.

저자의 경우 내과, 흉부외과 등의 다양한 전문의를 비롯해 한의사도 많이 포함되어 있었다. 또한 특정제품을 본인의 닉네임으로 정하고 홍보 활동을 하는 저자가 상당수 포함되어있다.

3.2 유효 콘텐츠 분류 모형

다음 <그림 2>의 모형은 온라인 Q&A 커뮤니티에서 데이터를 수집하고 전처리하는 과정을 거쳐 내용분석 및 필터링 시스템을 하는 것을 나타내고 있다.



<그림 2> 유효 콘텐츠 분류 모형

3.2.1 폐암 Q&A 콘텐츠

네이버 지식인은 국내 대표적인 온라인 Q&A 커뮤니티로 성장하였고 실제로 2015년 11월 9일 하루 기준으로 11,532건의 질문이 기

록되고 12,427건의 답변이 새로이 생성됨으로서 평균 1.1건의 질문에 따른 답변이 생성 있다.

네이버 지식검색 서비스는 이용자가 궁금한 것을 질문하고 다른 이용자 및 전문가가 이에 답변을 하는 일련의 과정을 끊임없이 반복함으로써 하나의 집단지성을 이루어 수많은 지식소비자들에게 유용한 혜택을 주었다. 집단지성은 지식 구축을 위한 혁신적인 모델로 평가를 받고 있지만 그 지식에 대한 전문성 결여, 정확성, 신뢰성 등 다양한 문제에 직면하고 있다.

이에 우리는 네이버 지식인의 수많은 영역 중에 건강 관련 Q&A 콘텐츠를 활용하여 온라인 집단지성의 유용성을 증명하고자 한다.

3.2.2 데이터 수집

네이버 지식인에서 생성되고 있는 폐암 관련 집단지성의 행태를 파악하고 실질적으로 계측하고자 질문, 답변을 수집하였고 더불어 일부 동일 저자로부터 편향된 정보가 업로드 되는 것을 확인하고자 저자와 날짜를 추가 수집하여 기술통계에 활용하고자 한다.

수집된 데이터에 대한 평가는 의과대학 연구자 3명이 2015년 8월 1일부터 10일까지 열흘간 진행되었고 분석대상이 되는 784개의 질문과 1542개의 답변을 수집하였다. 폐암 관련 질문, 답변의 경우 평균 하나의 질문에 1.96개의 답변이 달린 것을 확인하였다.

3.2.3 전처리

3명의 연구자가 함께 네이버 지식인 Q&A 페이지를 나누어 데이터 수집을 했지만 일부 중복의 우려가 있어 취합 후에 중복을 제거하였으며 질병과 관련이 없거나 한 줄 이내의 성의 없는 답변의 경우 Garbage Data로 인식하여 제거 하였다.

3.2.4 전문가 검토 및 분류

데이터 전처리를 마치고 질문 유형과 답변의 유형에 대한 기준을 정의하고 이에 맞게 분류하는 작업을 진행하였다. 모든 질문 및 답변 분류 기준은 병원 의료진의 질병 주기 별 단계에 근거하여 마련을 하였다.

다음 보는 바와 같이 질문의 유형은 의심증상, 치료, 예후, 관리 분류하였고 환자의 사회적 보상, 의료비 지원 등 질병의 지원적 활동을 기타로 정의 하여 총 5가지 카테고리로 나누었다.

<표 1> 질문 유형 정의

질문 유형	정의
의심 증상	환자들이 느끼는 주관적인 증상이나 행위 이후 느끼는 불안감에 생기는 질문이므로 의심과 증상을 같은 범주로 통합
치료	약물, 방사선, 수술 등 모든 치료를 포함할 수 있으며 의학적 타당성을 떠나 면역치료, 생식치료, 한방치료에 관한 질문
예후	생존율, 재발율, 전이여부 등의 모두 예후
관리	특정 질병에 관해 좋은 음식, 운동 등
기타	환자의 사회적 보상, 의료비 지원, 사회적 관리 등은 질병의 주된 활동 외적인 것으로 따로 분류함

다음은 답변의 유형이다. 답변의 내용은 좀 더 상세하게 나누어졌으며 해당 질병의 본원적 답변(치료, 증상, 진단, 관리 등) 보다는 비전문가로부터의 조언이나 종교 및 특정 제품을 추

천하는 등의 글이 발생하는 것을 확인하였으며 이와 더불어 치료 방법을 소개하듯 하며 영리를 위한 광고 글이 다수 포함되어 있었다.

답변의 경우 하나의 질문에 한 가지만 답하는 일문일답 형식이 아니라 추가적인 내용을 소개하기 때문에 답변 유형은 중복을 허용하여 분류하였다.

<표 2> 답변 유형 정의

답변 카테고리	정의
치료	약물, 방사선, 수술 모든 치료를 포함할 수 있는 답변의 경우 치료라 정의
증상	질병 초기, 매 시기마다의 증상, 치료받으며 생기는 통증, 현상 등을 포함하는 경우 증상이라 정의
진단	질문자의 증상에 대해 답변자가 임의로 판단하여 설명하는 내용의 경우 진단이라 정의
한방	의학적 근거 유무를 떠나 한방치료를 권유하거나 한의원을 방문할 것을 추천하는 답변의 경우 한방이라 정의
조언	답변자가 질문자를 위로, 공감 하거나 본인의 경험을 예로 들어 치료를 위한 가이드를 제시할 경우 조언이라 정의
종교	질병 치료차원에서 과학적 근거는 없지만 기적의 치료법이나 특정 종교를 거론하며 지지하는 답변의 경우 종교라 정의
관리	질병에 좋은 음식, 운동법, 요양 방법 등은 모두 관리에 포함되므로 관리라 정의
기타	질문에 전혀 부합하지 않는 주제이거나 질병에 관한 비용, 산업재해 등을 거론할 경우 기타라고 정의
광고	특정 제품을 홍보하거나 특정 홈페이지 접속을 유도하는 글의 경우 광고라 정의

답변의 분류는 다음 <표 2>와 같이 정의하였고 유형 간 중복을 허용하기 위하여 특정 카테고리에 해당이 될 경우 숫자 ‘1’ 로 표기를 하였고 그렇지 않은 경우 모두 숫자 ‘0’으로 표기하여 구분하였다.

추가적으로 폐암 관련 답변의 의학적 수준을 구분하기 위해 의료진이 분류 기준을 정의하였고 <표 3>과 같이 상, 중, 하로 분류하였다.

<표 3> 답변의 의학적 수준 정의

답변의 의학적 수준 (상, 중, 하)	정의
상	의학적으로 폐암 관련 중요한 정보가 시의 적절하게 구성된 경우
중	타당한 의학적 정보가 일부 있는 경우
하	타당한 의학적 근거가 일부 있다 하더라고 선전이나 개인의 사견이 들어간 경우

병원 의료진과 본 논문의 연구자들의 공동 연구 진행으로 최종 분류 된 데이터를 활용하여 토픽 추출을 하게 되었다.

3.2.5 토픽 추출

우리는 답변의 유형을 분류하였고 확인한 결과 광고, 종교 영역에서 과학적으로 입증되지 않았거나 일반인들을 호도할 가능성이 있다고 판단하였고 이를 자동 분류 할 수 있는 모듈을 앞서 제안하였다. 따라서 순수 정보 영역과 광고 및 종교 영역에서 상당수 출현하고 있는 단어들을 파악하기 위해 텍스트 마이닝 툴인 R의 KoNLP 패키지를 활용하여 단어의 빈도 분석을 하였다. 광고와 종교에 해당하는 유형과의 조합에 의해 생성된 답변에서 주요 특징을 답변의 길이, 종교, 음식, 치료법, 광고 단어 등 5가지로 정하고 이를 활용하여 다른 영역과의

분류 정확도를 확인하기 위한 분류 실험을 진행하였다.

3.2.6 필터링

온라인 커뮤니티의 집단지성을 통해 우리는 의료 정보를 매우 쉽고 빠르게 제공 받을 수 있게 되었다. 하지만 일반인 혹은 사업가들로부터 근거 없는 의학적 정보나 영리를 위해 반복적으로 기재되는 광고 글을 확인 할 수 있었으며 이러한 잘못된 의료 정보는 분명히 해당 질병을 앓고 있는 환자에게 심각한 결과를 초래할 수 있는 가능성이 있다. 이에 우리는 토픽 추출을 통해 얻은 광고 및 종교 관련 글에 대한 특징을 활용하여 문서 분류 알고리즘인 SVM 변수로 사용하였다.

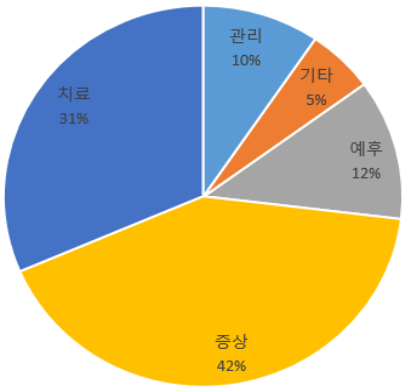
Ⅳ. 실험 및 결과

4.1 질문, 답변, 의학적 수준 분류 결과

4.1.1 질문 유형 분류

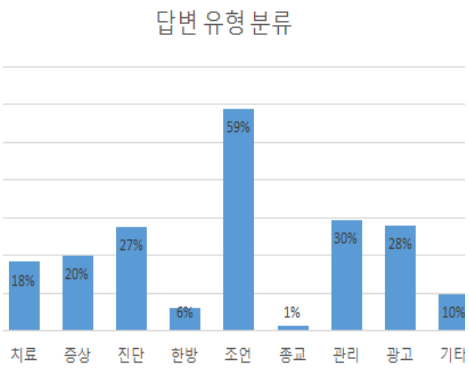
네이버 지식인의 의료 관련 Q&A 콘텐츠 분석 결과 질문의 경우 '증상' 카테고리가 42%로 가장 많은 비율을 차지하였고 '치료', '예후', '관리', '기타' 순이었다.

증상의 경우 본인이 느끼고 있는 질환에 대해 병원을 찾기 전 특정 질병이 맞는지 아닌지에 대해 불안함을 느끼며 네티즌들의 의견을 묻는 질문이 많았고 증상이 맞다는 가정 하에 해당 질병을 치료하기 위한 방법을 묻는 식의 질문이 주를 이루었다.



<그림 3> 질문 유형 분류 결과

4.1.2 답변 유형 분류



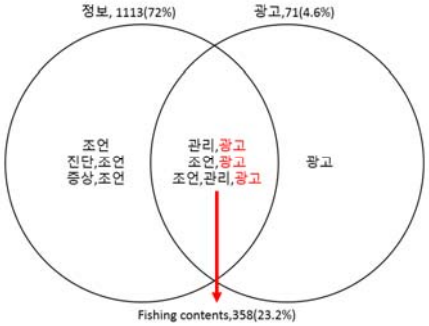
<그림 4> 답변 유형 분류 결과

답변의 유형을 분류 한 결과 중복 허용 시 대부분의 글에 공감 및 위로, 조언 등의 내용이 포함되어 있었고 질병을 관리하기 위한 음식, 운동법 등에 대한 소개도 주를 이루었다. 다음으로 광고가 28%를 차지하고 있는데 이는 개인의 영리나 홍보를 위한 목적이기 때문에 반드시 의학적 근거를 확인하여 지식 소비자의 건강상의 문제를 초래할 수 있음을 재차 확인해야 할 것이다.

<표 4> 답변 유형 구분표

구분	변수 구분
정보영역	광고와 종교를 제외한 나머지 유형간의 조합 영역
혼재영역	광고와 종교를 포함하는 유형간의 조합 영역
광고영역	순수 광고에 해당하는 영역

우리는 답변을 다시 정보 영역과 광고영역 그리고 둘이 합쳐진 혼재 영역으로 구분을 지었으며 구분표는 <표 4>와 같으며 그에 대한 비율은 <그림 5>에 다음과 같이 나타냈다.



<그림 5> 대분류에 따른 답변의 분류 결과

이 그림을 보면 정보영역의 경우 진단, 증상과 같은 질병 본원적 내용과 더불어 조언이 함께 쓰이는 것을 알 수 있었으며 혼재 영역의 경우 특정 질병에 좋은 음식, 운동법을 소개하는 관리와 함께 광고 답변이 함께 출현하는 것을 알 수 있었다.

의료진의 판단에 의해 확인한 결과 대부분의 영리를 위해 판매하고 있는 제품의 경우 해당 질병에는 별로 도움이 되지 않거나 의학적 타당성이 부족하다고 결론을 지었다. 하지만 광고 답변 중 글 서두에 선행되는 질병에 대한 부분에서는 과학적 근거를 제시함으로 인해 상당부분 혼동을 겪을 가능성이 있다고 지적하였다.

4.1.3 의학적 판별

광고 답변에서 일부 과학적 근거가 있다는 것을 착안하여 영역에 따라 의학적 근거 및 수준 정도를 알아보고자 한다. 정보 영역, 혼재영역, 순수광고 영역에서의 의학적 수준은 다음과 같다.

<표 5> 영역별 의학적 수준 분석 결과

의학적 수준	정보(#, %)	광고(#, %)	혼재(#, %)
상	12, 1%	0, 0%	0, 0%
중	246, 22%	0, 0%	1, 1%
하	287, 26%	1, 1%	75, 20%
의학 근거	545, 49%	1, 1%	76, 21%
의학 근거 없음	568, 51%	70, 99%	282, 79%
전체(답변 개수)	1113	71	429

정보 영역은 1113개의 답변 중 49%가 의학적 근거가 있는 글이며 의학적 수준은 앞서 정의 하였듯이 '상'은 1%, '중' 22%, '하' 26%의 비율을 차지하였다. 광고는 71개 중 오로지 1개의 '하' 답변만이 의학적 근거를 갖고 있었으며 혼재 영역의 경우 전체 답변 대비 21%가 의학적 근거가 있다고 판별이 되었다. 순수 광고 영역의 경우는 헬스케어 및 질병 관련 홈페이지를 링크를 따라 들어오라는 방식의 방문을 유도하는 글이 많아 상당부분 의학적 내용과는 상관이 없거나 근거 없는 내용이었다.

하지만 혼재 영역의 경우 일정부분 의학적 근거를 담고 있으며 이를 적절히 활용하여 지

식 소비자들로 하여금 특정 제품을 홍보하는 내용의 답변이 많았다. 실제로 일반인들이 보면 제품이 특정 환자들에게 효과가 있는 듯 보여 잘못하면 치명적인 결과를 실제로 낳을 수 있는 가능성이 제기 되는 부분이기도 하다. 정보 영역이라 할지라도 50%미만의 답변만이 의학적 근거를 갖고 기술을 한 내용이었고 그중의 1%만이 '상'에 해당하는 의학적 수준을 갖고 있었다. 이에 의료진이 측정한 헬스케어 및 의료 관련 데이터의 신뢰수준이 상당히 떨어지는 것을 알 수 있었다. 특히, 혼재 영역과 순수 광고 영역을 전체적으로 묶어 광고로 분류를 하고 이를 분류할 수 있는 자동 필터링 실험을 진행하였다.

<표 6> 필터링 실험 대상 구분표

구분	필터링 대상 영역	특징 추출 답변 유형
정보영역	필터링 대상 아님	없음
혼재영역	필터링 대상	종교, 광고
광고영역		순수광고

필터링의 대상은 <표 6>에서 보는바와 같이 구분하였으며 혼재영역의 종교, 광고 답변 유형에서 실험에 필요한 특징을 추출하였다.

4.2 유효 정보 필터링

문서 분류의 대표적인 데이터 마이닝 기법은 스팸 필터링이다. 스팸 메일을 걸러내기 위한 방법으로 기존에 연구되어지고 있는 문서 분류, 필터링 등의 데이터 마이닝 기법들이 사용되고 있다. 이러한 마이닝 기법을 활용하여 네이버 지식인의 유효한 정보와 유효하지 않은 정보를 자동 분류하기 위한 다음과 같은 과정을 수행하였다.

4.2.1 특징 추출(Feature Extraction)

선행연구에서 pang et al.(2002)는 용어, 빈도수, 단어의 존재여부를 활용하여 기계학습 기반의 분류 알고리즘을 수행하였다.

따라서 본 논문에서는 SVM 분류 알고리즘을 수행하기 위한 입력 변수 및 특징을 다음 <표 7>과 같이 정의하였다.

<표 7> SVM 실험을 위한 특징 정의

특징	정의
답변 길이	답변의 문자 수(광고, 종교 답변의 경우 상당히 길게 서술하는 경향을 보임)
종교 단어 빈도수	종교 관련 단어 중 가장 많은 빈도수를 갖는 제외어
음식 단어 빈도수	광고 답변에 포함되어 있는 음식 관련 단어 중 가장 많은 빈도수를 갖는 제외어
치료법 단어 빈도수	광고 답변에 포함되어 있는 치료법 관련 단어 중 가장 많은 빈도수를 갖는 제외어
광고 단어 빈도수	전체 답변 중 오로지 광고 영역에만 존재하는 단어의 빈도수

입력변수를 만들기 위해 정보 영역에서 가장 많이 출현하는 단어 3000개를 추출하였고 광고 영역과 혼재 영역에서 가장 많은 빈도를 보이고 있는 단어 3000개를 추출하여 일치하는 단어를 제외하고 남은 나머지 광고와 혼재 영역에서 주로 활용되고 있는 단어를 다음 <표 8>과 같이 특징을 재분류한 것이다.

혼재, 광고 영역에서 자주 출현하는 종교, 음

식, 치료법의 경우 의료진의 검토 결과에 따라 의학적 근거가 없으며 제외되어야 할 단어라고 판단하여 빈도순으로 단어를 추출하였다. 마지막으로 전체 답변에서 많은 빈도수를 보이는 단어에서 오로지 광고 답변에만 출현하는 단어를 선정하여 하나의 변수로 선정하여 실험을 진행하였다.

<표 8> 사전 예시 단어

사전	예시 단어	단어 개수
종교	하느님, 성서, 예수님, 부처님,기적	62
음식	건강식품, 발표식품, 차가버섯, 양배추	109
치료법	Adenosine, BBRC, Calebin	293
광고 단어	대체의학, 민간요법, 자연치료	1436

추가적으로 문서의 길이를 문서 분류를 위한 변수로 넣은 이유는 일반 지식 공유자들의 경우 문장의 길이가 짧고 간결한 경향을 보였지만 사업적 영리 목적을 가진 지식 공유자의 경우는 질병의 원인, 유명한 병원, 치료법등을 무척 상세하게 설명하는 경향을 보였기 때문이다. 5가지의 특징을 변수로 정하여 SVM 실험을 진행하였다. 해당 연구는 의료분야의 도메인 지식이 있는 의료진이 참여하여 어휘빈도수 기반의 키워드를 추출하여 변수를 구성하였기 때문에 보다 정교하고 신뢰성이 있다고 말할 수 있겠다.

4.2.2 SVM 실험 결과

우리의 실험 목적은 일반인들로부터 생성되는 네이버 지식인 답변이 환자 및 이를 활용하는 사람들을 호도하는지 혹은 그렇지 않은지를 분류하기 위함이다. 이번 실험을 위해 광고(혼재포함) 150개, 정보 150개의 무작위 표본 추출을 했고 테스트 데이터와 검증을 위한 데이터를 나누어 진행하였다. 알고리즘의 성능은 정확도, 재현율, F-measure (Kim,Jeong & Ghani,2014) 등으로 측정이 되고 본 연구에서는 전체 문서에서 정확히 분류된 문서의 비율에 대한 정확도만을 갖고 측정을 하였다. 실험에 대한 결과는 <표 9>와 같다. 분류기준은 32가지의 조합이 가능하나 특정 기준 이상의 정확도를 갖는 결과는 아래와 같이 9가지 실험이었다.

<표 9> SVM 실험 성능(정확도)

실험	분류 기준					정확도 (%)		
	광고	길이	종교	음식	치료법	최소	최대	평균
1	V					66.7	86.7	76.0
2		V				50.0	70.0	59.3
3	V	V				80.0	90.0	84.0
4	V		V	V	V	66.7	86.7	76.0
5	V	V	V	V	V	76.7	86.7	81.0
6	V		V			76.7	86.7	81.3
7	V			V		73.3	86.7	81.0
8	V				V	76.7	90.0	82.3
9	V	V			V	80.0	86.7	83.3

우리가 주목해야 할 실험은 3번째 경우이다. 실험 2의 길이만을 변수로 활용한 분류 정확도의 경우 59.3%로 가장 낮은 성능을 보였지만 길이와 광고 변수를 함께 사용한 실험이 평균 84%로 가장 높은 분류 정확도를 보였다. 실험 8,9의 광고+치료법의 실험과 광고+길이+치료법을 조합한 실험의 경우도 각각 82.3%, 83.3%

로 실험 3의 경우보다 낮은 성능을 보였다.

이번 실험을 통해 다른 변수의 사용 없이도 문서의 길이와 광고 성향을 보이는 1436개의 단어를 통해 84%의 성능으로 광고 답변을 자동 분류 할 수 있게 되었고 성능을 보다 높이기 위한 노력이 필요할 것이다.

실험을 위해 R의 “tm”, “KoNLP 그리고 SVM을 위한 e1091 패키지를 사용하였다.

Ⅴ. 결 론

네이버, 위키피디아, 야후 등과 같은 온라인 커뮤니티의 사용이 계속되고 있다. 이로 인해 일반인들도 지식을 생성하고 소비함으로써 집단지성이 현실화되었고 그에 따른 전문성의 결여로 인해 발생하는 문제가 연구를 통해 드러났다. 이에 우리는 헬스케어 및 의료 관련 정보의 전문성 결여 시 발생할 수 있는 문제의 심각성을 인지하고 콘텐츠 분석, 문서분류 등 다양한 분석을 시도했다.

먼저 온라인상에 건강 관련 질문과 답변이 어떻게 이루어져있는지 질문과 답변의 유형을 분류하여 기술통계로 확인하였다. 질문 유형은 의심증상, 치료, 예후, 관리, 기타로 나누어졌고 답변 유형은 치료, 증상, 진단, 한방, 조언, 종교, 관리, 기타, 광고 등 9가지로 분류되었다. 이렇게 분류 된 답변의 유효성을 구분하기 위해 실제로 현업에 종사하고 있는 의료진을 통해 답변의 의학적 근거를 판별하고 의학적 근거가 있는 답변의 경우 의학적 수준을 상, 중, 하 3점 척도로 구분하여 분류하였다.

분류한 결과 전체 질문 중 증상과 치료가 각각 42%, 31%로 가장 많은 비율을 차지하였고 답변은 1542개 중 622개의 답변(40.3%)만이 의학적 근거를 갖고 있다고 판명이 났으며 광고 답변이 27.8%를 차지하였다. 이에 우리는 광고 답변이 환자나 이러한 지식을 소비하는 사람에게 치명적일 가능성이 있다고 판단하여 이를 자동분류 할 수 있는 방법을 제안하였다.

본 연구의 목적은 첫째, 온라인 Q&A 커뮤니티에서 지식 소비자들을 자칫 호도할 수 있는 답변과 유효한 답변을 자동 필터링 할 수 있는 방법론을 제안하는 것이었고 둘째, 우리가 사용한 언어적 특징들을 활용해 얼마나 유용한 분류 정확도를 보이는지를 확인하는 것이었다. 분류 모형은 데이터 수집, 전처리, 전문가 검토/분류, 토픽추출, 유효 콘텐츠 분류 등으로 구성이 되어 있다. SVM 필터링 알고리즘은 답변의 길이, 종교 단어의 빈도수, 음식 단어 빈도수, 치료법 단어 빈도수, 광고 단어 빈도수와 같은 언어적 특징을 추출하여 변수로 활용하였다.

우리는 네이버 Q&A에서 폐암을 키워드로 검색한 결과를 바탕으로 질문 784개, 답변 1542개, 날짜, 저자를 수집하였고 기간은 2012년 1월 5일부터 2015년 6월 9일까지의 내용을 포함하고 있다. 우리는 이 데이터에서 유효성 답변 150개, 유효하지 않은 답변 150개 총 300개를 임의적으로 표본을 추출하여 SVM 알고리즘에 의한 실험에 활용하였다.

그 결과 광고와 문서의 길이를 조합한 변수가 10 Fold Cross Validation 실험에서 84%의 가장 높은 분류 정확도를 보였다. 우리는 이번 실험을 통해 답변 문서의 길이가 독자적인 변수로 쓰일 때의 정확도가 그리 높지 않다는 것을 알게 되었으며 다른 변수와의 결합을 통해

더 높은 성능을 보인다는 것을 증명하였다.

우리가 제안한 연구가 첫째로 의학정보를 이용하는 지식 소비자들에게 유용한 정보를 제공할 것이며 둘째로 답변 분류를 하는데 있어서 사용했던 언어적 특성 변수들이 유효하다는 것을 확인하는데 기여할 수 있을 것이라 기대한다. 또한 이러한 분석을 통해 우리는 온라인상에 건강 관련 질문과 답변이 어떻게 이루어지고 있는지에 대한 행태를 실증적으로 계측할 수 있었고 필터링 방법론을 제안함으로서 향후 온라인커뮤니티의 집단지성의 역기능을 조금이나마 해소하고 자동분류 시스템을 마련하는데 참고 논문으로 활용되기를 기대한다.

이번 연구의 한계점으로는 의학 분야를 대상으로 연구를 진행하긴 했지만 폐암이라는 아주 작은 분야에 초점을 맞추었기 때문에 우리가 제안한 언어적 특성 변수가 일반화되기는 어렵다. 또한, 유의하지 않은 답변을 필터링 하기 위해 언어적 특성만을 사용했지만 향후 연구에는 답변 채택수, 조회수, 저자등 다양한 변수를 활용하여 유용한 정보를 찾는 지식 소비자들에게 많은 도움을 줄 수 있을 것이다.

참고문헌

김달숙, 박아현, 강남준, “컴퓨터 분석프로그램을 적용한 암환자의 투병수기 분석”, 대한간호학회지, 제44권, 3호, 2014, pp.328-338.

김덕주, 박건우, 이상훈, “QualityRank : 소셜 네트워크 분석을 통한 Q&A 커뮤니티에서 답변의 신뢰 수준 측정”,정보과학회, 제37권, 6호, 2010, pp.343-350.

김태원, 김상욱, “집단지성 플랫폼으로서의 소셜미디어:커뮤니케이션 유형별 실험 분석”, 한국IT서비스학회, 제12권, 3호, 2013, pp.127-149.

김현준, 정재은, 조근식, “가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템”, 정보과학회, 제31권, 8호, 2004, pp.1092-1100.

송태민, “국내 건강정보 웹 사이트 현황 분석”, 보건복지포럼, 2006, pp.61-76.

연종흠, 심준호, 이상구, “확장된 나이브 베이즈 분류기를 활용한 질문-답변 커뮤니티의 질문 분류”, 정보과학회, 제16권, 1호, 2010, pp.95-99.

이종철, 오진아, “인터넷 지식검색 프로그램상 사회・역사 관련 지식 정보의 정확도에 관한 분석적 고찰 네이버(NAVER)「지식iN」과 다음(DAUM)「지식」의 실태를 중심으로”, 조사연구, 제15권, 2호, 2014, pp.149-186.

이태원, 홍태호,“Support Vector Machine을 이용한 온라인 리뷰의 용어기반 감성분류 모형”,Information Systems Review, 제17권, 1호, 2015, pp.49-64.

장종인, “지식 생산 및 전달 양식의 변화: NAVER 지식검색 서비스에서 찾아본 건강지식 사례분석”, 정보통신정책, 제18권, 16호, 2006, pp.1-18.

조수영, “인터넷 건강 정보의 정보원 유형과 상업 링크 유무, 질병의 심각성에 따른 설득 효과 차이”, 한국언론학보, 제55권, 3호, 2011, pp.123-152.

최항섭, “레비의 집단지성 : 대중지성을 넘어 전문가지성의 가능성 모색”, 사이버커뮤니케이션 학보, 제26권, 3호, 2009, pp.287-322.

Adams, S. A., “Revisiting the online health information reliability debate in the wake of “web 2.0”: An inter-disciplinary literature and website review”, International Journal of Medical Informatics, Vol.79, 2010, pp.391-400.

Delany, S. J., Buckley, M., & Greene, D., “SMS spam filtering: Method and data”, Expert Systems with Applications, Vol.39, 2012, pp.9899-9908.

Fichman, P., “A comparative assessment of answer quality on four question answering sites”, Journal of Information Science, Vol.37, No.5, pp.476-486.

Gao, Q., Tian, Y., and Tu, M., “Exploring factors influencing Chinese user’s perceived credibility of health and safety information on Weibo, Computer in Human Behavior, Vol.45, 2015, pp.21- 31.

Joo, J., and Normatov, I. R., “Relationships between Collective Intelligence Quality, Its Determinants, and Usefulness: A Comparative Study between Wiki Service and Q&A Service in Perspective of Korean Users”, Asia Pacific Journal of Information System, Vol.22, No.4, 2012, pp.75-99.

Lee, Y. M., and Moon, S. J., “Communication Technology and Network Information in food and Nutrition”,The Korean Journal of Nutrition, Vol.30, No.7, 1997, pp.870-878.

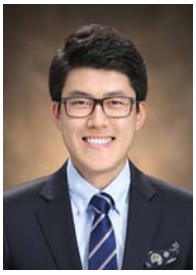
Pang, B., Lee, L., and Vaithyanathan, S., “Thumbs up? Sentiment Classification using Machine Learning Techniques”,In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2002, pp.79-86.

Savolainen, R., “Judging the Quality and Credibility of Information in Internet Discussion Forums”, Journal of the American Society for Information Science and Technology, 2011, pp.1-14.

Sohn, D. N., Lee, J. T., and Rim, H. C., “The contribution of stylistic information to content-based mobile spam filtering”, In Proceedings of the ACL/AFNLP 2009 conference short papers, pp.321–324.

Vapnik, V., The Nature of Statistical Learning Theory, Springer, 1995.

김 태 윤(Kim, Taeyun)

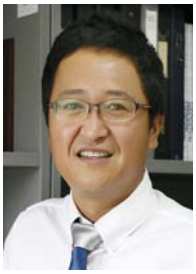


충북대학교 경영정보학사, 석사를 취득하였다. 현재 퍼니웍 선임컨설턴트로 재직하고 있으며, 주요 관심분야는 텍스트마이닝, 인공지능(AI) 시스템 등이다.

김 유 신(Kim, Yoosin)

국민대학교 정보관리학 학사, 경영정보학 석사와 박사학위를 취득하였다. 현재 서울시립대학교 빅데이터분석학 전공 객원교수로 재직하고 있으며, 주요 관심분야는 빅데이터와 텍스트마이닝, 인공지능(AI) 시스템 등이다

최 상 현(Choi, Sang Hyun)



한양대학교 산업공학사, KAIST 산업공학석사, 경영공학박사 학위를 취득하였다. 현재 충북대학교 경영정보학과 교수로 재직하고 있으며, 주요 관심분야는 빅데이터, 데이터마이닝, 지능형 안전 등이다

김 도 훈(Kim, Dohun)

계명대학교 의학사, 의학석사와 성균관대학교 박사 학위를 취득하였다. 현재 충북대학교 의과대학 교수로 재직하고 있으며, 주요 관심분야는 종양의 빅데이터적 접근, 흉부종양, 기흉, 다한증 등이다

장 유 진(Chang, Youjin)

인제대학교 의학사와 울산대학교 의학석사를 취득하였다. 현재 인제대학교 의과대학 교수로 재직하고 있으며, 주요 관심분야는 호흡기질환, 급성폐손상, 패혈증이다.

<Abstract>

Development of Classification Model for Healthcare Contents on the Online Community

Kim, Tae-Yun · Kim, Yoo-Sin · Choi, Sang-Hyun · Kim, Do-Hun · Chang, You-Jin

Purpose

In this paper we verified the reliabilities of healthcare-related information provided by various users on the site of Naver Jisikin, a Korean typical search platform. Based on Q&A contents we validated answers' reliabilities to the asked questions about a lung cancer with the help of professors at a medical school.

Design/methodology/approach

The content analysis includes that the types of questions are classified into symptom/diagnosis, therapy, prognosis, after-management and so on. The answers contains advice, advertisement, oriental medicine, and religion as well as the above 5 question categories. The validation results of medical evidence about each answer show that only 49% among all answers have medical grounds.

Findings

We classified the medical grounded answers into three levels; high, medium and low. Among all answers we need to find out the answers including advertisement because the answers can be harmful to patients. We found the method to select the answers containing advertisement contents with the help of text mining research. The selection model presents high performance as 84% classification accuracy.

Keyword: On-line Community, Content Validation, Healthcare, Q&A

* 이 논문은 2017년 10월 10일 접수, 2017년 11월 6일 1차 심사, 2017년 11월 28일 게재 확정되었습니다.