논문 17-42-02-25 The Journal of Korean Institute of Communications and Information Sciences '17-02 Vol.42 No.02
https://doi.org/10.7840/kics.2017.42.2.493

텍스트 분석의 신뢰성 확보를 위한 스팸 데이터 식별 방안

현 윤 진◆, 김 남 규°

Detecting Spam Data for Securing the Reliability of Text Analysis

Yoonjin Hyun◆, Namgyu Kim°

요 약

최근 뉴스, 블로그, 소셜미디어 등을 통해 방대한 양의 비정형 텍스트 데이터가 쏟아져 나오고 있다. 이러한 비정형 텍스트 데이터는 풍부한 정보 및 의견을 거의 실시간으로 반영하고 있다는 측면에서 그 활용도가 매우 높아, 학계는 물론 산업계에서도 분석 수요가 증가하고 있다. 하지만 텍스트 데이터의 유용성이 증가함과 동시에 이러한 텍스트 데이터를 왜곡하여 특정 목적을 달성하려는 시도도 늘어나고 있다. 이러한 스팸성 텍스트 데이터의 증가는 방대한 정보 가운데 필요한 정보를 획득하는 일을 더욱 어렵게 만드는 것은 물론, 정보 자체 및 정보 제공 매체에 대한 신뢰도를 떨어뜨리는 현상을 초래하게 된다. 따라서 원본 데이터로부터 스팸성 데이터를 식별하여 제거함으로써, 정보의 신뢰성 및 분석 결과의 품질을 제고하기 위한 노력이 반드시 필요하다. 이러한 목적으로 스팸을 식별하기 위한 연구가 오피니언 스팸 탐지, 스팸 이메일 검출, 웹 스팸 탐지 등의 분야에서 매우 활발하게 수행되었다. 본 연구에서는 스팸 식별을 위한 기존의 연구 동향을 자세히 소개하고, 블로그 정보의 신뢰성 향상을 위한 방안 중 하나로 블로그의 스팸 태그를 식별하기 위한 방안을 제안한다.

Key Words : Text Analysis, Text Mining, Topic Modeling, Spam Detection

ABSTRACT

Recently, tremendous amounts of unstructured text data that is distributed through news, blogs, and social media has gained much attention from many researchers and practitioners as this data contains abundant information about various consumers' opinions. However, as the usefulness of text data is increasing, more and more attempts to gain profits by distorting text data maliciously or nonmaliciously are also increasing. This increase in spam text data not only burdens users who want to obtain useful information with a large amount of inappropriate information, but also damages the reliability of information and information providers. Therefore, efforts must be made to improve the reliability of information and the quality of analysis results by detecting and removing spam data in advance. For this purpose, many studies to detect spam have been actively conducted in areas such as opinion spam detection, spam e-mail detection, and web spam detection. In this study, we introduce core concepts and current research trends of spam detection and propose a methodology to detect the spam tag of a blog as one of the challenging attempts to improve the reliability of blog information.

◆ First Author : Kookmin University The Graduate School of Business Information Technology, yoonjin0630@kookmin.ac.kr, 학생회원
° Corresponding Author : Kookmin University School of MIS, ngkim@kookmin.ac.kr, 정회원
논문번호 : KICS2017-01-009, Received January 9, 2017; Revised February 13, 2017; Accepted February 13, 2017

Ⅰ. 서 론

인터넷이 빠른 속도로 발전하고 스마트 기기가 대중화되는 등 일상생활 자체가 디지털화 되어감에 따라 매일 셀 수 없이 많은 양의 데이터가 쏟아져 나오고 있다. 이처럼 데이터의 양 자체가 문제의 일부분이 되는 빅데이터 이슈는 2012년 세계 경제 포럼을 비롯하여 이코노미스트(2011)[1], 맥킨지(2011)[2], 가트너(2011)[3] 등 주요 기관들에서도 그 중요성을 언급하였으며, 현재까지도 빅데이터 분석 기술에 대한 수요와 관심이 꾸준히 이어져 오고 있다. 특히, 최근에는 뉴스, 블로그, 소셜미디어 등을 통해 유통되는 비정형 텍스트 데이터에 대한 관심이 높아지고 있으며, 이러한 비정형 텍스트 데이터는 풍부한 정보와 사용자의 의견을 거의 실시간으로 반영하고 있다는 측면에서 그 활용도가 높아 많은 연구자들에게 각광을 받고 있다. 이에 국내 기업들 역시 텍스트 형태의 비정형 데이터를 분석함으로써 기존에는 파악하지 못했던 유용하고 새로운 정보를 얻어내고자 많은 노력을 기울이고 있다. 대표적으로 다음소프트의 소셜미디어 분석 솔루션인 '소셜 메트릭스'는 문맥 중심의 텍스트 마이닝 작업을 통해 각 데이터의 출현 원인 및 데이터들 사이의 관계를 도식화하여 제공하고 있으며, 와이즈넛은 트위터, 페이스북, 블로그, 카페 등에 올라온 대선 후보 관련 버즈(Buzz)를 분석하기 위한 '버즈인사이트바이럴 지수(BVI)'를 개발하여 제공하였다.

하지만 텍스트 데이터를 활용하여 할 수 있는 일이 많아짐과 동시에 이러한 텍스트 데이터를 악의적 혹은 비악의적으로 왜곡하여 특정 효과를 달성하려는 시도도 늘어나고 있다. 이러한 스팸(Spam)성 텍스트의 증가는 많은 양의 정보들 중 유용한 정보를 얻고자 하는 사용자들에게 불편함을 줄 뿐만 아니라, 정보 자체 및 정보 제공 매체에 대한 신뢰도를 떨어뜨리기 때문에 이를 미연에 방지할 수 있는 방안에 대한 연구가 반드시 필요하다.

이러한 스팸 데이터는 이메일(E-mail), SNS(Social Network Service), 블로그(Bolg) 등 다양한 유형으로 나타나고 있다. 이메일 스팸의 경우, 가장 고전적인 스팸 유형으로 정크 메일(Junk Mail) 혹은 벌크 메일(Bulk Mail)이라고도 불리며, 커뮤니티 사이트나 게시판 등에 게재되어 있는 이메일 주소를 수집하거나 단어나 숫자를 조합하여 수신자 이메일 주소를 생성해 원치 않는 상업적 이메일을 전송하는 방식으로 이루어진다. 이러한 이메일 스팸은 사용자가 원하는 메일을 손쉽게 찾는 것을 방해할 뿐만 아니라, 무분별한 스팸 메일 전송으로 인한 수신 서버 과부하 등의 문제를 불러일으킬 수 있다. 한편, SNS 스팸의 경우, 일반적으로 SNS만의 특징인 멘션(Mention)이나 해시태그(Hashtag) 등을 사용하여 스팸을 확산시키는 경우가 대부분이지만, 최근에는 스패머(Spammer)들 사이의 조정된 게시 활동을 통해 스팸을 확산시키거나, 사용자가 특정 키워드를 검색했을 때 스팸을 동시에 노출시키는 패시브 스팸(Passive Spam)[4] 등 더욱 어렵고 진화된 방법으로 스팸의 생성 및 확산이 이루어지고 있다. 이러한 SNS 스팸은 온라인을 통해 자유로운 의사소통을 하고자 하는 사용자들이 원치 않는 광고성 글에 노출되게 함은 물론, 스패머들과의 원치 않는 팔로잉(Following)을 통해 사용자들의 활동이 불특정 다수에게 노출됨으로써 여러 부작용을 낳고 있다. 이에 따라 대표적인 SNS 서비스인 트위터의 경우, 악의적인 링크를 게시하거나 최신 인기 주제로 해시태그를 작성하는 등의 스팸을 막기 위해 사용자가 직접 스팸을 신고하는 시스템을 도입하고 있으며, 이때 스팸 계정으로 확인이 될 경우 계정을 정지시키는 등의 시도가 이루어지고 있다. 또한 최근 페이스북에서는 뉴스피드 랭킹에 사용자가 기사를 읽는데 소요되는 시간을 반영하는 방식으로 알고리즘을 변경하여, 제목과 실제 내용이 부합하지 않거나 내용이 부실한 낚시성 기사에 불이익을 줌으로써 스팸을 방지하고자 하는 시도가 이루어지고 있다. 특히, SNS와 비슷한 기능을 하는 블로그 역시, 각종 상품이나 서비스에 대한 허위성 글을 포스팅하거나 본인이 작성한 글을 불특정다수에게 노출시키기 위해 허위 태그를 붙이는 방식으로 스팸이 생성되고 있다. 이로 인해 사용자들은 태그를 통해 블로그에 접속하였으나 본문 내용이 태그와 부합하지 않아 정보 검색을 위한 시행착오를 겪게 되며, 이 과정에서 해당 블로그 자체에 대한 만족도와 신뢰도 역시 떨어지게 된다.

이처럼 스팸글의 활성화로 인한 여러 부작용을 막기 위해 다양한 연구가 이루어져 왔으며, 대표적으로 오피니언 스팸(Opinion Spam), 이메일 스팸, 웹 스팸(Web Spam) 탐지 관련 연구 등이 있다[5]. 하지만 이러한 연구들에서 스팸을 인식하고 정의하는 기준에는 다소 차이가 존재하며, 이러한 기준은 크게 3가지로 나눌 수 있다. 우선 첫 번째는 (1) 글 작성자의 신원이나 사용 패턴을 파악함으로써 해당 작성자가 스패머로 인식될 경우, 해당 작성자가 남긴 모든 글을 스팸으로 인식하는 경우이다. 두 번째는 (2) 사용자가 의도하지 않았거나 관심이 없음에도 불구하고 사용자에게 노출되는 글을 스팸으로 인식하는 경우이며, 이메

일 스팸이나 허위 링크 등이 이에 해당된다. 마지막 세 번째는 (3) 악의적 혹은 비악의적으로 사실과 다르게 작성된 글을 스팸으로 인식하는 경우이다. 위의 (1)은 사용자의 접속 기록이나 인구통계 정보 등을 활용할 수 있다는 점에서 SNS를 통한 연구가 비교적 활발하게 이루어지고 있으며[6, 7, 8, 9, 10], 스팸을 차단하는 다양한 어플리케이션도 개발되어 상용화되고 있다. (2)의 경우 역시 스팸 탐지의 가장 고전적인 분야로, 주로 불특정 다수에게 발송되는 이메일 검출이나 SNS에서의 Embedded URL을 통한 스팸 검출 등에 사용되고 있다[11]. 이처럼 (1)과 (2)의 경우 작성된 글 이외에도 추가적인 메타 정보를 활용할 수 있기 때문에 비교적 많은 연구와 스팸 검출 방법이 고안되었으나, (3)의 경우에는 글의 내용만을 가지고 글의 사실 여부를 판단해야 한다는 어려움으로 인해 상대적으로 연구가 미흡한 실정이다.

예를 들어, 내용과 다른 태그를 설정하여 사용자를 유인하는 태그 스패밍의 경우 포스팅 내용에도 문제가 없고 태그 자체에도 문제가 없음에도 불구하고, 포스팅 내용에 부합되지 않는 태그의 연결이 스팸의 문제를 야기할 수 있다. 이는 그림 1을 통해 보다 자세히 설명된다.

그림 1에서 (a) Health 관련 포스팅에 “Exercise”와 “Diet”, (b) Movie 관련 포스팅에 “Avengers”, “Black Widow”, “Scarlett Johansson”이라는 태그가 연결되어 있으며, 이 경우 모든 태그의 연결은 적절한 것으로 보인다. 하지만 (c), (d)에서와 같이 (a)의 “Diet” 태그가 (b)의 Movie 관련 포스팅에, (b)의 “Scarlett Johansson” 태그가 (a)의 Health 관련 포스팅에 연결되는 경우를 가정하자. 이 때 포스팅이나 태그 자체는 스팸이 아닐지라도, 서로 부합하지 않는 포스팅과 태그의 연결로 인해 사용자에게 스팸성 정보로 인식될 수 있다. 이러한 유형의 스패밍은 글의 내용만을 가지고 진위를 판단해야 할 뿐만 아니라, 태그 전체가 아닌 태그의 일부가 스팸인 경우까지 감안해야 한다는 점에서 해결이 매우 어렵다. 따라서 본 연구에서는 블로그의 스팸성 태그 식별을 위해, 포스팅과 태그 간 연결 고리에 초점을 맞추어 글의 내용 자체를 분석함으로써 내용과 부합하지 않는 스팸 태그를 검출하는 방법론을 제안한다.

본 연구의 구성은 다음과 같다. 다음 장인 2장에서는 텍스트 마이닝과 스팸 탐지에 대한 연구들을 간략

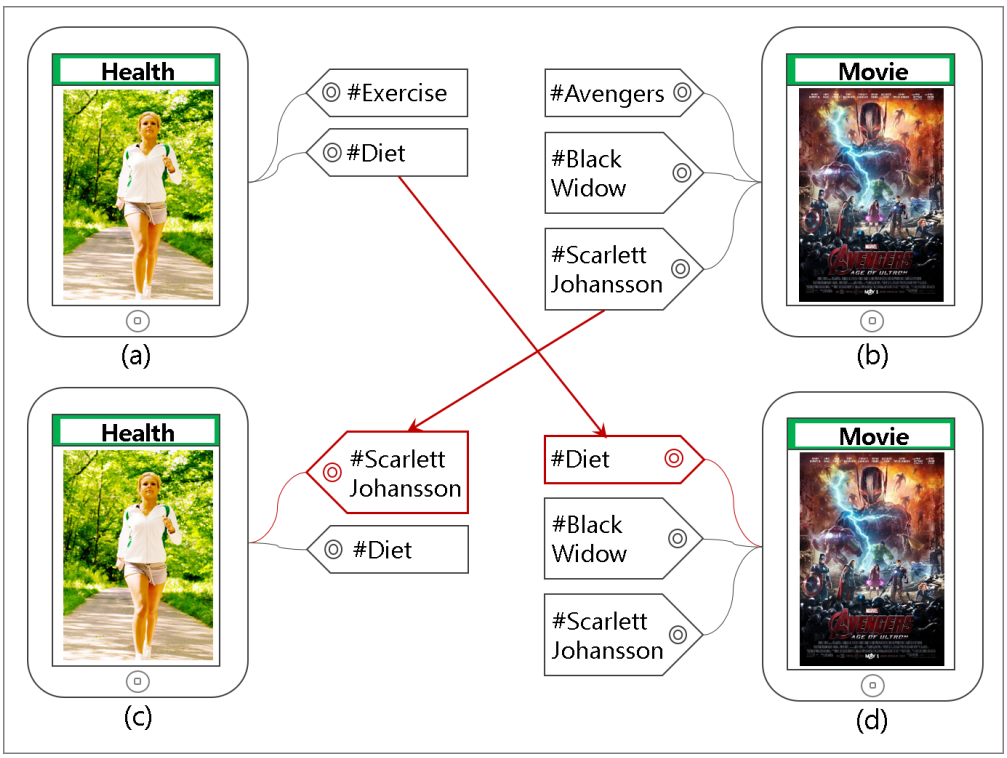

그림 1. 태그 스패밍의 예
Fig. 1. Example of Tag Spamming

하게 요약하고, 3장에서는 블로그 스팸 태그 탐지 방법론을 새롭게 제안하고 제안 방법론의 적용 사례를 소개하며, 마지막 4장에서는 본 연구의 기여와 향후 연구 방향을 제시한다.

Ⅱ. 관련 연구

2.1 텍스트 마이닝

구조화된 정형 데이터에 대한 분석을 통해 새로운 지식을 창출하기 위한 기존의 데이터 마이닝[12]과 달리 최근에는 뉴스 기사, 웹 게시물, 소셜미디어 등을 통해 유통되는 다량의 비정형 텍스트 데이터로부터 새로운 지식 및 유용한 패턴을 발견하기 위한 연구가 텍스트 마이닝이라는 이름으로 활발히 이루어지고 있다. 텍스트 마이닝은 다량의 텍스트에 대한 분석을 통해 의미 있는 정보를 추출하는 과정으로, 정보 추출, 정보 검색, 자연어 처리, 텍스트 요약, 자동분류, 토픽 추적 등 여러 분야의 기술을 종합적으로 활용한다[13, 14]. 이를 통해 기존의 데이터 마이닝 분야에서 다루어 온 전통적 주제[15-17]뿐 아니라, 더욱 다양하고 폭넓은 주제에 대한 분석이 이루어지고 있다[18,19].

텍스트 마이닝의 다양한 응용 중 여러 분야에서 가시적인 성과를 내며 가장 활발하게 활용되고 있는 대표적인 응용 기술은 토픽 분석(Topic Analysis)이다. 토픽 분석은 각 문서에 포함된 용어의 빈도수에 근거하여 유사 문서를 그룹화한 뒤 각 그룹을 대표하는 주요 용어들을 추출하여 해당 그룹의 토픽 키워드 집합을 제시하는 방식으로 이루어진다. 이 때 문서는 문서, 제목, 요약, 본문, 댓글 등을 포함하는 넓은 개념을 의미하며, 토픽 분석의 주요 이론적 배경은 벡터공간모델(Vector Space Model)[20,21]과 TF-IDF(Term Frequency-Inverse Document Frequency)[22]이다. 텍스트를 표현하는 기본적인 수단인 벡터공간모델은 분석 목적에 따라 행렬, 계층 벡터 등의 다양한 형태로 표현이 가능하다. 또한 TF-IDF는 여러 문서에서 자주 출현하는 일반적인 단어에 대해 가중치를 낮게 부여하고 특정 문서에서만 출현하는 특수한 단어에 대해 가중치를 높게 부여하는 계산 방식으로, 각 문서는 용어 수만큼의 차원과 TF-IDF를 값으로 갖는 벡터로 표현된다. 이 때, 문서 내에 존재하게 되는 용어의 수가 지나치게 많아지기 때문에 차원 축소 과정이 반드시 필요하며, SVD(Singluar Value Decomposition) 등의 차원 축소 방법이 널리 활용된다[19]. 토픽 분석의 결과로 하나의 문서는 여러 토픽에 동시에 대응될 수 있으며, 이는 전통적인 군집분석과 다른 특성이다.

최근에는 토픽 분석을 활용하여 다양한 분야의 문제를 해결하기 위한 시도가 활발하게 이루어지고 있다. 김지은 외(2014)[23]와 현윤진 외(2015)[24]에서는 토픽 분석을 통해 도출된 이슈의 수가 매우 방대한 경우, 클러스터링 기법을 활용하여 상위 이슈를 도출할 수 있음을 보였으며, 이 때 단순히 이슈의 유사성이 아닌 관점에 따라 상이한 군집화를 수행할 수 있음을 보였다. 또한 최성이 외(2015)[25]는 토픽 분석을 통해 사용자의 관심 이슈를 식별하고, 이를 활용하여 추천 시스템의 성능을 향상시키는 방법론을 제안하였다. 이 외에도 토픽 분석을 통한 사용자의 관심 기반 고객 세분화 방법론[26], 과학기술 이슈에 대한 여론 분석[27], 이슈의 동적 변이 과정[28] 등에 대한 연구가 활발히 이루어지고 있다.

2.2 스팸 탐지

인터넷의 발전과 함께 무방비하게 노출된 스팸의 위험을 극복하기 위한 연구가 스팸 탐지라는 이름으로 다양한 분야에 걸쳐 활발히 이루어져 왔다. 이들 연구는 크게 이메일 스팸, 웹 스팸, 오피니언 스팸 탐지의 세 가지 영역으로 구분되어 수행되어 왔으며, 그 중에서도 스팸 이메일 검출에 대한 연구가 활발하게 이루어져 왔다. 베이지안 방법론을 사용한 스팸 이메일 필터링 연구[29,30], SVM과 유의어 사전(Thesaurus Dictionary)의 결합을 통한 스팸 이메일 검출 연구[31], 2001년 회계부정 사건으로 유명한 Eron사의 이메일 데이터를 활용한 스팸 이메일 검출 연구[30,32] 등을 스팸 이메일 검출의 대표적 예로 들 수 있다.

한편, 웹 스팸의 경우에는 크게 링크 스팸(Link Spam)과 컨텐츠 스팸(Content Spam)의 두 가지 유형으로 나눌 수 있으며[5], 이를 방지하기 위한 연구가 지속적으로 이루어져 왔다. 대표적으로 TrustRank 알고리즘을 적용하여 산출된 웹 그래프의 신뢰점수를 통해 페이지별 점수를 부여함으로써 스팸 페이지를 필터링하는 연구[33], 웹 링크 구조를 통해 스패머의 링크 스패밍을 식별하는 연구[34], 스팸 페이지의 컨텐츠 분석을 통한 웹 스팸 탐지 연구[35] 등이 있다. 하지만 웹 스팸은 그 양이 기하급수적으로 증가할 뿐만 아니라, 변화의 속도 역시 너무 빠르기 때문에 스팸 탐지에 많은 어려움이 있다. 오피니언 스팸의 경우, 사회 정치적 혹은 특정 상품이나 서비스 등에 대해 사실과 다른 의견 혹은 리뷰 등을 작성하여 특정 목적을 달성하고자 하는 것을 의미하며, 이를 탐지하기 위해서는 사람이 해당 컨텐츠를 직접 확인함으로써 해당 글의 진위 여부를 파악해야 한다는 어려움이 존재한다. 예를 들

어 상품 리뷰에 대한 스팸 탐지의 경우, 스팸글을 쓴 작성자가 실제로 해당 상품을 사용했는지 여부의 확인이 불가능하기 때문에 사실상 스팸글을 정확하게 검출하기란 거의 불가능하다. 이러한 한계로 오피니언 스팸은 위의 두 스팸 탐지에 비해 상대적으로 연구가 미흡한 실정이며, 아직도 많은 과제가 남아있는 분야이다[5]. 이와 동시에 최근 SNS와 블로그 등이 많은 사람들의 삶의 일부분으로 자리 잡게 되면서, 이를 통해 노출된 스팸을 탐지하기 위한 시도가 활발히 이루어지고 있다. 대표적으로 사용자의 비정상적인 행동 패턴이나 사회적 태도를 분석하여 스팸을 검출하는 연구[6,7], 사용자의 계정이나 팔로우(Follow) 관계, 주고받은 메시지의 네트워크 구조 등을 활용한 연구[8-10], Tweet-Embedded URLs의 분류를 통한 스팸 탐지 연구[11] 등이 있다.

특히, 최근에는 SNS와 블로그를 통해 일반적으로 사용되고 있는 태그에 대한 스팸 탐지 연구의 필요성이 대두되고 있다. 태그는 사용자가 작성한 글을 보다 많은 사람들과 공유하기 위한 목적으로 사용되며, 사용자가 임의로 정의하여 자유롭게 사용할 수 있다는 특징을 갖는다. 이러한 태그는 정보 획득을 하고자 하는 사용자들이 원하는 정보에 보다 쉽게 접근할 수 있는 수단이 되기도 하지만, 내용과 무관한 스팸성 태그의 무분별한 사용으로 인해 사용자가 본인의 관심과 무관한 글에 노출될 수 있다는 부작용을 낳기도 한다. 따라서 이를 방지하기 위한 연구들이 이루어지고 있으며, 대표적 예로 해시태그의 유형을 분석한 연구[36], 특정 도메인에서 많이 사용되는 해시태그 추천을 통한 스팸 트윗의 영향력을 분석한 연구[37], 의미 있는 태그 선별을 위해 키워드 태그를 분석한 연구[38] 등을 들 수 있다. 이러한 시도들에도 불구하고 여전히 스팸 태그 탐지에 대한 연구는 미흡한 실정이며, 대부분의 연구가 트위터 데이터의 스팸 태그 탐지에 집중되고 있다. 이러한 현상은 트위터의 영향력 및 스팸으로 인한 파급 효과가 매우 크기 때문이기도 하지만 사용자 계정, 사용자 사이의 네트워크 구조, 사용자 정보 등 메타 데이터를 활용한 스팸 탐지가 가능하다는 점에 기인한 측면이 있다. 하지만 이러한 접근법을 따르는 연구는 메타 데이터를 충분히 활용할 수 없는 컨텐츠의 스팸 식별에는 적용되기 어렵다는 한계를 갖는다. 따라서 본 연구에서는 스팸으로 인한 파급 효과가 큰 매체 중 메타 데이터 기반 스팸 탐지 기법의 적용이 용이하지 않은 대표적 분야인 블로그에 대해, 블로그 스팸 태그 탐지 방법론을 제안하고자 한다.

Ⅲ. 블로그 스팸 태그 탐지 방법론 제안 및 적용

3.1 블로그 스팸 태그 탐지 방법론

본 절에서는 블로그 스팸 태그 탐지 방법론을 제안한다. 스팸이란 용어는 관점에 따라 다양한 의미로 정의될 수 있으며, 본 연구에서 스팸 태그는 태그가 속해있는 문서의 주제와 부합하지 않는 태그, 즉 글의 주제와 직접적인 연관이 없는 태그를 나타내는 것으로 정의한다.

그림 2는 본 연구의 전체 개요도를 나타내고 있으며, 원통형으로 표시된 부분은 블로그 포스트 데이터(Blog Post), 블로그 본문 데이터(Post Body), 해당 태그 정보(Post Tag) 등의 데이터 소스를 나타낸다. 또한 직사각형으로 표시된 부분은 주요 프로세스를 나타내며, 점선으로 표시된 도형은 각 프로세스의 산출물을 나타낸다.

제안 방법론은 Module1의 단일 문서 기반 스팸 탐지 방법과 Module2의 문서그룹 기반 스팸 탐지 방법으로 구성되어 있다. 우선 Module1은 포스트 본문을 대상으로 (1) Text Parsing을 통해 포스트별 주요 용어를 추출하고, (2) 도출된 포스트별 주요 용어와 포스트 태그를 비교한다. 이후, (3) 포스트 태그의 포스트 본문 내 출현 빈도수를 활용하여 문서 기반의 스팸 태그를 탐지한다. 하지만 이 방법은 포스트 태그가 우연히 본문 내에 포함되거나 우연히 누락된 경우에 대해 오작동할 수 있는 가능성이 있다는 한계를 갖는다. Module2의 문서그룹 기반 스팸 탐지 방법은 이를 극복하기 위해 고안한 방법으로, Module1에서 도출된 포스트별 주요 용어를 대상으로 (4) 클러스터링을 수행하여 포스트 그룹을 도출한 후, (5) 도출된 포스트 그룹별 주요 용어와 Post Tag를 비교분석한다. 이후, (6) 포스트 태그의 포스트 그룹 내 출현 빈도수를 활용하여 문서그룹 기반의 스팸 태그를 탐지한다. 두 기법의 성능 비교는 (7) Comparison 모듈에서 이루어진다.

Module1은 단일 문서 기반 스팸 태그 탐지 방법을 나타내며, 이는 그림 2의 프로세스 (1) ~ (3)에 해당된다. 본 방법론은 포스트 본문에 출현한 주요 용어와 해당 포스트 태그의 비교를 통해 스팸 태그를 탐지하는 것을 기본으로 하며, 이를 위해 포스트 본문에 출현한 주요 용어를 추출하는 과정이 우선적으로 선행된다. 포스트 본문을 대상으로 텍스트 파싱을 수행하여 각 포스트별 출현 용어를 추출하며, 이 때 분석의 품질을 향상시키기 위해 Stop List를 구축하여 적용한다. 이렇게 추출된 포스트별 주요 용어는 포스트별 유효 태그의 식별에 사용되며, 각 태그가 해당 포스트

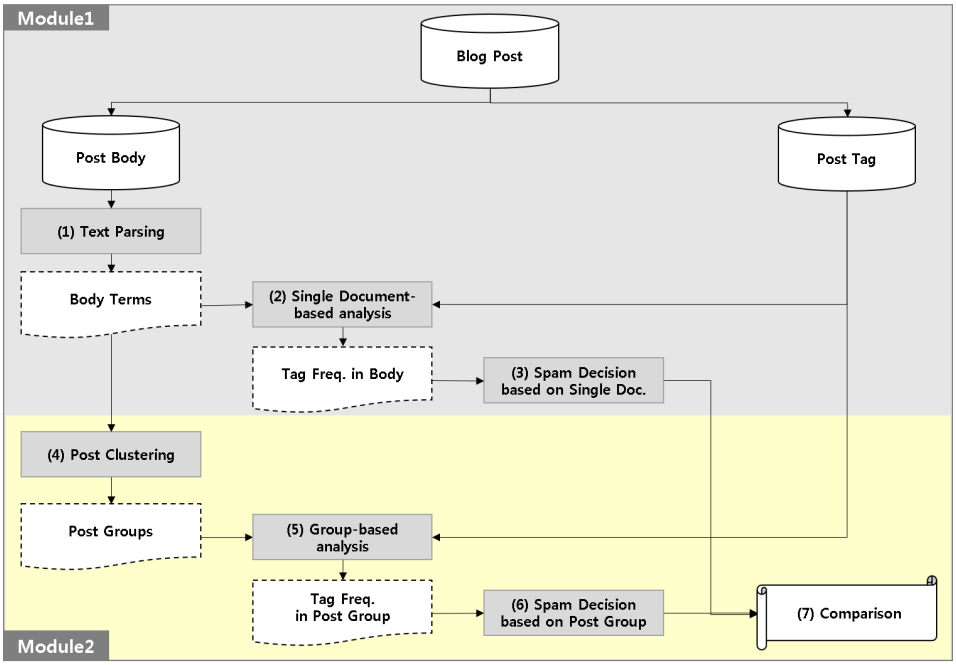

그림 2. 전체 연구 개요
Fig. 2. Research Overview

내에 출현한 빈도수를 기반으로 스팸 태그 탐지가 이루어진다. 즉, 특정 태그가 해당 포스트의 주요 용어와 전체 혹은 일부와 일치하면 해당 태그가 포스트 본문과 부합하는 것으로 인정하며, 이와 반대로 특정 태그가 해당 포스트의 주요 용어와 전혀 일치하지 않을 경우 해당 태그는 포스트의 내용과 부합하지 않는 스팸 태그로 처리한다.

하지만 Module1의 단일 문서 기반 스팸 태그 탐지 방법론은 포스트 본문의 내용과 관련이 있음에도 본문에 출현하지 않는 태그가 존재할 경우, 해당 태그는 포스트 본문 내용과 부합하지 않는 것으로 간주되어 스팸 태그로 오인된다는 한계를 갖는다. 따라서 이러한 문제를 해결하기 위해 본 연구에서는 Module2 문서그룹 기반 스팸 태그 탐지 방법론을 제안하며, 이는 그림 2의 프로세스 (4) ~ (7)에 해당된다. 문서그룹 기반 스팸 태그 탐지 방법은 특정 태그가 해당 포스트 본문에는 출현하지 않았으나 해당 포스트와 유사한 다른 포스트의 본문에 출현한 경우 해당 태그를 유효한 태그로 인정한다. 이를 통해 Module1에서 발현된 한계, 즉 포스트 본문과 관련이 있음에도 본문에 직접 출현하지 않아 스팸으로 오인된 태그를 햄(Ham) 태그로 인식함으로써 보다 정확한 스팸 탐지가 이루어질 수 있다. 이를 위해, Module2에서는 Module1의 프로세스 (1)을 통해 식별한 포스트별 주요 용어를 활용하여 클러스터링을 수행함으로써 포스트 그룹을 도출한다. 이후 Module1과 유사한 방식으로 포스트 그룹별 주요 용어 집합을 도출하고, 이들 집합과 포스트별 태그의 비교를 통해 스팸 태그를 탐지한다. 제안 방법론의 상세 내용은 다음 절의 실험을 통해 소개한다.

3.2 블로그 스팸 탐지 방법론의 적용

3.2.1 실험 데이터

제안 방법론의 실제 적용 가능성을 알아보기 위해 실제 블로그 데이터를 대상으로 실험을 수행하였다. 본 연구에서는 한국의 대표적인 포털사이트 'N'사의 블로그 포스트 중 태그를 포함하고 있는 블로그 포스트 8,000건을 수집하였다. 이 가운데 일상적인 내용이나 이미지가 주를 이루는 포스트들이 실험 결과에 미치는 영향을 최소화하기 위하여 토픽분석을 통해 샘플링을 수행하고, 주요 토픽에 미치는 영향이 큰 포스트만을 추출하여 분석에 사용하였다. 토픽분석을 통해 샘플링 작업을 수행한 결과 각 포스트의 토픽 가중치가 0.2 이상인 포스트 1,719건을 추출하였으며, 최종적으로 추출된 포스트 1,719건과 그에 해당하는 태그 8,397건을 분석에 사용하였다.

3.2.2 Module1: 단일 문서 기반 스팸 태그 탐지

우선, 수집된 포스트 본문 1,719건을 대상으로 SAS Enterprise Miner 14.1의 텍스트 마이닝 모듈을 사용하여 텍스트 파싱을 수행하였으며, 그 결과 총 211,093개의 용어(명사)가 도출되었다. 하지만 이 과정에서 Stop List를 적용하였음에도 불구하고, 영어에 최적화되어 있는 SAS 패키지의 특성상 형태소 분석이 제대로 이루어지지 않아 의미 없는 용어가 다수 추출되었다. 따라서 추가 정제 작업을 통해 한 글자 용어, 특수문자 등을 제거하여 최종적으로 총 194,207개의 포스트별 주요 용어를 추출하였다.

이후, 포스트 태그와의 비교를 수행하기에 앞서, 포스트 태그의 정제작업을 수행하였다. 일반적으로 태그는 사용자가 직접 정의하고 자유롭게 사용할 수 있기 때문에 그 길이가 너무 짧거나 길 수 있으며, 그 범위 또한 매우 포괄적이거나 지엽적일 수 있다. 이러한 부작용을 완화하기 위해 본 연구에서는 길이가 10 이하인 태그만을 추출하여 분석에 사용하였으며, 그림 3은 추출된 포스트별 주요 용어와 태그의 일부를 나타낸다.

이렇게 도출된 포스트별 주요 용어와 포스트 태그를 비교하여 표 1과 같이 각 태그의 포스트 본문 내 출현 빈도수를 추출한 후 문서 기반의 스팸 탐지를 수행하였다. 스팸 판정을 위한 임계값(Threshold)변화에 따른 분석 결과가 표 2와 그림 4에 나타나있다.

표 2는 정상 태그로 인식되기 위한 출현 빈도의 임계값, 해당 임계값 이하의 출현 빈도를 나타내어 스팸으로 분류되는 태그의 수, 그리고 전체 태그 수에 대한 스팸 태그 수의 비율을 나타내고 있다. 한편 그림 4는 표 2의 스팸 태그 비율을 그래프로 나타낸 결과이다. 하지만 본 연구에 쓰인 실험 데이터는 스팸 데이터가 거의 존재하지 않는 데이터임에 유의해야 한다. 따라서 위의 실험 결과는 출현 빈도의 임계값 변화에 따라 스팸으로 오분류되는 정상 태그의 비율에 대한 분포를 파악하는 정도의 의미만을 갖는 것으로 이해해야 한다. 제안 방법론에 따른 스팸 탐지의 정확성 평가를 위해서는 향후 명시적인 스팸 데이터를 추가 수집하거나 인위적으로 스팸 태그를 삽입하여 판별 비율을 살펴볼 필요가 있다.

PostID	Body Term
293	발색
293	살사
293	살사레드
293	뷰티블로거
293	뷰티스타그램
371	메이크업
371	디올
371	아이브로우
462	꿀추출물
462	립스틴
1474	경상북도
1474	배달
1551	율왕리해수욕장
1551	영종도
1551	중국음식
1694	황제해물문어보쌈
1694	해물라면

(a) 도출된 포스트별 주요 용어(일부)

PostID	Post Tag
293	미샤살사레드
293	맥철리져멈이
293	살사레드
293	미샤
371	메이크업포에버
371	파운데이션추천
371	아이브로우추천
462	더블유랩워너비립스틱
462	명품저렴이립스틱
1446	계산택지맛집
1474	경주중국집
1474	경주탕수육
1474	백리향
1551	율왕리중국음식
1551	율왕리맛집추천
1694	황제해물문어보쌈
1694	천안맛집

(b) 정제된 포스트 태그(일부)

그림 3. 도출된 포스트별 주요 용어 및 포스트 태그(일부)
Fig. 3. Extracted Key Terms and Post Tag for each Post (Part)

표 2. 문서 기반의 스팸 탐지 결과
Table 2. The Result of Single Document based Spam Detection

Threshold of Freq. of Post Tag	Num. of Spam Tag	Ratio of Spam Tag
0	1345	16.02%
1	3596	42.82%
2	5572	66.36%
3	7255	86.40%
4	7910	94.20%
5	8229	98.00%
6	8345	99.38%
7	8372	99.70%
8	8389	99.90%
9	8392	99.94%
10	8396	99.99%
11	8397	100.00%

3.2.3 Module2: 문서그룹 기반 스팸 태그 탐지

표 1에서 “가로수길맛집” 태그의 경우, 포스트 본문 내용과 관련이 있음에도 불구하고 포스트 본문에

표 1. 포스트 태그별 포스트 본문 내 출현 빈도 (일부)
Table 1. Frequency of Post Tag in Post Body (Part)

PostID	Post Tag	Freq.(inPost)	PostID	Post Tag	Freq.(inPost)	PostID	Post Tag	Freq.(inPost)
1772	핑크립스틱	2	816	신사동맛집	3	7752	데이트장소	0
1772	궁금해!	0	816	가로수길	2	7752	서현맛집	0
1772	립스틱	1	816	다온초밥	2	8136	청주피자맛집	2
293	살사레드	3	1289	가로수길맛집	0	8136	새로운피자	1
293	매트립루즈	2	1289	가로수길초밥	1	265	가로수길	2
293	미샤	0	1289	다온초밥	2	265	맛집	1

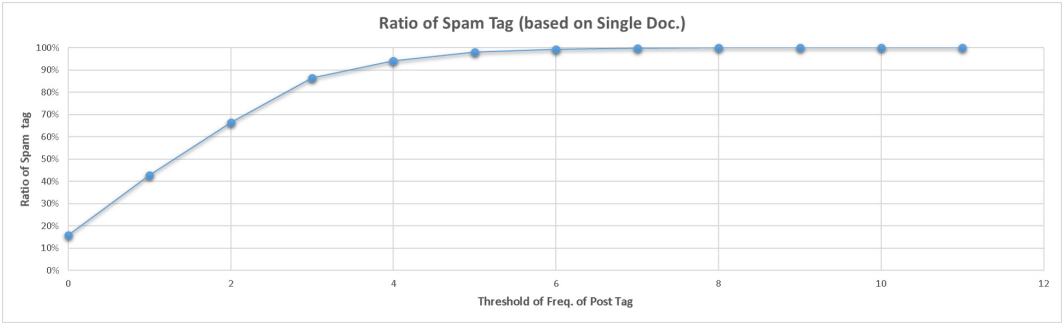

그림 4. 스팸 태그 비율 (단일 문서 기반)
Fig. 4. The Ratio of Spam Tag (based on Single Document)

출현하지 않아 스팸 태그로 오인된 예이다. 이처럼 스팸이 아님에도 불구하고 해당 포스트 본문에 출현하지 않아 정상 태그가 스팸으로 오인되는 부작용을 완화하기 위해 문서그룹 기반의 스팸 태그 탐지 분석을 수행하였다. 이를 위해 Module1에서 추출한 포스트별 주요 용어를 사용하여 클러스터링을 수행하였으며, 그 결과 총 20개의 포스트 그룹을 도출하였다. 표 3은 해당 결과의 일부를 보여주고 있으며, 표에서 Freq.(inPost)는 각 태그가 특정 포스트 그룹의 모든 포스트 본문 내에서 출현한 빈도를 나타낸다.

이렇게 추출된 포스트 그룹별 주요 용어와 포스트 태그에 대한 비교 분석을 통해 문서그룹 기반의 스팸 탐지 분석을 수행하였다. 표 4는 포스트별 태그의 포스트 본문 내 출현 빈도수, 각 태그가 속한 포스트 그룹 내에서 해당 태그가 출현한 문서 수, 해당 포스트 그룹의 총 문서 수, 그리고 해당 태그가 속한 포스트 그룹의 총 문서 수에 대한 태그 출현 문서 수의 비율을 나타내고 있다. 요약하면 Module1의 방법론은 Freq.(inPost)의 값에 따라 스팸 태그를 식별하지만, Module2의 방법론은 Ratio of appearance(inPG)의 값에 따라 스팸 태그를 식별한다. 이러한 차이로 인해 표 3의 단일 문서 기반 방법론에서 스팸으로 처리되었던 "가로수길맛집" 태그는 표 4의 문서그룹 기반 방법론에서 정상 태그로 인정받게 됨을 알 수 있다. Module2에 의해 스팸 태그를 식별한 결과가 표 5와 그림 5에 나타나있다.

표 4. 포스트 태그별 포스트 그룹 내 출현 비율 (일부)
Table 4. Ratio of Appearance for each Post Tag in Post Group (Part)

Post ID	Post Tag	Freq. (in Post)	Num. of Post	Num. of Post (for each PG)	Ratio of Appearance (in PG)
1772	핑크립스틱	2	1	70	1.43%
1772	궁금해!	0	0		0.00%
1772	립스틱	1	2		2.86%
293	살사레드	3	8		11.43%
293	매트립루즈	2	2		2.86%
293	미샤	0	2		2.86%
816	신사동맛집	3	10	57	17.54%
816	가로수길	2	5		8.77%
816	다온초밥	2	38		66.67%
1289	가로수길맛집	0	35		61.40%
1289	가로수길초밥	1	28		49.12%
1289	다온초밥	2	38		66.67%
7752	데이트장소	0	0	116	0.00%
7752	서현맛집	0	0		0.00%
8136	청주피자맛집	2	1		0.86%
8136	새로운피자	1	1		0.86%
265	가로수길	2	1		0.86%
265	맛집	1	3		2.59%

마지막으로 표 6은 동일한 태그가 두 가지 서로 다른 방법론에 의해 스팸 또는 햄으로 상이하게 판별되는 예를 보인다. 표 6은 포스트 그룹 PG1에 대한 내용을 다루고 있으며, 세 개의 태그를 각각 Ham, Spam, Ham으로 판별하였다. 포스트 그룹은 포스트 본문에 대한 토픽 모델링을 통해 유사 내용을 다룬 포스트를 그룹화한 것이므로, PG1에 속한 모든 포스트

표 3. 포스트 그룹별 주요 용어 (일부)
Table 3. Key Terms for each Post Group (Part)

PG_ID	PostID	Post Tag	Freq.(inPost)	PG_ID	PostID	Post Tag	Freq.(inPost)	PG_ID	PostID	Post Tag	Freq.(inPost)
1	1772	핑크립스틱	2	2	816	신사동맛집	3	8	7752	데이트장소	0
	1772	궁금해!	0		816	가로수길	2		7752	서현맛집	0
	1772	립스틱	1		816	다온초밥	2		8136	청주피자맛집	2
	293	살사레드	3		1289	가로수길맛집	0		8136	새로운피자	1
	293	매트립루즈	2		1289	가로수길초밥	1		265	가로수길	2
	293	미샤	0		1289	다온초밥	2		265	맛집	1

표 5. 문서그룹 기반의 스팸 탐지 결과
Table 5. The Result of Post Group based Spam Detection

Threshold of Ratio of Appearance	Num. of Spam Tag	Ratio of Spam Tag
0%	1264	15.05%
1%	4218	50.23%
2%	6309	75.13%
3%	7049	83.95%
4%	7386	87.96%
5%	7571	90.16%
6%	7644	91.03%
7%	7759	92.40%
8%	7800	92.89%
9%	7826	93.20%
10%	7882	93.87%

들은 유사 내용을 다루고 있는 것으로 가정할 수 있다. 따라서 각 태그에 대해 PG1에 속한 모든 포스트들은 동일한 판정을 내리는 것이 합리적이며, 이러한 측면에서 표 6의 맨 우측에 제시된 Module2에 따른 판정은 바람직한 특성을 갖고 있다고 할 수 있다. 하지만 이와 달리 Decision based on single Doc. 열에 제시된 Module1에 따른 판정은 각 태그의 판정이 여러 포스트에서 Spam 또는 Ham으로 상이하게 나타남을 알 수 있다. 예를 들어 "파운데이션추천" 태그가 371번 포스트에서는 Spam으로, 3547번 포스트에서는 Ham으로 식별되는데, 두 포스트가 내용이 유사하여 동일한 포스트 그룹인 PG1에 속해 있음을 감안하면 이는 바람직한 결과라고 할 수 없다. 따라서 표 6은 문서그룹 기반 스팸 태그 탐지가 단일 문서 기반 스팸

표 6. 단일 문서 기반과 문서그룹 기반 스팸 탐지 결과 비교 (일부)
Table 6. Single Document-based vs. Post Group-based Spam Detection (Part)

PG_ID	Post Tag	Post ID	Freq. (in Post)	Decision based on Single Doc. (Threshold < 2)	Ratio of Appearance (in PG)	Decision based on Doc. Group (Threshold ≤ 0.05)
1	파운데이션추천	371	1	Spam	7.14%	Ham
		3547	2	Ham		
		5583	2	Ham		
		6567	1	Spam		
		6691	1	Spam		
	쿠션파운데이션	6887	1	Spam	4.29%	Spam
		8367	2	Ham		
		9226	2	Ham		
		938	1	Spam		
		2347	2	Ham		
	카인다섹시	3127	1	Spam	8.57%	Ham
		3686	2	Ham		
		4397	1	Spam		
		4415	2	Ham		

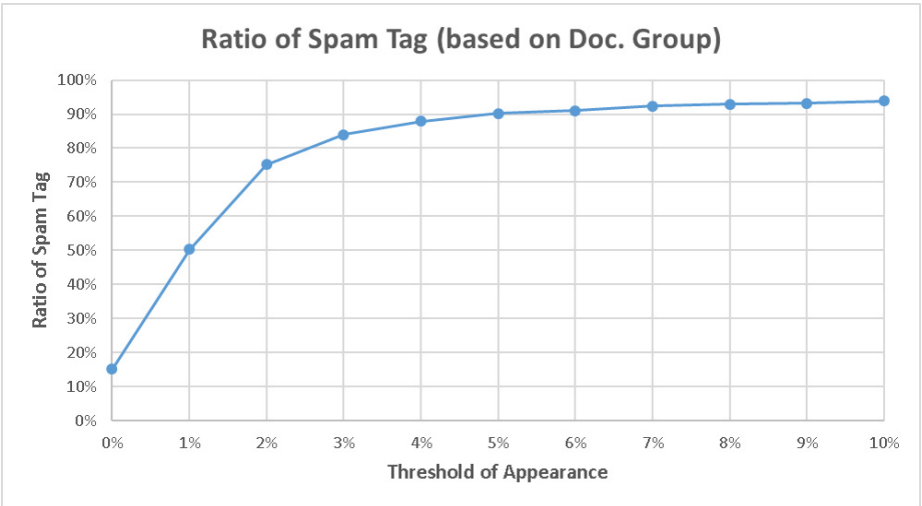

그림 5. 스팸 태그 비율 (문서그룹 기반)
Fig. 5. The Ratio of Spam Tag (based on Document Group)

태그 탐지에 비해 보다 바람직한 특성을 가질 수 있음을 나타낸다.

Ⅳ. 결 론

원본 데이터로부터 스팸성 데이터를 식별하여 제거함으로써 정보의 신뢰성 및 분석 결과의 품질을 제고하기 위한 연구가 이메일 스팸 검출, 웹 스팸 탐지, 오피니언 스팸 탐지 등의 분야에서 매우 활발히 수행되어 왔다. 본 연구에서는 스팸 식별을 위한 기존의 연구 동향을 소개하고, 블로그 정보의 신뢰성 향상을 위한 스팸 태그 식별 방안을 새롭게 제안하였다. 제안 방법론은 포스트 본문 또는 태그 자체가 아닌 본문과 태그의 연결의 적합성에 따라 스팸 여부를 판정하며, 태그와 단일 문서의 관련성 뿐 아니라 태그와 문서그룹의 관련성을 동시에 고려했다는 점에서 기존 연구와 차별성을 갖는다. 향후 연구에서는 제안 방법론의 객관적 성능 평가를 위해, 스팸 여부가 명시된 포스트 태그를 활용한 실험이 이루어져야 한다.

References

[1] Economist Intelligence Unit, *Big Data Harnessing a Game-Changing Asset*, The Economist, 2011.

[2] McKinsey Global Institute, *Big Data: The next Frontier for Innovation, Competition, and Productivity*, McKinsey and Company, 2011.

[3] Gartner Inc., *2012 Hype Cycle for Emerging Technologies*, Gartner Inc., 2011.

[4] C. Chen, J. Zhang, Y. Xiang, and W. Zhou, "Spammers are becoming "Smarter" on twitter," *Browse J. & Mags.*, vol. 18, no. 2, 2016.

[5] B. Liu, *Sentiment analysis and opinion mining, syntehesis lectures on human language technologies #16*, Morgan & Claypool Publisiers, 2012.

[6] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks," in *Proc. Ann. Netw. Distrib. Syst. Security Symp.*, San Diego, CA, 2013.

[7] J. Song, S. Lee, and J. Kim, "Spam filtering in twitter using sender-receiver relationship. Recent advances in intrusion detection," *Int. Workshop on Recent Advances in Intrusion Detection*, pp. 301-317, Heidelberg, Berlin, Sept. 2011.

[8] S. Yarde, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting spam in a twitter network," *First Monday*, vol. 15, no. 1, Jan. 2010.

[9] A. H. Wang, "Don't follow me: Spam detection in twitter," *IEEE SECRYPT*, pp. 1-10, Athens, Greece, Jul. 2010.

[10] Y. Ma, Y. Niu, Y. Ren, and Y. Xue, "Detecting spam on sina weibo," *CCIS*, Oct. 2013.

[11] S. Lee and J. Kim, "Warningbird: A near real-time detection system for suspicious URLs in twitter stream," *IEEE Trans. Dependable and Secure Comput.*, vol. 10, no. 3, pp. 183-195, Jan. 2013.

[12] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd Ed., Morgan Kaufmann Publishers, 2011.

[13] R. J. Mooney and R. Bunescu, "Mining knowledge from text using information extraction," *ACM SIGKDD Explorations Newsletter - Natural Lang. Process. and Text Mining*, vol. 7, no. 1, pp. 3-10, Jun. 2006.

[14] C. J. V. Rijsbergen, *Information Retrieval*, 2nd Ed., Butterworth, London, 1979.

[15] K. Kim and H. Ahn. "Development of web-based intelligent recommender systems using advanced data mining techniques," *J. Inf. Technol. Appl. Management*, vol. 12, no. 3, pp. 41-56, Sept. 2005.

[16] J. Hur and J. W. Kim, "Characteristics on inconsistency pattern modeling as hybrid data mining techniques," *J. Inf. Technol. Appl. Management*, vol. 15, no. 1, pp. 225-242, Mar. 2008.

[17] I. Hwang, "A study on dynamic query expansion using web mining in information retrieval," *J. Inf. Technol. Appl. Management*, vol. 11, no. 2, pp. 227-237, Jun. 2004.

[18] T. N. Phan and M. Yoo, "Facebook fan page evaluation system based on user opinion mining," *The J. Korean Inst. Commun. and*

Inf. Sci., vol. 40, no. 12, pp. 2488-2490, Dec. 2015.

[19] J. Moon, I. Jang, Y. C. Choe, J. G. Kim, and G. Bock, "Case study of big data-based agri-food recommendation system according to types of customers," *The J. Korean Inst. Commun. Inf. Sci.*, vol. 40, no. 5, pp. 903-913, May 2015.

[20] R. Albright, *Taming Text with the SVD*, SAS Institute Inc., 2006.

[21] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol. 18, no. 11, pp. 613-620, Nov. 1975.

[22] S. M. Weiss, N. Indurkhya, and T. Zhang, *Fundamentals of Predictive Text Mining*, Springer, 2010.

[23] J. Kim, N. Kim, and Y. Cho, "User-perspective issue clustering using multi-layered two-mode network analysis," *J. Intell. Inf. Syst.*, vol. 20, no. 2, pp. 93-107, Jun. 2014.

[24] Y. Hyun, N. Kim, and Y. Cho, "A multi-dimensional issue clustering from the perspective consumers' interests and R&D," *J. Inf. Technol. Serv.*, vol. 14, no. 1, pp. 237-249, Mar. 2015.

[25] S. Choi, Y. Hyun, and N. Kim, "Improving performance of recommendation systems using topic modeling," *J. Intell. Inf. Syst.*, vol. 21, no. 3, pp. 101-116, Sept. 2015.

[26] Y. Hyun, N. Kim, and Y. Cho, "Interest-based customer segmentation methodology using topic modeling," *J. Inf. Technol. Appl. & Management*, vol. 22, no. 1, pp. 77-93, Mar. 2015.

[27] D. Kim, W. X. S. Wong, M. Lim, C. Liu, N. Kim, J. Park, W. Kil, and H. Yoon, "A methodology for analyzing public opinion about science and technology issues using text analysis," *J. Inf. Technol. Serv.*, vol. 14, no. 3, pp. 33-48, Sept. 2015.

[28] M. Lim and N. Kim, "Investigating dynamic mutation process of issues using unstructured text analysis," *J. Intell. Inf. Syst*., vol. 22, no. 1, pp. 1-18, Mar. 2016.

[29] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz, "A bayesian approach to filtering junk e-mail," in *AAAI Workshop on Learning for Text Categorization*, vol. 62, pp. 98-105, Jul. 1998.

[30] X. Jia, K. Zheng, W. Li, T. Liu, and L. Shang, "Three-way decisions solution to filter spam email: An empirical study," *Int. Conf. Rough Sets and Current Trends in Comput*., pp. 287-296, Heidelberg, Berlin, Aug. 2012.

[31] I. Joe and H. T. Shim, "A SVM-based spam filtering system for short message service (SMS)," *J. KICS*, vol. 34, no. 9, pp. 908-913, Sept. 2009.

[32] B. Klimt and Y. Yang, "Introducing the enron corpus," *CEAS 2004, First Conf. Email and Anti-Spam*, California, USA, Jul. 2004.

[33] Z. Gyongyi, H. Garcia-Molina, and J. Pedersen, "Combating web spam with trustrank," *VLDB '04*, pp. 576-587, Toronto, Canada, Aug. 2004.

[34] Z. Gyongyi, P. Berkhin, H. Garcia-Molina, and J. Pedersen, "Link spam detection based on mass estimation," *VLDB '06*, pp. 439-450, Seoul, Korea, Sept. 2006.

[35] A. Ntoulas, M. Najork, M. Manasse, and D. Retterly, "Detecting spam web pages through content analysis," in *Proc. 15th Int. Conf. World Wide Web*, pp. 83-92, Edinburgh, Scotland, May 2006.

[36] P. Xanthopoulos, O. P. Panagopoulos, G. A. Bakamitsos, and E. Freudmann, "Hashtag hijacking: What it is, why it happens and how to avoid it," *J. Digital & Social Media Marketing*, vol. 3, no. 4, pp. 353-362, Feb. 2016.

[37] S. Sedhai and A. Sun, "Effect on spam on hashtag recommendation for tweets," in *Proc. 25th Int. Conf. Companion on World Wide Web*, pp. 97-98, Québec, Canada, Apr. 2016.

[38] J. Jung and M. Yoo, "Tag search system using the keyword extraction and similarity evaluation," *The J. Korean Inst. Commun. Inf. Sci.*, vol. 40, no. 12, pp. 2458-2487, Dec. 2015.

현 윤 진 (Yoonjin Hyun)

2013년 2월 : 국민대학교 경영정보학과 학사
2015년 2월 : 국민대학교 비즈니스IT전문대학원 석사
2015년 3월~현재 : 국민대학교 비즈니스IT전문대학원 박사과정

<관심분야> 데이터마이닝, 텍스트마이닝, 감성분석

김 남 규 (Namgyu Kim)

1998년 2월 : 서울대학교 컴퓨터공학과 학사
2000년 2월 : 한국과학기술원 경영공학 석사
2007년 2월 : 한국과학기술원 경영공학 박사
2007년 3월~현재 : 국민대학교 경영정보학부 교수

<관심분야> 텍스트마이닝, 데이터마이닝, 데이터베이스