韓國IT서비스學會誌
第13卷 第4號
2014年 12月, pp.275-287

Journal of Information Technology Services
http://dx.doi.org/10.9716/KITS.2014.13.4.275

거시적 이슈 트래킹의 한계 극복을 위한 개인 관심 트래킹 방법론

류 신* · 김남규**

Individual Interests Tracking : Beyond Macro-level Issue Tracking

Chen Liu* · Namgyu Kim**

■ Abstract ■

Recently, the volume of unstructured text data generated by various social media has been increasing rapidly; consequently, the use of text mining to support decision-making has also been growing. In particular, academia and industry are paying significant attention to topic analysis in order to discover the main issues from a large volume of text documents. Topic analysis can be regarded as static analysis because it analyzes a snapshot of the distribution of various issues. In contrast, some recent studies have attempted to perform dynamic issue tracking, which analyzes and traces issue trends during a predefined period. However, most traditional issue tracking methods have a common limitation : when a new period is included, topic analysis must be repeated for all the documents of the entire period, rather than being conducted only on the new documents of the added period. Additionally, traditional issue tracking methods do not concentrate on the transition of individuals' interests from certain issues to others, although the methods can illustrate macro-level issue trends. In this paper, we propose an individual interests tracking methodology to overcome the two limitations of traditional issue tracking methods. Our main goal is not to track macro-level issue trends but to analyze trends of individual interests flow. Further, our methodology has extensible characteristics because it analyzes only newly added documents when the period of analysis is extended. In this paper, we also analyze the results of applying our methodology to news articles and their access logs.

Keyword : Interests Tracking, Issue Tracking, Text Mining, Topic Analysis

Submitted : October 26, 2014 1st Revision : December 10, 2014 Accepted : December 13, 2014

* 국민대학교 비즈니스IT전문대학원 석사과정
** 국민대학교 경영정보학부 부교수, 교신저자

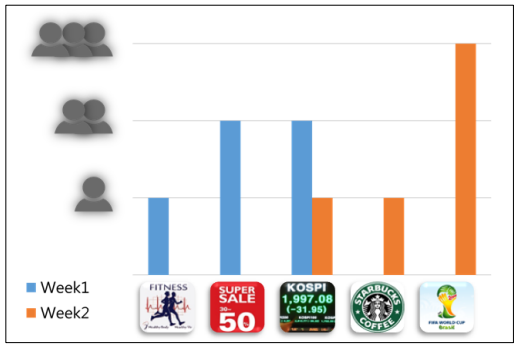
1. 서 론

다양한 스마트기기와 클라우드 기술을 포함한 정보통신기술이 급속도로 발달함에 따라 사람들은 장소와 공간에 상관없이 인터넷에 쉽게 접근할 수 있게 되었으며, 이로 인해 매우 방대한 데이터가 빠른 속도로 생성되고 있다. 이처럼 그 양이 방대하여 기존의 방법이나 도구로는 수집, 저장, 검색, 분석, 시각화가 어려운 정형 또는 비정형 데이터를 빅데이터(Big Data)라고 한다(Mckinsey, 2011). 특히 사용자들이 실시간으로 참여하여 사회적 이슈나 개인의 관심사에 대해 의견을 표현할 수 있는 다양한 소셜미디어가 활성화되고 있으며, 이러한 현상은 비정형 텍스트 처리 기술의 발전과 더불어 빅데이터 분석의 수요 증가를 견인하고 있다. 즉, Facebook, Twitter 등의 SNS(Social Network Service)를 통해 유통되는 텍스트 데이터의 양이 급증함에 따라, 이러한 글에 대한 분석을 통해 사회적 이슈의 흐름 또는 특정 제품이나 서비스에 대한 의견을 파악하기 위한 시도가 활발하게 이루어지고 있다.(Lee et al., 2010, Xu et al., 2013).

이러한 시도 중 비정형 텍스트 분석을 통해 의미를 도출하기 위한 각종 알고리즘 및 방법론을 다루는 분야를 텍스트 마이닝(Text Mining)이라 하며, 다양한 텍스트 마이닝의 세부 분야 중 특히 다수의 문서로부터 핵심 주제를 식별해내는 토픽 분석(Topic Analysis)은 이미 여러 분야에서 가시적인 성과를 내고 있다(Huang et al., 2013; Yin et al., 2012). 토픽 분석에 주로 사용되는 원본(Source) 데이터는 크게 Facebook, Twitter 등의 SNS 데이터와, 뉴스 또는 인터넷 게시판 등의 인터넷 게시물을 들 수 있다. SNS 데이터는 다양한 의견을 실시간으로 들을 수 있다는 장점이 있지만 정제되지 않은 표현이 너무 많고 주제가 특정 분야에 한정되지 않아 분석이 어렵다는 한계를 갖는다. 한편 뉴스 데이터는 비교적 정제된 표현과 주제의 집중도로 인해 분석이 용이하다는 장점을 갖지만, 사회 전체적인 이슈의 흐름을 나타낼 뿐 이 이슈와 개인들과의 관계를 파악하기는 어렵다는 한계를 갖는다.

토픽 분석이 특정 시점의 이슈의 분포를 파악하기 위한 정적인 분석의 성격을 갖는 것과 달리, 최근에는 여러 시점에 걸친 이슈의 변화를 분석하고 추적하기 위한 이슈 트래킹에 대한 연구가 다수 이루어지고 있다(Aggarwal et al., 2014). 이슈 트래킹은 대상 기간에 게시 또는 접속된 문서들로부터 주요 이슈를 추출한 뒤, 대상 기간을 세부 기간으로 나누어 각 세부 기간별 주요 이슈의 분포를 분석하는 것을 목적으로 한다. 하지만 전통적인 이슈 트래킹은 사회 전체 관점(Macro-level)에서의 이슈의 추적할 뿐, 각 이슈에 관심을 갖는 개인들의 관심의 흐름을 추적하지는 못한다는 한계를 갖는다. 이러한 한계는 <Figure 1>~<Figure 3>을 통해 설명된다.

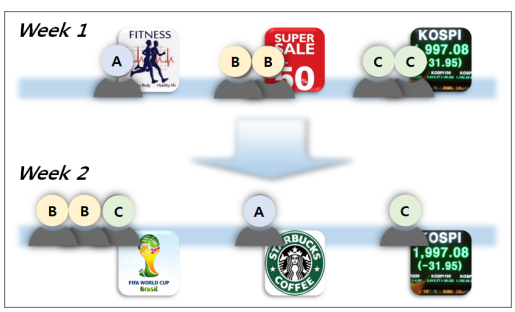
〈Figure 1〉 Individuals' Interests During Week1 and Week2



〈Figure 2〉 Traditional Issue Tracking

<Figure 1>은 Week1과 Week2에 걸친 주요 이슈 및, 각 이슈에 관심을 갖고 있는 개인 수의 변화를 나타내고 있다. 그림에서 개인의 수는 해당 이슈에 대응되는 글을 작성하거나 조회한 사람의 수로 이해될 수 있다. 그림에서 첫 주에는 건강, 쇼핑, 금융의 이슈가 주요 이슈이며, 둘째 주에는 월드컵, 외식, 금융의 이슈가 주요 이슈임을 알 수 있다. 이러한 이슈의 변화는 <Figure 2>의 그래프로 도식화하여 나타낼 수 있다. <Figure 2>에서 건강과 쇼핑의 이슈는 소멸되었으며, 금융의 이슈에 대한 관심이 감소하고, 이러한 이슈들에 대한 관심은 새로 생성된 이슈인 외식과 월드컵으로 옮겨갔음을 알 수 있다.

물론 이러한 분석을 통해 주요 이슈의 흐름을 파악함으로써 이에 따라 마케팅 전략을 수립하는 것도 매우 의미있지만, 전체 이슈의 변화가 곧 개인의 관심 변화를 나타낸다고 보기는 어렵기 때문에 분석 결과의 활용 범위는 제한적일 수밖에 없다. 이는 대부분의 마케팅 전략은 궁극적으로 고객(개인)을 대상으로 하는 것에 반해, 전통적인 이슈 트래킹의 결과는 이슈의 흐름만 분석할 뿐 개인에 대해서는 충분한 정보를 주지 못하고 있기 때문이다. 예를 들어 <Figure 3>은 <Figure 1>의 상황에서 개인별 관심 추세의 변화를 파악할 수 있다고 가정한 상황을 나타내고 있다. 즉 Week2의 월드컵 이슈는 Week1의 금융과 쇼핑에 관심을 갖던 개인들을 흡수하였으며, 건강에 관심을 갖던 개인은 외식으로 관심이 변경되었음을 알 수 있다.



〈Figure 3〉 Flow of Individuals' Interests

본 연구에서는 이와 같은 전통적 이슈 트래킹의 한계를 극복하기 위한 개인별 관심 트래킹 방안을 제시하고자 한다. 관심 트래킹을 통해 개인들이 어떤 이슈에서 어떤 이슈로 관심이 변화하는지를 추적할 수 있고, 이를 통해 개인의 관심 변화를 예측하고 이에 대해 선제적으로 대응하는 마케팅 전략을 수립함으로써 새로운 가치창출을 극대화할 수 있을 것이다. 또한 제안 방법론은 분석 대상이 되는 전체 기간의 문서를 한꺼번에 분석해야 하는 기존의 이슈 트래킹과 달리, 새로운 기간의 문서가 추가되었을 때 전체 기간의 문서가 아닌 새로 추가된 문서에 대해서만 토픽 분석을 수행하고 이를 기존의 분석 결과와 통합할 수 있으므로 기간의 확장성 측면에서도 바람직한 특성을 가질 수 있다.

본 논문의 이후 구성은 다음과 같다. 다음 장인 제 2장에서는 텍스트 마이닝, 토픽 분석, 이슈 트래킹 등 본 연구와 관련된 기존의 연구를 요약한다. 제 3장에서는 제안 방법론인 개인 관심 트래킹의 개념과 과정을 소개한다. 그리고 제 4장에서는 제안 방법론을 적용한 실험 결과를 소개하며, 마지막 장인 제 5장에서는 본 연구의 결론, 기여 및 한계, 그리고 향후 연구방향을 제시한다.

2. 관련 연구

2.1 텍스트 마이닝

세상에서 텍스트가 없는 곳은 없다(Provost and Fawcett, 2013). 텍스트는 의료, 고객 항의, 제품 문의, 제품 수리, 의견 표출 등의 일상 생활에서 컴퓨터가 아닌 사람의 소통을 위해 사용된다. 특히 최근에는 정보통신의 발달과 다양한 소셜미디어 서비스의 활성화로 인해 무수히 많은 양의 텍스트가 유통되고 있다. 즉 웹 2.0의 목표인 사용자 참여 및 소통을 지원하기 위해 다양한 기술의 발전이 이루어졌으며, 그 결과로 더욱 많은 텍스트 컨텐츠들이 생성되고 유통되고 있다. 따라서 이러한 텍스트

데이터에 대한 분석을 통해 더욱 의미있는 가치를 창출하고자 하는 텍스트 마이닝(Text Mining)에 대한 관심이 급증하는 것은 매우 당연하다고 할 수 있다.

텍스트 마이닝이란 방대한 양의 텍스트 문서로부터 의미있는 정보를 추출하는 일련의 과정을 의미한다. 텍스트 마이닝은 정보 검색, 데이터 마이닝, 기계 학습(Machine Learning), 통계학, 컴퓨터 언어학 등이 결합된 학제적(Interdisciplinary) 분야(Han et al., 2011)의 성격을 갖는다. 구체적으로 텍스트 마이닝은 문서 분류(Classification), 문서 군집화(Clustering), 정보 추출 등에 활용되며, 일반적으로 행렬, 계층, 벡터 등의 형식으로 정형화된 뒤 이후 분석 작업이 이루어진다(Weiss et al., 2010). 특히 가장 기본적인 변환 방식으로 각 문서에 용어의 빈도를 요약하는 벡터공간 모델(Vector Space Model)(Albright, 2006; Salton et al., 1975)이 주로 사용된다.

텍스트 마이닝 응용 중 가장 대표적인 것으로 토픽 분석(Topic Analysis)을 들 수 있다. 토픽 분석은 벡터공간모델과 TF-IDF(Salton and McGill, 1983)에 기반하여 수행되며, 주로 파싱(Parsing), 필터링(Filtering)을 거친 후에 이루어진다(Hong et al., 2014; Kim et al., 2014). 토픽 분석은 유사한 주제를 갖는 문서들을 묶는다는 점에서는 전통적인 군집화와 유사하지만, 하나의 문서가 다수의 토픽에 대응될 수 있다는 점에서 군집화와는 구별되는 특징을 갖는다.

2.2 이슈 트래킹

한 시점의 주요 이슈를 파악하는 정적인 분석인 토픽 분석과 달리, 여러 시점에 걸친 주요 이슈의 변화를 추적하는 분야를 이슈 트래킹(Issue Tracking)이라 한다. 이슈 트래킹은 인터넷 뉴스 기사, 소셜미디어 게시물 등에서 사람들이 자주 언급하는 정치, 경제, 연예, 스포츠 등 사회 전반에 걸친 이슈들을 추출, 발견하는 것을 목표로 한다(Ding and Chen, 2014). 이슈 트래킹을 통해 대상 기간의 주요 이슈를 파악할 수 있음은 물론, 기간의 변화에 따라 어떤 이슈가 새로 생성되고 또 얼마나 지속되는지 파악할 수 있다.

주로 해외에서 활발하게 연구되며 사용되고 있는 이슈 트래킹은 주로 뉴스 데이터에 대한 분석(Ma et al., 2014; Jin et al., 1999)이 주를 이루며, 이미 시스템으로 제공되어 다양한 응용에 사용되고 있다. 국내의 경우에도 최근 트위터 데이터를 대상으로 한 이슈 트래킹 연구(Bae et al., 2014)가 활발하게 수행되고 있으며, 분석을 통해 대선 후보별 이슈를 분석한 연구(Bae et al., 2013)가 자주 인용되고 있다.

하지만 기존의 이슈 트래킹은 분석 대상이 되는 전체 기간의 문서를 한꺼번에 분석해야 하기 때문에 확장성이 낮다는 한계를 갖고 있다. 또한 기존의 이슈 트래킹 결과는 특정 기간에 어떤 이슈에 대한 관심이 높았는지는 보여줄 수 있지만, 개인들의 관심이 어떤 이슈에서 어떤 이슈로 이동했는지는 보여주지 못한다는 한계를 갖는다.

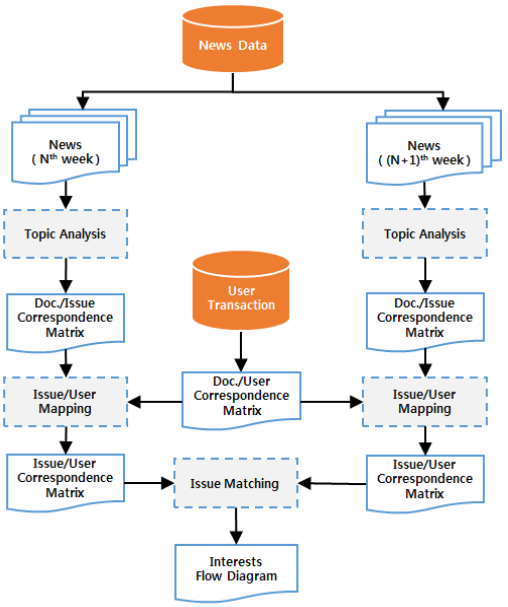
3. 개인 관심 트래킹 방법론

3.1 연구 모형

본 장에서는 개인 관심 트래킹의 개념 및 방안을 제시한다. 제안 방법론의 범위 및 세부 과정이 <Figure 4>에 요약되어 있다. 그림에서 원통형 도형은 외부 데이터를, 점선 직사각형은 프로세스를 나타낸다. 그 외의 도형은 중간 및 최종 산출물을 나타낸다.

제안 방법론을 구성하는 각 세부 과정에 대한 간략한 설명은 다음과 같다. 우선 분석 대상 기간에 속하는 전체 뉴스를 분석 단위인 세부 기간별로 구분한다. 간략한 설명을 위해 <Figure 4>에서는 세부 기간을 N주와 (N+1)주의 두 기간으로만 구분하였으나, 실제로는 보다 많은 세부 기간에 대한 분석이 필요하다. 다음 단계에서는 각 세부

기간별 뉴스에 대해 각각 토픽 분석을 별도로 수행하며, 각 토픽 분석의 결과로 문서/이슈간 대응 매트릭스가 도출된다. 한편 개인의 뉴스 접근 기록이 저장된 트랜잭션으로부터 문서/개인간 대응 매트릭스를 도출한다. 이렇게 도출된 두 매트릭스, 즉 문서/이슈간 매트릭스 및 문서/개인간 매트릭스를 병합하여 이슈/개인간 대응 매트릭스를 도출한다. 이렇게 도출된 두 개의 이슈/개인간 대응 매트릭스를 개인 관점에서 통합하여 최종 산출물인 각 개인 단위의 관심 흐름도를 도출한다. <Figure 4>의 주요 과정에 대한 세부 내용은 본 장의 이후 절에서 소개한다.



〈Figure 4〉 Research Overview

3.2 세부 기간별 토픽 분석

제안 방법론의 독창적인 부분 중 하나는 세부 기간별로 독립적인 토픽 분석을 통해 각 기간별 이슈를 도출하는 과정이다. 즉 각 기간별로 이슈를 도출하고, 추후 각 기간의 이슈별 매핑을 통해 전체 기간의 이슈의 흐름을 파악하고자 한다. 또한 이러한 이슈의 매핑이 개인 관점에서 이루어지기 때문에, 이슈의 흐름은 개인별 관심의 흐름을 나타내는 것으로 파악될 수 있다.

이러한 분석을 위한 첫 단계는 세부 기간별 이슈를 담고 있는 문서 집합을 준비하는 것이다. 본 연구에서는 게시일이 명확할 뿐 아니라 매우 정제된 표현만을 담고 있는 뉴스 기사를 분석 대상으로 사용한다. 따라서 분석의 첫 단계에서는 수집된 뉴스 기사를 각 세부 기간별로 나누는 작업이 수행된다. 본 장에서는 설명의 용이성을 위해 2주간의 뉴스를 수집하여 N주와 (N+1)주의 두 세부 기간으로 나누는 것을 가정하였으나, 실제로는 더욱 많은 세부 기간이 분석에 사용되어야 한다.

다음으로 이렇게 분할된 세부 기간별 뉴스에 대해 각 기간별로 토픽 분석을 실시한다. 토픽 분석은 각 문서에 포함된 용어의 빈도수에 근거하여 문서를 계량화한 뒤, 문서간의 거리에 기반하여 유사 문서를 그룹화하는 방법이다. 토픽 분석에서의 빈도수는 단순 빈도수가 아닌 TF-IDF(Term Frequency-Inverse Document Frequency) 기반 상대 빈도수가 사용되며, 토픽 분석의 결과로는 대표 키워드로 표현되는 각 토픽과 해당 토픽에 대응되는 문서 리스트가 생성된다. 본 장의 설명을 위해 간단한 가상 문서 집합에 대해 토픽 분석을 실시한 결과의 예가 <Figure 5>에 나타나있다.

Nth Week

이슈ID	주요 키워드	대응 문서
T1_1	정치, 국회, 국정감사, 대통령, 지방자치	D1_1, D1_2, D1_3
T1_2	연예, 걸그룹, 공연, 한류, 비보이	D1_2, D1_4, D1_5
T1_3	복지, 무상급식, 인권, 교육감, 세금	D1_1, D1_3
T1_4	주식, 경제, 선물, 금융, 투자	D1_3, D1_4, D1_5

(N+1)th Week

이슈ID	주요 키워드	대응 문서
T2_1	북한, 핵, 전쟁, 도발, 포격	D2_1, D2_2, D2_4, D2_5
T2_2	대선, 선거, 여론, 후보, 정당	D2_2, D2_3, D2_4
T2_3	종교, 교황, 기독교, 신앙, 가톨릭	D2_4, D2_5
T2_4	날씨, 태풍, 폭우, 호우, 재해	D2_1, D2_5

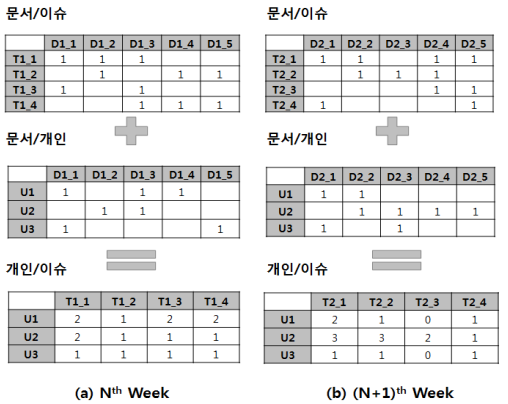
〈Figure 5〉 Example of Topic Analysis

<Figure 5>는 N주의 문서 5개(D1_1~D1_5)와 (N+1)주의 문서 5개(D2_1~D2_5)에 대해 각각 가

상 토픽 분석을 실시한 결과를 보여주며, 분석을 통해 N주의 토픽 4개(T1_1~T1_4)와 (N+1)주의 토픽 4개(T2_1~T2_4)가 도출되었음을 알 수 있다. 위의 결과에서 각 토픽은 다수의 문서를 포함하며, 어떤 문서는 둘 이상의 토픽에 대응됨을 확인할 수 있다. 토픽 분석의 세부 과정은 이미 많은 전문 서적 및 논문에 충분히 소개되어 있으므로, 본 논문에서는 자세히 소개하지 않는다.

3.3 개인별 관심이슈 도출

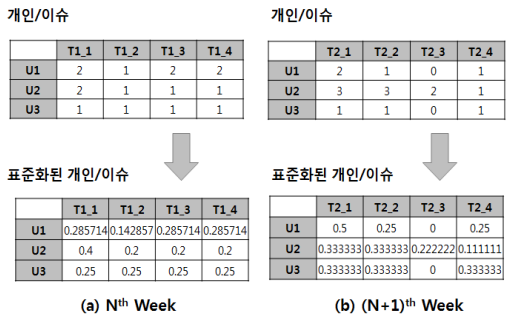
본 연구는 인터넷 뉴스를 조회하는 각 개인을 분석의 최소 단위로 사용하기 때문에, 분석을 위해서는 각 개인의 인터넷 뉴스 조회 기록이 반드시 필요하다. 본 절에서는 <Figure 5>에서 도출한 문서/이슈간 대응 매트릭스와 개인의 인터넷 뉴스 조회 기록으로부터 도출한 문서/개인간 대응 매트릭스를 통해 개인별 관심이슈를 도출하는 과정을 소개한다. 이를 위한 전체 과정은 <Figure 6>에 소개되어 있으며, <Figure 6>에서 각 개인은 U1~U3로 나타난다.



〈Figure 6〉 Individual/Issue Matrix

<Figure 6(a)>는 N주의 개인별 이슈 도출 과정을, <Figure 6(b)>는 (N+1)주의 개인별 이슈 도출 과정을 보여주고 있다. 최상단 매트릭스는 <Figure 5>의 표를 재정리한 것으로, 각 이슈별 대응 문서를 보여준다. 또한 가운데의 매트릭스는 개인의 인터넷 뉴스 조회기록을 요약한 것으로, 어떤 개인이 어떤 뉴스 기사를 조회했을 경우 대응되는 셀의 값은 '1'로 나타난다. 최하단 매트릭스는 개인별 이슈를 요약한 것으로, 셀 내의 값은 해당 개인이 조회한 기사 중 해당 이슈에 속한 기사의 수를 의미한다. 예를 들어 행 U1과 열 T1_1이 교차하는 셀의 값 '2'는 개인 U1이 이슈 T1_1에 속한 뉴스 두 개(D1_1, D1_3)를 조회한 결과로 나타난다.

이렇게 도출된 개인/이슈 매트릭스는 각 개인별로 표준화된 뒤 이후 과정에 사용된다. 즉 본 연구에서는 각 개인이 다양한 이슈에 대해 갖는 총합을 일정하게 유지함으로써, 인터넷 뉴스 조회 빈도가 높은 개인이 그렇지 않은 개인에 비해 모든 이슈에 대한 관심이 높게 나타나는 왜곡 현상을 방지하고자 한다. 이러한 개인별 표준화는 <Figure 7>의 과정을 통해 수행될 수 있다. <Figure 7>에서 상단 매트릭스는 <Figure 6>의 개인/이슈 매트릭스를 나타내며, 여기서 각 셀의 값을 각 행의 총합으로 나눔으로써 표준화된 개인/이슈 매트릭스를 도출할 수 있다.



〈Figure 7〉 Standardized Individual/Issue Matrix

3.4 이슈 매칭을 통한 개인 관심 트래킹

본 절에서는 제안 방법론의 마지막 단계인 이슈 매칭을 통한 개인 관심 트래킹 과정을 소개한다. 본 과정에서는 기본적으로 <Figure 7>의 최종 산출물인 두 개의 표준화된 개인/이슈 매트릭스를 사

용한다. 본 과정은 모든 개인에 대해 각 개인의 두 이슈에 대한 관심의 가중합을 구함으로써 이루어진다(식 (1)). 식 (1)에서 Match(T_a, T_b)는 이슈 T_a와 이슈 T_b의 대응도를 나타내며, n은 개인의 수를, U_i^a와 U_i^b는 각각 개인 U_i가 이슈 T_a와 이슈 T_b에 대해 갖는 관심도를 의미한다.

$$Match(T_a, T_b) = \sum_{i=1}^{n} (U_i^a \times U_i^b) \quad (1)$$

위의 식에 의하면 두 이슈 T_a와 이슈 T_b에 동시에 높은 관심을 가진 개인이 많을수록 두 이슈의 대응도인 Match(T_a, T_b)는 높게 나타난다. 이와 같은 방식으로 N주의 모든 이슈와 (N+1)주의 모든 이슈간의 대응도를 위 식에 의해 산출해야 하며, 이 과정은 행렬 곱에 의해 쉽게 구현될 수 있다. <Figure 7>에서 나타난 N주와 (N+1)주의 각각의 이슈간 대응도를 계산하는 과정이 <Figure 8>에 소개되어 있다.

	U1	U2	U3
T1_1	0.285714	0.4	0.25
T1_2	0.142857	0.2	0.25
T1_3	0.285714	0.2	0.25
T1_4	0.285714	0.2	0.25

(a) 표준화된 개인/이슈 (N^{th} Week)

	T2_1	T2_2	T2_3	T2_4
U1	0.5	0.25	0	0.25
U2	0.333333	0.333333	0.222222	0.111111
U3	0.333333	0.333333	0	0.333333

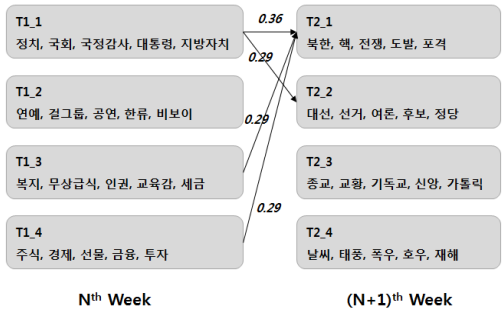
(b) 표준화된 개인/이슈 ($(N+1)^{th}$ Week)

	T2_1	T2_2	T2_3	T2_4
T1_1	0.36	0.29	0.9	0.20
T1_2	0.22	0.19	0.4	0.14
T1_3	0.29	0.22	0.4	0.18
T1_4	0.29	0.22	0.4	0.18

(c) 이슈간 대응 매트릭스

〈Figure 8〉 Issue Correspondence Matrix

<Figure 8(a)>는 <Figure 7(a)>의 최하단 매트릭스의 전치행렬(Transpose Matrix)를 나타내며, <Figure 8(b)>는 <Figure 7(b)>의 최하단 매트릭스와 동일하다. 이 두 매트릭스의 행렬 곱을 통해 <Figure 8(c)>를 도출할 수 있으며, 이 결과가 N주와 (N+1)주의 이슈간 대응 매트릭스가 된다. 이 슈간 대응 매트릭스에 대해 임의의 임계값을 설정하고, 이 임계값을 상회하는 대응 관계만을 도식화함으로써 N주와 (N+1)주간에 나타난 개인들의 관심의 흐름을 <Figure 9>와 같이 파악할 수 있다.



〈Figure 9〉 Flow of Individuals' Interests from Week N to Week(N+1)

<Figure 9>에서 N주에 T1_1, T1_3, T1_4의 이슈에 관심을 가졌던 개인들은 대체로 (N+1)주에는 T2_1의 이슈에 관심을 갖게 되었음을 알 수 있다. 또한 N주에 T1_1의 이슈에 관심을 가졌던 개인들은 (N+1)주에는 T2_1과 T2_2의 이슈로 관심이 분산되었음을 알 수 있다.

본 장에서는 제안 방법론의 전체 과정을 간략한 가상 시나리오에 대한 설명을 통해 소개하였다. 하지만 설명에 사용된 예는 매우 단순할 뿐 아니라 가상 예이기 때문에, 본 장의 예를 통해 제안 방법론의 실제 적용 가능성을 판단하기에는 무리가 있다. 따라서 충분한 양의 실제 뉴스 기사에 대해 제안 방법론을 적용해 볼 필요가 있으며, 이는 다음 장인 제 4장에서 다루기로 한다.

4. 실험

이전 장에서는 개인 관심 이슈 트래킹 방법론을 제안하고, 간단한 가상 예를 사용하여 제안 방법론을 설명하였다. 본 장에서는 인터넷 포털 사이트 뉴스 기사와 개인의 뉴스 접근 기록에 대하여 제안 방법론을 적용한 실험 과정 및 분석 결과를 제시한다.

4.1 데이터 소개

본 실험에서는 국내 한 인터넷 사이트 순위 분석 전문 업체로부터 제공받은 패널 5,000명의 2012년 7월부터 2013년 6월까지의 웹 사용 기록 약 1억 5천만 건을 사용하였다. 전체 사용 기록 중 해당 기간 내에 대형 인터넷 뉴스 포털 사이트를 방문한 사용자는 총 4,308명이었으며, 이들은 동일 기간 동안 234,776건의 뉴스 기사를 총 337,786번 방문한 것으로 나타났다. 기간별 이슈 분석을 위해 본 실험에서는 방문 기록에 나타난 뉴스 기사 원문 234,776건을 크롤링을 통해 수집하였다.

시행착오를 줄이기 위한 파일럿 실험에서는 세부 기간을 월 단위로 구분하여 사용하였다. 즉 2012년 7월과 2012년 8월에 게시된 뉴스 기사에 대한 분석을 통해 관심 트래킹을 시도하였다. 하지만 월 단위의 분석을 통해서는 두 기간 사이의 공통 이슈를 찾기가 매우 어려운 것으로 나타났다. 이는 이슈의 생성 및 변화 속도가 매우 빠른 현상에 기인한 것으로 해석될 수 있다. 따라서 실제 실험에서는 세부 기간을 주 단위로 구분하여 사용하였으며, 구체적으로는 2012년 7월 1일부터 2012년 7월 7일까지, 그리고 2012년 7월 8일부터 2012년 7월 14일까지의 2주간의 기록을 사용하였다. 전체 방문 기록 중 이 기간에 포함된 데이터는 사용자 1,112명의 방문 기록 1,506 건이었다. 또한 해당 기간에 게시되어 분석에 사용된 뉴스는 총 8,474건으로 나타났다. 수집 데이터와 분석 데이터에 대한 요약이 <Table 1>에 나타나있다.

〈Table 1〉 Summary of Experimental Data

항목	수집 데이터	분석 데이터
기간	2012.07.01 ~ 2013.03.31	2012.07.01 ~ 2013.07.14
사용자 수	4,308명	1,112명
방문 건수	337,786건	1,506건
뉴스 기사 수	234,776건	8,474건

4.2 세부 기간별 토픽 분석

본 절에서는 제 3.2절에서 제시된 과정인 세부 기간별 토픽 분석에 대한 결과를 소개한다. 우선 분석 데이터를 2012년 7월 1일부터 2012년 7월 7일까지의 첫 주(W1)의 기사 4,071건과 2012년 7월 8일부터 2012년 7월 14일까지의 둘째 주(W2) 기사 4,403건으로 분할하였다. 다음으로 분할된 두 기간의 뉴스 기사 각각에 대한 독립적인 토픽 분석을 통해 각 기간의 주요 이슈 및 각 이슈에 대응되는 기사를 파악하였다.

토픽 분석은 SAS Enterprise Miner 12.1의 Text Miner 모듈을 사용하여 파싱, 필터링, 토픽 분석의 순으로 진행하였으며, 각 기간별 토픽의 수는 25개로 제한하였다. 토픽 분석 결과로 나타난 문서/이슈 매트릭스 일부에 대한 스냅샷이 <Figure 10>에 나타나있다. <Figure 10(a)>는 W1의 기사에 대한 토픽 분석 결과를, <Figure 10(b)>는 W2의 기사에 대한 토픽 분석 결과를 보여준다. 각 그림에서 최상단 행은 이슈의 주요 키워드를, 다음 행은 해당 이슈의 번호를 나타낸다. 또한 맨 좌측 열은 문서(뉴스 기사)의 번호를 나타내며, 특정 문서가 특정 이슈에 해당되는 경우 해당 문서 번호와 이슈 번호가 교차하는 셀의 값을 '1'로 나타낸다.

	위원장,의원,경선,캠프,대선	의원,검찰,저축은행,회장,수사	갤럭시,애플,삼성전자,넥서스,구글	폭격,신사,드라마,장동건,도진	텔레콤,서비스,가입자,요금,유플러스	나지원,프록터,선수,두산,경기
Doc. No.	Topic4_2	Topic4_3	Topic4_5	Topic4_6	Topic4_8	Topic4_9
17	0	0	0	0	0	0
215	0	0	0	1	0	0
260	0	0	0	0	0	0
406	0	0	0	0	0	0
1028	0	0	0	0	0	0
1029	1	0	0	0	0	0
1030	0	0	0	1	0	0
1143	0	0	0	0	0	0
1192	0	0	0	0	0	0
1231	0	0	0	0	0	0

(a) W1의 문서/이슈 매트릭스

	시장,부동산,아파트,분기,경제	갤럭시,스마트폰,애플,제품,삼성전자	대선,경선,위원장,출마,새누리당	드라마,장동건,엘범,영화,엔터테인먼트	의원,협의,저축은행,검찰,수사	박지성,선수,감독,맨유,시즌
Doc.No.	Topic_2	Topic_4	Topic_5	Topic_6	Topic_8	Topic_10
19	1	0	1	0	0	0
128	0	1	0	0	0	0
301	0	0	0	0	0	0
773	1	1	0	0	0	0
962	0	0	0	0	0	0
1025	0	0	0	0	0	0
1026	0	0	1	0	0	0
1027	0	0	1	0	0	0
1053	0	0	0	0	0	0
1229	0	0	0	0	0	0

(b) W2의 문서/이슈 매트릭스

〈Figure 10〉 Document/Issue Matrix for Each Period (Part)

4.3 개인별 관심이슈 도출

본 절에서는 앞에서 도출한 세부 기간별 문서/이슈 매트릭스와 개인의 뉴스 조회 기록으로부터 개인/이슈 대응 매트릭스를 도출한 실험 결과를 소개한다. W1과 W2의 개인/이슈 매트릭스 도출과정은 서로 동일하므로, 본 절에서는 W1에 대한 결과만을 소개한다.

U_ID	Doc. No.
1	13089
1	21299
1	98498
5	43592
25	27536
25	52091
25	64357
25	76930
25	89166

(a) 원 데이터

U_ID	Doc. No.							
	13089	21299	27536	43592	52091	64357	76930	89166	98498
1	1	1	0	0	0	0	0	0	1
5	0	0	0	1	0	0	0	0	0
25	0	0	1	0	1	1	1	1	0

(b) 매트릭스 변환

〈Figure 11〉 Document/Individual Matrix(Part)

<Figure 11>은 W1 기간 내에 이루어진 개인의 뉴스 조회 기록, 즉 문서/개인 매트릭스를 보이고 있다. 지면 관계상 본 그림에서는 U_ID 1, 5, 25번 개인이 Doc. No. 100,000번 이내의 문서를 조회한 기록만을 보인다. <Figure 11(a)>는 조회 기록에 대한 원(Raw) 데이터이며, 이를 매트릭스로 변환한 형태는 <Figure 11(b)>와 같다. <Figure 10(a))>의 W1의 문서/이슈 매트릭스와 <Figure 11(b)>의 W1의 문서/개인 매트릭스에 대해 본문 제 3.3절에 소개된 방법을 적용하여 도출한 W1의 개인/이슈 매트릭스가 <Figure 12(a)>에 나타나있다. 한편 <Figure 12(b)>는 이와 동일한 방법에 의해 도출한 W2의 개인/이슈 매트릭스를 나타낸다. 이렇게 도출된 개인/이슈 매트릭스는 개인별 표준화 과정을 거쳐 이후 프로세스에 사용된다.

U_ID	위원장,의원,경선,캠프,대선	의원,검찰,저축은행,회장,수사	갤럭시,애플,삼성전자,넥서스,구글	폭격,신사,드라마,장동건,도진	텔레콤,서비스,가입자,요금,유플러스	나지완,프로토,선수,두산,경기
	Topic_2	Topic_3	Topic_5	Topic_6	Topic_8	Topic_9
1	0	0	0	3	0	1
25	0	0	0	2	0	0
29	1	2	0	0	0	0
30	5	7	0	3	1	0
32	0	0	2	0	2	0

(a) W1의 개인/이슈 매트릭스

U_ID	시장,부동산,아파트,분기,경제	갤럭시,스마트폰,애플,제품,삼성전자	대선,경선,위원장,출마,새누리당	드라마,장동건,엘범,영화,엔터테인먼트	의원,혐의,저축은행,검찰,수사	박지성,선수,감독,맨유,시즌
	Topic_2	Topic_4	Topic_5	Topic_6	Topic_8	Topic_10
1	0	0	0	0	0	1
4	0	0	0	0	1	0
28	1	0	1	0	1	5
30	2	1	4	1	3	0
34	5	2	1	0	1	0

(b) W2의 개인/이슈 매트릭스

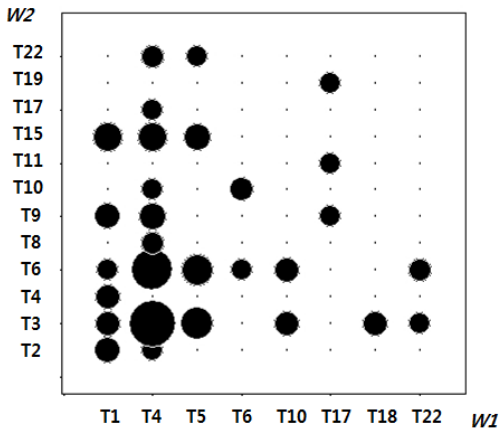
〈Figure 12〉 Individual/Issue Matrix for Each Period(Part)

4.4 이슈 매칭을 통한 개인 관심 트래킹

본 절에서는 앞 절에서 도출된 W1와 W2의 표준화된 개인/이슈 매트릭스를 통해 각 개인의 관심 이슈를 매칭한 결과를 소개한다. 이슈 매칭은 기본적으로 행렬 곱 연산에 의해 이루어지며, 자세한 과정은 제 3.4절에 소개되어 있다. 이슈 매칭 결과로부터 의미를 도출하기 위해 임계값을 설정해야 한다. 이 때 임계값이 너무 낮으면 전체 결과와 유사한 결과가 나타나며, 임계값이 너무 낮으면 당연한 결과만 얻게 될 수 있다. 따라서 다양한 임계값을 적용하는 시도를 통해 적정 임계값을 찾아야 하며, 본 실험에서는 1.5의 임계값을 적용하였다. 이슈 매칭 결과 전체에서 1.5 이상의 값을 갖는 셀을 회색으로 표시하고, 각 행이나 열 중 1.5 이상의 값을 하나도 갖지 못하는 행이나 열을 제거한 결과가 <Figure 13>에 나타나있다. 또한 1.5 이상의 값만을 버블 차트(Bubble Chart)로 도식화한 결과가 <Figure 14>에 나타나있다.

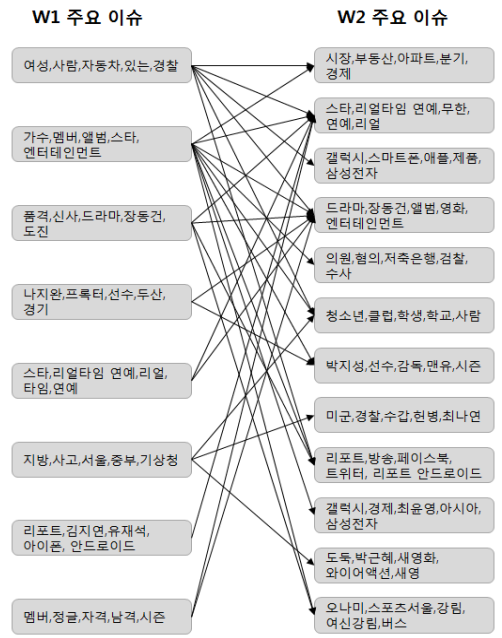
		W2										
		T2	T3	T4	T6	T8	T9	T10	T11	T15	T17	T19	T22
W1	T1	2.22	2.04	2.16	1.58	1.12	2.29	1.32	0.99	2.88	0.82	1.16	0.61
	T4	1.62	7.5	1.12	5.72	1.72	2.53	1.63	1.37	3	1.54	1.27	1.82
	T6	0.6	3.51	0.45	3.35	0.71	1.15	0.86	0.34	2.44	0.4	0.24	1.53
	T9	0.54	0.84	0.41	1.62	0.55	0.85	1.88	0.25	0.51	0.23	0.5	0.73
	T10	0.62	2.12	0.36	2.13	0.39	0.56	0.91	0.15	1.07	0.24	0.19	0.84
	T17	1.4	0.55	0.77	0.48	1.31	1.57	0.32	1.59	1.22	0.39	1.66	0.33
	T18	0.34	2.13	0.22	1.32	0.3	0.71	0.65	0.09	0.68	0.22	0.09	0.45
	T22	0.43	1.6	0.2	1.73	0.57	0.66	0.39	0.24	0.95	0.23	0.22	0.76

〈Figure 13〉 Results of Issue Matching



〈Figure 14〉 Plot Matrix for the Results of Issue Matching

마지막으로 <Figure 15>는 <Figure 13>에서 나타난 결과 중 임계값 이상의 대응도를 갖는 이슈 쌍을 도식화한 결과이다. 본 그림에서 이슈간 화살표는 W1의 해당 이슈에 관심을 갖던 개인들 중 많은 수가 W2의 해당 이슈로 관심을 옮겨갔음을 의미한다. 따라서 이러한 결과로부터 W1과 W2 사이의 개인들의 관심을 트래킹할 수 있다. 예를 들면 W1 기간에 “리포트, 김지연, 유재석, 아이폰, 안드로이드”라는 이슈에 관심을 갖고 있던 개인들은 W2 기간에는 “스타, 리얼타임 연예, 무한, 연예, 리얼”이라는 이슈에 관심을 보이는 경향이 있음을 알 수 있었다.



〈Figure 15〉 Flow of Individuals’ Interests During W1 and W2

본 장에서는 제안 방법론을 실제 뉴스 기사 및 해당 기사를 개인들이 조회한 기록에 대해 적용한 실험을 수행하였다. 실험 결과 각기 독립적으로 수행된 두 기간의 토픽 분석 결과로부터 주요 이슈를 추출하고, 이를 개인의 관점에서 대응시킴으로써 해당 기간 개인 관심의 흐름을 추적할 수 있었다. 하지만 토픽 분석의 결과 일부 정제되지 않은 이슈가 형성되고, 유사한 이슈가 둘 이상의 이슈로 구분되어 존재하는 등의 한계가 발견되었다. 향후 전처리 과정의 보강을 통해 보다 정제된 결과를 얻기 위한 노력이 반드시 필요하다.

5. 결 론

최근 다양한 소셜미디어 및 인터넷 매체를 통해 유통되는 텍스트 데이터로부터 주요 이슈를 발굴하기 위한 토픽 분석 및 여러 시점에 걸친 이슈의 변화를 분석하고 추적하기 위한 이슈 트래킹에 대한 연구가 활발하게 수행되고 있다. 하지만 기존의 이슈 트래킹은 특정 기간에 어떤 이슈에 대한 관심이 높았는지는 보여줄 수 있지만, 개인들의 관심이 어떤 이슈에서 어떤 이슈로 이동했는지는 보여주지 못한다는 한계를 갖고 있다. 또한 기존의 기법은 분석 대상이 되는 전체 기간의 문서를 한꺼번에 분석해야 하기 때문에 기간의 확장성이 낮다는 한계를 갖고 있다.

본 연구에서는 이러한 두 가지 한계를 극복하는 개인 관심 트래킹 방법론을 제안하였다. 즉 거시적 이슈의 흐름이 아닌 개인별 관심의 흐름을 파악할 수 있는 방안을 제시하였으며, 새로운 기간의 문서가 추가되었을 때 전체 기간의 문서가 아

닌 새로 추가된 문서에 대해서만 추가 분석을 수행할 수 있는 기간 확장 방안을 제시하였다. 학술적 관점에서 문서 자체의 정보가 아닌 관련 정보를 간접적으로 활용하여 이슈의 흐름을 분석하는 방법은 향후 다양한 후속 연구에서 활용될 수 있을 것으로 기대한다. 또한 실제 인터넷 뉴스 기사 및 해당 기사에 방문한 접속 기록에 대해 제안 방법론을 적용함으로써, 제안 방법론의 실무 적용 가능성을 분석하였다. 실무적 측면에서 제안 방법론을 통해 개인별 관심의 흐름을 파악할 수 있고, 이를 다양한 마케팅 전략 수립 과정에 활용할 수 있을 것으로 기대한다. 또한 내용만으로는 관련이 없는 것으로 보이는 이슈들 간의 선후관계를 파악함으로써, 하나의 이슈에서 다른 이슈로 관심이 옮겨가는 현상에 주목하고 그 원인을 파악하는 데에 단초를 제공할 수 있을 것이다.

하지만 본 연구가 제안하는 방법론이 실무적 성과를 얻기 위해서는 다음과 같은 후속 연구가 수행되어야 한다. 우선 양질의 이슈를 도출하기 위해서는 양질의 용어사전 및 불용어사전이 필요하므로, 이들 사전을 구축 또는 확보하기 위한 노력이 필요하다. 또한 본 연구의 실험에서는 세부 기간을 1주일 단위로 설정하고, 2주간의 관심 변화만을 추적하였다. 향후 개인의 관심 흐름을 보다 정교하게 추적하고 패턴을 발견하기 위해서는, 다양한 세부 기간에 대해 장기간에 걸친 분석이 반드시 이루어져야 한다. 또한 본 연구의 성과를 극대화하기 위해서는 제안 방법론의 최종 결과물인 이슈간 대응 매트릭스와 개인 관심 흐름도의 품질을 측정하기 위한 정량 지표의 개발이 후속 연구에서 다루어져야 하며, 이를 활용한 성과 평가가 반드시 수행되어야 한다.

References

Aggarwal, A., G. Waghmare, and A. Sureka, "Mining Issue Tracking Systems Using Topic Models for Trend Analysis, Corpus Exploration, and Understanding Evolution", *Proceedings of the 3rd International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering*, 2014, 52-58.

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

Bae, J., N. Han, and M. Song, "Twitter Issue Tracking System by Topic Modeling Techniques", *Journal of Intelligence and Information Systems*, Vol.20, No.2, 2014, 109-122.

Bae, J., J. Son, and M. Song, "Analysis of Twitter for 2012 South Korea Presidential Election by Text Mining Techniques", *Journal of Intelligence and Information Systems*, Vol.19, No.3, 2013, 141-156.

Ding, W. and C. Chen, "Dynamic Topic Detection and Tracking : A Comparison of HDP, C-word, and Cocitation Methods", *Journal of the Association for Information Science and Technology*, Vol.65, No.10, 2014, 2084-2097.

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

Hong, J., N. Kim, and S. Lee, "A Methodology for Automatic Multi-Categorization of Single-Categorized Documents", *Journal of Intelligence and Information Systems*, Vol.20, No.3, 2014.

Huang, S., W. Peng, J. Li, and D. Lee, "Sentiment and Topic Analysis on Social Media : a Multi-task Multi-label Classification Approach", *Proceedings of the 5th Annual ACM Web Science Conference*, 2013, 172-181.

Jin, H., R. Schwarts, S., Sista, F. Walls, "Topic Tracking for Radio, TV Broadcast and

Newswire", *Proceedings of DARPA Broadcast News Workshop*, 1999.

Kim, J., N. Kim, and Y. Cho, "User-Perspective Issue Clustering Using Multi-Layered Two-Mode Network Analysis", *Journal of Intelligence and Information Systems*, Vol.20, No.2, 2014.

Lee, K.M., H. Namgoong, E.H. Kim, G.Y. Lee, and H.K. Kim, "Analysis of Multi-dimensional Interaction among SNS Users", *Journal of Korean Society for Internet Information*, Vol.12, No.2, 2010, 113-122.

Ma, J., Y., Wang, H., Zhu, and Y Shen, "Research on Method of Adaptive Topic Tracking Based on Evolution of Public Opinion Ontology", *ACEEE International Journal on Information Technology*, Vol.4, No.1, 2014.

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

Provost, F. and T. Fawcett, Data Science for Business, O'Reilly, 2013.

Salton, G. and M.J. McGill, Introduction to Modern Information Retrieval, McGraw Hill, 1983.

Salton, G., A. Wong, and C.S. Yang, "A Vector Space Model for Automatic Indexing", *Communications of the ACM*, Vol.18, No.11, 1975, 613-620.

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

Xu, P., Y. Wu, E. Wei, T. Peng, S. Liu, J.H. Zhu, and H. Qu, "Visual Analysis of Topic Competition on Social Media", *IEEE Transactions on Visualization and Computer Graphics*, Vol.19, No.12, 2013, 2012-2021.

Yin, Z., L. Cao, and J. Han, "Latent Community Topic Analysis : Integration of Community Discovery with Topic Modeling", *Journal of ACM Transactions on Intelligent Systems and Technology*, Vol.3, No.4, 2012.

About the Authors

Chen Liu (liuchen@kookmin.ac.kr)

Chen Liu received the B.A. degree in Management Information Systems from Chungbuk University in 2012. He is a Master Candidate in Business IT at Kookmin University. His research interests include text mining and data mining.

Namgyu Kim (ngkim@kookmin.ac.kr)

Professor Namgyu Kim received the B.S. degree in Computer Engineering from Seoul National University in 1998 and Ph.D. degree in Management Engineering from Korea Advanced Institute of Science and Technology (KAIST) in 2007. He has been working for Kookmin University since then. His current research interests include text mining, data mining, and data modeling.