Scientia Iranica B (2015) 22(6), 2198–2208

SCIENTIA
IRANICA

Sharif University of Technology
Scientia Iranica
Transactions B: Mechanical Engineering
www.scientiairanica.com

PA-DHK: Polarity analysis for discovering hidden knowledge

J.-D. Kim[a,*], J. Son[a], H. Peter In[a], S.-H. Hwang[b], H. Lee[b] and D.-K. Baik[c]

a. *Department of Computer and Radio Communications Engineering, Korea University, Seoul, Republic of Korea.*
b. *Department of Computer Science & Engineering, Sun Moon University, Asan, Republic of Korea.*
c. *Graduate School of Convergence IT, Korea University, Seoul, Republic of Korea.*

Received 22 May 2014; received in revised form 15 June 2015; accepted 5 October 2015

KEYWORDS
Social network services;
Polarity analysis;
Formal concept analysis;
Knowledge discovery;
Twitter content.

Abstract. In a Social Network Service (SNS), a large amount of data with a variety of characteristics is generated through voluntary participation of users. These data are called "Big Social Data." Big social data can identify not only content registered on the web but also the relations of the friends of users. One of the most representative studies on SNS is analysis of the characteristics of social content and social relations, because SNS users tend to add people who are in close contact with them and have similar interests to their list of friends. Finding new knowledge from these large amounts of big social data can be very useful. This paper proposes a polarity analysis method for discovering hidden knowledge based on formal concept analysis in SNSs called PA-DHK. Further, we show, via experiments, that our data analysis approach can be applied to knowledge discovery using association rules.

1. Introduction

Social data analysis, in which various web resources are crawled in order to collect the required information, is a hot research topic. In a Social Networking Service (SNS), the collected information is used to gain more knowledge and the right approach to particular issues [1]. One of the most representative studies on SNSs is analysis of the characteristics of the social content and social relation, because SNS users tend to add people who are in close contact with them and have similar interests to their list of friends [2].

Finding hidden knowledge from these large amounts of big social data is very important [3]. Further, as the utility of social data analysis becomes more recognized, extensive studies are being conducted to actively analyze data in SNSs [4,5]. An SNS is a platform for building social networks or social relations among users that is able to generate and share large volumes of information in real time [6]. Most SNSs are web-based and provide means for users to interact over the internet. As the number of SNSs continues to increase, increasingly, more content is being created by users. Consequently, a large amount of data with a variety of characteristics is generated through voluntary participation of users in an SNS. These data are also called "Big Social Data" [7]. Social data analysis research on finding new knowledge from a large amount of data in big social data can identify not only content registered on the web, but also the author of that content as well as friends of that author. That is, big social data created in an SNS can be divided largely into "relational information between people" and "contents created by users". In particular, "relational information between people," which is a unique feature of SNSs, is highly appropriate

[*] *Corresponding author.*
E-mail addresses: kjd4u@korea.ac.kr (J.-D. Kim); redfunky07@korea.ac.kr (J. Son); hohin97@korea.ac.kr (H. Peter In); shwang@sunmoon.ac.kr (S.-H. Hwang); mahyun91@sunmoon.ac.kr (H. Lee) baikdk@korea.ac.kr (D.-K. Baik)

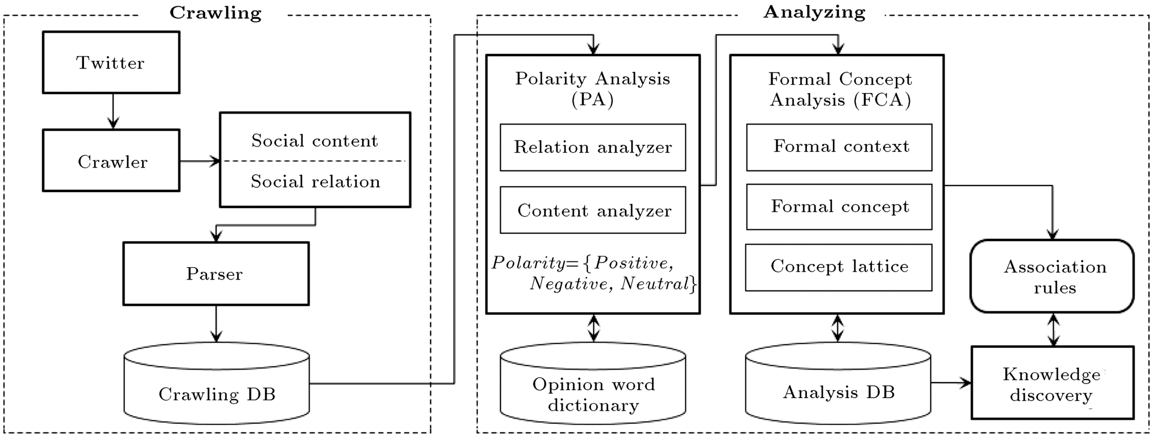


Figure 1. Overview of the PA-DHK model.

information for personalized services, which can be utilized in search and recommendation services. Thus, data generated via an SNS contains richer information than existing web data, which requires more complex data analysis processing [1].

As a result, SNS content (such as big data) has emerged as a new issue. Further, when a user's social network is available, the preferences of the user's related people can be utilized to assist in obtaining the user's preferences, assuming closely related people have similar interests. This is the main assumption when user interests and preferences are predicted based on the preferences of similar persons [6].

In a mountain of social content, it is becoming increasingly difficult for users to identify content in which they are interested. To solve this problem, we introduce Formal Concept Analysis (FCA) as the basis for a practical and well-founded methodological approach for web data analysis that identifies conceptual structures among datasets [8,9]. FCA is a method mainly used for the analysis of data, i.e. for investigating and processing explicitly given information. FCA classifies data based on an ordinary set into concept units consisting of objects and attributes that those objects have in common. More specifically, FCA extracts formal concepts from a given data table, grasps conceptual structures between concepts, and constructs a conceptual hierarchy. FCA has been applied to various domains, such as medicine, bioinformatics, social sciences, data mining, ontology, and software engineering [10].

In this paper, we propose a polarity analysis method based on FCA for discovering hidden knowledge in SNS. The proposed method is called PA-DHK. In addition, we show, via experiments, that our proposed approach can be applied for knowledge discovery.

The remainder of this paper is organized as follows: Section 2 gives an overview of the model for the proposed PA-DHK based on FCA. Section 3 outlines the experiments conducted and presents the results of polarity analysis and association rules for the proposed PA-DHK. Finally, conclusions and future study plans are summarized in Section 4.

2. Proposed approach

In this section, we give an overview of the model developed for the proposed PA-DHK and apply the complete model to analysis of user polarity and frequency in Twitter content. In addition, we describe the characteristics of the social content and describe the basic notions for understanding FCA.

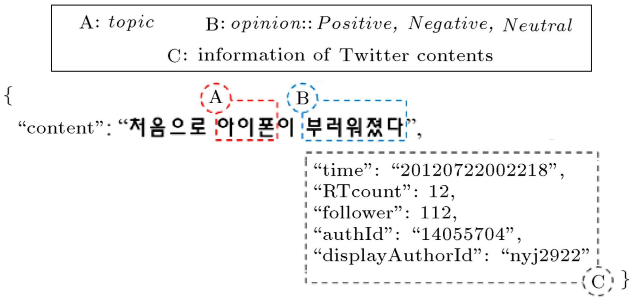
2.1. System overview

The proposed PA-DHK consists of two main parts: (1) *Crawling* of Twitter data such as social content and social relations, and (2) *Analyzing*, which is further separated into two sub-parts, *Polarity Analysis* (PA) and *FCA*.

Figure 1 gives an overview of the structure of the proposed PA-DHK. 1) *Crawling*: Twitter data are collected (social datasets such as social content and friendships) using the streaming API, parsed, and then stored in the *Crawling DB*. 2) *Analysis:* In this part, various preprocessing steps are performed. The PA is generated by the *Relation Analyzer* (RA) and *Content Analyzer* (CA). We developed our own Opinion Word Dictionary (OWD) and expanded SentiWordNet for analysis of Korean content in the Twitter dataset. Social content is analyzed via polarity using OWD. In addition, we analyze the intimacy and similarity of friends and friendship levels using RA. The PA results from analysis of Twitter data using OWD, and works in conjunction with FCA to provide association rules for discovering new knowledge through the polarity of the social data analyzed earlier.

2.2. Crawling

One of the most representative studys on Twitter is analysis of the characteristics of *Social Content* and *Social Relations*. *Social Content* is composed of


A: *topic*
B: *opinion*:: *Positive, Negative, Neutral*
C: information of Twitter contents
A
B
C

```
{
 "content": "처음으로 아이폰이 부러워졌다",
 "time": "20120722002218",
 "RTcount": 12,
 "follower": 112,
 "authId": "14055704",
 "displayAuthorId": "nyj2922" }
```

Figure 2. Example of contents crawled from Twitter.

combinations such as *contents*, *written time*, *RTcount*. *Social Relations* indicate the relationships between users, friends, and relations.

The crawler is collected using the streaming API, parsed using the JSON data style parser, and then registered in the *Crawling DB*. Figure 2 shows a part of the raw social content.

In our study, we crawled Korean content related to the smartphone domain iPhone, Galaxy, Optimus, Vega, Blackberry, HTC, etc. and then found all users who have content in our crawled data. Then, we analyzed the polarity in order to find topics (Figure 2-A: *iPhone)*, opinions (Figure 2-B: envy::positive), and information (Figure 2-C: *time, Rtcount, follower, authId*, etc.) characteristics of Twitter content.

Figure 3 shows a part of the raw social contents collected using a crawler in Twitter data. Figures 4 and 5 describe a part of the processed social contents data and the friendship relations of user data, respectively. In Figure 4, social content data has various pieces of information about a single content; for example: author, retweeted count, and time written. In Figure 5, user ID "14055704" possesses the social relations "17093617", "3108351", "18479513", and "15907720."

2.3. Analyzing

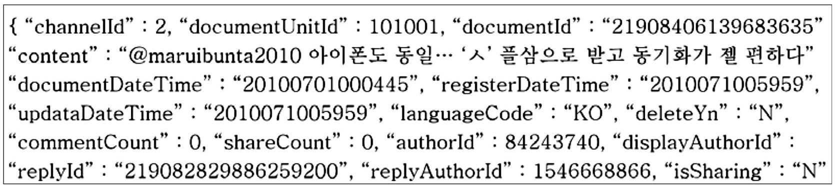
The proposed model consists of two main parts: PA and FCA. PA constitutes analysis of topics and opinions from Twitter content. FCA analyzes the associated rules for hidden knowledge.

2.3.1. Polarity Analysis (PA)

PA constitutes analysis of the polarity of topics such as *Positive*, *Negative*, and *Neutral* in Twitter content. We use SentiWordNet [11,12], a lexical resource for opinion mining that is associated with three sentiment scores in each WordNetsynset [13], in the PA. In other words, the method relies on training a set of ternary classifiers, each of which is capable of deciding whether a synset is *Positive, Negative*, or *Objective* [14].

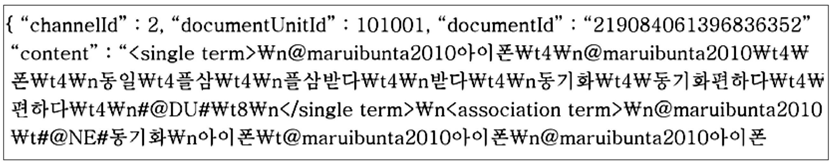
Definition 1. The SentiWordNet method defines L as the union of three seeds (i.e., training) sets, L_p, L_n, and L_o, of known *Positive, Negative*, and *Objective* synsets, respectively.

Each ternary classifier is generated using the semi-supervised method presented by Esuli and Sebastiani [15]. A semi-supervised method is a learning process whereby only a small subset, $L \subset T_r$ of the training data, T_r, has been manually labelled. Initially, the training data in $U = T_r - L$ are unlabeled. The process itself labels them, automatically, using L



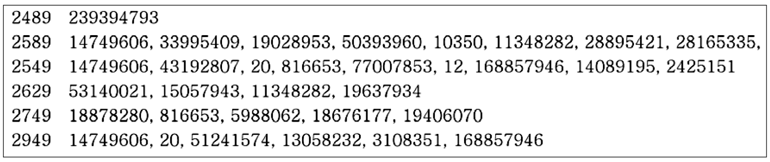
{ "channelId" : 2, "documentUnitId" : 101001, "documentId" : "21908406139683635"
"content" : "@maruibunta2010 아이폰도 동일… 'ㅅ' 플삼으로 받고 동기화가 젤 편하다"
"documentDateTime" : "20100701000445", "registerDateTime" : "2010071005959",
"updataDateTime" : "2010071005959", "languageCode" : "KO", "deleteYn" : "N",
"commentCount" : 0, "shareCount" : 0, "authorId" : 84243740, "displayAuthorId" :
"replyId" : "219082829886259200", "replyAuthorId" : 1546668866, "isSharing" : "N"

Figure 3. Part of the raw social contents in crawling DB.



{ "channelId" : 2, "documentUnitId" : 101001, "documentId" : "219084061396836352"
"content" : "<single term>₩n@maruibunta2010아이폰₩t4₩n@maruibunta2010₩t4₩
폰₩t4₩n동일₩t4플삼₩t4₩n플삼받다₩t4₩n받다₩t4₩n동기화₩t4₩동기화편하다₩t4₩
편하다₩t4₩n#@DU#₩t8₩n</single term>₩n<association term>₩n@maruibunta2010
₩t#@NE#동기화₩n아이폰₩t@maruibunta2010아이폰₩n@maruibunta2010아이폰

Figure 4. A part of the processed social contents in Analysis DB.



2489	239394793
2589	14749606, 33995409, 19028953, 50393960, 10350, 11348282, 28895421, 28165335,
2549	14749606, 43192807, 20, 816653, 77007853, 12, 168857946, 14089195, 2425151
2629	53140021, 15057943, 11348282, 19637934
2749	18878280, 816653, 5988062, 18676177, 19406070
2949	14749606, 20, 51241574, 13058232, 3108351, 168857946

Figure 5. A part of the processed friend relations in Analysis DB.

Table 1. OWD for smartphone domain.

Polarity	Opinions of contents	#
Positive	Good, convenience, simple envy, impress, intense, buy, fast, luxurious, recommend, strong, useful, and so on	384
Negative	Bad, inconvenience, slow, annoying, non-buy, difficult, non-recommended, fussy, countrified, complicated, and so on	509

(with the possible addition of other publicly available resources) as input.

L_p and L_n are two small sets, which we defined by manually selecting the intended synsets for 14 "paradigmatic" *Positive* and *Negative* terms (e.g., the positive terms good, nice, excellent, positive, fortunate, correct, and superior; the negative terms bad, nasty, poor, negative, unfortunate, wrong, and inferior) which were used as seed terms by Turney and Littman [16].

In addition, the polarity of Twitter content is analyzed using OWD. In this paper, we develop our own OWD by exploiting SentiWordNet. Using SentiWordNet, we find representative Korean vocabulary representing *Positive* and *Negative*, then add and modify the Twitter dataset; content with ambiguous opinion are classified as *Neutral*. Thus, Twitter content is classified into *Positive*, *Negative*, and *Neutral*. Table 1 shows a part of the OWD for Korean PA in the smartphone domain.

2.3.2. Formal Concept Analysis (FCA)

FCA is primarily used to analyze data, i.e. to investigate and process explicitly given information. Such data are structured into units that are formal abstractions of concepts of human thought, allowing meaningful comprehensible interpretation. FCA was introduced as a mathematical theory for modeling the concept of a "concept" in terms of lattice theory [8,9]. This approach arose independently of ontologies, resulting in a different formalization of concepts. FCA consists of *Formal Context*, *Formal Concept*, and *Concept Lattice*.

FCA starts with a *Formal Context* comprising a set of objects, a set of attributes, and a relation describing which objects possess which attributes. In the formal definition, the set of objects is denoted by O, and the set of attributes is denoted by A.

Definition 2. A formal context is a triple $(O,\ A,\ R)$, where O is a set of objects and A is a set of attributes, and $R \subseteq O \times A$ is a binary relation between O and A. In order to express that an object, o, is in a relation with an attribute, a, we write $(o, a) \in R$ and read it as "the object o has the attribute a".

The central notion of FCA is the *Formal Concept*. Objects from a context share a set of common attributes, and vice versa. Concepts are pairs of objects and attributes which are synonymous and thus characterize each other. Concepts can be imagined as maximal rectangles in the context table. If we ignore the sequence of rows and columns, we can identify even more concepts. A formal definition of the concept is given in the following:

Definition 3. *Let $(O,\ A,\ R)$ be a context. A formal concept is a pair (X, Y) with $X \subseteq O$ is called extension, $Y \subseteq A$ is called intension, and $(X = extent(Y)) \wedge (Y = intent(X))$:*

In other words, a concept is a pair consisting of a set of objects and a set of attributes which are mapped into each other by the Galois connection. The set of all concepts of the context, $C = (O,\ A,\ R)$, is denoted by $B(C)$ or $B(O,\ A,\ R)$, i.e., $B(C) = \{(X,Y) \in 2^O X 2^A | X = extent(Y) \wedge (Y = intent(X))\}$.

The set of formal concepts is organized by the partial ordering relation $\leq$ to be read as "is a sub-concept of" as follows.

Definition 4. *For a formal context $C = (O,\ A,\ R)$ and two concepts $c_1 = (O_1,\ A_1)$, $c_2 = (O_2,\ A_2) \in B(C)$ the sub-concept/super-concept relation is given by $(O_1,\ A_1) \leq (O_2; A_2) \Leftrightarrow O_1 \subseteq O_2 (\Leftrightarrow A_1 \supseteq A_2)$.*

In the Formal Concept Lattice, a relationship shows that dualism exists between attributes and objects of concepts. A concept, $c_1 = (O_1,\ A_1)$, is a sub-concept of concept $c_2 = (O_2, A_2)$ iff the set of its objects is a subset of the objects of c_2, or, an equivalent expression is iff the set of its attributes is a superset of the attributes of c_2. That is, a sub-concept contains fewer objects and more attributes than its super-concept. The set of all formal concepts of context C with the *sub-concept/super-concept realtion* is always a complete lattice, called the *(formal) concept lattice* of C, and denoted by $L := (B(C), \leq)$.

A *Concept Lattice* can be represented graphically using line diagrams (such as Hasse diagrams). These structures are composed of nodes and links. Each node represents a concept with its associated intentional description. The links connecting nodes represent the sub-concept/super-concept relation between them. This relation indicates that the parent's extension is a superset of each child's extension. Attributes propagate along the edges to the bottom of the diagram and dual objects propagate to the top of the diagram. More abstract or general nodes occur higher in the hierarchy, whereas more specific ones occur at lower levels. Herein, we can summarize the above considerations as a brief algorithm to construct the concept lattice in Algorithm 1.

An *Association Rule* extraction is one of the most

```
INPUT: a formal context $C := (O,\ A,\ R)$
OUTPUT: concept Lattice $L := (B(C), E_{\leq})$
for all $o \in O$ do
    $B(C) \leftarrow B(C) \cup (extent\ (intent(o)), intent(o))$;
end for
for all $c \in B(C)$ do
    for all $o \in (O - extent(c))$ do
        $X \leftarrow extent(c) \cup \{o\}$;
        if $(extent\ (intent(X)), intent(X) \notin B(C)$ then
            $B(C) \leftarrow B(C) \cup (extent\ (intent(X)), intent(X))$;
        end if
    end for
for all $c_1 \in B(C)$ do
    for all $c_2 \in B(C) - \{c_1\})$ do
        if $(c_1 \leq c_2) \wedge (\nexists c_3 \in B(C) - \{c_1, c_2\}[(c_1 \leq c_3) \wedge (c_3 \leq c_2)])$ then
            $E_{\leq} \leftarrow E_{\leq} \cup \{(c_1, c_2)\}$;
        end if
    end for
end for
```

Algorithm 1. Generate concepts and build *Concept Lattice.*

actively researched areas in data mining. It aims to extract the association relationship between two data groups. In this paper, the formal definition of an association rule is expressed in terms of FCA as follows.

Definition 5. Given that the user has defined minimum support, *minsup* $\in [0,\ 1]$, and minimum confidence, $minconf \in [0,\ 1]$, thresholds, a formal context, $(O,\ A,\ R)$, satisfies the association rule, $P \Rightarrow Q$, with $P, Q \subseteq M$, called the antecedent and consequent of the rule, respectively, if $sup(P \Rightarrow Q) = |P_R \cap Q_R|/|O| \geq minsup$ and $conf(P \rightarrow Q) = |P_R \cap Q_R|/|P_R| \geq minconf$.

The ratios, $sup(P \Rightarrow Q)$ and $conf(P \Rightarrow Q)$, are called the *Support* and the *Confidence* of the rule, $P \Rightarrow Q$, respectively. The *support* is the probability of an object satisfying both P and Q. The *confidence* is the number of objects satisfying both P and Q, divided by the number of objects satisfying the attribute set, P.

3. Experiments and result

In this paper, one of the most important experiments is to determine the characteristics of the opinion of topic, and its preferences. Furthermore, the most important activity is analysis for discovering hidden knowledge. In this experiment, we utilize the two analysis models, PA and FCA, and present the results of the experiment for hidden knowledge discovery.

PA constitutes analysis of explicit information such as topic, opinion, time written, and frequency of topic in Twitter content. FAC is used to analyze implicit information/hidden knowledge, because FAC can automatically search for association rules. In this experiment, we used the social content of the Korean Twitter dataset collected over a period of one month, from July 1, 2012 to July 31, 2012. In addition, we used OWD to analyze the polarity of contents. The overall size of the content data obtained was 105.4 GB. Table 2 summarizes the experimental Twitter content dataset.

Table 2. Dataset of Twitter contents.

Timescale		**2012-07-01 ~ 2012-07-30**
Contents	Domain	Smartphone
	Topics	*iPhone, Galaxy, Optimus,* *Vega, Blackberry,* HTC
	# of contents	259,176
Users	All users	25,249
	Active users	100

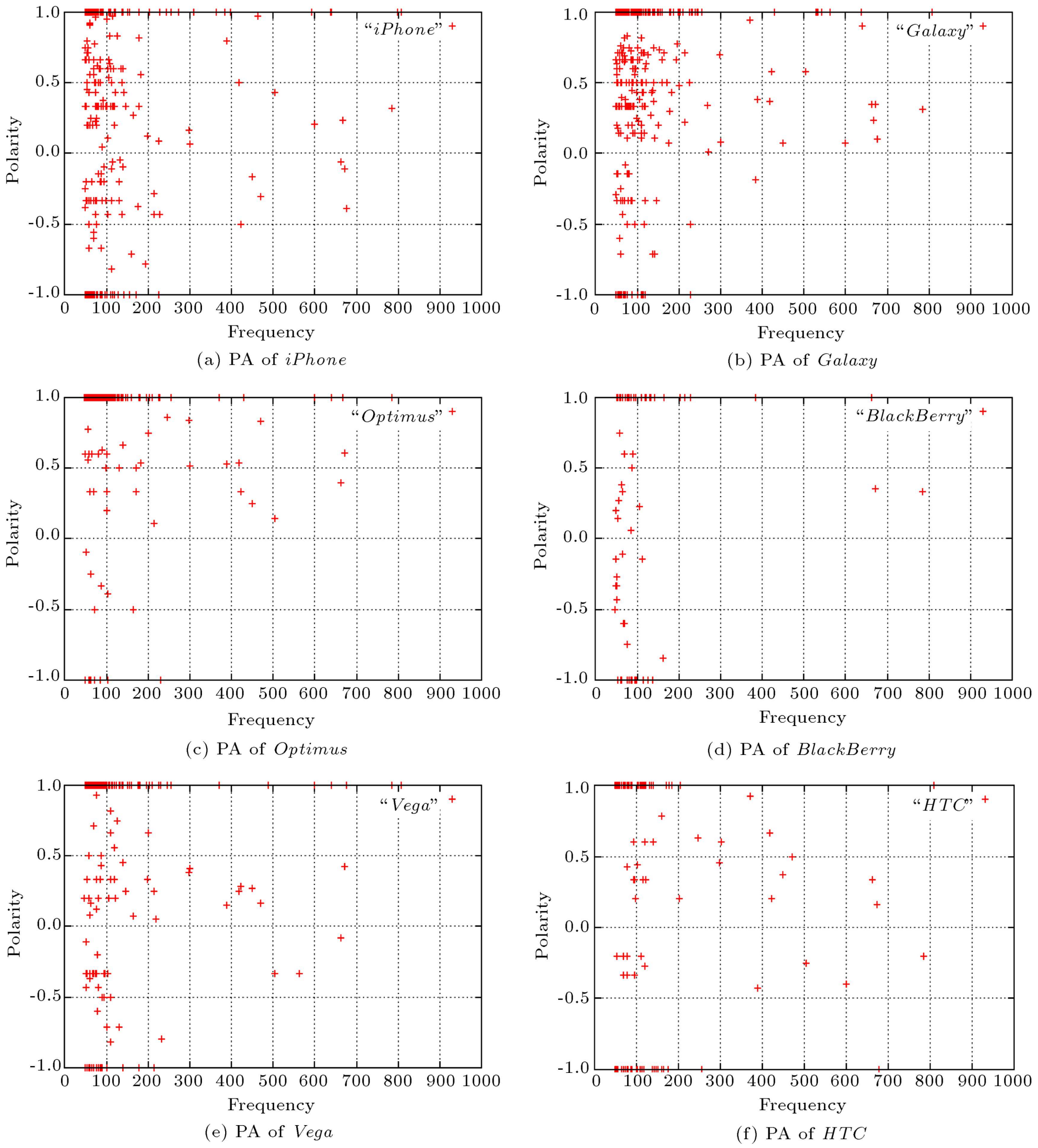
In the experiment, we crawled Twitter contents related to the smartphone domain. Then, we found all topics (*iPhone, Galaxy, Optimus*, and so on) and opinions of users with content in our crawled data. The polarity of the Twitter content was analyzed using OWD. OWD consists of 384 *Positive* words and 509 *Negative* words. The Twitter contents were classified into *Positive, Negative*, and *Neutral* for opinion analysis in this experiment. The *Positive* and *Negative* cases in Twitter content were clear and easily analyzed. However, ambiguous content is not easy to classify accurately. Therefore, ambiguous content was classified as neutral.

First, the results for polarity and frequency in the smartphone domain were obtained from the PA analysis model. In PA, we analyzed the topics, polarity and frequency of polarity of 25,249 registered users, over a period of one month. Figure 6 shows the results of PA in the form of the Gnuplot visualization. Gnuplot is a command line interface software that depicts data in 2D and 3D data. Further, Gnuplot is freeware and supports various operating systems and widely used scientific data expressions in academia [17].

Figure 7 shows the overall results of frequency relating to the polarity (*Positive* and *Negative*) of 25,249 users in Twitter content. In Figure 7, the x-axis indicates the polarity of the topics in the smartphone domain, and the y-axis represents the frequency of the polarity. In addition, Figure 7 represents a value greater than zero, indicating a *Positive* polarity; polarity values less than zero are shown as *Negative.*

The topic of *iPhone* showed 13,166 positive and 8,620 negative polarities. It was 1.53 times more positive than negative. The topic of *Galaxy* showed 114,983 positive and 3,481 negative polarities, which is 4.29 times more positive than negative. As shown in Figure 7, the *Galaxy* has more positive preferences than the *iPhone*; i.e. when the amount of negative polarity is *Galaxy: iPhone* = 3,481 : 8,620 (that is, about 1: 2.48).

The results for hidden knowledge, using association rules in a smartphone, were found using the FCA analysis model. Table 3 shows the results for the analysis topic of the smartphone domain using the formal context with polarities. Based on this formal context, we built a concept lattice. Figure 8



(a) PA of *iPhone*

(b) PA of *Galaxy*

(c) PA of *Optimus*

(d) PA of *BlackBerry*

(e) PA of *Vega*

(f) PA of *HTC*

Figure 6. Results of PA in smartphone domain.

Table 3. Formal context of polarity from Twitter dataset.

User id	**Topics**																	
	iPhone			***Galaxy***			***Optimus***			***Vega***			***Blackberry***			***HTC***		
	Pos.	***Neg.***	***Neu.***	***Pos.***	***Neg.***	***Neu.***	***Pos.***	***Neg.***	***Neu.***	***Pos.***	***Neg.***	***Neu.***	***Pos.***	***Neg.***	***Neu.***	***Pos.***	***Neg.***	***Neu.***
163299788	x	x	x	x	x	x		x	x			x	x		x			x
282578719	x		x	x	x	x	x		x	x	x	x	x		x	x		x
185527079	x		x	x		x												
612435148			x	x		x	x	x		x	x	x				x	x	x
525332693	x		x	x		x				x		x				x		
214716076	x		x															
96033417	x	x	x	x	x	x	x		x	x		x	x	x	x	x	x	x
125025519	x	x	x	x	x	x				x		x			x		x	x
477843013	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
521165154	x	x	x	x	x	x	x		x									

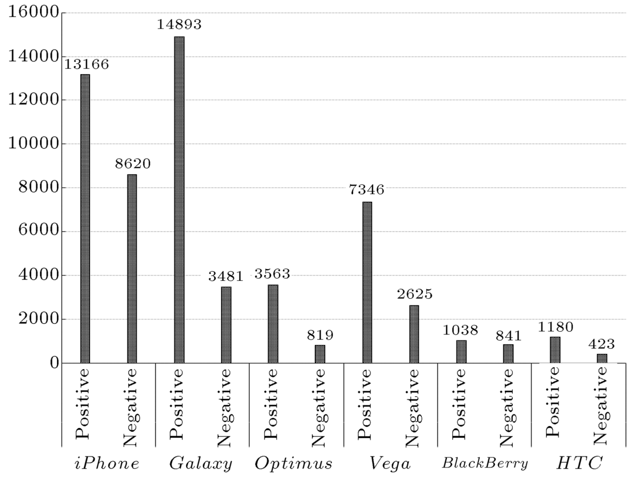


Figure 7. Overall results of frequency for polarity in smartphone domain.

shows the concept hierarchy from Table 3. The concept lattice allows us to easily identify the sets of preferences (or nonpreferences) that are best suited to be taken into account for the definition of user opinion.

In the FCA experiments, the results showed *Positive, Neutral*, and *Negative.* The topics mentioned intermittently are not suitable for preference analysis. Therefore, for the results of the FCA model, we conducted the experiment using 100 active users with significant amounts of content among 25,249 users related to the smartphone domain.

In the experiment, formal contexts were generated as users, and the polarity of their attributes in the Twitter content analyzed. Association Rules (AR) were then extracted from the concept lattice created earlier, when the number of users is ten. From the *Concept Lattice*, we automatically inferred AR such as those shown in Figure 9. In the figure, *users: concepts* for ten users from the experiment are displayed. It can be seen that a total of 65 ARs were extracted: among them 30 ARs with 100%, two ARs with 90%, one AR with 86%, seven ARs with 83%, nine ARs with 80%, one AR with 78%, six ARs with 75%, five ARs with 67%, and one AR with 50% confidence rates.

The AR allows us to identify a number of opinions associated with users who share common interests. In FCA, association rule $A \rightarrow B$ means that every object possessing each attribute from A also has each attribute from B. Two basic metrics, *Support* and *Confidence*, are used to find the sets of interest defining user opinions in the Twitter dataset:

Support: This FCA denotes the proportion of users who expressed their interest in a set of attributes, such as $Support(A, B) = P(A \cap B)$. For example, association rule 29 (in Figure 9) indicates that six users have interesting opinions on *iPhoneN* and *HTCN*, rule 31 indicates that ten users have an interesting opinion

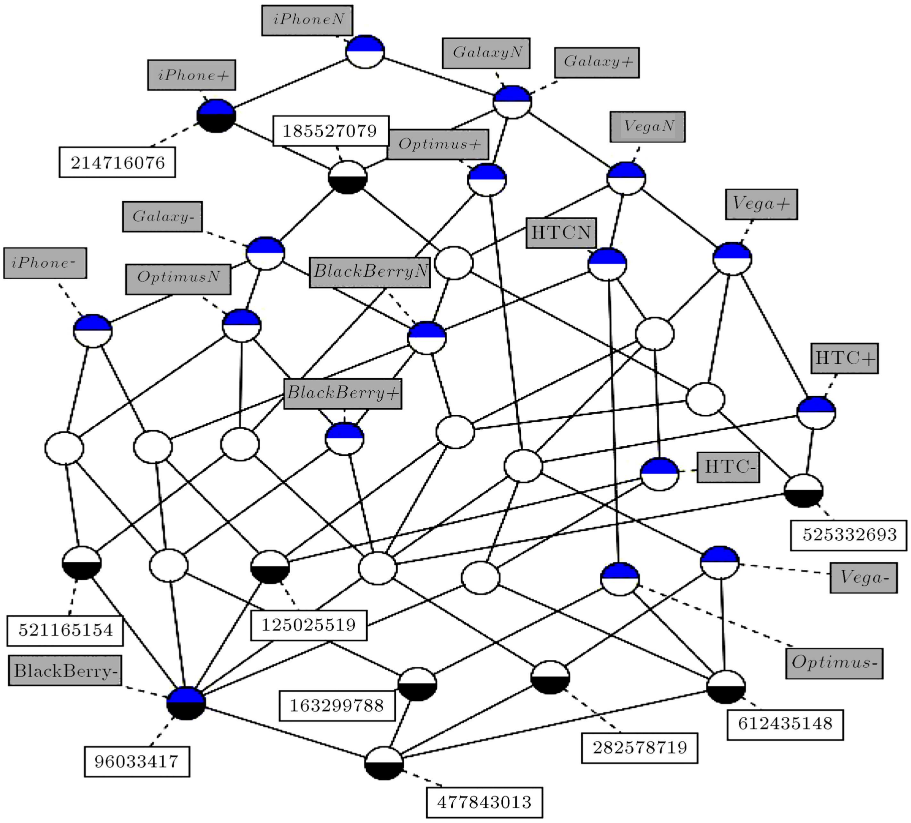
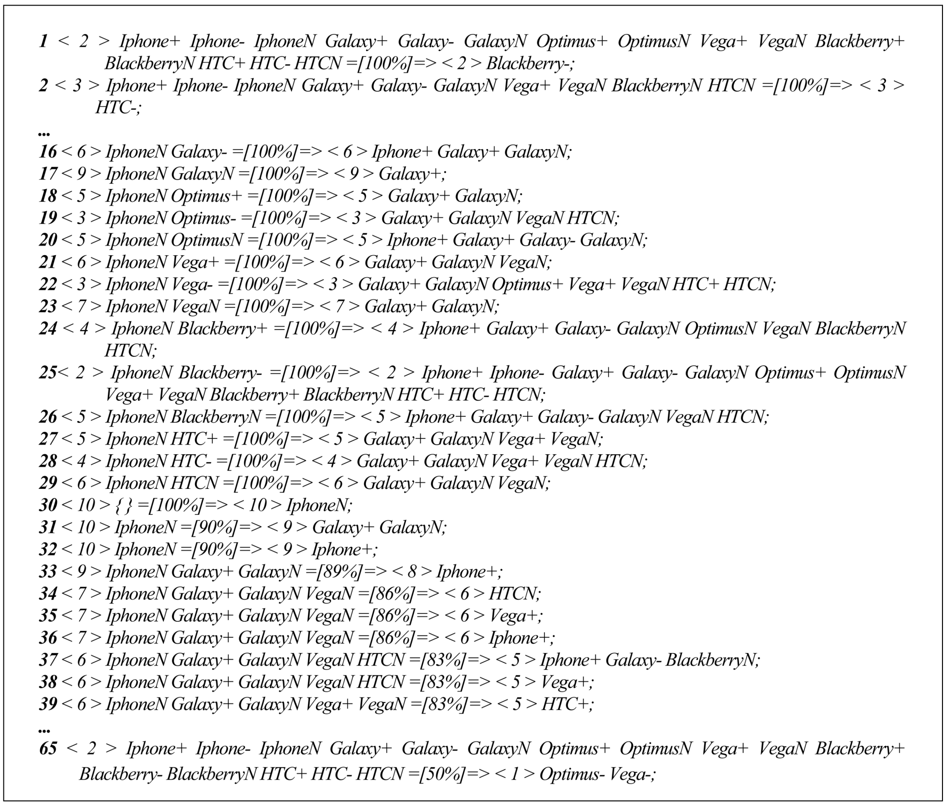


Figure 8. Concept lattice from Table 3.



```
1 < 2 > Iphone+ Iphone- IphoneN Galaxy+ Galaxy- GalaxyN Optimus+ OptimusN Vega+ VegaN Blackberry+ BlackberryN HTC+ HTC- HTCN =[100%]=> < 2 > Blackberry-;
2 < 3 > Iphone+ Iphone- IphoneN Galaxy+ Galaxy- GalaxyN Vega+ VegaN BlackberryN HTCN =[100%]=> < 3 > HTC-;
...
16 < 6 > IphoneN Galaxy- =[100%]=> < 6 > Iphone+ Galaxy+ GalaxyN;
17 < 9 > IphoneN GalaxyN =[100%]=> < 9 > Galaxy+;
18 < 5 > IphoneN Optimus+ =[100%]=> < 5 > Galaxy+ GalaxyN;
19 < 3 > IphoneN Optimus- =[100%]=> < 3 > Galaxy+ GalaxyN VegaN HTCN;
20 < 5 > IphoneN OptimusN =[100%]=> < 5 > Iphone+ Galaxy+ Galaxy- GalaxyN;
21 < 6 > IphoneN Vega+ =[100%]=> < 6 > Galaxy+ GalaxyN VegaN;
22 < 3 > IphoneN Vega- =[100%]=> < 3 > Galaxy+ GalaxyN Optimus+ Vega+ VegaN HTC+ HTCN;
23 < 7 > IphoneN VegaN =[100%]=> < 7 > Galaxy+ GalaxyN;
24 < 4 > IphoneN Blackberry+ =[100%]=> < 4 > Iphone+ Galaxy+ Galaxy- GalaxyN OptimusN VegaN BlackberryN HTCN;
25< 2 > IphoneN Blackberry- =[100%]=> < 2 > Iphone+ Iphone- Galaxy+ Galaxy- GalaxyN Optimus+ OptimusN Vega+ VegaN Blackberry+ BlackberryN HTC+ HTC- HTCN;
26 < 5 > IphoneN BlackberryN =[100%]=> < 5 > Iphone+ Galaxy+ Galaxy- GalaxyN VegaN HTCN;
27 < 5 > IphoneN HTC+ =[100%]=> < 5 > Galaxy+ GalaxyN Vega+ VegaN;
28 < 4 > IphoneN HTC- =[100%]=> < 4 > Galaxy+ GalaxyN Vega+ VegaN HTCN;
29 < 6 > IphoneN HTCN =[100%]=> < 6 > Galaxy+ GalaxyN VegaN;
30 < 10 > { } =[100%]=> < 10 > IphoneN;
31 < 10 > IphoneN =[90%]=> < 9 > Galaxy+ GalaxyN;
32 < 10 > IphoneN =[90%]=> < 9 > Iphone+;
33 < 9 > IphoneN Galaxy+ GalaxyN =[89%]=> < 8 > Iphone+;
34 < 7 > IphoneN Galaxy+ GalaxyN VegaN =[86%]=> < 6 > HTCN;
35 < 7 > IphoneN Galaxy+ GalaxyN VegaN =[86%]=> < 6 > Vega+;
36 < 7 > IphoneN Galaxy+ GalaxyN VegaN =[86%]=> < 6 > Iphone+;
37 < 6 > IphoneN Galaxy+ GalaxyN VegaN HTCN =[83%]=> < 5 > Iphone+ Galaxy- BlackberryN;
38 < 6 > IphoneN Galaxy+ GalaxyN VegaN HTCN =[83%]=> < 5 > Vega+;
39 < 6 > IphoneN Galaxy+ GalaxyN Vega+ VegaN =[83%]=> < 5 > HTC+;
...
65 < 2 > Iphone+ Iphone- IphoneN Galaxy+ Galaxy- GalaxyN Optimus+ OptimusN Vega+ VegaN Blackberry+ Blackberry- BlackberryN HTC+ HTC- HTCN =[50%]=> < 1 > Optimus- Vega-;
```

Figure 9. Results of knowledge discovery from Table 2: Association rules.

Table 4. Overall results about the number of association rules.

	Confidence					
# of users	**100%**	**99%~90%**	**89%~80%**	**79%~70%**	**69%~60%**	**Less than 59%**
# 100	121	547	59	10	3	1
# 90	113	481	53	9	3	1
# 80	111	462	52	9	2	1
# 70	100	405	37	5	2	2
# 60	89	292	36	6	3	1
# 50	70	171	21	4	3	2
# 40	52	81	17	4	3	2
# 30	42	47	19	2	3	2
# 20	39	5	33	7	2	1
# 10	30	2	20	7	5	1
# 5	19	0	2	3	4	4

on *iPhoneN*, and rule 39 indicates that six users have an interesting opinion on *iPhoneN, Galaxy+, GalaxyN, Vega+*, and *VegaN*. The support values are, therefore, 45%, 98%, and 25%, respectively.

Confidence: This FCA represents the proportion of users who have an interest in consequent rules, given that they have an interest in antecedent rules, such as $\mathit{Confidence}\ (A, B) = P(B|A) = P(A \cap B)/P(A)$. AR number 29 in Figure 9 states that six out of six users who have interesting opinions on *iPhoneN* and *HTCN* have interesting opinions on *Galaxy+, GalaxyN*, and *VegaN*. Therefore, the confidence in this case is equal to 100%.

The experimental results of clustering and classification of the AR are summarized in Table 4 and Figure 10. Table 4 shows overall results about the number of association rules for which the confidence rates are 100%, 99~90%, 89~80%, 79~70%, 69~60%, and less than 59%, respectively. In Figure 10, the *x*-axis indicates the confidence ratio, and the *y*-axis

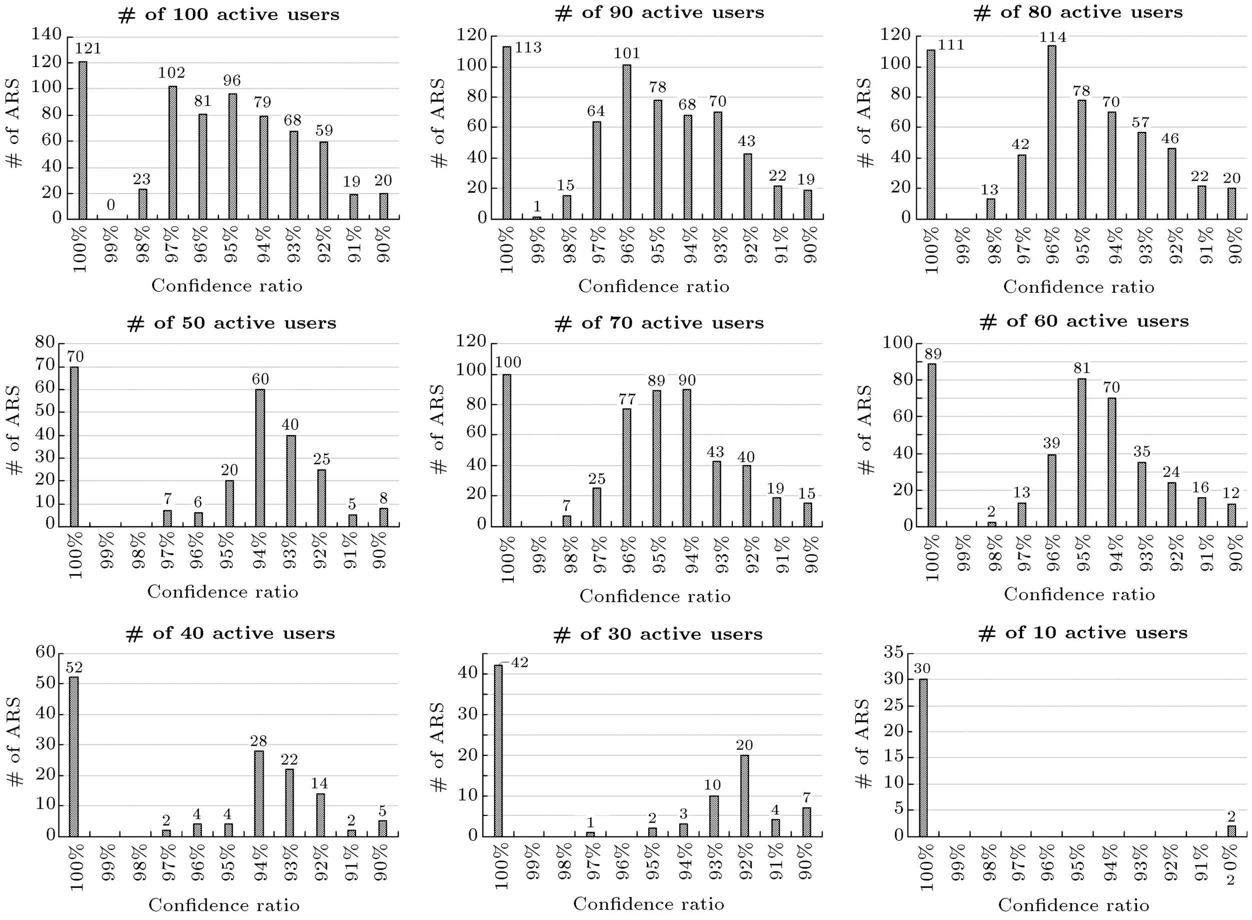


Figure 10. Results of association rules greater than 90% confidence.

represents the number of APs for each active user, such as 10, 30, 40, 50, 60, 70, 80, 90, and 100.

4. Conclusion

This paper proposes the PA-DHK method that identifies conceptual structures among Twitter content. The FCA-based data analysis approach in PA-DHK consists of two parts: PA and FCA. PA is used to analyze the polarity of web data using the extended OWD, while FCA, in fact an FCA-based analysis module, is used to discover new knowledge, such as association rules, through the polarity of the web data analyzed earlier. A key feature of the proposed data analysis method is that it supports clustering and extracting association rules using the polarity of terms from social relations among users. In addition, the experiments conducted showed how our data analysis method can be applied for knowledge discovery from Twitter datasets. We have designed a novel approach to efficiently represent evolving user preferences and interests. The proposed approach will help service providers to provide personalized content and service, and will contribute to increased satisfaction with opinion mining services.

However, the proposed approach still faces many challenges, particularly, in areas such as usability and visualization capabilities. In the future, we plan to improve usability and visualization and also to extend its capability to various domains.

Acknowledgments

This research was supported by a Korea University Grant, and the Basic Science Research Program, through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (2015R1C1A1A02036442), and the Next-Generation Information Computing Development Program through the NRF funded by the Ministry of Science, ICT & Future Planning (2012M3C4A7033346). The authors thank Daumsoft, for assistance with experiment data (Twitter data crawling).

References

1. Liu, B. "Web data mining: Exploring hyperlinks, contents, and usage data", 2nd Edn., Springer Publishes (2011).
2. Seol, K.S., Kim, J.D., Shin, H.N. and Baik, D.K. "In-

timacy measurement method and experiment between social network service users", *J. of KIISE.*, **39**(4), pp. 335-341 (2012).

3. Jiang, M., Cui, P., Liu, R., Yang, Q., Wang, F., Zhu, W. and Yang, S. "Social contextual recommendation", *21st ACM Int. Conf. on Information and Knowledge Management*, Hawaii, USA, pp. 45-54 (2012).

4. Golbeck, J. "Generating predictive movie recommendations from trust in social networks", *4th Int. Conf. on Trust Management*, Pisa, Italy, pp. 93-104 (2006).

5. Guy, I., Zwerdling, N., Ronen, I., Carmel, D. and Uziel, E. "Social media recommendation based on people and tags", *33rd Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Geneva, Switzerland, pp. 194-201 (2010).

6. Wasserman, S. and Faust, K., *Social Network Analysis Methods and Applications*, Cambridge Univ. Publishers, NY (1994).

7. Cambria, E., Rajagopal, D., Olsher, D. and Das, D., *Big Data Computing-Big Social Data Analysis*, Cambridge University Publishers, NY (2014).

8. Ganter, B. and Wille, R., *Formal Concept Analysis: Mathematical Foundations*, Springer Publishers (1999).

9. Birkhoff, G., *Lattice Theory*, American Mathematical Society Coll., Publishers (1940).

10. Aufaure, M.A. and Grand, B.L. "Advances in FCA-based applications for social networks analysis", *J. of Conceptual Structures and Smart Applications*, **1**(2), pp. 73-89 (2013).

11. Esuli, A. and Sebastiani, F. "SENTIWORDNET: A publicly available lexical resource for opinion mining", *5th Conf. on Language Resources and Evaluation*, Genoa, Italy, pp. 417-422 (2006).

12. Ohana, B. and Tierney, B. "Sentiment classification of reviews using SentiWordNet", *9th IT & T Conf.*, Dublin, Ireland (2009).

13. Andreevskaia, A. and Bergler, S. "Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses", *EACL-06, 11rd Conf. of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, pp. 209-216 (2006).

14. Esuli, A. and Sebastiani, F. "Determining the semantic orientation of terms through gloss classification", *CIKM-05, 14th ACM Int. Conf. on Information and Knowledge Management*, Bremen, Germany, pp. 617-624 (2005).

15. Esuli, A. and Sebastiani, F. "Determining term subjectivity and term orientation for opinion mining", *EACL-06, 11th Conf. of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, pp. 193-200 (2006).

16. Turney, P.D. and Littman, M.L. "Measuring praise and criticism: Inference of semantic orientation fromassociation", *ACM Transactions on Information Systems*, **21**(4), pp. 315-346 (2003).

17. Gnuplot, "Portable command-line driven graphing utility", http://www.gnuplot.info/ (2015).

18. Seol, K.S., Kim, J.D. and Baik, D.K. "Common neighbor similarity-based approach to support intimacy measurement in social networks", *Journal of Information Science*, pp. 1-10 (2015).

Biographies

Jeong-Dong Kim received his MS degree in Computer Science and a PhD degree in Computer Engineering, in 2008 and 2012, respectively, from Korea University, Seoul, Korea, where he is currently Research Professor in the Department of Computer and Radio Communications Engineering. His research interests include learning management systems, metadata-based integration, semantic Web, ontology, social computing, and access control.

Jiseong Son received her BS degree in Computer Engineering from Seoul Women's University, Seoul, Korea, in 2007, and an MS degree in Computer Engineering, in 2009, from Korea University, Seoul, Korea, where she is currently a PhD student in the Department of Computer and Radio Communications Engineering. Her research interests include semantic web data management, relational databases, query translation, and ontology engineering, e-Learning, learning management systems, and access control.

Hoh Peter In received his PhD degree in Computer Science from the University of Southern California (USC), and is currently Professor in the Department of Software Technology and Enterprise at Korea University, Seoul. He was also Assistant Professor at Texas A&M University and earned the most influential paper award for 10 years in ICRE 2006. His primary research interests include WBAN, embedded software engineering, social media platform and services, and software security management. He has published more than 100 research papers.

Suk-Hyung Hwang received a BS degree in Computer Science from Kangwon National University, Korea, in 1991, and ME and PhD degrees in Information and Computer Science from Osaka University, Japan, in 1994 and 1997, respectively. He is currently Professor in the Department of Computer Science and Engineering, Sun Moon University, Korea. His research interests include object-oriented analysis, design and programming, ontology, semantic web, and formal concept analysis, etc.

Hyun Lee received BE and MS degrees from Sun Moon University, Asan, Korea, in 1998 and 2002, respectively, and his PhD degree in the Department of Computer Science and Engineering at the University of Texas at Arlington, USA. He is currently Assistant Professor in the Department of Computer Science and Engineering at SunMoon University, Korea. His research interests include issues related to sensor fusion techniques and the integration of heterogeneous sensors and RFID systems for improving the reliability of contextual information in pervasive computing areas.

Doo-Kwon Baik received his BS degree in Mathematics from Korea University, Seoul, Korea, in 1974, and his MS and PhD degrees in Computer Science from Wayne State University, USA, in 1983 and 1986, respectively. He is currently Full Professor in the Department of Computer and Radio Communications Engineering, Korea University, Seoul, Korea. His research interests include modeling, simulation, data, and software engineering.