Neural Comput & Applic (2014) 25:1603–1614
DOI 10.1007/s00521-014-1642-9

ORIGINAL ARTICLE

Study of collective user behaviour in Twitter: a fuzzy approach

Xin Fu · Yun Shen

Received: 15 October 2013 / Accepted: 24 May 2014 / Published online: 28 June 2014

Abstract The study of collective user behaviours in social networking sites has become an increasing important topic in social media mining. Understanding such behaviours has its potential to extract actionable patterns that can be beneficial to develop effective marketing strategies, optimise user experiences and maximise website revenues. With the rapid development of micro-blogging, Twitter has become a richer source of intelligence that can be used to study collective user behaviour, due to its efficient and meaningful user-to-user interactions. However, the classical statistical methods have some drawbacks in bridging the gap between user-generated data and human analysts who mostly use linguistic terms to analyse data and model/summarise knowledge learned. To address this gap, this work proposes a new approach, which employs the mass assignment theory-based fuzzy association rules algorithm (MASS-FARM), for the first time, to extract useful interaction behaviour of Twitter users. The influential factors (including activity time, number of friends/followers and the number of tweets) are represented as fuzzy granules, and the associations amongst are studied by employing MASS-FARM. The collective user behaviours are analysed in the *Reply category* and the *Non-Reply category*, respectively. The applicability and usefulness of the proposed method are demonstrated via an empirical study on a collected Twitter data set. The derived results are also discussed and compared with existing works.

Keywords Collective user behaviour · Twitter data analysis · Fuzzy association rule mining · Mass assignment

1 Introduction

The development of social networking sites (SNSs) is one of the key phrases when next generation of Internet is concerned. Web 2.0 provides a platform where new SNSs have emerged to facilitate people to build up virtual relationships online. Due to its convenience and easiness to use, users just need to set up their profiles, and these SNSs would facilitate their further interactions with other users. Such SNSs have become a rich source of user-generated data, and they are attracting growing interests in various research domains, including sentiment analysis, community detection, social recommendation and online user behaviour analysis. Mining SNSs' data to extract actionable patterns would be beneficial for better understanding new phenomena, providing better customer services, maximising profits and developing innovative marketing strategies. For example, such task can help to identify the influential users in SNSs, detect implicit user groups/communities and recommend products and friends based on users' SNSs' records.

Amongst the mainstream SNSs, the popularity of Twitter grows exponentially. Launched in 2006, Twitter has created a new social phenomenon and has accrued more than 500 millions of registered users as of 2012, posting 340 million tweets per day. Twitter mixes conventional blogging features with social networking features. Users (tweeters) in Twitter can post (tweeting)

X. Fu
Department of Management Sciences, School of Management,
Xiamen University, Xiamen 361005, China
e-mail: xfu@xmu.edu.cn

Y. Shen (✉)
Symantec Research Labs, Ballycoolin Business Park, Dublin 15,
Ireland
e-mail: yun_shen@symantec.com

messages (tweets) with up to 140 characters and can reply messages (replying messages starting with the '@' sign). Tweeters can follow other people (friends) and/or be followed by other users (followers). Once a user adds another user as a friend, his/her friend's tweets will be automatically linked into this user's home page. This feature enables users to have more efficient exchanges in Twitter than traditional blogging sites or SNSs.

Twitter has a very broad range of users—such as politicians, reporters, businessmen and students. This Web-based micro-blogging service has become a reliable platform of expressing their thoughts and opinions, demonstrating ideas, sharing social activities, (re)broadcasting breaking news, promoting products, etc. Today, research about large-scale human behaviour patterns is often based on user-generated data [17]. Accordingly, from the research point of view, the publicly available tweets are turning into a richer source of intelligence that can be used to study online user behaviour than traditional blogging sites (e.g. Blogspot) and social sites (e.g. Facebook), because of its meaningful and massive user-to-user exchanges.

Based on the Twitter platform, a wide variety of research issues in mining Twitter data have been investigated. The representatives include sentiment analysis and opinion mining [20, 22, 24], community detection [29], social recommendation [4, 23] and user profiling [18]. However, the study of collective user behaviour in Twitter is still at the very early stage. Collective behaviour refers to how users behave when they are exposed in a social network environment [31]. Within a social network, behaviours of individuals tend to be interdependent, and they can be influenced by the behaviours of other users, especially his/her friends. For example, Backstrom et al. [3] find that the more friends a user has in a community, the more likely he/she is to join, share and communicate. Hence, collective behaviour is not simply the aggregation of individuals' behaviour. The study of collective user behaviour in Twitter is driven by some challenges: (1) How can a user be heard? Everyone can speak, but users only listen to a few. (2) How can a user attract more friends/followers? (3) How can a user's tweets attract more replies? Answers to these questions are hidden in the collective user behaviour mining. The study of collective behaviour gives the opportunity to uncover, understand and predict behaviours of users in SNSs. This would help advertisers to find the influential people to maximise the reach of their products, assist sociologists to study in-group and out-group behaviours of users.

Existing work has used statistical methods [6, 19, 26, 30] to study collective user behaviours in Twitter. However, it is hard to extract useful user behaviour knowledge from various parameters in these statistical models. Since user behaviours are naturally described by linguistic terms, such statistical models can hardly bridge the gap between data and human analysts who mostly use linguistic terms to analyse data and model/summarise knowledge learned. Consider a concept "large number of users…" as an example, an analyst may use subjective criteria, depending on certain SNS being analysed, to define properties of such concept accordingly.

Fuzzy set theory [35], particularly Zadeh's paradigm of computing with words [38], was especially developed for the task of representing human linguistic concepts in terms of a mathematical object, a fuzzy subset [14, 15]. However, limited attention was paid to apply fuzzy theory to social networking analysis (SNA). A computationally intelligent and scalable method is still desirable to automatically analyse social network data, particularly for large-scale data sets. The motivation of this work is to use data mining techniques, especially fuzzy association rules mining algorithm, to extract useful collective user behaviour knowledge from a large collection of public Twitter data set.

This paper proposes a new approach which employs an extended fuzzy association rules algorithm based on mass assignment theory (MASS-FARM), for the first time, to automatically extract useful knowledge of Twitter users' collective behaviour. From the interpretation point of view, the extracted knowledge is represented by linguistic terms, this ensures its transparency and it can be easily understood by non-expert users. Besides this, from the computation point of view, the proposed approach is quite efficient in handling large-scale data sets, because the MASS-FARM algorithm scales linearly with respect to the size of the given data set. Particularly, the approach itself is generic, it can be readily expanded to explore the collective user behaviours in other SNSs. The applicability and utility of the proposed approach are demonstrated and validated by using a collected Twitter data set. The correlations amongst different influential factors (e.g. the activity time, the number of posted tweets, the number of friends/followers) are investigated.

The rest of the paper is organised as follows. Section 2 introduces the related work in Twitter data analysis. Section 3 presents a fuzzy association rules mining algorithm based on mass assignment theory (MASS-FARM). Section 4 demonstrates that it is hard to extract useful knowledge from statistical model and also conducts a detailed study of user behaviour in Twitter by applying the proposed MASS-FARM algorithm. Section 5 concludes the paper and points out future works.

2 Related works

The Twitter data analysis has attracted growing attentions in recent years, and a wide range of features and aspects for analysing Twitter data has been researched with promising results. The representatives loosely relating to the collective user behaviour analysis include sentiment analysis and opinion mining, community detection, social recommendation and user profiling. Sentiment analysis and opinion mining aim to automatically extract opinions expressed in the user-generated content, and it allows businesses to understand product sentiments, brand perception, new product perception and reputation management. Kontopoulos et al. [24] propose an original ontology-based techniques towards a more efficient sentiment analysis of Twitter posts. Each post receives a sentiment grade for each distinct notion and results in a detailed analysis of post opinions regarding a specific topic. Khan [22] presents an algorithm for tweets classification based on a hybrid approach. Compared to similar classification techniques, the classification accuracy is greatly improved in [22]. For community detection, users in the same community tend to interact with each other more frequently than with those outside the community. A community can be observed via connections in SNSs, but the main challenge is that the SNSs are highly dynamic, and thus, the communities can shrink or dissolve in such dynamic environment. For example, Newman [29] employs a modularity matrix to detect community structure in networks. For social recommendation, different from traditional recommendation systems which recommend items based on aggregated rating of products from users' historical purchases, social recommendation also can make use of buyer's SNSs-related information and social influence from his/her friends within SNSs in the recommendation. Examples of such social recommendation systems include book recommendations based on SNS friends' reading lists on Amazon and friend recommendations on Twitter and Facebook [4, 23]. For user profiling, this technique employs diverse dimensions to cluster Twitter users, and it is particularly helpful for marketing segmentation. For instance, Ikeda et al. [18] propose a hybrid text-based and community-based method for the demographic estimation and then profiling the Twitter users.

In comparison with the above research, the study of collective user behaviour in Twitter is, due to its recent emerge, less scientifically explored. Since vast amounts of user-generated content are created on Twitter every day, this trend is likely to exponentially continue, in turn, providing rich data source for collective user behaviour analysis. With respect to the analysis of users' behaviour on Twitter, some interesting researches have been conducted. Bollen et al. [20] find out that measurements of collective mood posses more practical implications, as it could be correlated with the value of the Dow Jones Industrial Average (DJIA) overtime. Dodds et al. [12] propose a new metric to uncover and explain temporal variations in happiness and information level over different timescales, by examining expressions that user made on the Twitter. Moreover, Golder and Macy point out in [16] that the mood of Twitter users has diurnal and seasonal patterns which could be correlated with work, sleep and day length.

Existing research regarding the collective user behaviour analysis is all framed by either proposing statistical metric to measure user's mood or employing statistical methods to model user behaviour. However, in practice, domain experts often use linguistic terms to describe user behaviours. Some existing research has employed fuzzy techniques in SNA. Yager [33] argues that there is a natural connection between graph theory and granular computing, and theoretically discussed several concepts associated with SNA by modelling social networks as fuzzy graphs [9]. In addition, Zhang et al. [39] and Krajci et al. [25] consider clustering techniques; Zhang et al. [39] use fuzzy *c*-means to identify overlapping community structure, and the work of [25] uses one-sided fuzzy concept lattice and especially modified Rice-Siff's algorithm to form user cluster. Adjacency relation has been discussed in [8], which applies fuzzy logic and OWA operators [32] to interpret fuzzy *m*-ary adjacency relations in SNA. Interestingly, it is a natal fit to use association rules to model the interdependent user behaviour in the social networking sites. To the best knowledge of the authors, currently there is no research efforts employing fuzzy association rule mining to automatically extract meaningful collective user behaviour yet, which is the motivation and the focus of this paper.

3 Mass assignment-based fuzzy association rules mining

3.1 Association rule mining

Association rule mining [2] was initially proposed to mine rules from transaction databases. The aim of association rule mining is to determine whether or not there are links between two disjoint subsets of items. For example, do customers generally buy biscuits and cheese when beer and wine are bought?

Notationally, let $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ be a set of items and $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions. Note that any transaction can be represented as $t_i \subseteq \mathcal{I}$. Let x and y be two non-overlapping subsets of $\mathcal{I}$, $\mathcal{X}$ and $\mathcal{Y}$ be two sets of transactions containing items x and y, respectively. The goal of association analysis is to search for interesting associations or correlations between $\mathcal{X}$ and $\mathcal{Y}$. That is, if the items in x appear in a transaction, it is likely that the items in y will also appear (i.e. it is not an implication in the formal logical sense).

There are two useful measures for the association analysis: *support* and *confidence*. The *support* of a rule is the fraction of transactions in which both $\mathcal{X}$ and $\mathcal{Y}$ appear, and the *confidence* of a rule is an estimate (based on the samples) of the conditional probability of $\mathcal{Y}$ given $\mathcal{X}$. The mathematical definitions are given below:

$$\mathrm{Supp}(\mathcal{X},\mathcal{Y}) = \frac{|\mathcal{X}\cap\mathcal{Y}|}{|\mathcal{I}|}, \tag{1}$$

$$\mathrm{Conf}(\mathcal{X},\mathcal{Y}) = \frac{|\mathcal{X}\cap\mathcal{Y}|}{|\mathcal{X}|}. \tag{2}$$

Typically, association rule mining algorithms discover all item associations (or rules) in the data set that satisfy the user-specified *minimum support (minsup)* and *minimum confidence (minconf)* constraints. *Minsup* controls the minimum number of data cases that a rule must cover. *Minconf* controls the predictive strength of the rule. A association rule is *correct* if it satisfies these two thresholds.

When analysing collective user behaviour within the context of social network research, it is natural and desirable to use linguistic labels to describe various factors in order to conduct a more insightful study to answer questions such as “whether or not users spending a long time in the Twitter are more likely to tweet, if so, is this behaviour associated with the number of friends or is it associated with the number of user’s followers?” in a more human understandable way. In this paper, it is argued that user behaviour should be better analysed from a subjective perspective and introduces a granular association analysis algorithm to intelligently extract user behaviour knowledge in a human-centric way to answer above questions. The rest of this section details the theoretical foundation of a mass assignment theory-based fuzzy association rules mining algorithm (MASS-FARM).

3.2 MASS-FARM

Fuzzy association analysis has been studied in recent years [7, 11, 13, 21]. In creating association rules within transaction databases (e.g. [1, 2], see also [13] for a clear overview), the crisp approach is to consider a table in which columns correspond to items and each row is a transaction. A column contains 1 if the item was bought, and 0 otherwise. In fuzzy association analysis, the (0,1) value of an item in a transaction is extended to a membership degree ranging within $[0, 1]$. The standard extension to the fuzzy case is to treat the (multi-) sets $\mathcal{X}$, $\mathcal{Y}$ as fuzzy set and find the intersection and cardinality using a *t*-norm and the *sigma* count, respectively. The *confidence* which represented in Equation (2) is transformed into:

$$\mathrm{Conf}(\mathcal{X},\mathcal{Y}) = \frac{\sum_{\mathrm{i}\in\mathcal{I}}\mu_{\mathcal{X}\cap\mathcal{Y}}(\mathrm{i})}{\sum_{\mathrm{i}\in\mathcal{I}}\mu_{\mathcal{X}}(\mathrm{i})} \tag{3}$$

and the *support* is defined in Eq. (4) as,

$$\mathrm{Supp}(\mathcal{X},\mathcal{Y}) = \frac{\sum_{\mathrm{i}\in\mathcal{I}}\mu_{\mathcal{X}\cap\mathcal{Y}}(\mathrm{i})}{|\mathcal{I}|} \tag{4}$$

where μ refers to the monadic fuzzy membership function. As pointed out by [13], using *min* and *sigma* count for cardinality can be unsatisfactory, because it does not distinguish between several tuples with low memberships and few tuples with high memberships. Using a fuzzy cardinality (i.e. a fuzzy set over the possible cardinality values) is also potentially problematic. Since the result is a possibility distribution over rational numbers, the extension principle [36] gives a wider bound than it should, due to neglect of interactions between the numerator and denominator in this expression.

Zadeh [37] argues that information can be processed on different levels of abstraction in which data objects are organised into meaningful granules[1] so as to convey a perception of information itself. This perception is in line with the “modelling with words” approach. Thus, it is necessary to study the criteria for deciding how “a clump of points (objects) are drawn together by indistinguishability, similarity, proximity or functionality [37]”. Interestingly, when working with large data sets, a natural approach is to group similar items/transactions into categories (or sets) and summarise the data in term of such categories. This is a key feature of human intelligence to discover knowledge. Granules provide a convenient way to describe a finite set of objects by specifying certain selection criteria. For example, instead of looking for an association between *potato crisps* and *beer*, it would make sense to find an association between *snack foods* and *alcoholic drinks*.

Once the granules are constructed, one may be interested in developing computational methods to study the relationships among these granules. It is important to note that this work deals with conjunctive fuzzy sets (monadic fuzzy relations). Mass assignment theory is normally applied to fuzzy sets representing possibility distributions, and the operation of finding the conditional probability of one fuzzy set given another is known as semantic unification [5]. This rests on the underlying assumption of a single-valued attribute. A different approach is required to find the conditional probability when dealing with set-valued attributes, and is discussed in the rest of this section.

[1] A fuzzy subset of the universe corresponds to a *granule* in this paper.

Table 1 Activity time versus the number of Tweets

ID	Activity time (hour)	No of Tweets
id_1	800	100
id_2	640	40
id_3	500	80
id_4	160	70

Take two granules (described by linguistic terms) into consideration, represented as monadic fuzzy relations $\mathcal{X}$ and $\mathcal{Y}$ on the same domain, and to calculate the degree of association between them. For example, consider a sample Twitter data as shown in Table 1, containing activity time and the number of tweets. The transactions can be categorised according to whether their activity time are *long*, *medium* or *short* and also according to whether their tweets figures are *high*, *moderate* or *low*. A mining task might be finding out whether the "*high tweet figure*" is associated with "*long activity time*".

It is important to note that the conjunctive fuzzy sets (monadic fuzzy relations) are dealt with herein. A relation is a conjunctive set of ordered *n*-tuples, i.e. it represents a conjunction of *n* ground clauses. For example, if *U* is the set of dice scores, then a predicate *differBy*4*or*5 can be defined on $U \times U$ as the set of pairs

$$[(1,6),(1,5),(2,6),(5,1),(6,1),(6,2)].$$

This is a conjunctive set in that each pair satisfies the predicate. By extending the classical relation, a fuzzy relation represents a set of *n*-tuples that satisfy a predicate to a specified degree. Thus, *differByLargeAmount* could be represented by

$$[(1,6)/1,(1,5)/0.6,(2,6)/0.6,(5,1)/0.6,(6,1)/1,(6,2)/0.6].$$

A mass assignment [5] is a way of representing uncertainty, which encompasses both probabilistic and fuzzy uncertainty. Probabilistic uncertainty is related to whether or not a specified event occurs, e.g. the score on a dice is equal to 2. Fuzzy uncertainty is connected with events and propositions for which no precise definition exists, e.g. "2 is a low score". Subject to certain conditions, mass assignment allows to handle both forms of uncertainty within a single framework and is computationally easy to implement.

Mass assignment theory is normally applied to fuzzy sets representing possibility distributions. The operation of finding the conditional probability of one fuzzy sets given another is known as semantic unification [5]. This rests on the underlying assumption of a single-valued attribute. A different approach is required to find the conditional probability when dealing with set-valued attributes. In the rest of the section, this work introduces an extension of the theory to computing the association between fuzzy categories that has been proposed in [27].

Considering Table 1, given the source granule as *long activity time* ($\mathcal{X}$) and the target granule as *high tweet figure*, the monadic fuzzy relations of $\mathcal{X}$ and $\mathcal{Y}$ can be defined as:

$$\mathcal{X} = [id_1/1, id_2/0.8, id_3/0.5, id_4/0.2],$$

and

$$\mathcal{Y} = [id_1/1, id_2/0.4, id_3/0.8, id_4/0.7].$$

It is important to note that these sets of values that all satisfy the related granule to a degree. Unless explicitly mentioned, all the sets in the rest of the section should be interpreted as conjunctive relations. The confidence in an association rule in such setting can be calculated as follows.

Step 1: Calculate the corresponding mass assignments of the source granule and the target granule.

Given a source granule:

$$\mathcal{X} = [x_1/\mu_X(x_1), x_2/\mu_X(x_2), \ldots, x_{|\mathcal{X}|}/\mu_X(x_{|\mathcal{X}|})],$$

and assume the set of distinct memberships in $\mathcal{X}$ be $\{\mu_X^{(1)}, \mu_X^{(2)}, \ldots, \mu_X^{(nx)}\}$, where $\mu_X^{(1)} > \mu_X^{(2)} > \ldots > \mu_X^{(nx)}$ and $n_X \leq |\mathcal{X}|$. Let $\mathcal{X}_1 = \{[x|\mu_X(x) = \mu_X^{(1)}]\}$ and $\mathcal{X}_i = \{[x|\mu_X(x) \geq \mu_X^{(i)}]\}$, $1 < i \leq n_X$, then the mass assignment corresponding to $\mathcal{X}$ is:

$$\{\mathcal{X}_i : m_X(\mathcal{X}_i)\}, \quad 1 \leq i \leq n_X,$$

where $m_X(\mathcal{X}_k) = \mu_X^{(k)} - \mu_X^{(k+1)} \quad (\mu_X^{(i)} = 0 \, if \, i > n_X)$.

For example, given source granule $\mathcal{X} = [id_1/1, id_2/0.8, id_3/0.5, id_4/0.2]$, it has the corresponding mass assignment:

$$M_X = \{[id_1] : 0.2, [id_1, id_2] : 0.3, [id_1, id_2, id_3] : 0.3, [id_1, id_2, id_3, id_4] : 0.2\}.$$

The corresponding mass assignment of the target granule can be obtained in the same manner.

Step 2: Calculate the *confidence* in the association between the granules $\mathcal{X}$ and $\mathcal{Y}$ using mass assignment theory

In general, this will be an interval as it is free to move mass (consistently) between elements of each $\mathcal{X}_i$ and $\mathcal{Y}_j$.

For two mass assignments,

$$M_X = \{\mathcal{X}_i : m_X(\mathcal{X}_i)\}, \quad 1 \leq i \leq n_X$$

and

$$M_Y = \{\mathcal{Y}_j : m_Y(\mathcal{Y}_i)\}, \quad 1 \leq j \leq n_Y.$$

The composite mass assignment is defined as:

$$M_C = M_X \bigoplus M_Y = \{S : m_C(S)\}, \tag{5}$$

where m_C is the composite mass allocation function subject to:

Table 2 Mass table

	$[id_1]$:0.2	$[id_1,id_3]$:0.1	$[id_1,id_3,id_4]$:0.3	$[id_1,id_3,id_4,id_2]$:0.4
$[id_1]$:0.2	id_1	id_1	id_1	id_1
$[id_1,id_3]$:0.1	id_1	id_1	id_1	id_1 id_3
$[id_1,id_3,id_4]$:0.3	id_1	id_1 id_3	id_1 id_3	id_1 id_3 id_2
$[id_1,id_3,id_4,id_2]$:0.4	id_1	id_1 id_3	id_1 id_3 id_2	id_1 id_3 id_4 id_2

$$\sum_{j=1}^{n_Y} M_C(ij) = m_X(\mathcal{X}_i) \text{ and } \sum_{i=1}^{n_X} M_C(ij) = m_Y(\mathcal{Y}_i).$$

It derives

$$\mathrm{Conf}(\mathcal{X},\mathcal{Y}) = \frac{\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} M_C(ij) \times |\mathcal{X}_i \cap \mathcal{Y}_j|}{\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y} M_C(ij) \times |\mathcal{X}_i|}. \tag{6}$$

When ranking association rules, it is preferable to have a single figure for confidence, rather than an interval, which can lead to ambiguity in the ordering.

In the given example, the mass table is derived in Table 2. According to Table 2 and Equations (6), the confidence value in the association between the granules $\mathcal{X}$ and $\mathcal{Y}$ is calculated as:

$$\begin{aligned}\mathrm{Conf}(S,T) = {} & (0.2 \times (0.2 + 0.1 + 0.3 + 0.4) \times 1 \\ & + 0.3 \times (0.2 + 0.1 + 0.3) \times 1 + 0.3 \times 0.4 \times 2 \\ & + 0.3 \times 0.2 \times 1 + 0.3 \times (0.1 + 0.3) \times 2 \\ & + 0.3 \times 0.4 \times 3 \\ & + 0.2 \times 0.2 \times 1 + 0.2 \times 0.1 \times 2 \\ & + 0.2 \times 0.3 \times 3 + 0.2 \times 0.4 \times 4)/2.5 \\ = {} & 0.744\end{aligned}$$

The result can easily be interpreted as “Given long activity time, the *confidence* of a user having high twee figure is 0.744”. Note that the nested structure of mass assignment enables to calculate association confidences with roughly $O(n)$ complexity, rather than $\mathrm{O}(n^4)$ where n is the number of focal elements in the source $\mathcal{X}$.

4 Tweeting behaviour study

In this section, a detailed empirical study of Twitter user behaviour is provided based upon the MASS-FARM method discussed in Sect. 3.

4.1 Twitter data set

Twitter provides a set of developer APIs for accessing publicly available data (if user’s profile is made public, user’s tweets and other information are classified as publicly available). The data set used in this study was collected by monitoring Twitter public timeline for a period of 5 weeks, which starting from Oct 7, 2009 to November 13, 2009. A set of 20 recent tweets and their associated user information, such as friends, followers, geo-location, timestamp, the number of tweets, were fetched from Twitter every 8–26 seconds. In total, there are 1,242,522 posts from 1,052,739 distinct users included in the data set.

There are two categories of tweets in the data set:

- *Reply category*: tweets start with ‘@’ following with a user name, i.e. a user (*User*) replies to another user (*InReplyToUser*) in Tweeter. There are 381,955 (out of 1,242,522) tweets in the *Reply category* representing 30.74 % of overall posts.
- *Non-Reply category*: posts without an ‘@’, i.e. posts as they are. There are 860,567 (out of 1,242,522) tweets in the *Non-Reply category* representing 69.26 % of overall posts.

4.2 Limitations of statistical model

Complementary cumulative distribution function (CCDF) is used for an example to demonstrate the limitation of statistical models. CCDF describes the probability distribution of a real-valued random variable X that is above a particular level, and is normally employed to describe the main features of a collection of data in quantitative terms and is usually employed to study Twitter user behaviour.

- **Example 1—Tweets Analysis**. Figure 1 shows CCDF that User & InReplyToUser in the *Reply category* publish more than a certain number of tweets (versus User in the *Non-Reply category*). Marginal distribution of User in the *Non-Reply category* is a Pareto with $k = 0.6554$ while User (respectively InReplyToUser) in *Reply category* has Pareto distribution with $k = 0.5251$ (respectively $k = 0.4841$). It is difficult to obtain further useful knowledge from parameter k in Pareto distribution.
- **Example 2—Friends/Follower Analysis in the *Reply category***. Figure 2 shows CCDF that User & InReplyToUser in the *Reply category* has more than a certain number of friends/followers. As evident from Fig. 2, *InReplyToUser* is likely to have more followers than *User*, while *InReplyToUser* is having similar number of friends to *User*. It is still hard for human analysts to obtain further information from this statistical analysis.

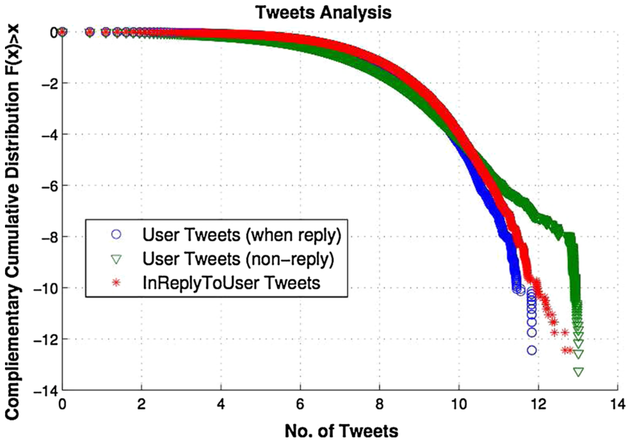

Fig. 1 Tweets analysis: Reply category versus Non-Reply category

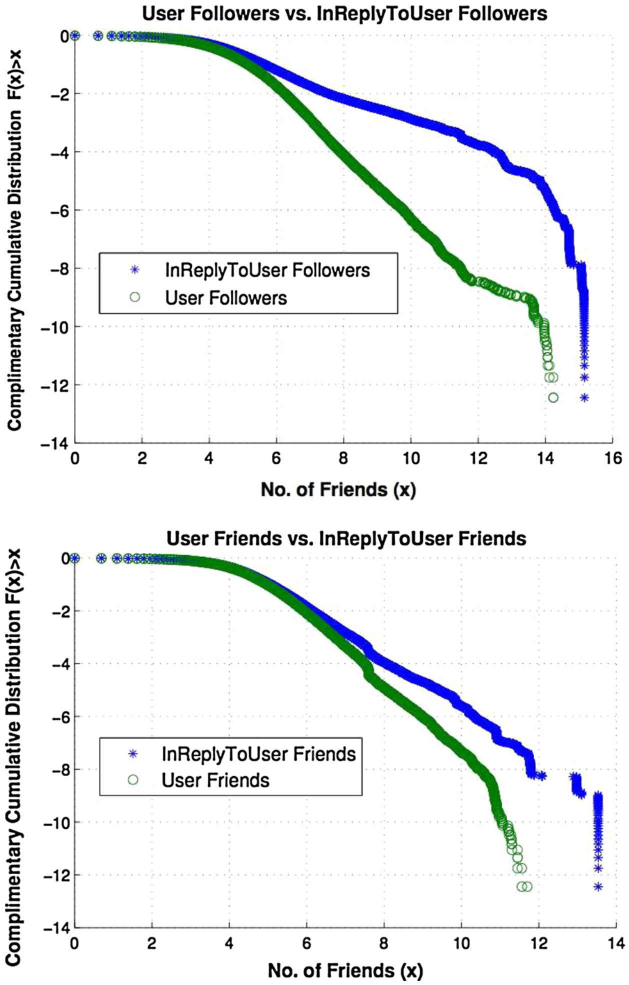

Fig. 2 Friends and follower analysis: User versus InReplyToUser

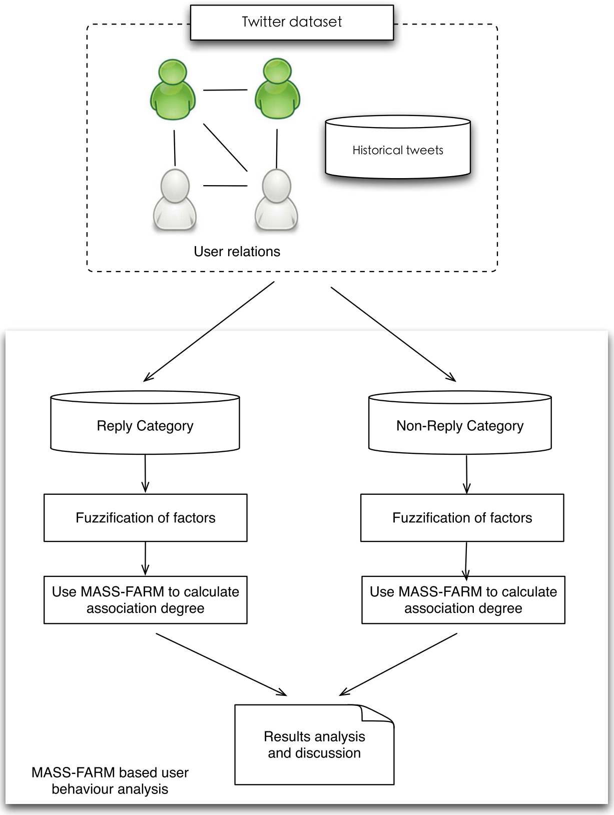
Twitter dataset
Historical tweets
User relations
Reply Category
Non-Reply Category
Fuzzification of factors
Fuzzification of factors
Use MASS-FARM to calculate association degree
Use MASS-FARM to calculate association degree
Results analysis and discussion
MASS-FARM based user behaviour analysis

Fig. 3 Overview of the proposed approach

4.3 Collective user behaviour analysis

As stated in Sect. 3, the goal of the association analysis is to search for interesting associations or correlations between $\mathcal{X}$ and $\mathcal{Y}$, by taking a “computing by words” approach to allow natural language concept to be utilised to qualify such relationship. The main processes of the proposed approach are outlined in Fig. 3.

4.3.1 Collective user behaviour in the Reply category

The interaction behaviour can be monitored via the replied posts. In other words, if user A replies user B’s tweet, or vice versa, then there is an interaction behaviour between these two users. This section aims at studying user interaction behaviour in order to reveal interesting associations between various user-specific factors, including activity time, the number of tweets and the number of friends/follower.

- Firstly, the *activity time* factor associated with user interaction behaviour is examined. The justification using activity time to qualify user’s experience is that "users must be active in Twitter in order to post reply tweets and the longer he/she uses Twitter, the more skilful the user is". Hence, the users are categorised as *novice*, *intermittent*, *frequent* and *loyal* in this study. The corresponding membership functions are shown in Fig. 4. By applying the MASS-FARM method to the given data set, the obtained results are shown in Table 3. The results reveal that in terms of the user behaviour of replying tweets, there is no much difference amongst the four user categories, as the association values in each column are quite similar. It is interesting to observe that all Twitter users are more likely to reply tweets posted by well-established *loyal* users, because relatively strong association values are presented at Column 5 in Table 3. This result would be appealing to Internet advertisement/campaign companies to spread "strong word of mouth" for certain products/candidates by targeting at these *loyal* Twitter users. In contrast, the tweets posted by *novice* users attract less replies from all types of Twitter users (as shown at the Column 2 in Table 3). One possible reason is that the *novice* users are inexperienced in posting tweets, which may attract impassioned discussions.

- Secondly, the *number of tweets* factor associated with user interaction behaviour is studied. The number of users' tweets is defined as *small*, *medium* and *large* in this work. The corresponding membership functions are shown in Fig. 5. Similar to the *activity time* factor, there is no much difference amongst these three user categories in their interaction behaviour. The obtained results are shown in Table 4. It is surprisingly to notice that all users are likely to reply to users with *small* number of tweets. Since there are strong association values shown at Column 2 in Table 4. This is quite different from the common expectation. In general, the more tweets a user posted, the more experiences he/she has to attract replies. Meanwhile, users with the *medium* number of tweets receive least attentions. One possible explanation is that the Twitter accounts which post a *large* number of tweets could be broadcasting user accounts. The main aim of such account is to spread breaking news and production promotions. As a result, users may just prefer to receive information rather than reply it. In addition, to achieve better broadcasting and promotion effects, the same content of tweets might be repeatedly posted. This result would also be interesting to automatic natural language parsing research [28] to further analyse tweets semantically and understand why the *small* number of tweets can trigger replies. Also, a closer examination will be taken into the user accounts with *large* number of tweets, and see how many of them are broadcasting accounts, so as to better explain and validate the derived results.

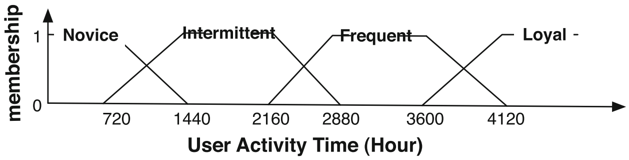

Fig. 4 Membership function: user activity time (Hour)

- Thirdly, the *number of friend/follower* factor that behind the user interaction in Twitter is analysed herein. For example, it would be desirable to investigate whether user *A* replies *B*, because there is strong association between the number of friends/followers that A has and the number of friends/followers that B has. It employs $small(s)$, $medium(m)$, $large(l)$ and $verylarge(vl)$ to qualify the number of friends/followers that a user has. The corresponding membership functions are shown in Fig. 6. Note that, the data preprocessing indicates that data distributions of the number of friends and the number of followers are quite similar. Hence, this work employs the same membership functions to model these two factors. The obtained results are recorded in Table 6.

Different from the previous two factors, the interaction behaviours of four user segments are quite diverse in this case. Strong association values are presented at Column 2 & 6 in Table 5 demonstrating that users tend to reply to users with *small* number of friends/followers, but such associations decline with the increasing number of users' own friends/followers. Most interestingly, there are very weak associations at

Table 3 Collective user interaction behaviour—activity time factor (*Reply category:* 381,955 Tweets)

	irtu_Novice	irtu_Intermittent	irtu_Frequent	irtu_Loyal
u_Novice	0.12767	0.16958	0.16549	**0.53726**
u_Intermittent	0.1247	0.16834	0.168	**0.53896**
u_Frequent	0.12429	0.17055	0.16946	**0.5357**
u_Loyal	0.12175	0.16794	0.16496	**0.54536**

u User, *irtu* InReplyToUser

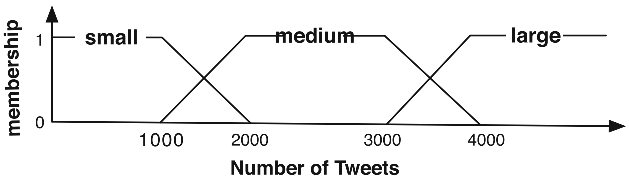

Fig. 5 Membership function: the number of user's Tweets

Table 4 Collective user interaction behaviour—the number of Tweets factor (*Reply category:* 381,955 Tweets)

	irtu_tweet(s)	irtu_tweet(m)	irtu_tweet(l)
u_tweet(s)	**0.53088**	0.2046	0.26452
u_tweet(m)	**0.52937**	0.20415	0.26648
u_tweet(l)	**0.52801**	0.20317	0.26882

u User, *irtu* InReplyToUser

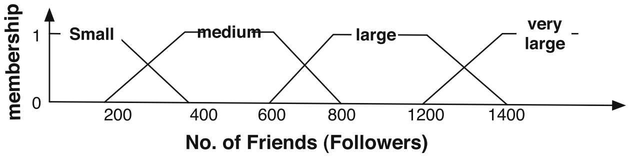

Fig. 6 Membership Function: the number of Friends (Followers)

Column 4 & 8 in Table 5. This shows that users seldom reply tweets posted users with *large* number of friends/followers. Researchers have noticed that spammers usually generate more data (including tweet replies) than legitimate users [10, 34]. One possible reason for the above observations is that users believe his/her reply would be more visible when the user has *small* number of friends/followers (includes fewer spammer friends/followers), so they are more willing to reply. In addition, similar to the previous case, the broadcasting user accounts tend to attract *large* number of followers, but such followers might prefer to only receive information rather than interact with the information provider. This well explains the derived results. Of course, in this study, the interpreted linguistic rules are heavily depend upon the pre-defined membership functions. Although the current membership assignments are based on statistical distributions of the given data set, further investigation at this point will be conducted in future work.

4.3.2 Collective user behaviour in the Non-Reply category

Twitter user behaviour in the *Non-Reply category* is investigated in this section. Since there is no interaction between users in this category, this section aims to explore the correlations between different influential factors. The factors (including *activity time*, *the number of friends/followers* and *the number of tweets*) and their associated membership functions (respectively shown in Figs. 4, 5 and 6) discussed above are reused herein.

Starting with the associations between the *activity time* and *the number of followers* factors, the obtained results are included in Table 6. The results show a significantly strong association at Column 2 in Table 6, but such associations decline with the increase of the activity time that user spends on Twitter. This indicates that regardless how long users have been using Twitter, they can only attract a *small* number of followers. In other words, the time spend on Twitter is not a driven factor to attract followers.

The above finding motivates to add *the number of tweets* factor and investigates that whether increasing number of tweets leads to an increase of the number of followers. In this case, two factors, the *activity time* and *the number of tweets*, are combined together to represent a joint factor. Then, an investigation on the association between the joint factor and *the number of followers* factor is conducted. The obtained results are shown in Table 7. Similar to the previous scenario, according to the Column 2 in Table 7, regardless the length of user's activity time and the number of posted tweets, only a *small* number of followers can be

Table 5 Collective user interaction behaviour—the number of friends/followers factor (*Reply category:* 381,955 Tweets)

	irtu_FOL(s)	irtu_FOL(m)	irtu_FOL(l)	irtu_FOL(vl)	irtu_FRI(s)	irtu_FRI(m)	irtu_FRI(l)	irtu_FRI(vl)
u_FOL(s)	**0.75457**	0.06512	**0.02795**	0.15235	**0.81356**	0.11282	**0.02681**	0.0468
u_FOL(m)	**0.62788**	0.14785	**0.06637**	0.15791	**0.61934**	0.21902	**0.06493**	0.09671
u_FOL(l)	**0.57398**	0.15168	**0.08113**	0.19321	**0.57181**	0.22761	**0.0782**	0.12238
u_FOL(vl)	**0.50585**	0.14121	**0.07507**	0.27787	**0.53147**	0.20226	**0.07924**	0.18702
u_FRI(s)	**0.75972**	0.06328	**0.02702**	0.14998	**0.82543**	0.1073	**0.02466**	0.04261
u_FRI(m)	**0.63359**	0.13114	**0.06018**	0.17509	**0.62967**	0.21474	**0.06242**	0.09318
u_FRI(l)	**0.57148**	0.1483	**0.07323**	0.20699	**0.54419**	0.23054	**0.08277**	0.14249
u_FRI(vl)	**0.48969**	0.14254	**0.07992**	0.28786	**0.45736**	0.20867	**0.09116**	0.24281

u User, *irtu* InReplyToUser, *FOL* followers, *FRI* friends attracted. Another interesting finding is that regardless the activity time, with the increase of the number of posted tweets, all types of users gradually increase the number of followers. In particular, the association values between the joint factor and the *medium/very large* number of followers jump substantially from the *medium* number of tweets to *l*arge number of tweets (as exemplified at Column 3 and Column 5 in Table 7). Moreover, regardless the length of user's activity time and the number of tweets, users failed to attract a *large* number of followers (as exemplified at Column 4 in Table 7). The underlying reason for this finding requires further investigations.

4.4 Discussion

The goal of the method is to search for interesting associations or correlations between $\mathcal{X}$ and $\mathcal{Y}$, by taking a "computing by words" approach to allow natural language concept to be utilised to qualify such relationship. Nevertheless, it requires sufficient evidence to support the association rules discovered in the above experiments. Equation (4) is employed to calculate the support values, and some results regarding the association rules between the *activity time* and *the number of followers* factors are reported in Table 8. Interestingly, Pearson correlation between these two facts is 0.04489. Normally, such result implies that there is no strong correlation between *activity time* and *the number of followers*. The proposed approach, however, identifies the patterns hidden away from the statistical approach and finds useful correlations with solid support values.

Table 6 Collective user behaviour—activity time versus no of follower factor (*Non-Reply category:* 860,567 Tweets)

	u_FOL(s)	u_FOL(m)	u_FOL(l)	u_FOL(vl)
u_Novice	**0.93575**	0.02785	0.01294	0.02346
u_Intermittent	**0.88734**	0.04452	0.02208	0.04606
u_Frequent	**0.88261**	0.04728	0.02	0.05011
u_Loyal	**0.79585**	0.08107	0.03569	0.08739

u Users, *FOL* followers

5 Conclusion

This paper has proposed a new approach which applies the mass assignment-based fuzzy association rules mining (MASS-FARM) algorithm to Twitter data analysis, for the first time, to automatically extract useful and meaningful knowledge from large-scale data set. The collective user behaviour has been analysed in the *ReplyCategory* and the *Non − Reply Category*, respectively. By applying the MASS-FARM algorithm, the association values amongst influencing factors (including *activity time*, *number of friends/followers* and *the number of tweets*) have been calculated.

Table 7 Collective user behaviour—activity time, no of Tweet versus no of follower factor (*Non-Reply category:* 860,567 Tweets)

	u_FOL(s)	u_FOL(m)	u_FOL(l)	u_FOL(vl)
Novice∧tweets(s)	0.96901	0.01576	0.006892	0.00833
Novice∧tweets(m)	0.79308	**0.07575**	0.03878	**0.09239**
Novice∧tweets(l)	0.51571	**0.18294**	0.08808	**0.21330**
Intermittent∧tweets(s)	0.95748	0.02022	0.00887	0.01342
Intermittent∧tweets(m)	0.82642	**0.06719**	0.04179	**0.06460**
Intermittent∧tweets(l)	0.55696	**0.15689**	0.07585	**0.21031**
Frequent∧tweets(s)	0.95374	0.02174	0.00929	0.01524
Frequent∧tweets(m)	0.83489	**0.06031**	0.02876	**0.07604**
Frequent∧tweets(l)	0.58815	**0.15906**	0.06348	**0.18931**
Loyal∧tweets(s)	0.92163	0.034101	0.01499	0.02928
Loyal∧tweets(m)	0.78305	**0.09095**	0.037178	**0.08882**
Loyal∧tweets(l)	0.51739	**0.18089**	0.08202	**0.21969**

u User, *FOL* followers

Table 8 Collective user behaviour—activity time versus no of follower factors (with support values)

	u_FOL(s)		u_FOL(m)		u_FOL(l)	
	Supp	Conf	Supp	Conf	Supp	Conf
u_Novice	**0.7397**	0.93575	0.0217	0.02785	0.0104	0.01294
u_Intermittent	**0.5976**	0.88734	0.0299	0.04452	0.0150	0.02208
u_Frequent	**0.5788**	0.88261	0.0310	0.04728	0.0132	0.02
u_Loyal	**0.7390**	0.79585	0.0736	0.08107	0.0335	0.03569

u Users, *FOL* followers

The *ReplyCategory* study mainly investigated the Twitter user interaction behaviours, and some interesting and insightful results have been derived. For example, regardless the length of activity time and the number of posted tweets, the interactive user behaviours of replying tweets are very similar. All Twitter users are likely to reply tweets posted by well-established *royal* users. In contrast, the tweets posted by *novice* users attract less replies. It is surprisingly to notice that all users are likely to reply to users with small number of tweets. The possible reasons for such observations have also been discussed. The *Non – Reply Category* study mainly concerns the correlations between different influential factors. The derived results reveal that the time spend on Twitter and the number of posted tweets play insignificant role in attracting Twitter followers. The finding of other influential factors (e.g. tweets context, posted timestamp, etc) require further investigations.

The proposed work is capable of extracting and representing collective Twitter user behaviours in linguistic terms. The effectiveness of the approach is compared against the traditional statistical models. Experimental results demonstrate that the statistical models have limited capability to bridge the gap between data and human linguistic terms. In contrast, by employing the proposed approach, the collective Twitter user behaviours can be intuitively represented. This ensures that the derived association rules are interpretable and explainable to non-expert users. This is essential in assisting the end-user to better understanding new phenomena, providing better customer services and developing more profitable marketing strategies based on the extracted linguistic association rules.

By extending this work, the promising results can be applied to further social network analysis, such as targeting the influential users in Twitter for advertisements. Also, it would be quite interesting to categorise Twitter users and study how different categorical users are influenced by different users/events. More specifically, the influential factors (i.e. activity time, the number of friends/followers and the number of tweets) employed in this work can be directly used as the segmentation attributes. Third, the temporal and spatial variations will be taken into account in future work. The correlations between the collective user behaviour and their diurnal and seasonal patterns (even cross social sites) will be explored.

Acknowledgments This work is supported by the National Nature Science Foundation of China (No. 71301133, 71371159) and Humanity and Social Science Youth foundation of Ministry of Education, China (Grant No. 13YJC630033). The authors are also grateful to the referees for their invaluable and insightful comments that have helped significantly to improve this work.

References

1. Agrawal R, Imieliński T, Swami A (1993) Mining association rules between sets of items in large databases. In: SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data, ACM, New York, pp 207–216
2. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of 20th international conference on very large Data Bases, VLDB
3. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: Membership, growth, and evolution. In: Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pp 44–54
4. Backstrom L, Leskovec J (2011) Supervised random walks: Predicting and recommending links in social networks. In: Proceedings of the fourth ACM international conference on Web Search and data mining, pp 635–644
5. Baldwin J, Lawry J, Martin TP (1996) Efficient algorithms for semantic unification. In: Information processing and the management of uncertainty
6. Benevenuto F, Rodrigues T, Cha M, Almeida VAF (2009) Characterizing user behavior in online social networks. In: Internet measurement conference, pp 49–62
7. Bosc P, Pivert O (2001) On some fuzzy extensions of association rules. In: IFSA World Congress and 20th NAFIPS international conference, 2001. Joint 9th, vol 2, pp 1104–1109
8. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8
9. Brunelli M, Fedrizzi M (2009) A fuzzy approach to social network analysis. In: ASONAM '09: Proceedings of the 2009 international conference on advances in social network analysis and mining. IEEE Computer Society, Washington, DC, pp 225–230
10. Cagman N, Citak F, Aktas H (2012) Soft int-group and its applications to group theory. Neural Comput Appl 21(1 Supplement):151–158
11. Chu Z, Gianvecchio S, Wang H, Jajodia S (2010) Who is tweeting on twitter: Human, bot, or cyborg? In: Proceedings of the 26th annual computer security applications conference, pp 21–30
12. Delgado M, Marn N, Snchez D, amparo Vila M (2003) Fuzzy association rules: general model and applications. IEEE Trans Fuzzy Syst 11:214–225
13. Dodds P, Harris K, Kloumann I, Bliss C, Danforth C (2011) Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter. PLoS ONE 6(12):e26752. doi:10.1371/journal.pone.0026752
14. Dubois D, Hllermeier E, Prade H (2006) A systematic approach to the assessment of fuzzy association rules. Data Min Knowl Disc 13(2):167–192
15. Fu X, Shen Q (2010) Fuzzy compositional modelling. IEEE Trans Fuzzy Syst 18(4):823–840
16. Fu X, Shen Q (2011) Fuzzy complex numbers and their application for classifiers performance evaluation. Pattern Recogn 44(7):1403–1417
17. Golder SA, Macy M (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333:1878–1881
18. Gundecha P, Liu H (2012) Mining social media: a brief introduction. In: Tutorials in operations research—new directions in informatics, optimization, logistics, and production (INFORMS), pp 1–17
19. Ikeda K, Hattori G, Ono C, Asoh H, Higashino T (2013) Twitter user profiling based on text and community mining for market analysis. Knowl Based Syst 51:35–47

20. Java A, Song X, Finin T, Tseng B (2007) Why we twitter: understanding microblogging usage and communities. In: WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, New York, NY, pp 56–65
21. Kacprzyk J, Zadrozny S (2003) Linguistic summarization of data sets using association rules. In: Fuzzy Systems, 2003. The 12th IEEE international conference on FUZZ '03, vol 1, pp 702–707
22. Khan FH, Bashir S, Qamar U (2013) Tom: Twitter opinion mining framework using hybrid classification scheme. Decision Support Systems (0)
23. Konstas I, Stathopoulos V, Jose JM (2009) On social networks and collaborative recommendation. In: Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval, pp 195–202
24. Kontopoulos E, Berberidis C, Dergiades T, Bassiliades N (2013) Ontology-based sentiment analysis of twitter posts. Expert Syst Appl 40(10):4065–4074
25. Krajci S, Krajciova J (2007) Social network and one-sided fuzzy concept lattices. In. In Proceedings of the 16th international conference on Fuzzy systems, pp 1–6
26. Krishnamurthy B, Gill P, Arlitt M (2008) A few chirps about twitter. In: WOSP '08: Proceedings of the first workshop on Online social networks. ACM, New York, NY, pp 19–24
27. Martin T, Shen Y, Majidian A (2010) Discovery of time-varying relations using fuzzy formal concept analysis and associations. J Intell Syst 25(12):1217–1248
28. Martin TP, Shen Y, Azvine B (2008) Incremental evolution of fuzzy grammar fragments to enhance instance matching and text mining. IEEE Trans Fuzzy Syst 16(6):1425–1438
29. Newman M (2006) Finding community structure in networks using the eigenvectors of matrices. Phys Rev E 74(3):1–22
30. Tang L, Liu H (2009) Scalable learning of collective behavior based on sparse social dimensions. In: CIKM '09: Proceeding of the 18th ACM conference on information and knowledge management. ACM, pp 1107–1116
31. Tang L, Liu H (2010) Towards predicting collective behaviour via social dimension extraction. IEEE Intell Syst 25:19–25
32. Yager RR (1988) On ordered weighted averaging aggregation operators in multicriteria decisionmaking. IEEE Trans Syst Man Cybern 18(1):183–190
33. Yager RR (2008) Intelligent social network analysis using granular computing. Int J Intell Syst 23(11):1196–1219
34. Yardi S, Romero D, Schoenebeck G, Boyd D (2009) Detecting spam in a twitter network. First Monday 15(1):1–4
35. Zadeh LA (1965) Fuzzy sets. Inform Contl 8:338–353
36. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning-III. Inform Sci 9(1):43–80
37. Zadeh LA (1997) Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets Syst 90(2):111–127
38. Zadeh LA (2000) From computing with numbers to computing with words—from manipulation of measurements to manipulation of perceptions. In: Intelligent systems and soft computing, pp 3–40
39. Zhang S, Wang R, Zhang X (2007) Identification of overlapping community structure in complex networks using fuzzy cc-means clustering. Phys A 374(1):483–490