Measuring user's influence in the Yelp recommender system

Andres Bejarano, Agrima Jindal and Bharat Bhargava
Department of Computer Science, Purdue University, West Lafayette, Indiana, USA

Received 8 February 2017
Revised 24 March 2017
Accepted 30 April 2017

Abstract

Purpose – Recommender systems collect information about users and businesses and how they are related. Such relation is given in terms of reviews and votes on reviews. User reviews gather opinions, rating scores and review influence. The latter component is crucial for determining which users are more relevant in a recommender system, that is, the users whose reviews are more popular than the average user's reviews.

Design/methodology/approach – A model of measure of user influence is proposed based on review and social attributes of the user. User influence is also used for determining how influenced has been a business being based on popular reviews.

Findings – Results indicate there is a connection between social attributes and user influence. Such results are relevant for marketing, credibility estimation and Sybil detections, among others.

Originality/value – The proposed model allows search parameterization based on the social attribute weights of users, reviews and businesses. Such weights defines the relevance on each attribute, which can be adjusted according to the search needs. Popularity results are then a function of weight preferences on user, reviews and businesses data join.

Keywords Reviews, Measure, Users, Influence, Recommender systems

Paper type Research paper

1. Introduction

Recommender systems let users share their reviews about products, places, establishments or services. Such kinds of reviews are found useful by other users who prefer real descriptions given by experience instead of advertisements that may not reflect the reality. It is typical of such systems to let users cast their votes and ratings to both products and existing reviews. If a product receives more votes or reviews, it becomes more popular among customers. It is expected that higher number of positive reviews is beneficial for marketing's success of the product.

Among the total number of reviews, there are some that have gained more attention among users given its content or user who wrote it. Such attention is reflected by the number of votes the review has received. Those votes can be understood as the number of users who agree with the review or find it interesting or useful. Reviews with a high number of votes are usually shown in the first locations of the total list of reviews of a product. The relevance and acceptance of such reviews make them significantly influential.

PSU Research Review
Vol. 1 No. 2, 2017
pp. 91-104
Emerald Publishing Limited
2399-1747
DOI 10.1108/PRR-02-2017-0016

Recommender systems also allow users to define a social network based on friendship or mutual interests. By defining a friendship link a user is indicating there's a personal contact with other person, usually based on personal or professional interaction. Recommender systems process such links for finding common features between both users (e.g. types of places both users have been before or rated with a good score). When a user posts a review, it is displayed on his contact's dashboard as long as the system finds it relevant to be showed to the respective contact.

Interactions based on user's followers are usually allowed by recommender systems. A user is considered a fan of another user when there's no personal relationship between them (e.g. friendship), but still the former wants to see the latter's posts in his or her own main page. This type of social link is based on interests and affinities with celebrities, special products, places or common interests.

When a review is posted, it is seen by the submitter's friends and followers, increasing its chances of readings about the reviewed product or place. The more social links a user has, the more persons will likely read the review. This social component has become an interesting aspect of recommender systems because it allows an increase of popularity of a product or place. Such popularity comes as a consequence from the satisfaction of the consumers. It is expected for satisfactory reviews to be honest and objective. For this reason, some businesses approach the received reviews seriously to keep good satisfaction rates or for increasing them.

Since some reviews received more votes than others, it can be stated that some reviews are numerically more relevant than others. Additionally, recommender systems keep the number of readings a review has got. These numeric records can be used for data analysis using modeling and statistical approaches. A review then is relevant when it has received a considerable amount of votes, and it has been read more often compared to other reviews. The relevance of a review is tied to its influence over people, which could be measured by the impact it had over users concerning the popularity of a product or place.

The social component of recommender systems also allows giving votes to users aside of their posted reviews. As expected, such votes are different than the review votes because they reflect personal preferences towards other users. Some systems let users to give and receive compliments based on appealing factors such as physical appearance or the writing style of the reviews. This mechanism collects a considerable amount of data concerning users' interests and their popularity in the system. In a similar way as the numeric data from reviews, the obtained social data reflects the influence of the users in the system as persons instead of reviewers. It is expected that more influential users have stronger influence with their reviews rather than regular users. Therefore, businesses are more impacted by the reviews of influential users than the reviews from less influential ones.

In this paper, we consider the fundamentals of the measure of influence on recommender systems. Influence in such systems is approached in the following contexts:

- the influence of a review;
- the influence of a user; and
- the influence over a product or place.

The proposed model is applied to the Yelp Dataset from the Yelp Dataset Challenge (Yelp, 2015).

The paper is organized as follows: Section 2 considers the previous work with respect of the measure of influence in users. Section 3 describes the nature of the Yelp Dataset Challenge. Section 4 describes of the proposed model. Section 5 explains the implementation of the data extraction and influence measurement. Section 6 describes the experimented results and the future work.

2. Previous work

User influence in recommender systems is a topic considered by its relevance in consumer psychology, data analysis and Sybil user detection. A considerable amount of research has been done by analyzing user data from Twitter. Cha *et al.* (2010) considered the influence as the flow of information over other people. Starting from sociology and viral market concepts, they analyzed the influence in Twitter from a threefold perspective: (1) in degree (regular tweets), (2) retweets and (3) mentions. They found that popular users with high degree of popularity are not necessarily influential. Additionally, most influential users can hold significant influence in a variety of topics. Finally, they found such influence is not gained immediately or accidentally but with some additional effort by the user such as limiting the number of tweets over certain topics. Liu *et al.* (2013) also considered Twitter as the source of information and user influence. They proposed a user-tweet integration model to describe the possible relationships between users and their tweets. Furthermore, it was introduced a time attenuation component over the tweets to give more relevance to recent postings and lowering it for older ones. The performed analysis is based on a network topology with users and tweets as nodes and actions such as mentioning and retweets as edges. They considered the Spearman's rank correlation for measuring the correlation between relevant tweets and influential users. The complexity of the data obtained from Twitter allows researchers to consider new data analysis alternatives besides the classical analysis tools. Mei *et al.* (2015) questioned the effectiveness of principal feature techniques for measuring user influence. They used entropy and rank correlation analysis methods for measuring user's influence. The analysis of hidden attributes was performed using the principal component analysis and stepwise multiple linear regression. It was found that besides mentions and retweets, there are other effective metrics such as the number of public lists, new tweets and follower to friend ratio that are effective for measuring user's influence. Authors also found that popularity, engagement and authority are the most important social attributes of influence in Twitter.

A comprehensive survey on user influence measurement proposals applied to Twitter was presented by Riquelme (2015). His findings go from simple metrics up to complex mathematical models. He found an important number of methods based on the PageRank algorithm, traditionally used for ranking pages on internet. Another subset of techniques based their analysis on timeline approaches. Finally, a third subset focus their attention to content analysis and specific topics. His review concludes with the measuring of activity and popularity as well as the computational complexity for the influence measurement context.

A different aspect for user's influence analysis is in terms of security. Resnick and Sami (2007) focused their attention on the attacker in recommender systems. They described an influence-limiting algorithm that can turn recommender systems into manipulation-resistant systems. The algorithm reduces the number of shills (Sybil attacks), bounding the number of attacks to a small amount. The system could use the information from honest sources and informative users. However, doing such type of classification affects a portion of good data. They describe the influence limits and the information loss incurred due to the limitations applied by the system. Following the same trend, Resnick and Sami (2008) focused on the combination of three ideas:

1. prediction market rewards;
2. user punishment when causing marginal changes to communal prediction; and
3. the use of online learning algorithms to limit the influence of individual users.

Authors argue that the combination of these approaches could "make recommender systems robust against manipulation while making good use of information from genuine raters" (Resnick and Sami, 2008). The goal is to bound the damage that an attacker can do while using a fixed number of Sybils. By doing this, it is possible to classify information and make effective use of the good ones from honest users.

User influence can also be stated as a problem of user credibility. Such approach is considered by Can *et al.* (2012) on the scenario of P2P systems. Credibility is a user attribute estimated by its past interactions and given recommendations. The proposed method considers the build of a trust network based on user's locality rather than global trust information. As in the Twitter case, interactions and recommendations are considered for evaluation, focusing on importance, recentness and peer satisfaction parameters.

The information gathered by recommendation systems is also useful for predictions using users' reviews and purchases. Following this approach, Nikulin (2014) proposed a hybrid model based on collaborative filtering (user past behavior) and content-based filtering (discrete characteristics of an item) approaches. The proposed model uses all of the businesses in the review data for training or testing the model. Additionally, data about the businesses is categorized. Such categories, as well as some additional data, are transferred to the users in the review data. The new data collection is analyzed using standard regression model with rectangular matrix of explanatory variable. For solving the high-dimensionality problem the author replaces ones in the binary matrix of features using their respective positive importance rating calculated with Random Forests. The model was developed for the ACM RecSys 2013 contest, granting the author the second best prize. This hybrid model was adapted later for predicting customer loyalty. Nikulin (2015) and Nikulin *et al.* (2015) focused on the problem of predicting users who will re-buy a product given their purchase history. The methods summarize customer transactions to periods of few months and transfer the data to a standard rectangular format suitable for classification or regression models. The proposed models focus on finding the intensity of the customer transactions during periods of time immediately before a product is offered. Such information is used for estimating how likely (loyal) is the user to buy the product again.

3. Yelp Dataset description

The Yelp Challenge (Yelp, 2015) is a contest offered by Yelp to students interested in data analysis. Up to December 2015, the data set contained information about 1.6 million reviews, 500,000 tips, 61,000 businesses and 366.000 users with 2.9 million social edges. The plain size of such information round about 1.5 GB.

The data set groups data into five main single object types: business, review, user, check in and tip. Only the first three objects contain information about reviews and social attributes. Such objects are the ones considered for the proposed influence measurement model.

3.1 The review object

The review object, named yelp_academic_dataset_review.json, contains the information about the reviews made by users about businesses in the Yelp system. For each review entry, the business id, the user id, the number of given stars by the user, the text of the review, the date when the review was posted and the different votes (cool, funny and useful) given by other users to the review are stored. For influence measurement purposes, it is considered the business id, the user id, the number of given stars, the date of the review, and the received votes:

```
{
   'type':'review',
   'business_id':(encrypted business id),
   'user_id':(encrypted user id),
   'stars':(star rating, rounded to half-stars),
   'text':(review text),
   'date':(date, formatted like '2012-03-14'),
   'votes':{(vote type):(count)},
}
```

3.2 The user object

The user object, named yelp_academic_dataset_user.json, contains the information about users in the system. For each user entry, the user id, the user's first name, the number of reviews written by the user, the average number of stars given by the user to the reviews, the number of votes (cool, funny and useful) the user's reviews have received from other users, the user's list of friends (other users' ids), the years when the user was an elite, the number of compliments (Thank You, Good Writer, Just a Note, Write More, Great Photo, You're Funny, Cute Pic, Hot Stuff, Like Your Profile, You're Cool, and Great Lists) given to the user in each compliment category and the number of fans following the user are stored. For influence measurement purposes, we only consider the user id, the number of reviews, the average number of given stars, the number of received votes, the length of the friends list, the number of years the user was an elite, the number of compliments and the number of fans:

```
{ 'type':'user',
  'user_id':(encrypted user id),
  'name':(first name),
  'review_count':(review count),
  'average_stars':(floating point average, like 4.31),
  'votes':{(vote type):(count)},
  'friends':[(friend user_ids)],
  'elite':[(years_elite)],
  'yelping_since':(date, formatted like '2012-03'),
  'compliments':{
   (compliment_type):(num_compliments_of_this_type),
  ...
  },
  'fans':(num_fans),
}
```

3.3 The business object

The business object, named yelp_academic_dataset_business.json, contains the information of businesses reviewed in the system. For each business entry, the business id, the business name, the neighborhoods where the business it is located, full address of the business, the city and state where the business is located, the localization coordinates of the business, the number of received stars given by the users, the number of received reviews by the users, the categories of the business, the open/close schedule and the attributes of the business are stored. For influence measurement purposes, we only consider the business id, the number of received stars and the number of received reviews:

```
{
   'type': 'business',
   'business_id': (encrypted business id),
   'name': (business name),
   'neighborhoods': [(hood names)],
   'full_address': (localized address),
   'city': (city),
   'state': (state),
   'latitude': latitude,
   'longitude': longitude,
   'stars': (star rating, rounded to half-stars),
   'review_count': review count,
   'categories': [(localized category names)]
   'open': True/False (corresponds to closed, not business hours),
   'hours': {
     (day_of_week): {
        'open': (HH:MM),
          'close': (HH:MM)
    },
    ...
   },
   'attributes': {
      (attribute_name): (attribute_value),
       ...
           },
}
```

4. Proposed model

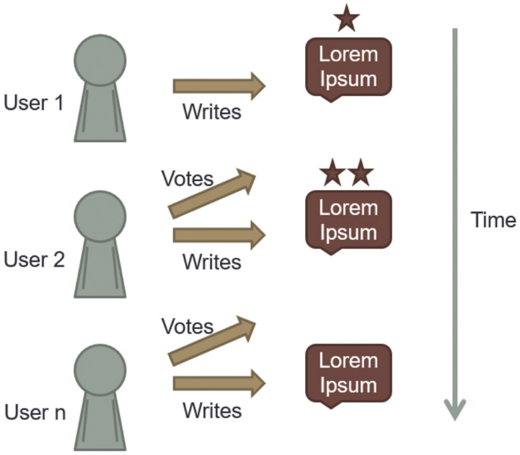
The proposed influence measurement model applied to the Yelp Dataset follows the work by Liu *et al.* (2013). By definition, influence is the power or capacity of causing an effect in indirect or intangible ways. On recommender systems, such effect is the reaction after reading a review of a product or by checking someone else's preferences.

A recommender system works in base of posted reviews and their interaction with users. Such interaction is presented in the user-review schema shown in Figure 1. It consists of the features identified between reviews posted by a user and the received votes casted by other users. In the scheme, it is considered two ways of interaction:

(1) *Posting a review*: A user posts a review about a product or place. Usually it is a written description of his own experience, following a five-star score which indicates the level of satisfaction.

(2) *Voting a review*: A user reads a review posted by other users. Sometimes a reader casts his vote to the respective review. The way such votes work varies from system to system. A general consensus is allowing a positive vote on the review. Users are usually allowed to vote a single time per review.

In the practical sense, a review in the Yelp system is a piece of text describing the experience of a user in a particular business. Each review contains a five-star score which indicates how pleasant the experience was, the greater the number of stars the better the experience. Only the numerical attributes of the dataset are considered for the context of influence measurement based on social metrics. A complementary analysis including text parsing



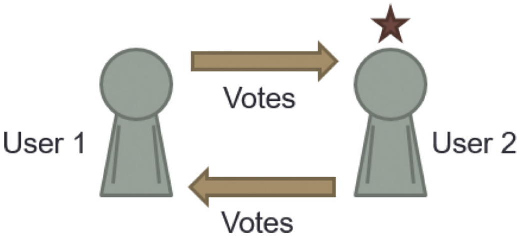
Note: Users post reviews about a business. In the following time, other users cast votes on already existing reviews

Figure 1. A user-review schema

techniques or machine learning approaches (e.g. sentiment analysis) over the piece of text is out of the scope of this paper.

A second type of interaction corresponds to user-user interaction (Figure 2). This case represents the personal preference towards another user in the system, usually seen as compliments. Each system defines its own user-user voting system including types of votes and allowed number of votes between two users. For celebrities and renowned users, it is included a fan counter which indicates the number of followers they have. As in the case of social networks, two regular users have a friendship relation if both users agree on such type of connection (following should be accepted by each user independently), whereas celebrities do not have to accept each following request. The user-user votes are then allowed for regular users as well for celebrities and their fans without major distinctions.

A third schema is about the influence given over a business in a period of time. The recommender systems store the timestamp of the review, allowing a time classification of reviews over a time range. In Figure 3, it is shown that reviews are given sequentially over time



Note: Two users with a kind of relationship can cast votes on each other

Figure 2. User-user interaction

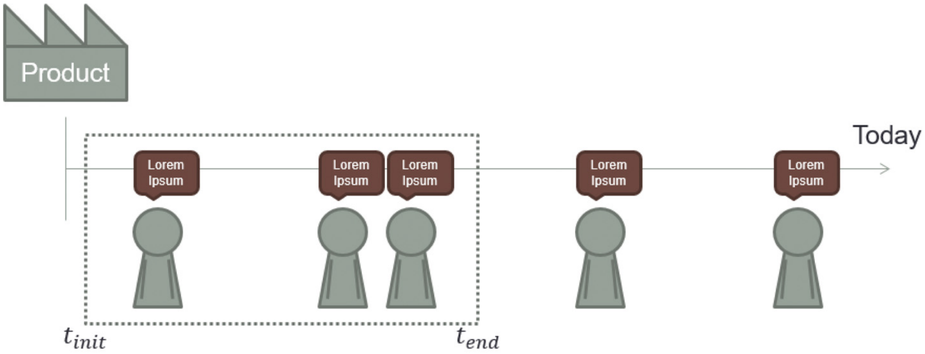


Figure 3. Influence over business in a period of time

Note: Information on a date window can be filtered and analyzed for influence measurement over a period of time

as well as the scores of the business. By defining an initial date and an end date, the reviews can be recollected in such date range, then influence on the time period is analyzed using the filtered information. This approach is crucial for a good understanding of the popularity of a business over time or during very specific periods of times (e.g. shopping seasons and holydays).

Following the information of the schemas we define the user-influence and business-influence models using the information stored by the Yelp system. The influence of a user u is the out degree of the user, which is the number of social metrics corresponding his or her own activity in the system. Such score is defined by the acceptance of the user's written reviews (reflected in the received votes by other users) and the compliments given by other users. Then, for a user u, influence is defined as:

$$Influence\,(u) = w_1 V + w_2 C + w_3 F + w_4 I + w_5 E$$

$$\sum_{i=1}^{5} w_i \leq 1,\; w_i \in [0,\,1]$$

where V is the total number of votes user u received on all its posted reviews, C is user u's total number of received compliments, F is the total number of fans following user u, I is the total number of friends' user u has in the system and E is the number of years user u was an elite. For each component, there is an associated weight w_i that represents the relevance of such feature in the model. Weights can be adjusted accordingly to represent different contexts (e.g. the number of friends is more relevant than the number of fans). To keep scores bounded, it is added the restriction of the sum of the score weights to be less or equal than 1.

In a general setting for any recommender system with different social attributes, the influence of a user u as the summation over all of the social attributes of u times the weight of each attribute can be measured:

$$Influence\,(u) = \sum_{i \in F} w_i i$$

$$\sum_{i \in F} w_i \leq 1,\; w_i \in [0,\,1]$$

where F is the set of the social attributes per user defined in the recommender system. The influence of the users is useful for measuring the influence over a business. Because we are interested in the influence given in a period of time, the influence over a business is defined as follows:

$$\mathit{Influence}\,(B, t_i, t_e) = TAC(t_e - t_i) \times \sum_{r \in B_R} \mathit{Influence}\,(r_u)$$

where B is the business over which the influence will be measured, t_i and t_e ($t_i < t_e$) are the initial and end dates defining a time period, time-effectiveness attenuation coefficient [(TAC) (Liu *et al.*, 2013)] is defined as $TAC(\Delta t) = 1/(1 + 2^{\Delta t})$, r is a reviews in the set of reviews B_R posted about business B and r_u is the user who posted the review r. The TAC coefficient is used as a mitigation mechanism for older reviews, which helps the model to give more relevance to recent reviews instead of considering all of them with the same influence weight score. The time range period defined by t_i and t_e could be defined for any time granularity. In our implementation we considered the number of days between the two given dates.

5. Data extraction and preprocessing

As mentioned before, the Yelp Dataset contains a considerable amount of plain data about businesses, users and reviews. It is required to preprocess the data before using it in the influence measurement model. The information of each review, user and business was extracted from the original dataset, keeping only the attributes indicated in Section 3. The influence measurement model can be applied directly by specifying the weights of the user's attributes.

Data preprocessing must be done in a specific order because the attribute values for user's influence measurement are not centralized in a single object. The following is the order implemented for extracting and preprocessing the information from the Yelp Dataset Challenge:

- Two data subsets were extracted from the user object. Single information and social network information. Single information corresponds to a simplification of the user ids because the originals are an encrypted 22-bit long string. For each user, it is assigned a unique integer number which corresponds to its order of appearance in the data set. Social network information corresponds to the attributes that reflect the user's interaction with other users in the system (e.g. votes, compliments, number of friends and number of fans). This latter subset is used for the measurement of the user's influence. Additionally, a temporary data structure is generated for storing the list of friends for each user. Because the Yelp Dataset stores a user's friends as a list of user ids then then such ids are replaced with the simple integer numbers given by the order of appearance.
- The information about the businesses is extracted from the business object. A single data subset is generated for this data, corresponding to the attributes for each business as explained in Section 3.3.
- The information about reviews is extracted from the review object. A single data subset is generated for this data, corresponding to the id of the business, the id of the user and the attributes explained in Section 3.1.

6. Results

The user influence measurement model was implemented using Python 2.7, following the provided examples in the Yelp's Academic Dataset Examples available on Yelp's GitHub account (https://github.com/Yelp/dataset-examples). The machine where data preprocessing and the model were executed is an Intel Core i7-3630QM CPU 2.40 GHz processor with 6.00 GB (5.90 GB usable) RAM memory running Windows 10 x64 Home Basic operating system.

Given that the information about users has no timestamps, their influence scores were calculated in a single run. The proposed model was used for finding the most influential user in the Yelp Dataset (up to December 2015) considering all of the attributes to have the same weight. The information below corresponds to the most influential user found using weight values $w_i = 0.2$:

```
Simple id: 33317
Name: Brian
Yelping since: 2008-12
Review votes: Cool: 31248 Funny: 24792, Useful: 32327
Elite: 0 (No)
Stars (avg): 4.28,
Reviews: 1628,
Friends: 63
Influence Score: 52908.6
```

The plain data corresponding to such user is the following:

```
{'yelping_since': '2008-12', 'votes': {'funny': 24792, 'useful':
  32327, 'cool': 31248}, 'user_id': 'ftm0zhX_fQDzAed8ESTlkA',
  'name': 'Brian', 'elite': [], 'type': 'user', 'compliments':
  {'profile': 9955, 'cute': 10473, 'funny': 10254, 'plain': 19019,
  'writer': 10167, 'list': 7805, 'note': 10033, 'photos': 53641,
  'hot': 16817, 'cool': 20407, 'more': 7373}, 'fans': 169, 'average_
 stars': 4.28, 'review_count': 1628, 'friends':
  ['eSAmN-5RAu8y1igSvz_yJw', 'IIWtH1LHdfWDtIKYgdrrrA',
  'zQ_FpDX-GDPuai4GOldCJQ', 'ynGxw3zZqAjahVou563zXQ',
  '1h2Zmu7R9IMiK9FFYj-yhw', '0-c7pC240GjXkcHrtsW8HA',
  'oupgo3fb2kl4AbJ35fhk3w', 'DdxfNRK6O8UWVkhT4Nsrcg',
  '7NaqQ5VmttoYJYc8rjPzWg', '9kRjgUNKvICfG2pH4BY6Eg',
  'Ia9N8RQ8rcXXEjhaQKR3Mg', '0mVScF9nfPVIkS2OlbvpNw',
  'JnEpFNLnj0TWOO9ZZ_Q4-w', 'nOA4lSY3D4xsUXJBT4H0pw',
  '0bPdkjTs4-_OBlTxya-qww', 'tz5imqXo5zrimU434i3u2Q',
  'TYdOKKLtA7LITxV7Y2_wjA', 'XIOTsKjvbMLp-i-p-x5RvA',
  'fgZrwWlAfcNQGTuSfWITuw', 'sLZYkJanCixT6Agd3zeGew',
 '6X2c1d4T8-qGGZ5PQ7BrcA', 'UjS6tMpdxoVHrT5XRAPVOA',
  'YN3sv32NMKfcNDA5BDRlHw', 'xPGVu0bynlz62GYMCvDLcw',
  'rxGMXiiZyxwnYUrWP1OJTQ', 'wk3ZgGeHzdMyKvL5NeOcAw',
  'u6VZAszfuylriytDU66gxw', 'fnNQwOOOSOSRLy9X2bYF1Q',
  'pgR0i5Mn8aLOKZPyxbs_Iw', 'o-Sir3Q4VWsoY01vjktu_Q',
  'ZpYfzg_GZa2CwWflRR2b5A', 'alSpPmNQ_143kpp-YpWGqw',
  '7eXEzO3qCRonF-MM-oQb6g', 'sQBMbqZWNvqII8QwW_6RPQ',
  'XPWJVAjAZwwOvOlz4K8OQA', '4zIz_XkhHPwsxcMoTLJZhw',
  '6DuBPBgCN6YMowFwrp_P5g', 'n9MY58inPus2NlbOrQoudw',
  'H4bi5dCXGnC4fReslhZ2ew', 'huzlXrlIn8Irqphw4_gIIA',
'HC2EZcawhitJs9-E02VcYg', '9cCTmiJ7hz35rHIdr8n9kA',
'bRLPoRvnbJbh8SiLjQRBlQ', 'SgDWDjBr8fV7id1UnFaAFg',
'DZO_4yihrCGurF6WpmXT2g', 'LusAw6vTDC7KAfbuClMReA',
'0bpqsRkCUBpu2_IjCSwdfA', 'ByDaM5gVaNUFIUy3rQP4bg',
'hdbVD3ksI6LfkfYccU6QVg', 'DNN-pOqu6ldnt1-UP5cHaA',
'FTSmBX-WvzxFpH_vOiRIkg', 'zTgQjEWRtbtxB-jdB3dL3Q',
'DAslVQqAFZ9YzuZxj-keUQ', 'n-0JFhD3V0FfN7sVmQs5lg',
'9ITZFb1M5RvnjACTA-WfuQ', '1zfwRlu3O6bhehSwgiAz5Q',
'CdMMq_o3QAJTXD0slw9C7A', '2HmHgW3hRYvXYFmQyQtLuw',
'VR_fkiVm9Yfz_OiDBrEapw', 'PqRPuVEWCxQfR7OTXp1efw',
'PhslBpqS6FWID_WDJh8THA', '6vEFaCaJ1w1yfOOm1knzzA',
'11QzsQ-nmRykVvD4agGRPw']}
```

The overall result for all users in the dataset (using the same weight values as described before) shows that less than 25 per cent of the users are significantly influential. In Table I, it is shown the results of minimum, Q1, Q2, Q3 and maximum scores over the entire data. An initial thought may suggest that influential users correspond to a small sample of the users.

These results correspond to the activity of the users in the Yelp system. It can be inferred that users are not equally active; therefore, a group of them will gather some attention because their contributions are more significant. As an example, the following is the plain data of the user in the median obtained with the proposed model with the same weight values as described before:

```
{'yelping_since': '2009-07', 'votes': {'funny': 1, 'useful': 5,
 'cool':  1},  'user_id':  '9xVrzbALv-smeh_7AAFUug',  'name':
 'Cecil', 'elite': [], 'type': 'user', 'compliments': {'plain': 1,
 'cool': 1}, 'fans': 0, 'average_stars': 4.29, 'review_count': 31,
 'friends': ['Prsk7SiPZNfgcPyygOdGHg']}
```

By comparing the most influential user with the user in the median, it can be seen that there is a significant difference in terms of received votes, number of friends, number of reviews and received compliments. Only a total of 1,677 users have an influence score higher than 1,000 points, representing 0.46 per cent of the population (using the same weight values as described at the beginning of the section).

For the case of influence over businesses, the special case of the most reviewed business in the data set was considered. We applied the model described in Section 4 over the two most reviewed businesses in the system up to December 2015: True Food Kitchen located in Phoenix and Pinball Hall of Fame located in Las Vegas. The results of both scores are shown in Figures 4 and 5, respectively.

It is seen that influence over a business is a varying indicator over time. This behavior could be explained by several factors not stored in the dataset (e.g. change in management, changes

Rank	Score
Minimum	0
Q1	0.4
Q2	2
Q3	8.4
Maximum	52,908.6

Table I.
Box plot information of all users influence

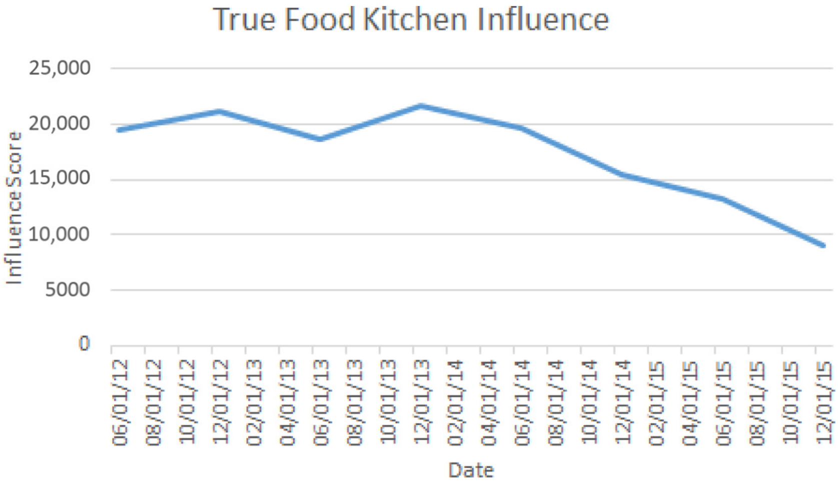


Figure 4.
Influence over time for True Food Kitchen

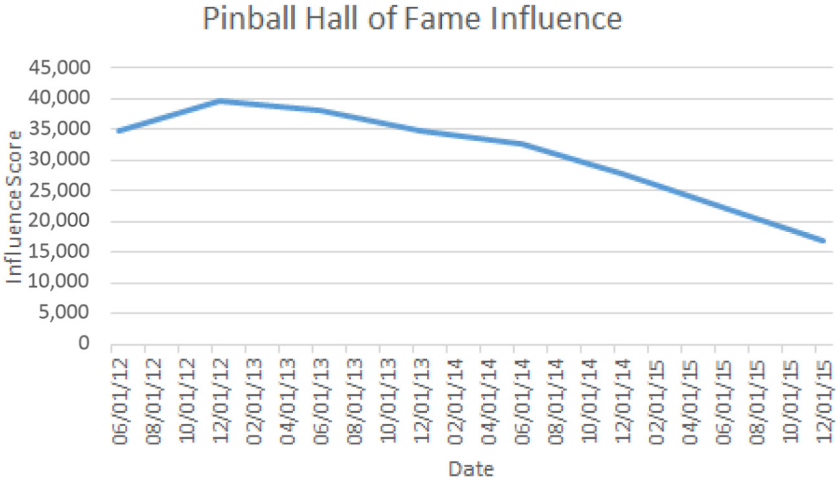


Figure 5.
Influence over Pinball Hall of Fame

in menu or services, changes in staff or changes on closed to public time periods, among others). Because the scores of the businesses are not stored with a timestamp, it is not possible to obtain historical records that bring support to the obtained influence scores. An analysis over the text information of the comments (e.g. String Parsing or Sentiment Analysis) might bring some help for understanding the changing values of the scores. However, such kind of analysis is out of the scope of the proposed model.

The decreasing tendency of the influence scores is worth mentioning. For the Yelp Dataset, all the reviews from the beginning up to the specified end date were considered; this is an accumulation of attribute values over the influence score. Having a decreasing tendency does not mean that the popularity of the business is going down; it means that by considering all of the reviews up to the specified date, the model estimates an overall score. A deeper statistical analysis of this trend as well as a correlation analysis between the involved variables is required for a better understanding of this lowering tendency.

7. Conclusions and future work

The obtained results indicate there is a link between users' popularity, users' reviews and the impact on businesses reviews' scores. As expected, the percentage of influential users is considerable small compared to the population of regular users. It is seen that influential users tend to review more places than the average number of users. Similarly, influence over a business is more affected by such influential users rather than regular users registered on the recommender system.

An actual influence score should reflect the popularity of a business for a very specific date period. Unfortunately, the Yelp Dataset Challenge lacks of such information. Obtaining historical records for reviews, users and businesses will improve the influence scores and therefore the correctness of the proposed model.

It is of our interest finding the influence of users and reviews on different systems that collect social attributes, such as Amazon and YouTube. We consider that the influence measurement model can be extended to any system that allows social interactions between users, as well as posting review content different than text (e.g. audio records, images or videos). Still, to keep record of influence variation over time, it is required the timestamp information on each review and publication.

Current plans for improving the proposed model includes adding more rigorous data analysis, as it is required for determining the real impact factor of each user, review and business attribute. The proposed model assumes the relation between attributes is linear; however, this should not be necessarily the case. Applying data mining approaches for improving the model is an undergoing task.

Finally, in terms of security, we are interested in the correlation of Sybils (both users and reviews) and their influence over products, services or places on recommender systems. In particular, our interest is whether such attacks are influential or not for the popularity of such products. This model can be used as the cornerstone for a Sybil detection model based on user's credibility.

References

Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K.P. (2010), "Measuring user influence in twitter: the million follower fallacy", 4th International AAAI Conference on Weblogs and Social Media (ICWSM).

Liu, D., Wu, Q. and Han, W. (2013), "Measuring micro-blogging user influence based on user-tweet interaction model", in Tan, Y., Shi, Y. and Mo, H. (Eds), *Advances in Swarm Intelligence*, Springer, Berlin, Heidelberg, Vol. 7929, pp. 146-153.

Mei, Y., Zhong, Y. and Yang, J. (2015), "Finding and analyzing principal features for measuring user influence on twitter", IEEE First International Conference on Big Data Computing Service and Applications (BigDataService), *Redwood City, CA*, pp. 478-486.

Nikulin, V. (2014), "Hybrid recommender system for prediction of the Yelp users preferences", in Perner, P. (Ed.), *Advances in Data Mining: Applications and Theoretical Aspects*, Springer, Cham Vol. 8557.

Nikulin, V. (2015), "On the method for data streams aggregation to predict shoppers loyalty", International Joint Conference on Neural Networks (IJCNN), *Killarney*, pp. 1-8.

Nikulin, V., Huang, T.H. and Lu, J.D. (2015), "Mining shoppers data streams to predict customers loyalty", 10th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), *Taipei*, pp. 26-33.

Resnick, P. and Sami, R. (2007), "The influence limiter: provably manipulation-resistant recommender systems", *Proceedings of the 2007 ACM Conference on Recommender Systems: RecSys '07*, ACM, Minneapolis, MN, pp. 25-32.

Resnick, P. and Sami, R. (2008), "Manipulation-resistant recommender systems through influence limits", *ACM SIGecom Exchanges*, Vol. 7 No. 3.

Riquelme, F. (2015), "Measuring user influence on Twitter: a survey", CoRR abs/1508.07951, available at: http://arxiv.org/abs/1508.07951

Yelp (2015), "Yelp dataset challenge", available at: www.yelp.com/dataset_challenge (accessed 19 December 2015).

Further reading

Can, A.B. and Bhargava, B. (2013), "SORT: a self-organizing trust model for peer-to-peer systems", *IEEE Transactions on Dependable and Secure Computing*, Vol. 10 No. 1, pp. 14-27.

Corresponding author
Andres Bejarano can be contacted at: abejara@purdue.edu