

Mining implicit data association from *Tripadvisor* hotel reviews

Vittoria Cozza
Department of Information
Engineering,
University of Padua
Padua, Italy
vittoria.cozza@dei.unipd.it

Marinella Petrocchi
IIT-CNR
Pisa, Italy
marinella.petrocchi@iit.cnr.it

Angelo Spognardi
Dipartimento di Informatica,
Sapienza Università di Roma
Rome, Italy
spognardi@di.uniroma1.it

ABSTRACT

In this paper, we analyse a dataset of hotel reviews. In details, we enrich the review dataset, by extracting additional features, consisting of information on the reviewers' profiles and the reviewed hotels. We argue that the enriched data can gain insights on the factors that most influence consumers when composing reviews (e.g., if the appreciation for a certain kind of hotel is tied to specific users' profiles). Thus, we apply statistical analyses to reveal if there are specific characteristics of reviewers (almost) always related to specific characteristics of hotels. Our experiments are carried out on a very large dataset, consisting of around 190k hotel reviews, collected from the *Tripadvisor* website.

1 INTRODUCTION

Social media, forums, and blogs are privileged vehicles for posting and spreading online reviews. Among the goods and services that are discussed every day on the Internet, we can find those belonging to the most disparate categories, like, e.g., food, clothes, music, toys, etc. Particularly, the practice of choosing and booking preferred destinations has been greatly eased by the possibility for users to consult previous feedback about hotels and restaurants. According to *comScore Media Metrix*¹, *Tripadvisor* is the world's largest travel e-advice site, providing advices as reported by actual travellers. *Tripadvisor* counts more than 87 million visitors per month².

Not only common users, but also service providers have strong motivations to analyse the myriads of posts, tweets, and comments available online. The latter will benefit by adjusting, e.g., their products lines and advertisement campaigns, while the former by relying on previous experiences for addressing their needs and matching their expectations. Furthermore, online reviews are a precious source of information, e.g., to unveil implicit and/or unexpected characteristics of the reviewers. As an example, in [13] the authors investigate if and how the words —and their use— in a review are linked to the reviewer's gender, country, and age.

In [8], the authors present a novel approach to build feature-based user profiles and item descriptions by mining user-generated reviews. Such additional information can be integrated into recommender systems to deliver better recommendations and an improved user experience.

In our previous work [9], we exploited a *Tripadvisor* dataset in order to investigate how subjectivity of reviewers affects the scores assigned to hotels. Thus, we leverage sentiment analysis

techniques to identify mismatches between the text and the score in online review platforms.

Since several aspects can influence the customer experience (e.g., the hotel price, or the presence of restaurants, cafe, discos in the hotel neighborhood, the connections with bus/train stations and airports, etc.), in this work we propose an automatic approach - based on association rules - to understand which factors most influence consumers' reviews. We consider a very large dataset consisting of around 190k hotel reviews collected from *Tripadvisor*, enriching the dataset by extracting a series of hotel-centric and reviewer-centric features. We leverage these features to list correlations among hotel properties, reviewer's characteristics, and the review score. The results are obtained applying association rules techniques to our dataset. Findings are both expected - such as that the hotels close to entertainment and food areas are ranked with the highest scores - and less intuitive - such as that those reviewers featuring a very low activity (measured with a lower bound in term of given reviews), considering their stay in a particular area, select - very often - hotels with a low number of transportation means in the neighbourhood.

We argue that, with our approach, sociologists and marketing experts could analyse the results of the association rules to better understand some extra reviewers's characteristics and connections with the reviewed service. This kind of analysis paves the way for surveying a larger segment of the population than that usually interviewed through standard polls.

2 DATASET

To conduct our study, we grounded it in a dataset composed of real reviews taken from the *Tripadvisor*³ website. In particular, our dataset contains all the reviews that can be accessed on the website between the 26th of June 2013 and the 25th of June 2014 - date of the newest extracted review - for hotels in New York, Rome, Paris, Rio de Janeiro, and Tokyo. With a straightforward approach, we were able to collect the following pieces of information for each review:

- the review date, text, and numeric score;
- the reviewer username, location, and *triptype*, being the type of trip, one among the following five categories: Family, Friends, Couple, Solo Traveler, and Businessman;
- the ID of the hotel which the review refers to.

In addition to the above elements, we collected from *Tripadvisor* all the hotels of the considered reviews and included in our review dataset some additional data regarding the reviewed hotels. In particular, leveraging the ID of the hotel which the review refers to, we have gathered

- the hotel name and full address (where full address includes the street address, the city, and the country);

¹<https://www.comscore.com/Products/Audience-Analytics/Media-Metrix> - All sites last accessed December 23, 2017.

²<https://www.comscore.com/Insights/Rankings> - Statistics updated to June 2017.

© 2018 Copyright held by the owner/author(s). Published in the Workshop Proceedings of the EDBT/ICDT 2018 Joint Conference (March 26, 2018, Vienna, Austria) on CEUR-WS.org (ISSN 1613-0073). Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

³<http://www.tripadvisor.com>

- the category of the hotel (number of stars);
- the number of guest pictures for the hotel.

It is worth noting like the above lists are not exhaustive, i.e., they do not represent all the information accessible from *Tripadvisor*. As an example, further information available for a review are the scores assigned by reviewers to specific aspects of a hotel, like location, cleanliness, sleep quality, rooms, and service. However, for the scope of the current work, we focus on those summarised for the reader’s convenience in Table 1. We exploited such pieces of information to further expand the dataset, with enriched features, as described in the next Section 2.1.

Basic information	
Review	Hotel
Date	Name
Text	Street address
Score	City
Reviewer username	Country
Reviewer location	Guest pictures
Triptype	
Hotel ID	

Table 1: Considered information in the basic dataset

We have discarded reviews by “Anonymous” users, since they represent users of the platform <http://www.daodao.com>—the Chinese version of *Tripadvisor*—where all the reviewers are indifferently grouped in this single virtual username. We have further limited our analysis on reviews whose textual part is in English, following the language identification and analysis approach presented in [5]. While the reviews accessible from *Tripadvisor* in the year under investigation are 353,167, after the pre-processing the resulting dataset is made up of 189,304 reviews in English, provided by 142,583 *Tripadvisor*’s registered users that reviewed 4,019 hotels. Table 1 recaps the information extracted from the dataset, while Table 2 shows the distribution of the reviews per given score value. As shown, the values distribution is highly unbalanced, being the highest score the most frequent in the dataset (reflecting indeed the distribution usually featured by review platforms).

Rating Value	Occurrences
1	6,504
2	8,826
3	24,627
4	64,949
5	84,398

Table 2: Distribution of the given scores in the dataset

Hereafter, we will refer to this dataset as the *basic* dataset. Indeed, in the following, we will extract hotel-centric and reviewer-centric features to enrich the basic set (see Section 2.1).

2.1 Hotel-centric and reviewer-centric features

Starting from the information collected in the basic dataset, we have augmented it performing some further elaboration. In particular, we enriched the data regarding the reviewed hotel with the following features:

- the *popularity*, defined as the number of reviews for a given hotel. While we have neither the list of actual bookings available, nor *Tripadvisor* requires the reviewer to show a proof to have been a guest in the hotel, this feature, when computed on a large number of reviews per hotel, could indirectly act as a quantification of the actual hotel clients;
- the *hotel triptype*, defined as the most frequent reviewer triptype for a given hotel (whereas triptypes are Families, Friends, Couples, Solo Travelers, and Businessmen);
- the geospatial coordinates (latitude and longitude);
- three *points of interest* (POI) features, defined as the number of transportation services, restaurants, and attractions, respectively, in a range of 300 meters around the hotel.

Popularity and *Hotel triptype* have been computed looking at how many and which kind of reviewers have reviewed the hotel. The geospatial coordinates have been calculated with Google Places APIs⁴, starting from the hotel name and full address. Then, latitude and longitude, together with the parameter “radius=300”, have been given as input to the Google Radarsearch API⁵ to find the number of points of interest (POI) related to transportation, food, and entertainment.

The data regarding a reviewer, instead, have been enriched with the following features:

- the reviewers’ *activity*, defined as the number of reviews they have written (under the observation period). Our intuition is that this feature could be useful to discriminate between frequent travelers and sporadic ones.
- the *gender* of the reviewer. This feature has been extracted with the *Namsor Onomastics*⁶ machine learning tool, able to recognise the language behind a name, thus identifying the gender according to that language vocabulary with high accuracy [4].

After cleaning the username from numbers and symbols and splitting it in two parts (where one is likely to be the name and the other one, when available, the surname), we have called the “onomastics/api/json/gendre” API. This service takes as input name and surname and returns the recognised gender. We have used regular expressions to clean the username from symbols and numbers and for splitting the username. This was possible since, in many cases, the name and surname were separated by a space, or the surname started with an uppercase letter. Some examples of username are: “Eldon S”, “MeganJones88”.

Unfortunately, for a subset of reviewers, it was not possible to derive the gender from their usernames. This happened for 9,507 reviewers (corresponding to 6% of the entire reviewers set), which wrote 12,653 reviews. Examples of usernames for which it was not possible to derive the gender are *Hope-and-Dreams*, *mistyrabbit*, *A TripAdvisor Member*, *R W, E A*, *Nickeykol*, *NawakRed*, *FreeTravel81*. We labeled with *unknown* the gender of such 9,507 reviewers.

⁴<https://developers.google.com/places>

⁵<https://maps.googleapis.com/maps/api/place/radarsearch>

⁶<http://api.namsor.com/onomastics/api>

Features	
Hotel	Reviewer
Popularity	Activity
Hotel Triptype	Gender
Geospatial Coordinates	
Points of Interest	

Table 3: Hotel-centric and reviewer-centric features augmenting the basic *Tripadvisor* dataset

It is worth noting that *Popularity*, *Hotel triptype*, and *Activity* have been calculated as the result of queries to the basic dataset, with the aim of making explicit some data that originally were implicit in the information at disposition. A story apart deserves the computation of the reviewer gender, the points of interest close to the hotel, and its geospatial coordinates. As above described, the latter have been computed relying on external data sources, namely the Google Points of Interest and the Namsor database, containing 800k names and statistical information about names in each country of the world.

Table 3 recaps the hotel-centric and reviewer-centric features we used to enrich the basic dataset.

3 ASSOCIATION ANALYSIS

Association rule mining is a well known and widely applied methodology for discovering frequent patterns, correlations, and causal structures in transaction and relational databases, as well as in other information repositories [12]. Thus, given a set of items (or *itemsets*), association rule mining allows to define rules predicting the occurrence of an item (or more), given the occurrence of other items in the same itemsets.

A popular application is basket data analysis, where itemsets are transactions, representing lists of items in the consumers' baskets. An example of transaction is: {Bread, Steak, Juice, Butter, Chips, Beer}. When several others are collected, e.g., in a large database, the methodology allows to automatically find associations like, e.g., {Bread} \Rightarrow {Steak} (steaks are often purchased with bread). Beside sales transactions, the basket analysis can be applied to other situations like click stream tracking, spare parts ordering and online recommendation engines - just to name a few⁷.

An association rule (AR) is generally defined as an implication expression of the form $X \Rightarrow Y$, where X and Y are disjoint itemsets. They represent, *resp.*, the condition and the consequence of the rule.

The strength of an AR is commonly measured through the two metrics *support* and *confidence*. Support gives the fraction of itemsets in the dataset that contains both X and Y . Confidence says how frequently items in Y appear in itemsets that contain X . As an example, we want to know the strength of the rule {Bread} \Rightarrow {Steak} in a dataset with 100 transactions, corresponding to 100 consumers' baskets. Suppose that itemset {Bread, Steak} occurs 30 times, and that itemset {Bread} occurs 40 times, then the support of the rule is equal to $\frac{30}{100}$, while its confidence is $\frac{30}{40}$.

As discussed in [3], rules with high values for confidence and support do not always correspond to meaningful ARs, especially when working with real datasets, due data can be unbalanced.

For example, one rule could have a very high confidence, but only due to the fact that the item in the consequence is very frequent. In this case, the rule is not relevant. Instead, one rule could have a low confidence, due to the fact that the item in the consequence is very unfrequent in general, but it could still be relevant. Considering the above observation, to evaluate the statistical significance of the ARs, two other metrics are often used: lift and conviction.

Lift is defined as the confidence divided by the support of the consequence:

$$lift(X \Rightarrow Y) = \frac{supp(X \cap Y)}{supp(X) * supp(Y)} \quad (1)$$

With respect to confidence, the lift measures the importance of the association considering also the dependence from the support of the consequence.

Conviction is defined by the ratio of the frequency of itemsets that don't contain the consequence, to the frequency of incorrect predictions:

$$conv(X \Rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \Rightarrow Y)} \quad (2)$$

Both lift and conviction values ranging over the (0,1) interval mean negative dependence, values above 1 mean positive dependence, and a value of 1 means independence.

When items are also divided according to different classes, it is possible to force the AR analysis to return a specific class in the consequence. The obtained rule is called "class association rule" (CAR). The CAR is an implication of the form:

$$X \Rightarrow y, \text{ where } X \subseteq I \text{ and } y \in Y \quad (3)$$

where I stands for the itemsets and Y for the classes. The definition of the aforementioned metrics holds also for CARs.

The *a priori* algorithm [2, 16] is one of the most popular algorithms to find *frequent* itemsets, i.e., itemsets whose support $\geq minsup$.

In this work, we apply the association rule mining to the hotel reviews scenario. Each itemset corresponds to a distinguished review, and it is a vector whose components are the values of the features extracted and detailed in Section 2. The same features are reported in Tables 4, 5, 6 for the reader's convenience, together with additional information that are useful here. CARs analysis can be applied when considering also the class, that in our scenario corresponds to the review score, a discrete value with a range between 1 and 5.

To enable the application of the *a priori* algorithm, we have first discretised those features that natively ranged over a large set of values. As an example, in Table 5, a *very low* label for Guest Pictures indicates a hotel with a number of pictures comprised from 0 to 11. Still in that table, a *medium* label for Popularity means a hotel that has been reviewed n times, where n ranges over [433, 1156]. The values in Table 6 should be read as follows: looking at the first line of the "Geo Food" part of the table, our review set contains 37,851 reviews about a hotel, which has a number of restaurants in the range [0, 37] within a radius of 300 mt. Indeed, many different reviews are on the same hotels, being the number of hotels reviewed equal to 4,019, see Section 2.

All the tables also report the *Frequency* indication, i.e., how many reviews correspond to those values for those features, with respect to the values and features in the tables (still quite obviously, the sum on the values in the Frequency column equals to the total number of reviews considered, 189,304).

⁷<http://pbpython.com/market-basket-analysis.html>

Activity		Gender	
value	frequency	value	frequency
up to 5 reviews	138,419	male	102,565
6 or more	50,885	female	74,086
		unknown	12,653

TripType		Month			
value	frequency	value	frequency	value	frequency
solo	9,795	January	14,280	July	18,096
couple	66,557	February	11,440	August	16,870
family	35,833	March	14,146	September	18,466
friends	19,621	April	16,120	October	18,616
business	23,600	May	18,978	November	14,149
unspecified	33,898	June	14,937	December	13,206

Table 4: Discretised features on reviewers

Stars		Guest Pictures	
value	frequency	label: range	frequency
1	694	very low: (0-11)	39,932
2	8,224	low: (12-104)	37,790
3	55,464	medium: (105-271)	37,882
4	83,584	high: (272-525)	37,840
5	15,456	very high: >=526	37,860
unspecified	25,882		

HotelPopularity		Hotel Trip Type		Country	
label: range	frequency	value	frequency	value	frequency
low: (0-432)	63,165	solo	974	Italy (it)	31,224
medium: (433-1156)	63,093	couple	139,860	United States (us)	83,605
high: >=1157	63,046	friends	672	Brasil (br)	3,631
		family	20,429	Japan (jp)	11,966
		business	9,112	France (fr)	40,621
		unspecified	18,257	unspecified	18,257

Table 5: Discretised features on hotels

Geo Food		Geo Entertainment		Geo Transport	
label: range	frequency	label: range	frequency	label: range	frequency
very low: (0-37)	37,851	very low: (0-3)	38,282	very low: (0)	34,471
low: (38-136)	37,921	low: (4-15)	37,602	low: (1-3)	38,079
medium: (137-197)	36,827	medium: (16-35)	37,814	medium: (4-11)	40,382
high: (198-199)	21,848	high: (36-63)	37,708	high: (12-18)	39,674
very high: >=200	54,857	very high: >=64	37,998	very high: >=19	36,698

Table 6: Discretised geolocation-based features

In order to find ARs and CARs, we applied the Weka framework [11] implementation of the *a priori*. The Weka *a priori* implementation allows to rank the rules according to different metrics. Among them, we rely on confidence, lift, and conviction. For AR analysis we generate a large number of rules with lift above 1. For CAR analysis, we generate a large number of rules with confidence above 0.2 and then we compute the lift (since, for CAR, Weka does not natively include the ranking based on lift). We finally select the rules with lift greater than 1.

Both for the generated ARs and CARs, we then manually select the most interesting rules, among those with the highest lift and

conviction. Table 7 and Table 8 report an excerpt of the results for both scenarios.

3.1 Discussion

Association analysis results are reported in Table 7 and Table 8, please notice we only consider those rules that lead to a lift and conviction greater than 1. It is worth noting like $|X \cap Y|$, when divided by the size of the dataset, corresponds to the support of the given rule.

We summarise the main findings, as follows. Rule *r1* states that those reviewers featuring a very low activity, considering their stay in France, select - very often - hotels with a low number of transportation means in the neighbourhood. The rule holds for 19,199 reviews, over a total of 29,837 reviews, with equal premises. Rule *r2* says that males visiting US prefer hotels with a high popularity. Rule *r7* says that, when the hotel has low transportation means in the neighbourhood, and the number of stars for that hotel is unknown (this may correspond to accommodation facilities like hostels), its rating is equal to 3. Rule *r10* states that Japanese people staying in a 3 stars hotels rate those hotels with a score equal to 4. Rule *r14* in Table 8 states that hotels close to entertainments, which are 37,998, are scored with the top score 5 the 50% of times.

This kind of study provides a general approach for a preliminary data exploration. While the explanation for certain rules is very intuitive, well-grounded justification for others is left to experts in the field. We argue that this kind of analysis corresponds to a preliminary step, useful for suggesting which extra-features could be exploitable to build an enhanced hotel recommendation system. Also, we acknowledge that the analysis is based on the available (direct or indirect) information, obtained from the *Tripadvisor*'s website. More detailed features could consider elements like price or number of guests. This would allow to obtain other interesting rules, which remain an exclusive prerogative of the hoteliers.

4 RELATED WORK

E-advice technology offers a form of “electronic word-of-mouth”, with new potential for gathering valid suggestions that guides the consumer’s choice. Extensive and nationally representative surveys have been carried out in the recent past, “to evaluate the specific aspects of ratings information that affect people attitudes toward e-commerce”. It is the case, e.g., of work in [10], which highlights how people, while taking into accounts the average of ratings for a product, still do not take care of the number of reviews leading to that average. Recent work showed that, instead of showing first to the users the reviews with the highest scores, a different order, based, e.g., on the user profile, could be considered [8]: that work integrates new features based on the user profile into recommender systems, to deliver better recommendations and provide an improved user experience. Similarly, in [19], the authors focus on score values given by previous contributors whose preferences are close to the user’s preference. Even almost one decade ago, the work in [1] applies text mining tools to online reviews to define rules sets, to identify contextual information in the texts, which goes beyond a mere order of numerical scores. Similarly to our work, they rely on *Tripadvisor*, focusing however on text analysis only.

However, the cited literature proposes systems that recommend a service based on the intrinsic characteristics of that service (e.g., characteristics of the hotel and its facilities). Other works, similar to ours, investigate if, and how, the review data hide social and/or economic information of the reviewers. One example is mining reviews to exploit them as a textual resource for sociolinguistic studies at a large-scale, as done in [13]. This work leverages the size of the reviews corpus as a more statistically solid base for the analysis, with respect to manually-collected corpora. Since reviews sites, such as *Trustpilot*⁸, may contain reviewer metadata like, e.g., age, gender and location, the work

⁸<https://www.trustpilot.com/>

highlights gender-specific lexical differences, the the distribution of regional markers, spelling variations and the use of grammatical constructions across the reviewers.

The work from [17], which focused on reviews manipulation, exploits reviewer-centric and hotel-centric features to identify outliers: the work compares hotels reviews and related features across different review sites, outperforming the detection of suspicious hotels with respect to check the reviews on sites in isolation. Relying on visualization tools, the authors of [6] highlight suspicious changes on reviews scores, while work in [7] proposes new score aggregators to let review systems robust with respect to injection of fake scores.

Research effort has also being spent to understand which are the factors that let a review perceived as useful: in [15], the authors highlight how the reviewer history is a dominant factor to let a review be voted as useful or not. In [14] propose to use the reviews as a source for demographic recommendations.

In this work we enhance the review dataset with additional features based on characteristics of the reviewer (e.g., gender) and the hotel (e.g., popularity and the neighbourhood). On the contrary, work in [18] studies how, independently from the type of service or the type of reviewer, the scores may be affected by external factors, such as the whether conditions and the daylight length of the service cities. We leverage an extensive experimental campaign, addressing around 190k real reviews, which leads to the provision of statistically sound results. Addressing a large scale of data has been done also in [13], which already has targeted users’ reviews as a rich source of information for sociolinguistic studies. While they achieve correlations between metadata in the reviewers’ profile and the review text to let writing styles emerge, we highlight association evidence among hotels and reviewers features and the reviewer’s attitude to score the hotel.

5 CONCLUSIONS

We focused on hotel reviews to investigate which factors could impact the scores that reviewers assign to hotels throughout the world. First of all we have enriched review data with with novel hotel-centric and reviewer-centric features, obtained for example through linked data information available from the web, then we have applied association rule mining to focus on these features possibly motivating the classification scores.

The approach can help both consumers and providers: the former could achieve a better awareness on how to read the reviews (consumers), the latter on how to improve their services (providers). The providers also can query a very large segment of population, in an automatic way and without relying on standard interviews.

The proposed technique is also applicable to a various range of services: accommodation, car rental, food services, to cite a few. Being association rule mining parametric with respect to the itemsets in input, the approach is easily extensible to further features not considered here, such as, e.g., the service price.

6 ACKNOWLEDGMENTS

This research is partly supported by the EU H2020 Program, grant agreement #675320 (NECS: *European Network of Excellence in Cybersecurity*). Funding has also been received by Fondazione Cassa di Risparmio di Lucca that partially finances the regional project *ReviewLand*. Vittoria Cozza is also supported by the Starting Grants Project DAKKAR (DATA benchmark for Keyword-based Access and Retrieval) promoted by University of Padua, Italy and

Rule	Condition	Confidence	X	X∩Y	Lift	Conviction
r1	{memberActivity=1 country=fr ==> geotransp=low}	0.64	29,837	19,199	3.2	2.24
r2	{gender=male country=us ==> hotelPopularity=very high}	0.59	44,703	26,505	1.78	1.64
r3	{memberActivity=1 country=us ==> guestPics=very high}	0.34	61,155	20,926	1.71	1.22
r4	{memberActivity=1 country=us ==> geoenter=high}	0.54	61,155	20,4316	1.68	1.48
r5	{gender=male country=us ==> hotelTripType=couple}	0.76	44,703	34,192	1.04	1.11
r6	{memberActivity=very low revtripType=family ==> hotelTripType=couple }	0.74	27,343	20,362	1.01	1.02

Table 7: Excerpt of ARs where user features are premises and the consequences the features of selected hotel, results are sorted by decreasing lift

Rule	Condition	Confidence	X	X∩Y	Lift	Conviction
r7	{stars=0 country=None guestPics=very low geotransp=low ==> rating=3}	0.25	9,007	2,214	1.89	1.15
r8	{stars=5 hotelPopularity=medium geofood=very high ==> rating=5}	0.76	2,582	1,962	1.70	2.31
r9	{memberActivity=very low gender=female guestPics=very high hotelTripType=couple geoenter=very high ==> rating=5}	0.7	2744	1918	1.57	1.84
r10	{ stars=3 country=jp ==> rating=4}	0.47	5,265	2,492	1.38	1.25
r11	{memberActivity star=3 guestPics=low ==> rating=4}	0.46	4,326	1,998	1.35	1.22
r12	{star=3 geofood=very high ==> rating=4}	0.44	4,312	1,901	1.28	1.17
r13	{country=jp hotelTripType=business ==> rating=4}	0.44	4,483	1,954	1.27	1.16
r14	{geoenter=very high ==> rating=5}	0.5	37,998	19,120	1.13	1.12

Table 8: Excerpt of CARs, the class is the review rating, results are sorted by decreasing lift

Fondazione Cariparo, Padua, Italy. The first author would like to thank Giorgio Maria Di Nunzio, for his helpful support.

REFERENCES

- [1] Silvana Aciar. 2009. Mining context information from consumer's Reviews. *Proceedings of the Context-Aware Recommender Systems (CARS) Workshop* (2009).
- [2] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.
- [3] Sergey Brin, Rajeev Motwani, Jeffrey D. Ullman, and Shalom Tsur. 1997. Dynamic Itemset Counting and Implication Rules for Market Basket Data. In *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data (SIGMOD '97)*. ACM, New York, NY, USA, 255–264.
- [4] Elian Carsenat. 2013. Onomastics and Big Data Mining. *CoRR* abs/1310.6311 (2013). <http://arxiv.org/abs/1310.6311>
- [5] Fabio Celli, F. Marta L. Di Lascio, Matteo Magnani, Barbara Pacelli, and Luca Rossi. 2010. Social Network Data and Practices: The Case of Friendfeed. In *Advances in Social Computing*. LNCS, Vol. 6007. Springer Berlin Heidelberg, 346–353.
- [6] Alessandro Colantonio, Roberto Di Pietro, Marinella Petrocchi, and Angelo Spognardi. 2015. Visual detection of singularities in review platforms. In *30th Annual ACM Symposium on Applied Computing*, 1294–1295.
- [7] Roberto Di Pietro, Marinella Petrocchi, and Angelo Spognardi. 2014. A Lot of Slots - Outliers Confinement in Review-Based Systems. In *Web Information Systems Engineering Part I*. 15–30.
- [8] Ruihai Dong and Barry Smyth. 2016. From More-Like-This to Better-Than-This: Hotel Recommendations from User Generated Reviews. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization (UMAP '16)*. ACM, New York, NY, USA, 309–310.
- [9] Michela Fazzolari, Vittoria Cozza, Marinella Petrocchi, and Angelo Spognardi. 2017. A Study on Text-Score Disagreement in Online Reviews. *Cognitive Computation* 9, 5 (01 Oct 2017), 689–701.
- [10] Andrew J. Flanagin, Miriam J. Metzger, Rebekah Pure, Alex Markov, and Ethan Hartsell. 2014. Mitigating risk in e-commerce transactions: perceptions of information credibility and the role of user-generated ratings in product quality and purchase intention. *Electronic Commerce Research* 14, 1 (2014), 1–23.
- [11] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter* 11, 1 (2009), 10–18.
- [12] Jochen Hipp, Ulrich Güntzer, and Gholamreza Nakhaeizadeh. 2000. Algorithms for Association Rule Mining: a General Survey and Comparison. *SIGKDD Explor. NewsL.* 2, 1 (June 2000), 58–64.
- [13] Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User Review Sites As a Resource for Large-Scale Sociolinguistic Studies. In *24th International Conference on World Wide Web (WWW '15)*. 452–461.
- [14] Nikolaos Korfiatis and Marios Poulos. 2013. Using online consumer reviews as a source for demographic recommendations: A case study using online travel reviews. *Expert Systems with Applications* 40, 14 (2013), 5507 – 5515.
- [15] Asher Levi and Osnat Mokryn. 2014. The Social Aspect of Voting for Useful Reviews. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. LNCS, Vol. 8393. Springer International Publishing, 293–300.
- [16] Bing Liu, Wynne Hsu, and Yiming Ma. 1998. Integrating Classification and Association Rule Mining. In *KDD*. 80–86.
- [17] Amanda J. Minnich, Nikan Chavoshi, Abdullah Mueen, Shuang Luan, and Michalis Faloutsos. 2015. TrueView: Harnessing the Power of Multiple Review Sites. In *24th International Conference on World Wide Web (WWW '15)*. 787–797.
- [18] Syeda. Rahman, Tazin Afrin, and Don Adjeroh. 2015. Determinants of User Ratings in Online Business Rating Services. In *Social Computing, Behavioral-Cultural Modeling, and Prediction*. LNCS, Vol. 9021. Springer International Publishing, 412–420.
- [19] Koji Takuma, Junya Yamamoto, Sayaka Kamei, and Satoshi Fujita. 2016. A Hotel Recommendation System Based on Reviews: What Do You Attach Importance To?. In *Fourth International Symposium on Computing and Networking, CANDAR 2016, Hiroshima, Japan, November 22-25, 2016*. 710–712.