TrueView: Harnessing the Power of Multiple Review Sites

AMANDA MINNICH, NIKAN CHAVOSHI, ABDULLAH MUEEN, SHUANG LUAN, MICHALIS FALOUTSOS

UNIVERSITY OF NEW MEXICO, USA
1. Who read hotel reviews before booking?
2. Were you concerned about fake reviews?

$611,000 fine as TripAdvisor gets bad review in Italy

By Barry Neild, CNN
Updated 1423 GMT (2223 HKT) December 23, 2014

The regulator complained that people reading TripAdvisor Italy were unable to distinguish between genuine and fake reviews posted on the site. It said both were presented by TripAdvisor as "authentic and genuine in nature."
Writing fake reviews is a business!

Need reviews on tripadvisor.com

This project received 7 bids from talented freelancers with an average bid price of $149 USD.

We can help your business grow with positive reviews of your products.

We're a professional team, employing article writers and bloggers ready to post reviews.

Buy 20 positive reviews for iphone app

This project was successfully completed by Arkhu for $128 USD in a day.

Buy positive reviews for your business – Google Place, Yellowpages, InsiderPages, etc.
Our goal: leverage multiple review-sites to combat review fraud
Our contribution: TrueView

- Collect massive multi-site dataset: 15.5 million reviews from 3 sites
- Introduce novel features for cross-site comparison
- Propose new combined metric: TrueView
- Conduct study to detect anomalous hotel behavior
How TrueView Could Be Used in Practice

Hotel A
- **TrueView score: 0.25**
- 1,630 reviews | #14 of 460 hotels in Florence
- #1 Just for You
  - Near Giotto’s Bell Tower, Centrally... [more]
- "Hidden in the Center of Florence" 05/08/2015
- "Silver wedding celebration" 05/08/2015

Hotel B
- **TrueView score: 0.89**
- 2,101 reviews | #1 of 460 hotels in Florence
- #2 Just for You
  - Offers free wifi, Near The Basilica of San... [more]
- "Hotel B, One of Europe’s Great Hotels" 05/09/2015
- "Believe what you read!" 05/08/2015

Hotel C
- **TrueView score: 1.0**
- 1,143 reviews | #74 of 460 hotels in Florence
- #3 Just for You
  - Upscale, Popular romantic hotel in Florence... [more]
- "Great in many ways" 05/06/2015
- "Nicely appointed, centrally located hotel in Florence!" 05/06/2015

Hotel B
- **TrueView score: 1.0**
- 2,101 reviews | #1 of 460 hotels in Florence
- #3 Just for You
  - Upscale, Popular romantic hotel in Florence... [more]
- "Great in many ways" 05/06/2015
- "Nicely appointed, centrally located hotel in Florence!" 05/06/2015

Hotel A
- **TrueView score: 0.25**
- 1,630 reviews | #14 of 460 hotels in Florence
- #1 Just for You
  - Near Giotto’s Bell Tower, Centrally... [more]
- "Hidden in the Center of Florence" 05/08/2015
- "Silver wedding celebration" 05/08/2015
Our Data, now publicly available

- 3,173,346 reviews
- 46,095 hotels
- 1,426,252 reviewers

- 9,050,133 reviews
- 155,764 hotels
- 1,020,054 reviewers

- 3,234,834 reviews
- 63,816 hotels
- 1,143,085 reviewers

Total
- 15.5 million reviews
- 270k hotels
- 3.5 million reviewers

Anonymized subset of data available at: cs.unm.edu/~aminnich/trueview
Roadmap

- Disambiguate hotel identity
- Introduce novel features
- Detect outlier hotels
- Calculate TrueView Score
Roadmap

- Disambiguate hotel identity
- Introduce novel features
- Detect outlier hotels
- Calculate TrueView Score
Is this the same hotel?

Hotel Atlantico Buzios Convention & Resort - Estrada da Usina, s/n, Humaita, Buzios, CEP 28950-000

vs.

Hotel Atlantico Buzios Convention and Resort, strada da Usina 294Morro do Humait, Buzios, RJ, 28950-000, Brazil
Challenges of hotel disambiguation

- Hotel names are not unique
- Address formats are not standard across websites
- Getting geo-location not easy
Our approach: Disambiguating hotels across sites

Text similarity of hotel names + Distance between hotels

- Text similarity alone gives high precision but small number of matches
- Adding in distance allows us to relax similarity requirement
Roadmap

- Disambiguate hotel identity
- Introduce novel features
- Detect outlier hotels
- Calculate TrueView Score
Features: 142 total, 52 novel cross-site

- **We use 142 features across many dimensions:**
  - Spatio-temporal
  - Distribution-based
  - Graph-based
  - Behavioral
  - Text-based

- **We introduce 52 novel cross-site features**
  - To capture variations of features across sites

Is the feature discrepancy anomalous?
Single-site suspicious feature behavior:
Burst in reviews increases overall rating
Single-site suspicious feature behavior:

Many 5-star reviews follow a 1-star review

Why?
First impressions matter!
Single-site suspicious feature behavior:
Many reviews in same zip code is suspicious

~70 reviews left in a concentrated area, mostly 4 and 5 star
Single-site suspicious feature behavior:
Many reviews on same date, with repetitive text, is suspicious

<table>
<thead>
<tr>
<th>Date</th>
<th>Location</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 22, 2006</td>
<td>Newark: SpringHill Suites Newark Liberty</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>International Airport: Great Hotel. Beautiful Rooms.</td>
<td></td>
</tr>
<tr>
<td>January 22, 2006</td>
<td>Rochester: SpringHill Suites Mayo Clinic Area/St. Mary's: Great Hotel. Beautiful Rooms.</td>
<td>4.5</td>
</tr>
<tr>
<td>January 22, 2006</td>
<td>Eagan: TownePlace Suites Minneapolis-St. Paul Airport/Eagan: Great Hotel to Stay In.</td>
<td>4.5</td>
</tr>
<tr>
<td>January 22, 2006</td>
<td>Burnsville: Fairfield Inn &amp; Suites Minneapolis</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>Burnsville: Great Hotel. Beautiful Rooms.</td>
<td></td>
</tr>
<tr>
<td>January 22, 2006</td>
<td>Minneapolis: Minneapolis Marriott West: Great hotel. Wonderful staff. Awesome service.</td>
<td>4.5</td>
</tr>
<tr>
<td>January 22, 2006</td>
<td>Wichita: Marriott Wichita: Great Hotel. Room was small for a Marriott.</td>
<td>4.5</td>
</tr>
<tr>
<td>January 22, 2006</td>
<td>Bismarck: Radisson Hotel Bismarck: Great Hotel. Beautiful Rooms.</td>
<td>4.5</td>
</tr>
<tr>
<td>January 22, 2006</td>
<td>Billings: SpringHill Suites Billings: Great Hotel. Beautiful Rooms.</td>
<td>4.5</td>
</tr>
<tr>
<td>January 22, 2006</td>
<td>Tucson: Sheraton Tucson Hotel and Suites: Great Hotel. Beautiful Rooms.</td>
<td>4.5</td>
</tr>
</tbody>
</table>

32 reviews left on same day across different states with repetitive text
Cross-Site Features:
Comparing review time-series - Seasonality

Summer Peaks

Lack of seasonal peaks for Hotels.com
Cross-Site Features:
Comparing feature distributions

Negative correlation between rating distributions is suspicious
Other Cross-Site Features

Scalar ratios
- Average ratings
- Number of 5 star ratings, etc.
- Number of reviews
- Number of empty reviews
- Bursts
- Oscillation
- Text similarity

Correlation
- Integer rating distribution
- Review per day time series
- Mann Whitney U test p-value
- Floating point rating distribution
- Length of review body
- Length of review title
Roadmap

- Disambiguate hotel identity
- Introduce novel features
- Detect outlier hotels
- Calculate TrueView Score
Use ensemble of outlier detection algorithms to capture anomalous behavior

1. Global Density-Based Score
   • Assumes normal hotels form core cluster
   • Calculates score using density connectedness

2. Local, Density-Based Score (LOF)
   • Allows for several normal clusters
   • Average ratios of local reachability of a point and its $k$-nearest neighbors

3. Global Distance-Based Score
   • Cluster data hierarchically to produce dendrogram
   • Any hotels in clusters below cutoff are outliers
Roadmap

- Disambiguate hotel identity
- Introduce novel features
- Detect outlier hotels
- Calculate TrueView Score
Calculating TrueView

1. Transform outlier scores to trustworthiness probabilities
   i. Logarithmic transform: \[ z = -\log S(x) / S_{\text{max}} \]
   ii. Gaussian scaling: \[ TV(x) = \max(0, \text{erf}\left(\frac{z - \mu_z}{\sqrt{2}\sigma_z}\right)) \]

2. Combine all three trustworthiness scores:

\[
\text{TrueView} = \frac{TV_{\text{global\_density}} + TV_{\text{LOF}} + TV_{\text{hierarchical}}}{3}
\]

Recap of TrueView

Novel Features

90 Single-site Features
- Hotel-centric
- Reviewer-centric
- Review-centric

52 Cross-site Features
- Normalized Ratio
- Pearson’s Correlation
- Mann Whitney U test

Outlier Detection: Three methods

Global Density-Based Score
Local Density-Based Score
Global Distance-Based Score

Calculating Trustworthiness: TrueView

Logarithmic transform
Gaussian scaling
Linear inversion
Average

TrueView score
Our cross-site features are important

We use spectral feature selection to identify important features.
Sanity check: TrueView captures “outlier features”

- Hotels with high TrueView scores have low numbers of outlier features
- Hotels with low TrueView scores have high numbers of outlier features

Top 100

Bottom 100
Case-study: TrueView in practice
Many anomalous features = low TrueView score

1. Discrepancy in rating distribution
2. Suspicious temporal bursts that increase average rating
3. Many empty reviews are suspicious

TrueView score: 0.26

Score breakdown (from 72 reviews)

<table>
<thead>
<tr>
<th>Rating</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Excellent</td>
<td>4</td>
</tr>
<tr>
<td>Very good</td>
<td>14</td>
</tr>
<tr>
<td>Average</td>
<td>7</td>
</tr>
<tr>
<td>Poor</td>
<td>7</td>
</tr>
<tr>
<td>Terrible</td>
<td>11</td>
</tr>
</tbody>
</table>

Comfort Inn & Suites, Statesboro, GA 30458
Summary of work

- **Collect massive multi-site dataset:** 15.5 million reviews from 3 sites spanning >5 years
- Introduce **novel features for cross-site comparison**
- Propose new combined metric: **TrueView**
- Conduct study to detect **anomalous hotel behavior**
Future Work

- Expand work to other well-defined and persistent services
  - Restaurants
  - Tourist attractions
- Generate a comprehensive validation set
- Improve hotel identity disambiguation
Thank You!

INFORMATION ABOUT THE PROJECT, DATA, AND CODE:
CS.UNM.EDU/~AMINNICH/TRUEVIEW
EMAIL: AMINNICH@CS.UNM.EDU