Text and Sentiment Analysis for Yelp Restaurants

Lau Wen Yang Aldred, Michelle Teo Sok Lee, Ng Hui Ying, Tan Yi Hao
Aldred.lau.2012@business.smu.edu.sg, michelleteo.2012@sis.smu.edu.sg, huiying.ng.2012@business.smu.edu.sg, yihao.tan.2012@sis.smu.edu.sg

Abstract – In this paper, text analysis was done on Yelp restaurant reviews to uncover latent subtopics. The purpose is to identify preferences and attributes desired by consumers within a large amount of reviews. These topics would then provide useful insights to restaurants on the taste and preference of Yelp users to increase their ratings and drive demand which is directly related to their revenue. The specific dataset used was the open dataset from the Yelp Dataset Challenge with approximately 1.1mil reviews. To uncover latent subtopics from reviews, we employed Latent Semantic Indexing to break down reviews into singular values to form our text corpora. We then present the distribution of latent topics over the reviews while assignment sentiment scores and extend our findings to correlate business star ratings. In summary, we have found several interesting insights which could prove to be beneficial to both existing and new business owners.

Index Terms – Restaurant Ratings, Text and Sentiment Analysis, Yelp Dataset Review, Latent Dirichlet Allocation

INTRODUCTION & PROBLEM OVERVIEW

Yelp is a ratings and recommendation portal for businesses to connect with their users. Our project uses Yelp's data amassed over 10 years to uncover latent subtopics within reviews.

Yelp began its operations in 2005 as a medium for consumers to provide reviews and ratings for local businesses. Businesses organize their information on the website providing basic information while users rate the businesses from 1-5 stars and provide their own reviews based on experiences. The purpose was to help consumers have a reliable source of information and testimony when deciding to patronize businesses. To further strengthen the interaction within the Yelp community, a meta system was developed where users can vote on helpful, cool or funny reviews.

After 10 years of operations, Yelp has amassed a huge amount of raw data on businesses. However, while businesses are able to view their ratings and reviews, as the amount of data increases and more businesses are added into the system, businesses become increasingly curious of how their businesses fare against competitors. How can they improve their businesses to match restaurants that are doing well? What are the most popular attributes which cause consumers to patronize again or award high ratings? Existing or new businesses are also curious to find out how their businesses have fared over time and are there new areas to explore and enter locations which have not been saturated. With this knowledge, a business owner may then know areas to improve in their business or find out the benchmarks within their competitors. They will also be able to identify new opportunity within different locations as well as introduce new items or services to attract more customers.
In this paper, we present our efforts to explore and dissect this raw information in order to analyse the data and provide actionable insights to businesses. Rather than predicting the possible rating of a business, we instead look at identifying key attributes which are important to consumers to help businesses improve their offerings. Specifically, we also observe how businesses have changed in the perceptions of consumers over time by identifying the change in the sentiment of reviews. The former tells us if the business is performing up to standards that consumers expect for a popular and good business. The latter describes the trend that the business has followed over the years and is able to compare against its competitors on how it has fared against its competitors.

A Yelp dataset has been used to try and learn attributes or features related to high, average and bad restaurants. The scope of the data has been limited to restaurants located within the following states: PA, NC, NV, WI and AZ. The provided raw data set is rather unstructured with many empty fields of various attributes.

The majority of the project focuses on text and sentiment analysis where we attempt to identify key attributes and features that consumers are looking out for in popular businesses and features which contribute to bad businesses. We describe two methods: manipulation and cleaning of the data set and sentiment analysis on user reviews. In our experimental section, we describe some of the key features which were important and the specific words used by consumers as well as how our classifier model performed against an in-built classifier within existing packages.

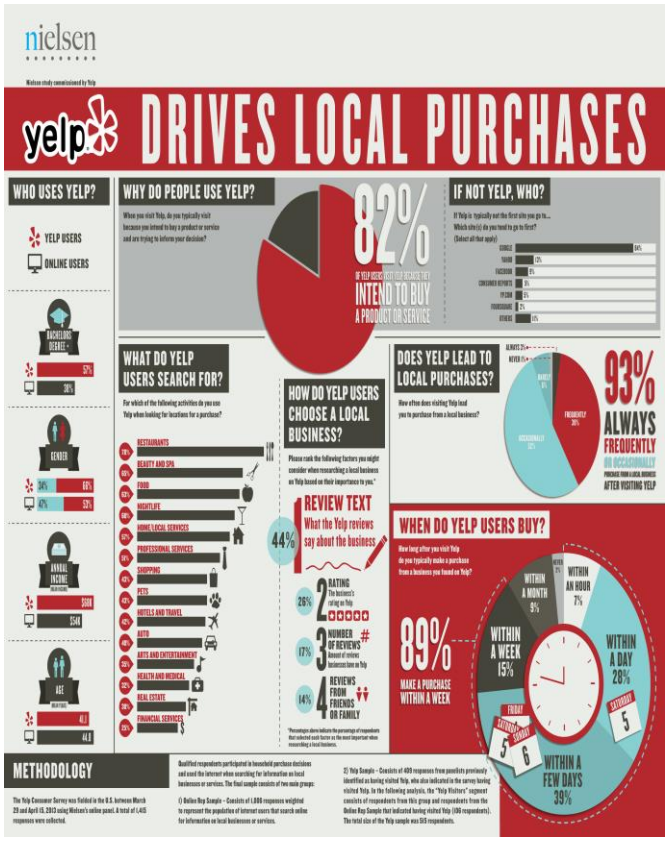


FIG 1. NIELSEN RESEARCH ON YELP

DATA DESCRIPTION

2.1 29136 Business Records
'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)]

2.2 1121776 Reviews
'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'),

GENERATED FEATURES

In this section, two methods that was used to generate additional features will be shared. Firstly, the process of data manipulation will be shared to create organized business categories. Next, the method to generate sentiment for the text reviews as well as how opinions were extracted to form into usable feedback which can be easily processed by businesses.

Business Categories

In the exploratory stages of the data set, it was revealed that there were 99 different restaurant categories which would prove to be too segmented to analyze. Hence, it was decided that the business categories would be re-categorized into 5 categories that would be as exhaustive while accounting for minimal overlap of business categories in order to identify distinctive preferred attributes for different restaurant categories. This would be useful in answering the question for business owners where they would be able to know the benchmarks within their restaurant category as well as compared their performance against competitors.

The final data set re-categorized the restaurant categories into the following:

1. BreakfastAndBrunch/Dinner
2. Bakeries
3. Bars
4. CoffeeAndTea
5. Dessert

Generating Review Features

Raw review text can be processed and utilized in many different ways - n-gram analysis, keyword associations and extractions, and sentiment analysis. For this project, the sentiment analysis described in [Hu and Liu 2004] was used. The objective was to mine the most frequently occurring keywords among all restaurant reviews and use counts of the sentiments for each of these keywords as features. The assumption is that restaurants with many positive statements about a particularly popular keyword would provide an insight into important features that distinguished their business from others.

Generating features from the review text began with cleaning up the text (spell checking, removing excessive punctuation, etc.). Next, each review was tokenized into sentences and then into words. Using the Python Natural Language Toolkit (NLTK), we then use part-of-speech (pos) tag for each token in the sentence.

Next, an algorithm that contained a feature was used to extract top occurring words and split into adjectives (labelled 'JJ'), nouns (labelled 'NN') and verbs amongst the different states and categories and similarities were merged. Words which had no meaning were deemed as stopwords (e.g. a, the) and were removed altogether.

Following that, the words were furthered passed through an algorithm to generate various bigrams and trigrams to

provide more specific feature identification which business owners could act upon.

Then, each bigram and trigram was assigned an over-arching feature which covered the latent topics within the text reviews after analysing the n-grams. The topics identified were:

1. Taste
2. Accessibility
3. Service
4. Ambience
5. Variety
6. Others

Training a Sentiment Classifier

Based on research, it was discovered that the in-built Naive Bayes classifier within the Python Textblob package was trained using the movie reviews data. As such, it was decided that it was essential to train a classifier based on food restaurant reviews as it was deemed to be more appropriate. Therefore, the Naive Bayes classifier within the Textblob package was utilized and the group proceeded to manually rate 5000 sentences as a training data set and 500 sentences as the test data. The training dataset consisted of 1000 reviews from the PA states from each restaurant category as it was decided that this would be sufficient for a training data set. The assumption was also that the expressions used amongst the different states would be similar. This was compared to existing in-built classifier within the package.

Key Attribute Extraction

After extracting a representative set of keywords used to describe restaurants, the number of positive and negative opinions about each keyword was assigned and a sentiment intensity score of between 1 and 3 was subsequently assigned as well in order to generate a topic attribute about reviews.

velvet cupcake 3	food name
delicious donuts 3	taste
white chapel 3	
lime cake 3	food name
best macarons 3	taste
best croissants 3	taste
awesome place 3	ambience
vegan options 3	variety
good french 3	taste
good bread 3	taste
thai chili 3	food name
best pastry 3	taste
chocolate raspberry 3	food name
great service 3	service

Manually Categorized into themes

```
Bigram,Freq,Category,Polarity,Score
happy hour,204,service,Pos,1
great service,49,service,Pos,2
nice staff,32,service,Pos,2
free wifi,30,service,Pos,1
good service,25,service,Pos,2
super fast,22,service,Pos,3
friendly staff,21,service,Pos,2
super friendly,19,service,Pos,3
friendly atmosphere,17,service,Pos,2
excellent service,16,service,Pos,3
friendly place,16,service,Pos,2
fast delivery,14,service,Pos,2
```

Rated 1-3 within themes

After this, the bigrams was used as a dictionary corpus and the process was reiterated dynamically to rate the trigrams in a similar way and the assign the attributes. These bigrams and trigrams were then used in a word cloud visualisation to help business owners identify the critical aspects of their businesses which could be improved or resonated with consumers.

Accuracy of Classifier Model

As mentioned in the training of a sentiment classifier, the purpose to generate a corpus of words and phrases using the restaurant reviews which would be more suitable and representative of the data set. However, the results of the model fell short of the accuracy of the in-built classifier.

	In-built Classifier	Trained Classifier
Hit Rate	79.656%	71.920%

The results revealed that the trained model was about 8% less accurate and the initial suspicion was that the training data set was much smaller compared to training data set for the in-built classifier. Hence, additional 1000 data points were rated to test our hypothesis. However, it revealed that the accuracy dropped to 70.856%. This suggested that there might have been many rows of data within the training data set that was similar in content and the broadness of training data set was not comparable to the training data for the in-built classifier. In order to improve accuracy beyond the in-built classifier, it might require more than 10,000 data points which was not possible given the fixed time frame for the project. As such, the in-built classifier was used to assign the sentiment for all the reviews and n-grams for analysis.

Sentiment Score Calculation and Bigram Importance

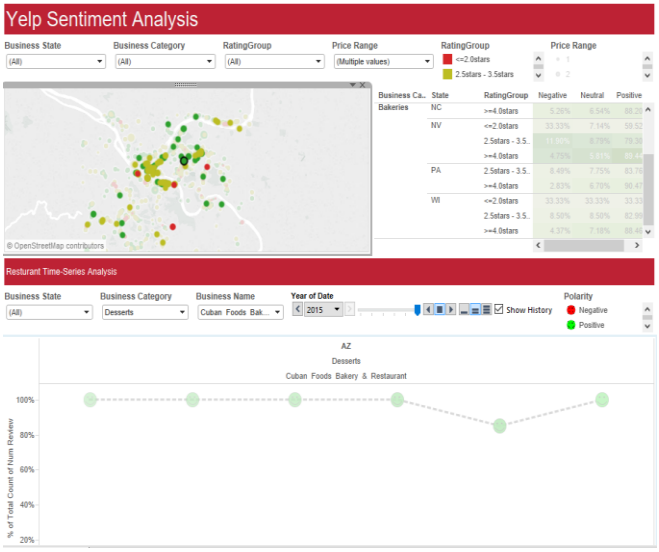
In order to improve the usefulness of the bigrams, the importance of the bigram was calculated using the following formula:

Importance = Frequency * Sentiment Intensity Score

This formula was calculated with the assumption that the more the bigram appears, and the higher the sentiment score, the more the importance of the bigram. With the importance score, the restaurant owners can accurately identify the bigrams with high positive and negative importance. This would be critical to identify, in customers' terms, what aspect are important in liking or hating a business.

VISUALIZATIONS

General Time Trend Analysis



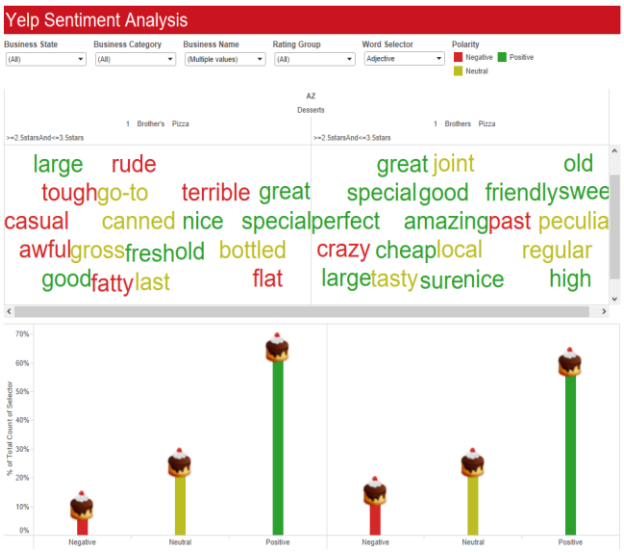
By creating the following dashboard for time trend analysis, it helps existing and new business owners identify areas of opportunities. On the geospatial visualization, the data points are restaurants and have been colour coded based on their ratings. This would help business owners identify popular restaurants and bad restaurants at a glance. The size of the data points represents the price range of restaurants as it was determined to be one of the most important factors of both owners and consumers when choosing restaurants [Matthews, 2011].

Using the dashboard, new businesses could choose to enter areas which are either less populated by popular restaurants or are lacking in a particular restaurant category. This would help them to fit in more easily and capture a different customer segment while minimizing competition.

The sentiment data table beside the map would also allow consumers and restaurants to determine at a glance which states and restaurant category are not performing as well based on the composition of sentiment reviews. A high percentage of reviews in negative reviews would imply an opportunity to be explored and vice versa.

It is also possible for business owners to make use of the 2nd segment of the dashboard to observe how their restaurant or competitors have performed since inception. The line graph displays the composition of positive, negative and neutral reviews over the entire period of operation where reviews were made. An improvement in the positive reviews percentage would signify an improving restaurant while a decrease would imply be an alert signal for restaurants.

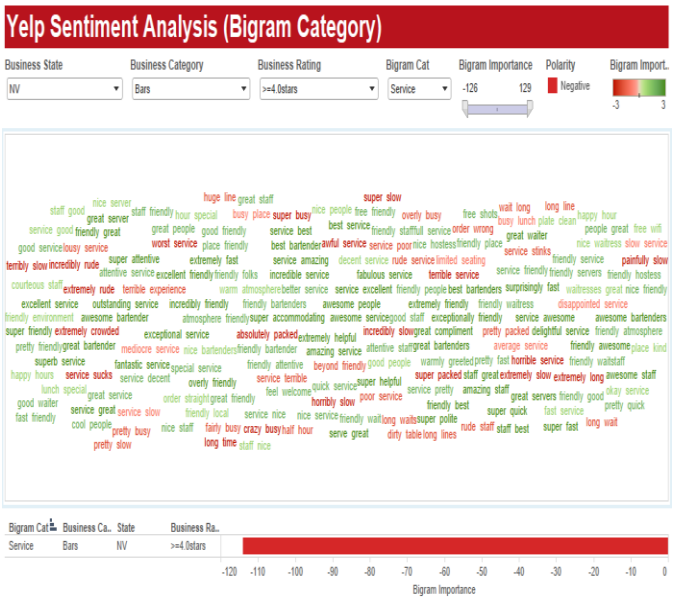
Competitor Analysis Visualization



The second dashboard reveals information for competitor analysis. By filtering on the different parameters of state, restaurant category and restaurants, business owners are able to do a side-by-side comparison of the distinguishing features between restaurants. The green represents positive while red represents negative and the neutral is yellow. Some words may appear to be classified wrongly because of the accuracy of the classifier and will be further discussed in future work.

The bar chart below the word cloud shows the composition of the sentiment of reviews for the restaurants and owners can base this to compare how there are performing with other restaurants.

Desired Common Attributes



Overview of Restaurant Categories in States

For business owners who are interested in looking at the restaurants from a more macro view, a dashboard that was created in our preliminary stages of exploratory can be used.

In this dashboard, restaurant owners could find out the top bigrams based on their frequency to find out specific attributes which are critical in the specific restaurant category within the specific states. This is further dissected into different business rating categories which will allow business owners to identify the difference between popular and less popular businesses and their distinguishing attributes.

FINDINGS

Common Desired Attributes

Accessibility

Good location, free/good parking, neighborhood location, wheelchair friendly, convenience

Ambience:

Bakeries: Beautifully decorated, great place/atmosphere, clean place, great concept, cozy place/atmosphere
Bars: Great outdoor/patio, live band, cool looking, beautifully decorated, great view
Breakfast & Brunch: Cozy atmosphere, clean, live jazz, great music, awesome décor
Coffee & Tea: free wireless, great vibe, live/nice music, comfortable chairs, nice/pleasant experience
Dessert: High class, great atmosphere, beautiful decoration, great outdoor/patio, free wireless

Service

Super friendly, super fast, super attentive, incredible service, great servers

Taste

Bakeries:
Food items: cupcakes, coffee, pastries, sandwich, cookie
Taste attributes: delicious, fresh, tasty
Bars:
Food items: drink (beer/draft/cocktails/wine), fish, wings, burgers, steak
Taste attributes: cooked perfectly, great quality
Breakfast & Brunch:
Food items: coffee, pancakes, sandwich
Taste attributes: delicious, French, perfectly cooked
Coffee & Tea:
Food items: chai, coffee (latte, espresso, cappuccino), chocolate, sandwich
Taste attributes: delicious, amazing, high quality
Dessert:
Food items: fish, drink, burger, pizza, chicken
Taste attributes: nice, done/cooked perfectly, best quality

Variety

Bakeries: Excellent/wide/huge selection/variety, gluten free, healthy options, daily special
Bars: Excellent/wide/huge selection/variety, great specials, creative/special/unique menu
Breakfast & Brunch: Great/huge choices/menu, healthy options, specials, gluten free
Coffee & Tea: Excellent/wide/huge selection/variety, healthy options, whole foods, daily special
Dessert: Wide array, extensive, assortment, delicious options

The common desired attributes were derived from the word cloud of bigrams. Given the importance score and the range selector for the importance score, one is able to derive common desired attributes using the positive scores.

Usefulness of Calculated Importance

However, bigrams with the top importance score might not always be the most useful to a business. Bigrams such as "great shop" or "amazing taste" might appear frequently and have a high sentiment score, however does not show any insight for a business to improve. Extracting the list of attributes required scrolling to the bigrams of medium importance.

Another finding is that it is important to accurately categorize the business categories. This is due to the limitation of many overlapping businesses with different restaurant categories. Regarding the desserts category, the food items mentioned were: fish, drink, burger, pizza and chicken, which are not very representative of desserts.

RELEVANT RESEARCH

Text and sentiment research builds upon an increasingly popular field of machine learning applied to complex, structured and unstructured datasets for predictive and prescriptive analytics. We describe a subset of these approaches that are directly related to, and inspired, our proposed work.

Using low-dimensional vectors as latent factors has been very common in recommendation systems. The task of suggesting items to users is usually viewed as matrix completion where the sparse rating matrix with users as rows and items as columns is completed with the prediction of business ratings. [Sarwar et al., 2000] showed how Singular Value Decomposition (SVD) can be used to decompose the rating matrix into low rank feature matrices to reduce dimensions of the rating matrix. This led to the creation of a matrix method [Koren et al., 2009] that hypothesize that the user and item factors can capture similarities between them, which can be used to predict ratings. This inspired us to use attributes as factors to distinguish the difference of ratings amongst businesses and offer recommendations based on specific attributes which businesses can work on to improve their ratings.

Another research which triggered out work is the Latent Dirichlet Allocation (LDA) factor model which utilizes the unsupervised learning of factors and topics for the Yelp restaurant review data. This model treats the probability distribution of each document over topics as a K-parameter hidden random variable rather than a large set of individual parameters (K is the number of hidden topics). Alternatively, there has been successful clustering of terms and text documents using non-negative matrix factorization techniques; such as the factoring of 90,000 terms in e-mails to 50 clusters. This would have been a different approach to a latent class model approach that we could have explored.

However, in our work, the focus will be to provide recommendations to businesses to improve their ratings based on feedback by users. Therefore, because text is unstructured data with no well-defined values like attributes, it has to be divided and converted into a structured representation. In this process, rather than using the words, the root forms of words are created using a technique called stemming (Porter M, 1980) where the goal is to create an umbrella representation behind similar words through categorization of n-grams by assigning attributes. The n-grams are then assigned with a value which is indicative of its importance or relevance. This term is the tf*idf weight (term-frequency times inverse document frequency).

$$w(t,d) = \frac{tf_{t,d} \log\left(\frac{N}{df_t}\right)}{\sqrt{\sum_i (tf_{t_i,d})^2 \log\left(\frac{N}{df_{t_i}}\right)^2}}$$

The tf*idf weight, w(t,d), of a term t in a document d is a function of the frequency of t in the document (tft,d), the number of documents that contain the term (dft) and the number of documents in the collection (N).

Nevertheless, this representation does not capture the context in which a word is used. It loses the relationship between words in the description. One variant on using words as terms is to use sets of contiguous words as terms. In order to achieve this relevance scale, one method that could be explored is the probabilistic text classification approach such as the Naive Bayesian classifier.

FUTURE WORK

After examining the relevant strengths and limitations of our work, we explore other areas of research which would help to further our research.

To increase the value of the attribute identification, this may be converted into variables for a latent model for prediction of business ratings as well as common topics discussed between users. These topics could be used as a means for user community clustering and using these topics to link up "nearest neighbours" not just based on "items" (restaurants visited and commented) but one level deeper by clustering the similarity based on their preference of attributes. This would provide an even more holistic element to recommender systems as this would delve into the psychology of consumers and understand their specific taste and preference in food or other aspects of the businesses such as ambience, service or variety which is difficult to capture in binary variables in the current Yelp dataset.

In order to create a more accurate classifier, there should also be considerations to continue training the review text using the n-grams generated as a based to assign the sentiment and create a large corpus for the classifier. The improvement in accuracy will also be a more accurate representation of customer's perceptions of businesses and will be reflected in our dashboards.

CONCLUSION

In this paper, we shared our approach to text and sentiment analysis of the restaurant reviews of the Yelp data set. We described the process to extract features which were generated from raw review text.

Based on our dashboards, we have been able to extract attributes which users care about and look out for in their reviews of restaurants. These common desirable attributes have been identified above, and can be looked into by new or existing restaurant owners. Competitor analysis can be done via the word cloud and bar chart, and trend analysis through observation can be done through the geospatial visualization and time chart.

The star ratings of businesses were also generally positively correlated with the overall sentiment of reviews. With sentiment scores assigned to n-grams and computed with the frequency, we were also able to discern which bigrams have a relatively higher importance in influencing the ratings of businesses.

With these ratings of specific subtopics that Yelp users care about, restaurants could sieve out insights on how to improve their businesses or explore new opportunities.

Through future works, we expect to explore more accurate and specific insights, possibly beneficial to other business categories. We also hope to enhance the accuracy of recommender systems of user-items by contributing to latent factor models.

REFERENCES

Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Application of dimensionality reduction in recommender system-a case study. Technical report, DTIC Document, 2000.

B. Murray, . ”Email Surveillance Using Non-negative Matrix Factorization”. Computational and Mathematical Organization Theory 2005

D. Blei, A. Ng, and M. Jordan. ”Latent Dirichlet Allocation.” Journal of Machine Learning Research, 3:9931022, January 2003.

M Hu and B Liu. 2004. Mining opinion features in customer reviews. In
Proceedings of the National Conference on Artificial . . . .

Matthews, J. (2011, August 31). 2011 Food & Health Survey: Consumer Attitudes Toward Food Safety, Nutrition & Health.

Paul et al. ,(1997) Recommender systems Communications of the ACM

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. Computer, 42(8):30–37, 2009.