Prediction and Analysis of Hotel Ratings from Crowd-sourced Data

Fátima Leal[1,3], Benedita Malheiro[2,3], and Juan Carlos Burguillo[1]

[1] EET/UVigo – School of Telecommunication Engineering, University of Vigo, Spain
[2] ISEP/IPP – School of Engineering, Polytechnic Institute of Porto, Portugal
[3] INESC TEC, Porto, Portugal
fatimaleal2@gmail.com, J.C.Burguillo@uvigo.es, mbm@isep.ipp.pt

Abstract. Crowdsourcing has become an essential source of information for tourists and the tourism industry. Every day, large volumes of data are exchanged among stakeholders in the form of searches, posts, shares, reviews or ratings. This paper presents a tourist-centred analysis of crowd-sourced hotel information collected from the Expedia platform. The analysis relies on Data Mining methodologies to predict trends and patterns which are relevant to tourists and businesses. First, we propose an approach to reduce the crowd-sourced data dimensionality, using correlation and Multiple Linear Regression to identify the single most representative rating. Finally, we use this rating to model the hotel customers and predict hotel ratings, using the Alternating Least Squares algorithm. In terms of contributions, this work proposes: (*i*) a new crowd-sourced hotel data set; (*ii*) a crowd-sourced rating analysis methodology; and (*iii*) a model for the prediction of personalised hotel ratings.

Keywords: Crowdsourcing, Expedia, Data Mining, Prediction.

1 Introduction

In the last decades, travelling has changed dramatically due to the evolution and popularisation of information and communication technologies (ICT) as well as mobile devices, namely, smartphones. These devices incorporate on board sensors together with significant computing and communication capabilities, enabling users to generate and share large volumes of data known as crowd-sourced data. Crowdsourcing is, according to Egger *et al.* [6], an outsourcing process supported by ICT and performed voluntarily by a large number of participants.

Not only tourists are continuously sharing on-line information regarding their travel experiences through ratings, reviews, comments, photos or videos, but the Web became the main source of tourism information (hotels, transportation, restaurants, attractions, *etc.*). Increasingly, tourists search on websites, wikis or social networks for information and rely on crowd-sourced information, *e.g.*, ratings, reviews or posts of tourists, for decision making. This crowd-sourced information or electronic Web of Mouth (eWoM), which classifies prior tourist experiences regarding tourism resources, is highly influential since it conditions the

behaviour of future tourist planning and decision making [5]. Additionally, according to Gula (2013) [10], the tourism industry has, not only, adopted Crowdsourcing as the major source of tourist feedback data, but relies heavily on crowd-sourced data analytics to define new business strategies.

In order to process and extract meaningful and timely information from the ever growing volume of tourism-related data, we present a tourist-centred analysis of Expedia hotel data through Data Mining methodologies to discover and predict trends and patterns which are relevant to the tourists and tourism businesses. This work comprised: (*i*) the creation of the Hotel Expedia data set, using the Expedia Application Programming Interface (API)[4]; (*ii*) the design of a data dimensionality reduction methodology, applying Multiple Linear Regression (MLR), to identify the single rating – overall rating – most representative of the guest profile; (*iii*) a best value for money analysis, involving the crowd-sourced overall rating, the official star rating and the room price; and (*iv*) the prediction of unknown hotel overall ratings, using Alternating Least Squares with Weighted-λ-Regularization Regularization (ALS-WR) matrix factorisation. Our contributions include: (*i*) the new crowd-sourced hotel data set; (*ii*) the hotel rating dimensionality reduction methodology; and (*iii*) the hotel rating prediction model.

This paper is organised as follows. Section 2 reviews related work on analysis and prediction of crowd-sourced data. Section 3 introduces Big Data analytics, describing current techniques and trends. Section 4 describes the algorithms used, the experiments performed and the results obtained. Finally, Section 5 provides the conclusions and discusses the outcomes of this work.

2 Related Work

The classification and recommendation of goods or services, taking into account the user preferences, is an important task of any on-line Business-to-Consumer (B2C) Web platform. The crowd-sourced feedback, which is volunteered by costumers typically in the form of ratings or reviews, is used by potential costumers to choose new goods or services and by businesses to suggest relevant products.

In the tourism domain, the crowd-sourced information is growing dramatically. This Big Data scenario has been addressed by several researchers: (*i*) Fuchs *et al.* [8] perform multi-criteria rating analysis from TripAdvisor using Linear Regression; (*ii*) Fang *et al.* [7], Han *et al.* and Wang *et al.* [22,23] apply regression models to textual reviews; and (*iii*) Jannach *et al.* [12] and Chen *et al.* [3] propose different recommendation approaches based on the user ratings.

As far as we know, the collection and processing of crowd-sourced information for profiling (supported by Big Data algorithms) and recommendation is a novel approach for mobile tourism applications whereby the related-work is sparse. While earlier work focused on ratings or reviews processing, here we analyse crowd-sourced data in order to forecast trends and patterns in the tourism

[4] `http://developer.expedia.com/directory`

domain, relying on Big Data approaches. Therefore, in this paper, we use crowd-sourced tourist hotel ratings to make personalised hotel recommendations by predicting unknown hotel ratings and identifying trends in the tourism industry.

3 Tourism Crowd-sourced Big Data Analytics

Big Data encompasses large volumes of data originated by technological development and stored disorderly over time. They are characterised by high volume (quantity of data), high velocity (data creation rate) and high variety (data heterogeneity) [9,24].

In the case of the tourism domain, the tourist behaviour has radically changed. On the one hand, technology provides ubiquitous access to endless collections of tourism-related Web services, *e.g.*, searching, booking or planning, allowing the tourist to autonomously plan a trip based solely on Web resources. On the other hand, tourists generate through crowdsourcing systems large volumes of tourism-related data. Tourism crowdsourcing includes sharing in the form of ratings and reviews (Expedia, Airbnb, Yelp or Booking), tourism wikis (Wikivoyage, Wikitravel or Tourpedia) or general purpose social networks (Facebook) [16]. For businesses, the information shared by tourists is of high relevance due to its influence in the tourist behaviour, namely in the decision making process. Moreover, Akerkar [1] states that Big Data allows companies to create better products and services by gathering information from numerous external sources, *e.g.*, travel companies portals, carriers or social networks. For tourism businesses, the Big Data approach makes comprehensible important travel patterns and, thus, promotes a substantial shift by empowering them with the ability to enhance and personalise the customer travel experience. The power to analyse, find and visualise the highlights underlying tourism-related crowd-sourced data offers businesses and tourists an insight of the market opportunities.

A wide variety of methodologies has been developed and adapted to store, aggregate, manipulate, analyse and visualise Big Data [17]. From the research perspective, Chen & Zhang [2], Chen *et al.* [4] and Kambatla *et al.* [13] provide a comprehensive state-of-the-art, including a detailed definition, problems and challenges to be addressed as well as the current trends, techniques and technologies developed and adopted in diverse Big Data domains. From the commercial perspective, McKinsey Global Institute describes the different Big Data related techniques and technologies currently in use [17]. Big Data techniques comprise Machine Learning, Data Mining, Statistics, Neural Networks, Social Network Analysis, Optimisation and Visualisation approaches.

This work applies Data Mining techniques to tourism crowd-sourced data to analyse and predict hotel ratings. Specifically, we apply regression analysis (MLR) to select the most representative rating (dimensionality reduction), and adopt (ALS-MR matrix factorisation) for rating prediction.

4 Experiments and Results

The off-line experiments were performed with Expedia data and involved the processing of hotel ratings, using *scikit-learn*[5].

4.1 Expedia Data Set

We built a crowd-sourced hotel data set, encompassing hotel ratings and reviews, through Expedia API, using the "Hotel Reviews" and "Hotel Search, Offers, and Info" services. While the "Hotel Reviews" service provides the reviews and ratings of the Expedia customers who stayed at a given hotel, the "Hotel Search, Offers, and Info" service provides the hotel data, such as name, address, geodetic location, images, policies, amenities, *etc.* The API, which implements both Representational State Transfer (REST) and Simple Object Access Protocol (SOAP) Web Service interfaces, is accessible via HyperText Transfer Protocol (HTTP) requests. The responses are returned in JavaScript Object Notation (JSON) in the case of REST or in eXtensible Markup Language (XML) in the case of SOAP. In terms of restrictions, although the API documentation refers download limitations, they are unspecified.

The resulting HotelExpedia data set[6] contains 6030 hotels, 3098 reviewers, including anonymous reviewers, and 381 941 reviews from 10 different locations. Although the data set includes anonymous reviewers and their reviews, we only used crowd-sourced data from the 1089 identified reviewers in all our analyses, *i.e.*, we discarded the anonymous users and their inputs. Each user classified at least 20 hotels and each hotel contains at least 10 reviews. Despite the large variety of information available through the Expedia API, we only collected the hotel and customer reviews data presented in Table 1[7]. This data set was built and used as a case study in the tourism domain.

Table 1. Expedia hotel and customer reviews data.

File	Features
Hotel	`hotelId`, `description`, `latitude-longitude`, `starRating`, `guestReviewCount`, `price`, `amenity`, `overall`, `recommendedPercent`, `cleanliness`, `serviceAndstaff`, `roomComfort`, `hotelCondition`
Reviews	`nickname`, `userLocation`, `hotelId`, `ratingOverall`, `ratingCleanliness`, `ratingHotelCondition`, `ratingService`, `ratingRoomComfort`, `reviewText` and `timestamp`

4.2 Rating Analysis

Expedia is a powerful platform, containing huge volumes of crowd-sourced hotel opinions. Each hotel has multiple customer reviews, including the overall, clean-

[5] `http://scikit-learn.org`

[6] `http://ave.dee.isep.ipp.pt/~1080560/ExpediaDataSet.7z`

[7] The hotel data set contains data from destinations selected according to the number of reviews and popularity level.

liness, hotel condition, service & staff and room comfort ratings, and textual reviews. Algorithm 1 summarises the implemented rating analysis [15].

Algorithm 1 Off-line crowd-sourced rating analysis.

Inputs	Crowd-sourced guest reviews Hotel data
Outputs	MLR regarding the single most representative guest rating (SMRGR) Best value for money analysis Crowd-sourced data temporal evolution Hotel rating prediction based on the SMRGR
Step 1	Identification of the most correlated crowd-sourced ratings and SMRGR
Step 2	MLR regression of most correlated crowd-sourced ratings and SMRGR
Step 3	Best value for money analysis
Step 4	Crowd-sourced data temporal analysis
Step 5	Personalised SMRGR prediction

Multiple Linear Regression (MLR) is typically applied to multivariate scenarios and predicts one or more continuous variables based on other data set attributes, identifying existing dependencies among variables [19]. In order to verify the relation between the different crowd-sourced ratings, we performed a correlation analysis. Then, we applied MLR to validate the correlation results and verify the dependency of the overall rating using the methodology proposed by Leal *et al.* (2016)[15]. This methodology estimates the value of a variable based on a set of other variables. Equation 1 displays the model of the MLR with k regression variables. The parameters β_i ($i = 1$ to k) are the partial regression coefficients, which represent the rate of change of one variable (Y) as a function of the changes of the other (X) [21].

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i \quad (1)$$

In this context, MLR was used to verify if the overall rating could be explained by other ratings, *i.e.*, using the overall rating as a dependent variable. First, we calculated the correlation among the ratings and selected the most relevant variables for the MLR. Table 2 presents the resulting correlation values between the crowd-sourced hotel ratings.

Table 2. Correlation among crowd-sourced ratings.

	Overall	**Cleanliness**	**Hotel Condition**	**Service & Staff**	**Room Comfort**
Overall		0.74	0.81	0.74	0.79
Cleanliness	0.74		0.75	0.61	0.71
Hotel Condition	0.81	0.74		0.63	0.72
Service & Staff	0.75	0.61	0.63		0.59
Room Comfort	0.79	0.71	0.72	0.59	

Then, we performed the MLR using the Ordinary Least Squares (OLS), which estimates the unknown parameters in a linear regression model, and minimising the differences between the observed responses in the data set

and the responses predicted by the linear approximation of the data [18]. The results showed that the cleanliness, hotel condition, service & staff and room comfort ratings were capable of explaining approximately 80 % of the overall rating (R-squared value). Based on these results, we chose the overall rating as the single rating which best represents the crowd-sourced customer feedback, *i.e.*, the guest profile, reducing the rating dimensionality to one.

Analysis of Crowd-sourced Hotel Data First, we compared the overall hotel rating with the hotel star rating, *i.e.*, the crowd wisdom with the official hotel quality classification, in an attempt to verify how disparate they are. Figure 1 plots both ratings (step 3 of Algorithm 1) for a cohort of 752 hotels in Barcelona. In particular, the results show that the crowd-sourced overall ratings tend, with the exception of five stars hotels, to be higher than official star ratings. All locations presented similar results.

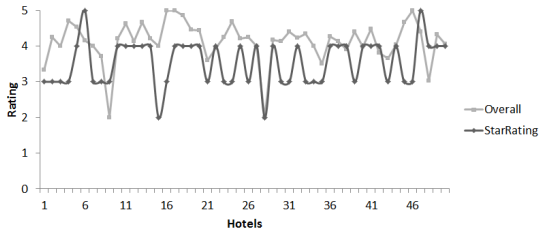

Fig. 1. Star and crowd-sourced overall ratings.

Moreover, we confronted the price with the overall and star ratings to find the best crowd-sourced value for money. Figure 2 depicts the crowd-sourced value for money in the case of Barcelona. This analysis identifies, on behalf of the tourist, the hotels with better value for money according to the crowd.

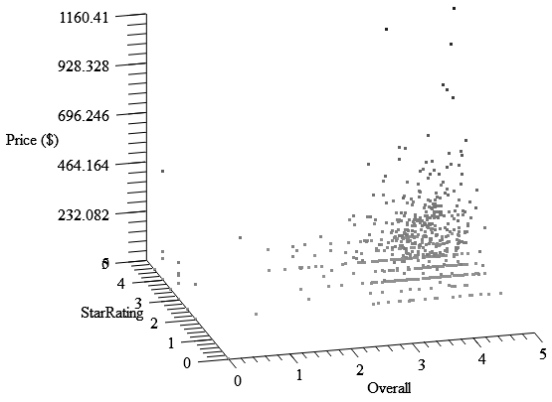

Fig. 2. Crowd-sourced best value for money.

Figure 3 illustrates the linear growth of the volume of crowd-sourced data provided by identified customers, regarding the same cohort of Barcelona hotels. It shows that these users have, steadily and increasingly, been volunteering hotel feedback, *i.e.*, directly influencing the decisions of prospective customers.

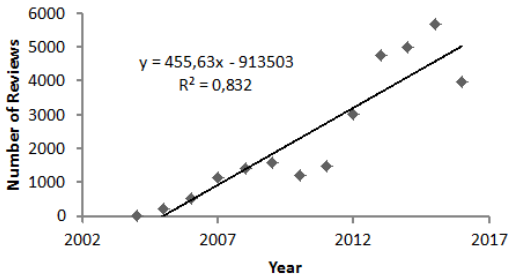

Fig. 3. Crowd-sourced data temporal growth.

4.3 Rating Prediction

The last step of our analysis was the prediction of the overall ratings of the hotels not yet classified by the users. The data set was randomly partitioned into training (80 %) and test (20 %). We implemented a collaborative recommendation filter based on the Alternating-Least-Squares with Weighted-λ-Regularization (ALS-WR) algorithm, using the training data. This algorithm, according to Takács and Domonkos (2012) [20] and Koren *et al.* (2009) [14], provides better results than other matrix factorisation algorithms despite its higher execution time. Zhou *et al.* (2008) [25] and Hu *et al.* (2008) [11] provide further details on the implementation of the ALS-WR algorithm. Algorithm 2 describes our ALS-WR implementation for the prediction of personalised ratings where P and Q represent the user feature matrix and hotel feature matrix, respectively. We defined the regularisation weight λ, the dimensionality of latent feature space (k) and the number of iterations (n) based on the above mentioned research works.

Algorithm 2 Alternating-Least-Squares with Weighted-λ-Regularization

Inputs	`userID`, `hotelID` and `Overall` ratings
Outputs	`userID`, `hotelID` and `Overall` rating predictions
Step 1	Matrix Factorisation with $\lambda = 0.1$, $k = 20$ and $n = 15$
Step 2	Create the P and Q latent matrices
Step 3	Fix Q and estimate P
Step 4	Apply ALS
Step 5	Fix P and estimate Q
Step 6	Apply ALS
Step 7	Calculate prediction matrix
Step 8	Calculate Root-Mean-Square Error (RMSE)

Once the predictions are made, we sort, for each user, the predicted ratings, and recommend the top five non-rated hotels located at the desired destination. Figure 4 plots the Root-Mean-Square Error (RMSE) of the predictions of the

training and test data partitions. In both cases the RMSE decreases monotonically and converges over time to approximately 0.14 (training) and 0.20 (test). The Mean Absolute Error (MAE) is approximately of 0.14 (training) and 0.21 (test). Table 3 presents the top five Barcelona hotel recommendations for a randomly selected user (User 15), including the actual and predicted overall ratings.

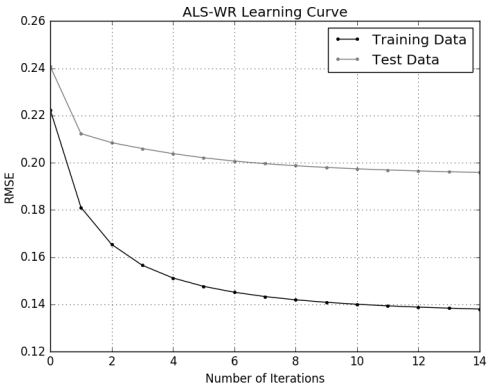

Fig. 4. RMSE with the training and test data.

Table 3. User 15 results

hotelID	Actual	Predicted
2342	5	4.51
3407	5	4.44
3965	4	3.93
2061	3	3.82
1502	4	3.71

5 Conclusions

The emergence of tourism crowdsourcing platforms, *e.g.*, Expedia, allows tourists to continuously produce and share large volumes of feedback data regarding tourism resources. This crowd-sourced information, which classifies prior tourist experiences, influences the behaviour of present and future tourists. However, in face of the resulting Big Data scenario, the tourist is unable to process, relate and visualise the available crowd-sourced information. To address this problem, *i.e.*, to process tourism crowd-sourced data, and to provide the tourist with relevant information for travel planning, we analysed existing Big Data techniques as well as the Expedia crowdsourcing platform. As a result, we designed a methodology which processes off-line crowd-sourced hotel data to forecast rating trends on behalf of the tourists. This paper exploits essentially tourism crowd-sourced information and, with the support of Big Data approaches, provides meaningful information to stakeholders.

In terms of contributions, we created, using the Expedia API, a new crowd-sourced hotel data set, holding hotel and customer reviews data, analysed the multiple crowd-sourced hotel ratings to choose the single most representative rating and implemented the collaborative filtering ALS-WR algorithm to provide personalised hotel recommendations based on the selected rating.

As future work, we intend to: (*i*) use the MLR results to create a combined rating in order to refine the recommendations; and (*ii*) build a reputation model of the data publishers to rate the quality of crowd-sourced contents.

6 Acknowledgements

This work was partially financed by: (*i*) the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness and Internationalisation - COMPETE Programme - within project «FCOMP-01-0202-FEDER-023151» and project «POCI-01-0145-FEDER-006961», and by National Funds through Fundação para a Ciência e a Tecnologia (FCT) - Portuguese Foundation for Science and Technology - as part of project UID/EEA/50014/2013; and (*ii*) ICT COST Action IC1406 High-Performance Modelling and Simulation for Big Data Applications (cHiPSet).

References

1. R. Akerkar. Big data & tourism. Technical report, Technomathmatics Research Foundation, 2012.
2. C. P. Chen and C.-Y. Zhang. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, 275:314–347, 2014.
3. J.-H. Chen, K.-M. Chao, and N. Shah. Hybrid recommendation system for tourism. In *e-Business Engineering (ICEBE), 2013 IEEE 10th International Conference on*, pages 156–161. IEEE, 2013.
4. M. Chen, S. Mao, and Y. Liu. Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209, 2014.
5. Y.-F. Chen and R. Law. A review of research on electronic word-of-mouth in hospitality and tourism management. *International Journal of Hospitality & Tourism Administration*, 17(4):347–372, 2016.
6. R. Egger, I. Gula, and D. Walcher. *Open Tourism: Open Innovation, Crowdsourcing and Co-Creation Challenging the Tourism Industry*. Springer, 2016.
7. B. Fang, Q. Ye, D. Kucukusta, and R. Law. Analysis of the perceived value of online tourism reviews: influence of readability and reviewer characteristics. *Tourism Management*, 52:498–506, 2016.
8. M. Fuchs and M. Zanker. Multi-criteria ratings for recommender systems: an empirical analysis in the tourism domain. In *International Conference on Electronic Commerce and Web Technologies*, pages 100–111. Springer, 2012.
9. A. Gandomi and M. Haider. Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, 35(2):137–144, 2015.
10. I. Gula. Crowdsourcing in the tourism industry—using the example ofideas competitions in tourism destinations. In *ISCONTOUR 2013: Proceedings of the International Student Conference in Tourism Research*, page 147. BoD–Books on Demand, 2013.
11. Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 263–272. Ieee, 2008.
12. D. Jannach, F. Gedikli, Z. Karakaya, and O. Juwig. Recommending hotels based on multi-dimensional customer ratings. In M. Fuchs, F. Ricci, and L. Cantoni, editors, *Information and Communication Technologies in Tourism 2012: Proceedings of the International Conference in Helsingborg, Sweden, January 25–27, 2012*, pages 320–331. Springer Vienna, Vienna, 2012.
13. K. Kambatla, G. Kollias, V. Kumar, and A. Grama. Trends in big data analytics. *Journal of Parallel and Distributed Computing*, 74(7):2561–2573, 2014.

14. Y. Koren, R. Bell, C. Volinsky, et al. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
15. F. Leal, J. M. Dias, B. Malheiro, and J. C. Burguillo. Analysis and visualisation of crowd-sourced tourism data. In *Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering*, C3S2E '16, pages 98–101, New York, NY, USA, 2016. ACM.
16. F. Leal, B. Malheiro, and J. C. Burguillo. Recommendation of tourism resources supported by crowdsourcing. In *ENTER 2016 PhD Workshop, International Conference on Information and Communication Technologies in Tourism 2016*, pages 18–25, 2016.
17. J. Manyika, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A. H. Byers. *Big Data: The Next frontier for Innovation, Competition, and Productivity.* McKinsey Global Institute, 2011.
18. M. Stone and R. J. Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society*, pages 237–269, 1990.
19. A. O. Sykes. An introduction to regression analysis. In E. A. Posner, editor, *Chicago Lectures in Law and Economics*. Foundation Press, New York, 2000.
20. G. Takács and D. Tikk. Alternating least squares for personalized ranking. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 83–90, New York, NY, USA, 2012. ACM.
21. M. Tranmer and M. Elliot. Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, 2008.
22. H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 783–792, New York, NY, USA, 2010. ACM.
23. H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis without aspect keyword supervision. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 618–626, New York, NY, USA, 2011. ACM.
24. H. J. Watson. Tutorial: Big data analytics: Concepts, technologies, and applications. *Communications of the Association for Information Systems*, 34(1):1247–1268, 2014.
25. Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan. Large-scale parallel collaborative filtering for the netflix prize. In *International Conference on Algorithmic Applications in Management*, pages 337–348. Springer, 2008.