A Novel Recommendation Framework for Micro-blogging based on Information Diffusion

Aaron R. Sun, Jiesi Cheng [*]and Daniel D. Zeng [†]

Abstract. Micro-blogging is increasingly extending its role from a daily chatting tool into a critical platform for individuals and organizations to seek and share real-time news updates during emergencies. However, extracting useful information from micro-blogging sites poses significant challenges due to the volume of the traffic and the presence of extensive irrelevant personal messages and spams. In this paper, we propose a novel recommendation framework to overcome this problem. By analyzing information diffusion patterns among a large set of micro-blogs who play the role of news providers, our approach selects a small subset as recommended emergency news feeds for regular users. We have evaluated our diffusion-based recommendation framework on Twitter during the early outbreak of H1N1 Flu. The preliminary results show that our method leads to more balanced and comprehensive recommendations compared to benchmark approaches.

1 Introduction

Micro-blogging is experiencing rapid growth and gaining explosive popularity worldwide. Compared to traditional blogging, micro-blogging allows a more instant and flexible form of communication. Micro-blogging sites typically restrict the length of posted messages. These messages can be published/received via a wide variety of means, including the Web, text messaging, instant messaging, and other third-party applications. Such a flexible and broad-based architecture significantly lowers the threshold for participation, and encourages users' frequent updates. Consequently, micro-blogging is widely adopted by the public to share/seek real-time information, especially during emergency events. For example, at the early stage of the recent H1N1 Flu (Swine Flu) outbreak, the volume of H1N1 Flu-related messages on Twitter - one of the most popular micro-blogging sites - has increased 1500 times over four days (Apr 24–Apr 27), and accounted for nearly 2% of all Twitter traffic in that time period [1]. Meanwhile, a large number of people turned to Twitter searching for latest updates of the outbreak, causing the keyword "Swine Flu" listed as the "top trending topic" on Twitter Search consistently.

On the other hand, it is becoming increasingly difficult to find contents of interest from micro-blogs generated by the exponentially-expanding micro-blogging community. During emergencies, the seeking of accurate and timely information can be even harder. In this paper, we formulate the task of navigating micro-bloggers to their desired information as a recommendation problem. There exist numerous micro-bloggers who play the role of "news reporters" (self-claimed or volunteering) during emergency events, by posting instant news stories on their micro-blogs. As such, instead of letting users actively perform searches, we aim to identify a small number of quality "news reporters" and recommend them to information seekers as emergency news feeds. We

[*] Aaron R. Sun and Jiesi Cheng are with Department of Management Information Systems, the University of Arizona (asun,chengj@email.arizona.edu)

[†] Daniel D. Zeng is with Department of Management Information Systems, the University of Arizona and holds a visiting position at the Institute of Automation, Chinese Academy of Sciences (zeng@email.arizona.edu)

demonstrate that such a task is distinctly different from standard content-based and link-based recommendation, as investigated in the blogging domain. A novel information diffusion-based framework is proposed to deal with the specific characteristics and requirements of micro-blogging recommendation. We empirically observed that these “news reporters” operate in social settings: they re-broadcast and refer to news stories from one another, facilitating rapid diffusions of news stories. Our hypothesis is that through an understanding of the news diffusion process, one can effectively measure the importance of each “news reporter” from various diffusion-related perspectives (e.g., volume and response time) and in turn make relevant recommendations. The rest of this paper is organized as follows. In Section 2, we briefly review recommendation techniques in a blogging context and highlight what is special about micro-blogging recommendation. A diffusion-based micro-blogging recommendation framework that utilizes information diffusion patterns is then proposed and its practical value is evaluated through a computational experimental study presented in Section 3. Conclusions and future directions are discussed in Section 4.

2 Information Diffusion-based Recommendation

To our knowledge, this paper represents the first attempt on micro-blogging recommendation. The closest body of literature is on blog recommender systems. There are two basic types of blog recommendation techniques: content-based and link-based [2]. In content-based approaches, a blog article is represented as a topic vector and a scoring system is used to calculate the degree of match between this article and user interest. Link-based recommendation relies on implicit or explicit network structures extracted from the blogging community. In a simple example of friendship network, two users sharing a broadly common set of friends can be considered as having similar interests and the articles viewed by one can be recommended to the other [2].

Information management concerning micro-blogs in emergency contexts is markedly different from that concerning regular blogs. (a) The lightweight design of micro-blogging tends to generate an overwhelming volume of message streams which are inefficient to process [7]. (b) A substantial proportion of these messages are of a personal conversational nature with little value to the general public. (c) There exist numerous “news reporters” who regularly post latest news stories on their micro-blogs, mostly on a voluntary basis. These “reporters” provide important information filtering and amplification services and can be effectively leveraged for recommendation. In the next section, we introduce our information diffusion-based recommendation framework for micro-blogging. By examining the dynamics of information flows, we construct a graph summarizing diffusion patterns and make recommendations using this graph.

2.1 Motivation and Problem Statement

Information diffusion through online social networks has recently become an active research topic [3]. In blog communities, the propagation of information from one blog to the next is frequently observed, as a result of low-cost information sharing and publishing [5]. Such diffusion patterns are also prevalent among micro-blogs. On Twitter, first-hand news stories normally originate from a limited number of professional news agencies (e.g., BBC and Reuters), though exceptions exist. These stories then spread across the community through the process of reposting/copying or commenting. In a news story’s diffusion process, any micro-blog that has posted this story is called a participant who “captures” the story. Now, if we raise the following question as to recommending no more than k micro-blogs as emergency news feeds (k is an exogenous parameter which is reasonably small to avoid information overload), these diffusion patterns could be helpful because they reveal how each micro-blog participated in the past diffusion processes. Intuitively, a micro-blog is more likely to be favored and recommended during emergencies if it captures news stories of interest more accurately and rapidly. More specifically, we use the following measures to quantify various aspects of this valuation process. (1) *Story Coverage* (SC): Multiple news stories regarding one broad topic could simultaneously spread. For example, when the H1N1 Flu outbreak occurred, "CDCEmergency" reminded the public that they would not get infected from eating pork, while "ForbesNews" was concerned about the Flu's impact on financial markets. The SC measurement describes the number of stories of interest captured by micro-blogs under study, and stronger recommendation is given to micro-blogs that have higher SC with other conditions being identical. (2) *Reading Effort* (RE): Certain micro-blogs can be crowded with messages. However, too many messages compromise readability and raise the cognitive effort to filter out irrelevant contents. The RE measurement describes the expected number of messages one has to read, in order to discover SC stories of interest from selected micro-blogs. (3) *Delay Time* (DT): The postings of one news story s on micro-blogs are time-stamped. The DT measurement describes the time passed from the first appearance of s in the community until s is captured by one of selected micro-blogs. A delayed capture is certainly undesirable in our application setting.

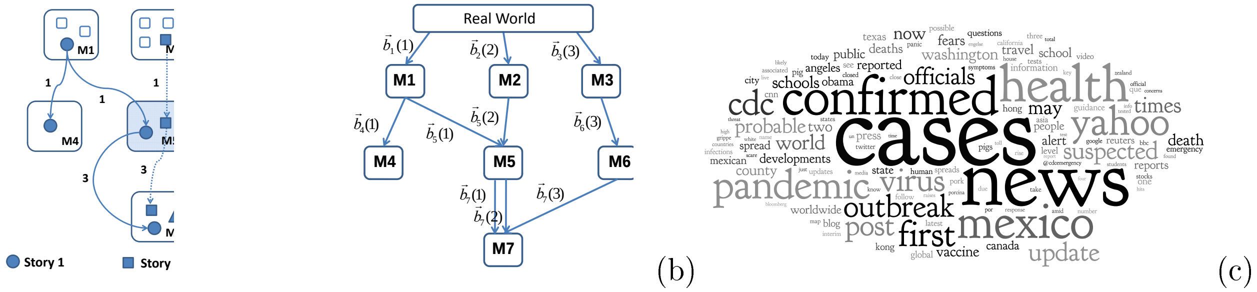


Figure 1: (a) Diffusion Paths (b) Diffusion Graph (c) Term Frequencies (larger font size indicating higher term frequency)

In this paper, we focus exclusively on these three basic measures. The above discussions are illustrated in Figure 1(a) where rounded rectangles represent micro-blogs, and solid/open shapes represent news stories relevant/irrelevant to the user interest. Directed links indicate the flow of a news story, and each link is associated with a numeric label representing the units of time spent. We define a diffusion path as the route through which a news story flows from its source to other micro-blogs. In Figure 1(a), three diffusion paths co-exist with each corresponding to a news story. In terms of recommending micro-blogs, many combinations turn out to have the highest SC, such as the sets $(M1, M2, M3)$, $(M3, M5)$, and $(M7)$. However, $(M1, M2, M3)$ is relatively weak in RE, and $(M7)$ has the longest DT in capturing each story. As a result, $(M3, M5)$ becomes our best choice. In real-world scenarios where a large number of diffusion paths exist, the recommendation task can be quite complex. In the next subsection, we will present an automated method.

2.2 A Diffusion Path-based Approach

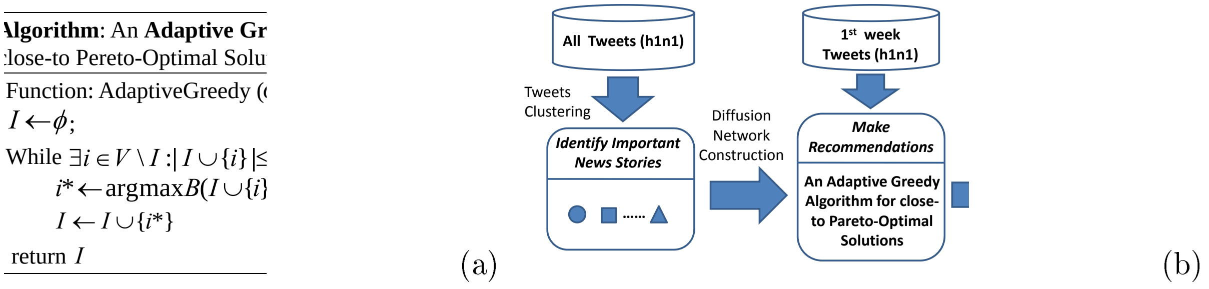
Inspired by the work of Krause et.al. [6] on outbreak detection in water distribution networks, we have developed a diffusion-based recommendation approach using diffusion graphs. A diffusion graph $G = (V, E)$ represents an aggregation of a set of diffusion paths. The node set V represents micro-blogs that have appeared on at least one diffusion path, and each directed edge in the edge set E indicates the flow of a news story. For simplification, we assume that each micro-blog has one and only one immediate predecessor from which a specific news story is obtained [8]. In case when such a predecessor is not traceable, we create a dummy "Real World" node [5] and view it as a general source. Since a diffusion graph can contain multiple diffusion paths, it is possible for two micro-blogs to be connected by more than one edge. Finally, if a news story s flows into node i via edge e, we assign a score vector $\overrightarrow{b_i(s)} = (b_{i(SC)}(s), b_{i(RE)}(s), b_{i(DT)}(s))$ to e indicating the benefit incurred after node i captures story s. As described in Section 2.1, $b_{i(SC)}(s) = 1$ as s is captured; $b_{i(RE)}(s)$ equals the average number of messages one has to read from micro-blog i in order to discover s, which can be estimated from i's historical messages. We assume that $b_{i(DT)}(s)$ drops exponentially with increasing DT, following $b_{i(DT)}(s) = e^{-\lambda \cdot DT}$. Figure 1(b) illustrates such a diffusion graph corresponding to the diffusion paths shown in Figure 1(a).

We now formulate the micro-blog recommendation problem in an optimization framework. Given a diffusion graph $G = (V, E)$, we aim to identify a set of at most k micro-blogs $I \subset V$ and $|I| \leqslant k$ to recommend. For each news story s captured by I, we define an associated benefit score vector as $\overrightarrow{b_I(s)} = (b_{I(SC)}(s), b_{I(RE)}(s), b_{I(DT)}(s)) = (\max\limits_{i \in I} b_{i(SC)}(s), \sum\limits_{i \in I} b_{i(RE)}(s), \max\limits_{i \in I} b_{i(DT)}(s))$. The objective is to recommend a set of nodes that optimizes the benefit score vector over all news stories diffused on the graph. As these three criteria might be in conflict, we use the *Pareto* optimality. A recommended set I is called Pareto-optimal on story s if there exists no another recommendation J such that $b_{J(M)}(s) \geqslant b_{I(M)}(s)$ for all measurements $M \in (SC, RE, DT)$, and $b_{J(M')}(s) > b_{I(M')}(s)$ for at least one measurement $M' \in (SC, RE, DT)$. One common approach for finding such Pareto-optimal sets is scalarization [4]. By choosing weights λ_{SC}, λ_{RE} and λ_{DT}, we can optimize an objective function $B(I) = \sum\limits_{s} b_I(s) = \sum\limits_{s}\sum\limits_{M} \lambda_M b_{I(M)}(s)$ instead. Any solution that optimizes $B(I)$ is guaranteed to be Pareto-optimal, and by adjusting λ_M, a recommendation can take full consideration of all aspects while allowing varying preferences. Finding exact Pareto-optimal sets is NP-hard, but an adaptive greedy algorithm was developed in [6] as an effective heuristic solution for finding close-to Pareto-optimal sets. Our method (Figure 2(a)) is fashioned after this algorithm, which starts with an empty set, and repeatedly adds a micro-blog attempting to maximize the benefit score. The algorithm stops once k micro-blogs are selected or the incremental benefit is less than a predefined small value. In the next section, we evaluate this recommendation method using a recent emergency scenario on Twitter.

3 An Empirical Study

We collected data from Twitter.com using its API from May 10 to May 16, 2009. We used keywords “swine flu” and “h1n1” to query Twitter every 15 minutes throughout the week. Each time Twitter search provided up to 1,500 latest published messages. We identified 1,034 unique accounts who had mentioned either keyword for more than 10 times during that week. In Twitter, each user u can maintain a list of friends and followers. u can follow the tweets of u's friends (a tweet is a message published by a Twitter user), and u's tweets are followed by u's followers. We then proceeded to retrieve each user's all available tweets (up to 3,200 historical tweets), and their lists of friends and followers. In our data set, for a majority of users, 3,200 tweets are more than adequate to cover two month-worth of postings, indicating that we were able to collect these users' near-complete tweets since the outbreak of H1N1 Flu (late April, 2009). In the end, a total of 1,308,800 tweets from 1,034 candidates were collected, among which 35,091 tweets contain keywords “swine flu”, or “flu”, or “h1n1.” We refer to these tweets as H1N1-related tweets thereafter.

Our detailed experimental design is shown in Figure **2**(b). We first divided all H1N1-related tweets into two groups by their published time (week 1 and weeks 2 & 3). For each group of tweets, we conducted a term frequency analysis after removing stopwords and stemming. Term frequencies for tweets in Group 1 are visualized in Figure 1(c). We only considered frequent terms



```
lgorithm: An Adaptive Gr
lose-to Pereto-Optimal Solu
Function: AdaptiveGreedy (
$I \leftarrow \phi$;
While $\exists i \in V \setminus I : |I \cup \{i\}| \leq$
    $i^* \leftarrow \mathrm{argmax} B(I \cup \{i\}$
    $I \leftarrow I \cup \{i^*\}$
return $I$
```

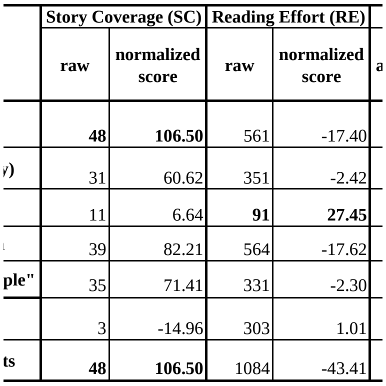
(a)
All Tweets (h1n1)
Tweets Clustering
Identify Important News Stories
Diffusion Network Construction
1st week Tweets (h1n1)
Make Recommendations
An Adaptive Greedy Algorithm for close-to Pareto-Optimal Solutions
(b)

Figure 2: (a)An Adaptive Greedy Algorithm (b)Experimental Design

that have appeared 5 times or more. Each tweet is then represented as a term frequency vector. We then applied hierarchical k-means clustering to segment these vectors into different clusters, with each generated cluster containing tweets describing one news story. The clustering results were quite reasonable owing greatly to the fact that tweets are short in length. We subsequently selected by hand "important" stories in each group which have been posted by 5 or more distinct users. For Group 1, 49 important stories were identified. For Group 2, 52 were identified. Each story was stored with a rich set of metadata: users who have posted this story, friend/follower relationships among these users, and time of posting. In the next step, we used these metadata to construct diffusion paths for each story and then aggregated them into a diffusion graph.

Various methods have been developed to identify information diffusion paths from the history of user interactions [3]. We adopted the following rules to approximate a diffusion path for each news story s. (a) Explicit referring: in Twitter, a tweet can include the use of "@*username*" as an indication that this tweet refers to one of *username*'s tweets before, or sometimes, the symbol @ is omitted or replaced by other symbols. As such, if user v posted story s after user u did, and v referred u's username explicitly, s is then assumed to flow from u to v. (b) Implicit referring: according to [3], if user v follows user u, and v frequently posted same stories after user u did, then we assume diffusions from u to v. (c) Unknown referring: when neither condition above is satisfied, we assume that v received the story from the dummy "Real World" node. As a result, we obtained a 167-node diffusion graph for news stories in group 1 containing 49 diffusion paths. We then calculated benefit scores for each edge on the graph as described in Section 2.2. All scores of SC, RE and DT were normalized by a Z score transformation. For the experiments reported here, the total benefit scores were calculated by placing equal emphasis on these three aspects.

Given this real-world diffusion graph, we selected a close-to Pareto-optimal set of $k = 3$ microblogs using the Adaptive Greedy Algorithm. Meanwhile, we also selected another 6 recommended sets using benchmark methods. The performance evaluations for all candidate sets using the news stories found in Group 2 are listed in Table 1. Our proposed diffusion-based method has obtained the highest benefit score, with a balanced performance in all three measures. As a benchmark, we built a friend/follower graph for all 1,034 candidates and made recommendations by the top 3 Authority/Hub scores generated by HITS algorithm. Although these "authority" and "hubs" were structurally important, they delivered only moderate performances partly because they were not as active in posting news. For example, "CDCEmergency" had more than half a million followers, but only posted two tweets per day. We next used Google Site Search and Twitter "Find People" to select top 3 ranked results using the query "swine flu." Both search engines performed reasonably well, but their results had relatively large delays for not considering the temporal factor. Lastly, we made recommendations by using two simple heuristics. From the 167 users in the diffusion graph, we first selected accounts by their number of followers. These accounts turned out to be traditional mass media that captured news stories very fast but performed poorly in coverage. Another set of users were selected by their number of first-week tweets. This set performed surprisingly well due to the highest volume of tweets, but this advantage was offset by the low RE scores.



	Story Coverage (SC)		Reading Effort (RE)		
	raw	normalized score	raw	normalized score	a
	48	**106.50**	561	-17.40	
y)	31	60.62	351	-2.42	
	11	6.64	**91**	**27.45**	
	39	82.21	564	-17.62	
ple"	35	71.41	331	-2.30	
	3	-14.96	303	1.01	
ts	**48**	**106.50**	1084	-43.41	

Table 1: Evaluation Results

4 Concluding Remarks and Future Directions

In this study, we propose a novel diffusion-based micro-blogging recommendation framework, aiming to recommend micro-blogs during critical events. We have developed a set of measures assessing the value of micro-blogs from a diffusion standpoint and a diffusion path-based optimization formulation. The preliminary evaluation using a real-world dataset indicates that this approach outperforms several other commonly-used approaches. Our ongoing research is focused on (a) an extensive evaluation of our approach with different parameter settings and additional real-world scenarios, and (b) efficient computation of Pareto optimal sets.

Acknowledgements

The authors wish to acknowledge research support from the US NSF (0428241, 0839990), US DHS (2008-ST-061-BS0002), NNSFC (60621001, 60875049, 70890084), MOST (2006AA010106), and CAS (2F07C01, 2F08N03).

References

[1] Swine flu news and concern dominates online buzz, http://blog.nielsen.com, April 2009.

[2] Zeinab Abbassi and Vahab S. Mirrokni. A recommender system based on local random walks and spectral methods. *In Proceedings of the 9th WebKDD*, 2007.

[3] Eytan Adar and Lada A. Adamic. Tracking information epidemics in blogspace. In *Proceedings of the 2005 IEEE Int'l Conf. on Web Intelligence*, 2005.

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge UP, March 2004.

[5] D. Gruhl, R. Guha, David Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proceedings of WWW*, 2004.

[6] Andreas Krause, Jure Leskovec, Carlos Guestrin, Jeanne VanBriesen, M. ASCE, and Christos Faloutsos. Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management*, 134(6):516-526, 2008.

[7] M. Kristina. Bursts of information: Microblogging. *The Reference Librarian*, 50:212-214, 2009.

[8] Aaron R. Sun and Daniel D. Zeng. Maximizing influence through online social networks. In *Proceedings of the 18th Workshop on Information Technologies and Systems*, 2008.

Do I Follow My Friends or the Crowd? Informational Cascades in Online Movie Rating

Young Jin Lee, Yong Tan
Foster School of Business, University of Washington, Seattle, WA
younglee@uw.edu ytan@uw.edu

Kartik Hosanagar
The Wharton School, University of Pennsylvania, Philadelphia, PA
kartikh@wharton.upenn.edu

Abstract

Online product review as a form of online Word of Mouth (WOM) and User-Generated Content (UGC) has attracted much attention recently. This study analyzes how online movie user ratings are generated through a complex interrelationship between product information, marketing effort, and social influences. In particular, we examine the effect of comparable WOM from the crowd or friends on user ratings. Our multilevel analysis indicates that, on average, higher predecessors' ratings increase the likelihood of a subsequent user providing a high rating, or, in other words, herding occurs. On the other hand, prior reviews by friends act to reduce this herding behavior. We show that the degree of herding behavior induced by the crowd's ratings can be significantly different across movies due to movie level heterogeneity.

Keywords: UGC, Recommendation system, Herding, Multilevel model, Latent variable

1. Introduction

Online user generated reviews can provide valuable information about products and can significantly enhance the "Buzz" effect surrounding certain products in a market. In recent years, firms have deployed new services and business models to leverage user-generated content. In March 2009, Netflix announced that it would be integrating with Facebook to let users link their accounts at the two sites and share movie user ratings (Tirrell 2009). Similarly, some startups have been developing new recommendation services by aggregating scattered user generated content across various online communities. For example, Nanocrowd analyzes user generated reviews and ratings data from various sources on the Internet and creates a list of customized movie recommendations for users (Glockner 2009).

Existing work in this area has analyzed the designs and performance of eBay and Amazon-like online recommendation systems (see a survey in Dellarocas 2003). Specifically, most studies on online reviews focus on the ex-post valence and dispersion of online reviews and their relationship with sales (Chevalier and Mayzlin 2006). Another school of research is about the motivation for user generated reviews on the Internet. Social psychologists have for long been studying emotional sharing (for reviews, see Rimé et al. 1998). The work on online reputation systems has primarily focused on the consequence of peer recognition (Jeppesen and Fredericksen 2006). Self-verification is another important driver of online contribution (Hornsey and Jetten 2004).

Regardless of the motivations of user-generated reviews and their relationship with product sales, intuitively, positive online review posts encourage other consumers to adopt products whereas negative opinions discourage them. However, prior reviews may not always lead subsequent online reviewers to generate similar reviews. People may be disgruntled by other reviews or overreact toward more positive (negative) reviews. As such, it is not clear how and to what extent prior reviews transfer information and private opinions about product to the subsequent reviewers. This becomes more complicated when each review is summarized into a rating and its underlying tone or reason cannot be easily revealed like text reviews. Therefore, online reviewers encounter a problem of rating under imperfect information from prior user ratings and their own private signals.

In this study, we characterize the user behavior in terms of review generation and examine the informational cascading in sequentially generated ratings. In particular, we use publicly available data from an online movie social networking community website, Flixster.com,[1] to construct measures of user-generated movie ratings. Movie characteristics also are available for us to control for relevant factors which could potentially affect user ratings.

Our multilevel analysis is designed to answer the following questions in user rating generation process: First, when a reviewer observes the average rating of prior reviewers for the same movie, how does the information affect the reviewer's own rating? Second, a reviewer is more likely to read the detailed text reviews of a "socially close" predecessor such as an online friend. As such, we examine how prior ratings by friends affect a subsequent reviewer's rating behavior. Third, we study the impact of other factors such as firms' marketing efforts and critics' ratings on user rating generation process and this in turn enables the effects of the social influences to be correctly identified. We use the average valence of others' ratings as a proxy for observable prevailing opinion about a movie before a subsequent reviewer generates a review. As a comparable piece of information, the average valence of friends' ratings on the site is a proxy for the observable local (i.e. within the social network) opinion for the movie. Following Zhang (2009), we define the former as Silence Word-of-Mouth (SWOM) and the latter as salient Friend Word-of-Mouth (FWOM).

By applying "generalized linear latent" approach with a multilevel model to address the heterogeneities of reviewers and movies, our findings suggest that, on average, subsequent reviewers tend to follow SWOM strongly while FWOM leads relatively less herding behavior. In other words, more favorable predecessors' ratings will lead successors to choose higher rating. We also show that the degree of herding behavior can be significantly different across movies due to movie level heterogeneity. We organize the rest of the paper as follow: First, we explain our data. Third, our models are presented and applied to the data. Fourth, the results of multilevel analysis with latent variables are used to explain user rating generating process. Finally, we conclude.

2. Data

Our data consists of movie level data and online user review level data. First, we collected movie specific data from several public websites[2] and sampled only movies released in 2007. Second, we gathered user level data from Flixster.com based on the sampled movies. It also provides friendship information among the users in the community. Hence, it provides a partially observed friendship network among reviewers on the sampled movies. We chose movies which have more than 1,000 user ratings within first four months periods after release to make sure there are large variations in user rating. Our final sample dataset contains 20,473 individual users who reviewed and rated at least one movie among 17 movies and 30,628 rating observations with variables listed in Table 1.

The valence of user rating for a movie is our dependent variable. Each user rating contains the time-stamp of when the rating was created and, this gives an advantage to keep track of the changes of user ratings overtime. In Flixster.com, the rating is on ten scales from 0.5 to 5.0 and we multiply rating by 2 to make values integer in our estimation.

For the independent variables in our model, we consider an information set based on other reviewers' ratings, firm's action, and critics rating. SWOM is the calculated average ratings of all others (including online friends) who have rated the same movie which a subsequent reviewer would rate at the next observation day. Similarly, FWOM were calculated by the ratings made by a corresponding reviewer's online friends. Therefore, zero FWOM indicates the reviewer has no friends who rated the same movie before. The information set also includes distributor's cumulative advertising spending and

[1] Flixster is one of the largest movie rating website with about 15 million unique visitors and about 2 billion movie user ratings.

[2] BoxOfficemojo.com, ImDB.com, Numbers.com, Metacritics.com, and Ad$pender.

critics' ratings.[3] In addition to the set, we also control for the observed relevant factors at the user and movie levels as shown in Table 1 in order to consistently estimate the effects of the elements in the information set.

Table 1. Data Descriptive Statistics

Level	Dimension	Variable	Description	Mean	Min	Max
Reviewers (20,473Obs.)	Demographic[4]	*Sex*	Dummy for sex (Female=0)	0.41	0	1
		Age	Reviewer's age	24.64	13	108
	Profile in Flixster.com	*MemberFor*	Membership days	657.92	237	1286
		ProV	Profile viewed by others	258.5	0	258,795
		NumF	The number of friends	33.17	0	830
		NumR	The number of ratings history	700.03	1	68,310
		NumRE	The number of reviews history	91.33	0	55,667
		Rat	Rating for movie (x2 to discretize)	7.90	1	10
		SWOM	Avg. rating of predecessors	8.09	6.99	9.17
	Ratings & Reviews	*FWOM*	Avg. rating of precedent friends	8.04	1	10
		NumFR	The number of preceding friend reviews	2.95	1	114
		%FR	% of friends who rated the same movie	0.01	0	1
Movies (17 obs.)	Characteristics	*RunT*	Running time in minutes	115.24	87	168
		CR	Metacritic.com's average critic rating	6.45	3.5	8.5
		NR	Weekly volume of Reviews	461.916	7	1608
	Advertising Performance	*Acc.Ad-Spending*	Acc. weekly adv. spending in million $	13.70	4.03	22.73
		LOWT	Log (opening weekend theaters)	8.12	7.13	8.38
		LWkth	Log(Weekend Theaters)	6.66	3.09	8.38
		Rank	Weekly Ranking	18.44	1	65

3. Model

We develop a latent response model to explain online movie ratings generation. The latent response $R^*_{i,j,t}$ is a true rating for movie i by reviewer j at time t for the error-prone observed rating $R_{i,j,t}$ due to the various noise factors that a reviewer has and the restrictive scale of ratings (Skrondal and Rabe-Hesketh 2004). We assume that the information set $\mathrm{A}_{i,j,t}$ directly affects reviewer's rating behavior. In addition to this, we include two random intercepts to explain individual heterogeneity and time related variation. This makes our model a three-level hierarchical model as:

$$R^*_{i,j,t} = X'_{i,j,t}\beta + \mathrm{A}'_{i,j,t}\delta + \zeta^{(2)}_{j,t} + \zeta^{(3)}_{t} + \varepsilon_{i,j,t}, \qquad (1)$$

$$R_{i,j,t} = \begin{cases} 1 & if\ R^*_{i,j,t} \le \kappa_1 \\ 2 & if\ \kappa_1 < R^*_{i,j,t} \le \kappa_2 \\ \vdots & \vdots \\ 10 & if\ \kappa_9 < R^*_{i,j,t} \end{cases}$$

[3] The correlations among SWOM, FWOM and critics rating in our dataset are low and this emphasizes the explanatory powers of our independent variables.

[4] Since there is 8% of missing sex values in our sample individuals, we exclude the individuals of missing sex. However, 25% of included individuals have still the missing values of age. Therefore we imputed missing age values.

where $\zeta^{(2)}_{j,t}$ is an individual level random intercept, $\zeta^{(3)}_{t}$ is a time (weeks since movie release) specific random intercept and $\varepsilon_{i,j,t}$ has a logistic distribution. We further assume $\zeta^{(2)}_{j,t} \sim N(0,\psi^{(2)})$ and $\zeta^{(3)}_{t} \sim N(0,\psi^{(3)})$. $X_{i,j,t}$ contains control variables. $A_{i,j,t}$ contains the information set. We also include interaction terms between each movie dummy and SWOM in $A_{i,j,t}$. This gives us two benefits that first it captures the direct effects of SWOM on each movie and second it greatly reduces the correlations of movie dummies with other movie specific covariates. As a consequence it addresses multicollinearity issue.[5] As mentioned earlier, $R_{i,j,t}$, are generated by the threshold model which represents cutoff values. We assume that κ_s, s=1,…, 9 are the same for all movies.[6]

4. Estimation and Results

We implement generalized linear latent variable and mixed models (GLLAMM) [7] to estimate our model parameters. This allows maximizing the likelihood of the conditional density of the response variable given the latent and explanatory variables with the prior density of the latent variables with adaptive quadrature.

Table 2. Regression Results for the Variances of Random Intercepts[8]

	Single Level		Two Level		Three Level	
Parameters	**Est.**	**(SE)**	**Est.**	**(SE)**	**Est.**	**(SE)**
Reviewer-level						
$\psi^{(2)}$	-	-	2.356***	(0.104)	2.355***	(0.214)
Time-level						
$\psi^{(3)}$	-	-	-	-	0.004	(0.003)
Log-Likelihood	-56679.912		-55768.037		-56513.934	

Note. Standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1.

Empirical evidence of unobserved heterogeneity

Using the techniques described in Skrondal and Rabe-Hesketh (2004), the estimated variances of random intercepts, $\hat{\psi}^{(2)}$ and $\hat{\psi}^{(3)}$ of a simpler model of equation (1) in three different specifications are shown in Table 2. Single-level represents single level ordinal response model without any random effect. Two-level includes only reviewer specific random intercept and three-level contains both random intercepts of reviewer and time. In the two-level model, $\hat{\psi}^{(2)}$ is estimated at 2.36 and significant. The log-likelihood becomes significantly greater than single-level. However, the estimated variance component for times is nearly zero and insignificant in the three-level specification[9]. Therefore, we only consider the unobserved heterogeneity in reviewers in our further estimation. Nevertheless, we includes a time variable (weeks since release) in the model to keep track of the effects over weeks.

Direct impacts of others' rating on a user rating

[5] The interactions between FWOM and movie dummies are not included in $A_{\mathrm{i,j,t}}$ due to multicollinearity and statistically insignificances.

[6] Rating scheme is fixed for all movies and therefore each threshold is homogenous in the sense that reviewers choose the thresholds in the fixed values for every movie.

[7] GLLAMMs are a class of multilevel latent variable models for (multivariate) responses of mixed type including continuous responses, counts, duration/survival data, dichotomous, ordered and unordered categorical responses and rankings (see Skrondal and Rabe-Hesketh 2004).

[8] The estimates of interaction terms and threshold parameters are omitted.

[9] We also tried other models in Table 2 and the results are virtually identical.

We run our model (1) with the different sets of control variables. In Table 3, Column 1 includes all available movie and reviewer specific variables which have no high correlations between the variables without movie dummies. Column 2 excludes all movie specific covariates but includes movie dummies. Column 3 has interaction terms between movie dummies and SWOM, reviewer variables, and two movie specific variables, weekly volume of reviews and critic rating, which are not highly correlated with other variables in this specification.

After we run Column 1, 2, and 3 specifications by ordered logistic regression without any random effect, Column 3 is the best-fit specification by log-likelihood (and also VIF[10] < 10). Column 3 is re-estimated with the random intercept of reviewer and the results are in Column 4. Again, $\hat{\psi}^{(2)}$ is significant and its log-likelihood is much greater than the others. Hence, Column 4 fits the model (1) better and the results are similar to Column 3.

Table 3. Regression Results for Ordered Logistic Model (1)[11]

	Column 1		**Column 2**		**Column 3**		**Column 4**	
	without		**with**		**with**		**Two-level of**	
Variables	**Movie Dummies**		**Movie Dummies**		**Interactions**		**Column 3**	
β_1 [*Sex*]	-0.575^{***}	(0.021)	-0.586^{***}	(0.021)	-0.584^{***}	(0.021)	-0.743^{***}	(0.033)
β_2 [*Age*]	-0.014^{***}	(0.001)	-0.015^{***}	(0.001)	-0.015^{***}	(0.001)	-0.021^{***}	(0.002)
β_3 [*Memberfor*]	-0.001^{***}	(0.000)	-0.001^{***}	(0.000)	-0.001^{***}	(0.000)	-0.001^{***}	(0.000)
β_4 [*Prov*]	0.002^{***}	(0.000)	0.002^{***}	(0.000)	0.002^{***}	(0.000)	0.002^{***}	(0.000)
β_5 [*NumF*]	-0.022^{***}	(0.004)	-0.022^{***}	(0.004)	-0.022^{***}	(0.004)	-0.038^{***}	(0.007)
β_6 [*NumFR*]	-0.022^{***}	(0.006)	-0.021^{***}	(0.006)	-0.022^{***}	(0.006)	-0.016^{**}	(0.007)
β_7 [*%F*R]	-1.453^{***}	(0.264)	-1.460^{***}	(0.264)	-1.475^{***}	(0.264)	-1.455^{***}	(0.345)
β_8 [*Weeks*]	-0.028^{***}	(0.007)	-0.010^{*}	(0.005)	-0.016^{**}	(0.007)	-0.009	(0.006)
δ_1 [*SWOM*]	1.087^{***}	(0.024)	0.523^{***}	(0.165)	1.361^{***}	(0.138)	1.679^{***}	(0.169)
δ_2 [*Acc.Ad-Spending*]	-0.016^{***}	(0.005)	-0.014^{*}	(0.007)	-0.005	(0.005)	0.000	(0.007)
δ_3 [*FWOM*]	0.018^{***}	(0.004)	0.018^{***}	(0.004)	0.018^{***}	(0.004)	0.026^{***}	(0.005)
δ_4 [*SWOM* × *Weeks*]	-0.003	(0.005)	0.000	(0.005)	0.000	(0.006)	0.003	(0.007)
δ_5 [*Acc.Ad-Spending* × *Weeks*]	0.003^{***}	(0.001)	0.002^{**}	(0.001)	0.001	(0.001)	0.003^{***}	(0.001)
δ_6 [*FWOM* × *Weeks*]	0.002^{**}	(0.001)	0.002^{**}	(0.001)	0.002^{**}	(0.001)	0.001	(0.001)
Variance($\psi^{(2)}$)							2.004^{***}	(0.096)
VIF	2.79		10.48		6.29		-	
Log-likelihood	-56012.415		-55994.909		-55999.691		-55290.616	

Note. Standard errors in parentheses. ***p<0.01, **p<0.05, *p<0.1.

The overall effect of SWOM is very positively related to subsequent reviewer rating ($\hat{\delta}_1$ =1.68 in Column 4). This indicates that the odds of high versus low ratings are about three to one. For example, a subsequent reviewer with higher SWOM is about 3 times more likely to generate ratings 5 or 4 versus 1 to 3, or ratings 2 to 5 versus 1 than a reviewer with lower SWOM. Therefore, we can interpret that a higher average prior rating of others would lead a higher subsequent reviewer's rating. This indicates strong tendency of following "consensus" of others in generating ratings.

Interestingly, even though FWOM is positively related to rating directly, the overall effect of FWOM is negatively related. *NumFR* takes the number of friends who rated the same movie and *%FR* shows the proportion of online friends who rated the same movie. Therefore, as more friends rated the same movie, a user tends to lower his rating. This indicates that herding in rating behavior becomes

[10] Variance Inflation Factors (VIF) are calculated after running OLS for each column.

[11] All other estimated coefficients of covariates are omitted due to space.

moderate if more friends rated the same movie. This is because it is easier for a user to observe the private signals of friends' ratings by text reviews and replies in this social networking site. As such, user learns not only from other rating choices but also from the private signals of friends, which results in reducing the herding (Vives 2008).

The effect of accumulative advertising spending is insignificant in Column 4. This demonstrates that the advertising does not impact the rating behavior of reviewers who have already been informed during the adoption process (buying a movie ticket) which, may turns out to be driven by marketing efforts rather than social contagion (Van den Bulte and Lilien 2001).

5. Conclusions and Future Research

We uncover how different social contagions can influence user ratings by applying multilevel analysis and latent variables approach to explain unobserved factors such as heterogeneity in movies and reviewers. Although there are differences across movies, our findings suggest that herding in online user rating can exist since the higher crowd's ratings increase a subsequent user's probability of choosing a higher rating versus a lower rating. However, more friends' ratings moderate the herding due to actual learning (private signal flows). Therefore, it might be an optimal policy for producers to put more effort into generating higher ratings in the early stage of movie release.

As we described earlier, observed rating may not convey perfectly the private signal of a reviewer. If we can capture the private information along with discrete rating to represent perceived quality, we would be able to fully explain how user ratings deviate from each other. Hence, our next step is to extract private information from text reviews to test whether its role is different with discrete scale ratings in recommendation systems.

References

Chevalier, J., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* **43**(3) 345–354.

Dellarocas, C. 2003. The digitization of word-of-mouth: Promise and challenges of online reputation mechanisms. *Management Sci.* **49**(10), 2003, pp. 1407-1424.

Glockner, P. 2009. Nanocrowd Has a New Take on Movie Recommendations. *Readwriteweb.com.* http://www.readwriteweb.com.

Hornsey, M. J., J. Jetten. 2004. The individual within the group: Balancing the need to belong with the need to be different. *Personality Soc. Psych. Rev*. **8**(3) 248–264.

Jeppesen, L. B., L. Fredricksen. 2006. Why do users contribute to firm-hosted user communities? The case of computer controlled music instruments. *Organization Sci.* **17**(1) 45–63.

Rimé, B., C. Finkenauer., O. Luminet., E. Zech., P. hilippot. 1998. Social sharing of emotion: New evidence and new questions. *European review of Soc. Psych*. **9** 145-189.

Skrondal, A., S. Rabe-Hesketh. 2004. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Boca Raton, FL: Chapman & Hall/ CRC Press.

Tirrell, M. 2009. Netflix, Facebook Link to Show Users' Film Ratings. *Bloomberg* http://www.bloomberg.com.

Van den Bulte, C., G. L. Lilien. 2001. Medical Innovation Revisited: Social Contagion versus Marketing Effort. Amer. J. Soc. **106**(5) 1409-1435.

Vives. X. 2008. Informational and Learning in Markets: The impact of market microstructure. *Princeton University Press*.

Zhang, J. 2009 The Sound of Silence: Observational Learning in the U.S. Kidney Market. *Marketing Sci*. Forthcoming.

Making Aggregate-Level Predictions in Recommender Systems Using Multi-Level Ratings

Akhmed Umyarov
New York University
aumyarov@stern.nyu.edu

Alexander Tuzhilin
New York University
atuzhili@stern.nyu.edu

Abstract

Aggregate-level ratings have been studied in recommender systems and have been shown to improve predictions of ratings of individual items for individual users. Similarly, individual-level ratings have also been used for the estimation of aggregate-level ratings for groups of items and users. In this paper, we combine these approaches and present a novel method for estimating unknown aggregate-level ratings from the known individual- and the aggregate-level ratings corresponding to different levels of the rating aggregation hierarchy. We show both theoretically and empirically that this combined approach outperforms the alternative methods that do not include the ratings from different levels of the rating aggregation hierarchy.

Keywords: recommender systems, predictive models, aggregate ratings

1 Introduction

Modern recommender systems, such as the ones deployed by Amazon.com and Netflix, typically recommend individual items to individual users. In contrast, aggregated ratings assigned to groups of users and items, such as the fact that graduate students rated adventure movies as 5.3 (out of 7), have not been fully utilized by recommender systems. However, in various critical recommendation applications, it is important to work with aggregate ratings which may allow to determine customer preferences at aggregate levels. For instance, consider the following example based on a true story. One of the major financial services companies was planning to launch a new group of mutual funds, and one of the objectives of these funds was to protect their customers against inflation. Clearly, senior management was interested in how different groups of customers would like various funds in that group. It later turned out that these funds were particularly popular among the economists. Using recommender systems terminology, the management was interested in estimating aggregate ratings that various groups of customers (such as the economists) would give to different types of mutual funds in that group. In particular, the management was interested to know that the economists would give high ratings to that type of mutual funds.

Although there has been some work done on incorporating aggregate rating information into the recommendation process that can be classified into the *top-down* and *bottom-up* approaches (Adomavicius et al. 2005), each of these methods offers a limited view on the use of aggregated ratings in the recommender systems. In the top-down approach (Bollen 2000; Masthoff 2003), aggregate rating information is only used for improving estimations of individual ratings. In contrast, the bottom-up approach (O'Connor et al. 2001; Jameson and Smyth 2007) uses individual ratings only in order to estimate the aggregate rating for a group. Nevertheless, even with this limited view, it was shown in (Umyarov and Tuzhilin 2007, 2008) both theoretically and empirically that the top-down approach can improve the quality of recommendations for different types of models and experimental conditions.

In this paper, we integrate all of these approaches and present a method that estimates aggregate ratings using a linear combination of the top-down, bottom-up and the traditional same-level approaches. This integration approach is novel to the recommender systems literature to the best of our knowledge, and the real-life performance of this method as well as a theoretical justification of the achieved superior performance is of the key interest for the field of recommender systems.

An example of such approach would be the task of estimating the unknown aggregated average rating that the females in the age category 18 to 24 give to the drama movies. In order to estimate it, we may want to use the known aggregate ratings information pertaining to the drama movies, college students, female college students and so on, as well as known individual ratings by individual females for individual drama movies.

We also provide a theoretical justification why the chosen approach of linear combination of all these methods provides the best rating estimation method vis-a-vis the cases when any of the three of these methods is missing. Further, we validate this theoretical result on the MovieLens and another dataset and empirically demonstrate that this is indeed the case. As a result, the proposed method can be particularly useful in those applications requiring accurate estimations of aggregate ratings, such as the aforementioned financial services application.

2 Prediction of multi-level ratings

2.1 *Model specification*

Assume that we have a set of N users and M items and let $r^{(0)}_{i_0 j_0}$ be a rating of *individual* user i_0 for *individual* item j_0 (either known or unknown). Also let $r^{(1)}_{i_1 j_1}$ be an aggregated average rating given by *group* i_1 of users to *group* j_1 of items. We refer to this aggregate rating as *a 1st-level aggregate rating* for the reasons that will be clear below. Assume also that group i_1 of users is a subgroup of a larger group i_2 of users and group j_1 of items is a subgroup of a larger group j_2 of items. Assume $r^{(2)}_{i_2 j_2}$ is an aggregated average rating given by group i_2 of users to group j_2 of items. We refer to $r^{(2)}_{i_2 j_2}$ as a *2nd-level aggregate rating* since it is a superset of 1st-level aggregate ratings. Further, we continue this process and define the m-th aggregation level inductively in terms of the aggregate ratings at level $m-1$ for all $m = 1, \ldots, M$. When $m > k$, we say that $\{r^{(m)}_{st}\}$ are *upper-level* ratings relative to rating $r^{(k)}_{ij}$; when $m < k$, we say that $\{r^{(m)}_{st}\}$ are *lower-level* ratings relative to rating $r^{(k)}_{ij}$; and when $k = m$, we say that the two ratings are at the *same level*. Also, we define the pair (i, j) as *a segment* when i is a group of users and j is a group of items.

One way to formally represent the resulting hierarchies (taxonomies) and the corresponding aggregate ratings is by using the OLAP-based approach (Adomavicius et al. 2005), where the corresponding OLAP cube would have only two dimensions Users and Items, and ratings r would be its measure.

For instance, continuing the example from Section 1, we can define a taxonomy of users (based on age and gender), taxonomy of movies (based on genres) and aggregate ratings according to these taxonomies. In this case, 0-level would be ratings by individual users i_0 of individual items j_0. At 0-level, each user forms her own group i_0 and each item its own group j_0. Further, the 1-level of aggregate ratings would constitute aggregated average ratings given by users of a certain gender and certain age category to certain sub-genres of movies. For instance, an example of group i_1 of users will be all males under 18 years and an example of group j_1 of items will be all comedy movies that form together a *segment* (i_1, j_1) consisting of ratings given by males under 18 years for comedy movies. In this case $r^{(1)}_{i_1 j_1}$ will be the aggregated average rating of all males under 18 years for all comedies. Similarly, the 2-level would be the aggregate ratings given by users in certain age categories to certain genres of movies and so on. At the top of the hierarchy (level M) would be one aggregate ("grand total") rating given by the whole population of users to all the movies.

Assume that we are given a set of known multi-level ratings $\{r^{(m)}_{st}\}$ for some segments (s, t) at some levels m. Then *our goal is to provide a method for estimating an <u>unknown</u> multi-level rating $r^{(k)}_{ij}$, given the set of <u>known</u> multi-level ratings $\{r^{(m)}_{st}\}$*, where some of these ratings $\{r^{(m)}_{st}\}$ may have $m < k$, some $m > k$ and others $m = k$. If the aggregation hierarchies are specified using the OLAP technology, then, intuitively, the goal of our method is to fill-in the entire OLAP hierarchy of aggregate ratings based on the set of known aggregate and individual ratings.

In conclusion, we need to produce the whole hierarchy of ratings for different levels of $m = 0, \ldots, M$, starting from the individual-level ratings for $m = 0$ and reaching the most-aggregated ("grand total") rating for $m = M$, with some of these multi-level ratings being known, while others remaining unknown. In the rest of the paper, we focus on estimating these unknown ratings.

2.2 *Our approach*

In our approach we propose to consider known multi-level ratings as *estimators* for unknown multi-level ratings and combine them in the optimal way. We, first, present the simplest estimation cases in this subsection and then describe the optimal combination strategy in Section 2.3.

Case 1: Only upper-level aggregate information is available. Assume that all we know is an upper-level rating $r_{st}^{(m)}$ given by a group of users s to a group of items t at level m, and we would like to estimate rating $r_{ij}^{(k)}$ at a lower level k ($k < m$) given by a smaller subgroup i of that group of users such that $i \subseteq s$ to a smaller set of items $j \subseteq t$.

Denote $\hat{r}_{ij}^{(k)}$ an estimator of the unknown quantity $r_{ij}^{(k)}$. A natural choice for that estimator is

$$\hat{r}_{ij}^{(k)} = r_{st}^{(m)}$$

i.e., we estimate the smaller subgroup aggregated average rating just as the larger group average rating. It turns out that, under certain probabilistic assumptions, this approach is the best, as the following theorem states, where by *the "best"* we assume here and throughout the paper the *estimator with the smallest value of the mean squared error (MSE)* of predictions.

Theorem 1. *Assume that all the ratings in the subsegment (i, j) are generated by independent random draws uniformly from the ratings in the segment (s, t). Then the best estimator of the unknown average rating in the subsegment (i, j) is the known average rating in the segment (s, t).*

Sketch of the proof. The statement of the theorem follows from the fact that the uniform independent random draws from segment (s, t) follow a multivariate hypergeometric distribution (Umyarov 2009). □

Case 2: Only lower-level aggregate information is available. Assume that we only know some aggregate rating $r_{st}^{(m)}$ for some subsegment (s, t) of a larger segment (i, j) and that we need to estimate aggregate rating $r_{ij}^{(k)}$ from $r_{st}^{(m)}$ for segment (i, j) that is at the upper-level ($k > m$). For example, assume that we know the aggregated average rating $r_{st}^{(m)}$ of all the males in the age group 18 to 29 for the movie "Madagascar" and we need to estimate the aggregated average rating $r_{ij}^{(k)}$ of all the males across all the age groups for that movie. Since we assumed that we have no other information except the $r_{st}^{(m)}$, it is natural to predict the unknown average rating for a larger group just as the known average rating for a smaller group. More formally, we define the estimator:

$$\hat{r}_{ij}^{(k)} = r_{st}^{(m)}$$

The following theorem shows that $\hat{r}_{ij}^{(k)}$ is the best under certain probabilistic assumptions, and its proof is similar to the proof of Theorem 1:

Theorem 2. *Assume that all the ratings in the subsegment (s, t) are generated by independent random draws uniformly from the ratings in the segment (i, j). Then the best estimator of the average rating in the large segment (i, j) is the average rating in the subsegment (s, t).*

Note also that Case 2 is not as limited as it seems to be. If we know *multiple* aggregate ratings at level $m < k$, then, within the assumptions of Theorem 2, this situation is exactly the same as if we knew one aggregate rating for the union of all the known segments.

Case 3: Only the same-level aggregate information is available. This case corresponds to the rating estimation methods used in the traditional recommender systems. In this case, we may define each group of users as some artificial user and each group of items as some artificial item and apply the traditional rating estimation approaches of recommender systems, such as collaborative, content-based filtering methods and their hybrids (Adomavicius and Tuzhilin 2005).

In this section, we considered only the basic types of estimators. In the next section, we show that when ratings at multiple levels are available, there is an optimal way to linearly combine them.

2.3 *Optimal combination of estimators*

We now show theoretically that, under certain assumptions, the linear combination of several individual estimators from Section 2.2 outperforms any of those single estimators or any of their subsets, thus demonstrating the power of combining all the estimators into a single model.

Theorem 3. *Assume $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$ are **linearly biased correlated** estimators of some unknown quantity x with the following properties:*

$$E\hat{x}_i = a_i x + b_i, \quad \mathit{Var}(\hat{x}_i) = v_i \quad \mathit{cov}(\hat{x}_i, \hat{x}_j) = c_{ij} \quad \forall i, j \in 1, \ldots, n \tag{1}$$

where $a_1, a_2, \ldots, a_n$, $b_1, b_2, \ldots, b_n$, $v_1, v_2, \ldots, v_n$ and all c_{ij} are known values, and Ex is the expected value of random variable x. Also, let estimator $\hat{x}$ be a linear combination of all estimators $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_n$, that is

$$\hat{x} = \alpha_0 + \alpha_1 \hat{x}_1 + \alpha_2 \hat{x}_2 + \cdots + \alpha_n \hat{x}_n \tag{2}$$

Then, there exist weights $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ such that

$$\mathit{Var}(\hat{x}) \leq \min(\mathit{Var}(\hat{x}_1), \mathit{Var}(\hat{x}_2), \ldots, \mathit{Var}(\hat{x}_n)) \tag{3}$$

Therefore, the combined estimator always performs better (not worse) than the best performing estimator in the set of estimators $\{\hat{x}_i\}$. The fact that lower variance leads to better predictions is well-known in statistical theory (Bishop 2006) since it can be shown that the best unbiased estimator in terms of MSE of predictions is the estimator that achieves the lowest variance.

When Theorem 3 is applied to our problem to estimate the unknown rating $x = r_{ij}^{(k)}$, the particular known aggregate ratings $r_{st}^{(m)}$ at different levels m can be used as estimators $\hat{x}_i$ as described in Section 2.2. Then, Theorem 3 states that there exists a linear combination of all these estimators that improves estimation of the unknown rating $x = r_{ij}^{(k)}$.

Unfortunately, Theorem 3 does not tell us how to find the optimal weights $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ when values $a_1, a_2, \ldots, a_n$, $b_1, b_2, \ldots, b_n$, $v_1, v_2, \ldots, v_n$ and c_{ij} are unknown. However, since the theorem claims the existence of such weights and better performance of the weighted estimator, we can apply Theorem 3 indirectly by estimating the weighting scheme $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ empirically by "letting the data speak for itself" to determine the best weighting scheme.

To summarize, in this section we described our approach from the theoretical point of view. In the following sections, we describe the datasets and training/testing procedures that we used to verify our theories and report the results that we obtained on the real-life rating datasets.

3 Experimental settings

To learn the weights $(\alpha_0, \alpha_1, \ldots, \alpha_n)$ and to compare performances of the linear estimator described in Section 2.3 with alternative approaches, we used the following datasets:

- *MovieLens Dataset* [Individual-level ratings] (http://www.grouplens.org). This dataset consists of more than 1 million individual ratings of 3900 movies provided by 6040 users. It also contains demographic information about users, such as their age and gender.
- *Movie Dataset from (Adomavicius et al. 2005)* [Individual-level ratings]. We also used the data from the study (Adomavicius et al. 2005) on 61 users that provided 1110 individual ratings for 62 movies. The dataset contains demographic information about users such as user's age, gender, and the home ZIP code.

- *IMDB Dataset* [Aggregate-level ratings] (http://www.imdb.com). In order to introduce multi-level rating information from the external sources into the individual-level rating datasets described above, we extracted the following average ratings for every movie in those datasets from the IMDB database: (1) total average rating for a movie, (2) average rating among males and average rating among females, (3) average rating among different age groups, (4) average rating among males of different age groups and average rating among females of different age groups.

As the *dependent variable* in equation (2), we have chosen the 1st-level aggregate rating, that is the gender-age level data from IMDB. An example of an instance of this variable is the average rating of males under 18 years for "Terminator 2" movie taken from IMDB.

As *independent variables* in equation (2), we use the following estimators of aggregate and individual ratings at different granularity levels of the ratings hierarchy:

- $\hat{x}_0 = 1$: constant term estimator
- $\hat{x}_1$: the lowest level estimator. This is the estimator based on average rating computed from the *external* individual-level data, such as, for example, the average rating computed from MovieLens dataset for all male users under 18 years old for "Terminator 2" movie.
- $\hat{x}_2$: the 2nd level estimator. This is the estimator based on gender-level data taken from IMDB, such as, for example, the average rating of all females for "Titanic" movie.
- $\hat{x}_3$: another (parallel) 2nd level estimator. This estimator is based on age-level data from IMDB; for example, the average rating of all users under 18 years for "Titanic" movie.
- $\hat{x}_4$: the 3rd level estimator. Based on the average across all users from IMDB for each movie.
- $\hat{x}_5$: the traditional user-based collaborative-filtering estimator (Adomavicius and Tuzhilin 2005) applied on all the 1st level ratings from IMDB.

The dataset was randomly split into 3 sets: 1) training set #1 (40% of the data); 2) training set #2 (40% of the data); 3) test set (20% of the data). The training set #1 was used to compute estimators $\hat{x}_1, \dots, \hat{x}_5$; the training set #2 was used *only* to learn the optimal weights $(\alpha_0, \alpha_1, \dots, \alpha_n)$ using the ordinary linear regression; and the test set was used to verify the resulting performance. The procedure was repeated 10 times to avoid dependencies on a particular sample.

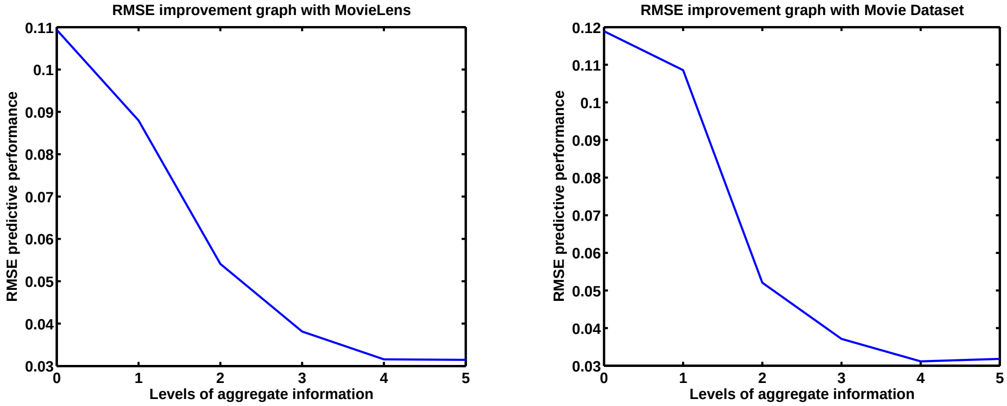
4 Experimental results

The results of our experiments are presented in Figures 1(a) and 1(b). On the x-axis, we plot the cumulative number of independent variables included in the prediction task. On the y-axis, we plot the root mean squared error (RMSE) predictive performance of each of those models. More specifically, the 0-th tick corresponds only to the $\hat{x}_0$ estimator included in the model, the 1st tick of x-axis corresponds to both $\hat{x}_0$ and $\hat{x}_1$ estimators, and so on up to the 5-th tick of x-axis that corresponds to all the 6 estimators of model (2). Figure 1(a) corresponds to the MovieLens dataset and Figure 1(b) to the Movie dataset.

As we can see from Figures 1(a) and 1(b), consecutive addition of extra levels of aggregate information indeed improves predictive performance of the linear model (2). This empirical result confirms the same theoretical finding, reported in Section 2.3. Moreover, as Figures 1(a) and 1(b) show, the combined estimator significantly outperforms the simplest model, such as predicting the aggregate rating using only the averaging of individual ratings of some external dataset.

5 Conclusions and future work

We proposed an approach of combining top-down, bottom-up and the-same-level methods that is novel in the field of recommender systems for estimation of aggregate average ratings. We showed empirically that the resulting combined approach outperformed each of the individual methods. Moreover, we provided the theoretical justification for this effect. Another important advantage of this new method is that exactly



(a) IMDB and MovieLens datasets

(b) IMDB and Movie datasets

Figure 1: RMSE improvement

the same logic can be used to compute *any* unknown aggregate (and also individual) rating in terms of the known aggregate and individual ratings, thus filling in the *entire* hierarchy of unknown aggregate ratings and thus allowing us to compute any individual and aggregated rating in the taxonomy. The proposed method can be used in various business applications requiring accurate predictions of aggregate ratings, including the financial services application described in Section 1.

As a future work, we plan to address the question of predicting the full aggregate distribution of ratings instead of predicting only aggregate average rating. For the prediction of full distribution, as demonstrated by our current experiments that are beyond the scope of this paper, the linear techniques do not fit as natural as for the task in this paper and more complicated non-linear models are required to fully formalize the task.

References

Adomavicius G, Sankaranarayanan R, Sen S, Tuzhilin A. Incorporating contextual information in recommender systems using a multidimensional approach. ACM Trans on Inf Systems 2005; 23: 103–145.

Adomavicius G, Tuzhilin A. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE 2005; 17: 734–749.

Bishop C. Pattern recognition and machine learning. Springer New York., 2006.

Bollen J. Group user models for personalized hyperlink recommendations. LNCS, 2000;

Jameson A, Smyth B. Recommendation to Groups. In Brusilovsky P, editor, The Adaptive Web: Methods and Strategies of Web Personalization. Springer, 2007.

Masthoff J. Modeling the multiple people that are me. In LNCS, 2003, Springer, pp. 258–262.

O'Connor M, Cosley D, Konstan J, Riedl J. PolyLens: A recommender system for groups of users. In Procs of the seventh conference on European CSCW, 2001, p. 218.

Umyarov A. Leveraging aggregate ratings for improving predictive performance of recommender systems. CeDER-08-03 Working paper, 2009.

Umyarov A, Tuzhilin A. Leveraging aggregate ratings for better recommendations. In Proceedings of the 2007 ACM conference on Recommender systems. ACM Press New York, NY, USA, 2007, pp. 161–164.

Umyarov A, Tuzhilin A. Improving Collaborative Filtering Recommendations Using External Data. In 8th IEEE International Conference on Data Mining, 2008, pp. 618–627.