**2.1.1 REVIEW SPAM DETECTION VIA TEMPORAL PATTERN DISCOVERY**

Online reviews play a crucial role in today’s electronic commerce. It is desirable for a customer to read reviews ofproducts or stores before making the decision of what orfrom where to buy. Due to the pervasive spam reviews,customers can be misled to buy low-quality products, whiledecent stores can be defamed by malicious reviews. We observe that, in reality, a great portion (> 90% in the data westudy) of the reviewers write only one review (singleton review). These reviews are so enormous in number that theycan almost determine a store’s rating and impression. However, existing methods did not examine this larger part of thereviews. Are most of these singleton reviews truthful ones?If not, how to detect spam reviews in singleton reviews? Wecall this problem singleton review spam detection.

To address this problem, we observe that the normal reviewers’ arrival pattern is stable and uncorrelated to theirrating pattern temporally. In contrast, spam attacks areusually bursty and either positively or negatively correlatedto the rating. Thus, we propose to detect such attacks viaunusually correlated temporal patterns. We identify andconstruct multidimensional time series based on aggregatestatistics, in order to depict and mine such correlations. Inthis way, the singleton review spam detection problem ismapped to a abnormally correlated pattern detection problem. We propose a hierarchical algorithm to robustly detectthe time windows where such attacks are likely to have happened. The algorithm also pinpoints such windows in different time resolutions to facilitate faster human inspection.Experimental results show that the proposed method is effective in detecting singleton review attacks. We discoverthat singleton review is a significant source of spam reviewsand largely affects the ratings of online stores.

**2.1.2 RANKING FRAUD DETECTION FOR MOBILE APPS: A HOLISTIC VIEW**

Ranking fraud in the mobile App market refers to fraudulent or deceptive activities which have a purpose of bumping up the Apps in the popularity list. Indeed, it becomes more and more frequent for App develops to use shady means, such as inflating their Apps’ sales or posting phony App ratings, to commit ranking fraud. While the importance of preventing ranking fraud has been widely recognized, there is limited understanding and research in this area. To this end, in this paper, we provide a holistic view of ranking fraud and propose a ranking fraud detection system for mobile Apps. Specifically, we investigate two types of evidences, ranking based evidences and rating based evidences, by modeling Apps’ ranking and rating behaviors through statistical hypotheses tests. In addition, we propose an optimization based aggregation method to integrate all the evidences for fraud detection. Finally, we evaluate the proposed system with real-world App data collected from the Apple’s App Store for a long time period. In the experiments, we validate the effectiveness of the proposed system, and show the scalability of the detection algorithm as well as some regularity of ranking fraud activities.

**2.1.3 RANK AGGREGATION VIA NUCLEAR NORM MINIMIZATION**

The process of rank aggregation is intimately intertwined with the structure of skew-symmetric matrices. We applyrecent advances in the theory and algorithms of matrix completion to skew-symmetric matrices. This combination ofideas produces a new method for ranking a set of items. The essence of our idea is that a rank aggregation describes a partially called skew-symmetric matrix. We extend an algorithm for matrix completion to handle skew-symmetric data and use that to extract ranks for each item. Our algorithm applies to both pairwise comparison and rating data. Because it is based on matrix completion, it is robust to both noise and incomplete data. We show a formal recovery result for the noiseless case and present a detailed study of the algorithm on synthetic data and Netix ratings.

**2.1.4 MINING PERSONAL CONTEXT-AWARE PREFERENCES FOR MOBILE USERS**

In this paper, we illustrate how to extract personal context-aware preferences from the context-rich device logs (i.e., context logs) for building novel personalized context-aware recommender systems. A critical challenge along this line is that the context log of each individual user may not contain sufficient data for mining his/her context-aware preferences. Therefore, we propose to first learn common context-aware preferences from the context logs of many users. Then, the preference of each user can be represented as a distribution of these common context-aware preferences. Specifically, we develop two approaches for mining common context-aware preferences based on two different assumptions, namely, context independent and context dependent assumptions, which can fit into different application scenarios. Finally, extensive experiments on a real-world data set show that both approaches are effective and outperform baselines with respect to mining personal context-aware preferences for mobile users.

**2.1.5 GETJAR MOBILE APPLICATION RECOMMENDATIONS WITH VERY SPARSE DATASETS**

The Netflix competition of 2006 has spurred significant activity in the recommendations field, particularly in approaches using latent factor models. However, the near ubiquity of the Netflix and the similar Movie Lens datasets 1 may be narrowing the generality of lessons learned in this field. At Get Jar, our goal is to make appealing recommendations of mobile applications (apps). For app usage, we observe a distribution that has higher kurtosis (heavier head and longer tail) than that for the aforementioned movie datasets. This happens primarily because of the large disparity in resources available to app developers and the low cost of app publication relative to movies. In this paper we compare a latent factor (Pure SVD) and a memory-based model with our novel PCA-based model,which we call Eigenapp.

We use both accuracy and varietyas evaluation metrics. PureSVD did not perform well dueto its reliance on explicit feedback such as ratings, which we do not have. Memory-based approaches that perform vector operations in the original high dimensional space overpredict popular apps because they fail to capture the neighborhood of less popular apps. They have high accuracy due to the concentration of mass in the head, but did poorly in terms of variety of apps exposed. Eigenapp, which exploits neighborhood information in low dimensional spaces, did well both on precision and variety, underscoring the importance of dimensionality reduction to form quality neighborhoods in high kurtosis distributions.

**2.1.6 EXPLOITING ENRICHED CONTEXTUAL INFORMATION FOR MOBILE APP CLASSIFICATION**

A key step for the mobile app usage analysis is to classify apps into some predefined categories. However, it is a nontrivial task to effectively classify mobile apps due to the limited contextual information available for the analysis. To this end, in this paper, we propose an approach to first enrich the contextual information of mobile apps by exploiting the additional Web knowledge from the Web search engine. Then, inspired by the observation that different types of mobile apps may be relevant to different real-world contexts, we also extract some contextual features for mobile apps from the context-rich device logs of mobile users. Finally, we combine all the enriched contextual information into a Maximum Entropy model for training a mobile app classifier. The experimental results based on 443 mobile users’ device logs clearly show that our approach outperforms two state-of-the-art benchmark methods with a significant margin.

**2.1.7 A TAXI BUSINESS INTELLIGENCE SYSTEM**

The increasing availability of large-scale location traces creates unprecedent opportunities to change the paradigm for knowledge discovery in transportation systems. A particularly promising area is to extract useful business intelligence, which can be used as guidance for reducing inefficiencies in energy consumption of transportation sectors, improving customer experiences, and increasing business performances. However, extracting business intelligence from location traces is not a trivial task. Conventional data analytic tools are usually not customized for handling large, complex, dynamic, and distributed nature of location traces. To thatend, we develop a taxi business intelligence system to explore the massive taxi location traces from different businessperspectives with various data mining functions. Since weimplement the system using the real-world taxi GPS data, this demonstration will help taxi companies to improve their business performances by understanding the behaviors of both drivers and customers. In addition, several identified technical challenges also motivate data mining people to develop more sophisticate techniques in the future.