IEEE 2017-18 Data Mining projects

1. CYBERBULLYING DETECTION BASED ON SEMANTIC-ENHANCED MARGINALIZED DENOISING AUTO-ENCODER

"**Cyberbullying**" is when a child, preteen or teen is tormented, threatened, harassed, humiliated, embarrassed or otherwise targeted by another child, preteen or teen using the Internet, interactive and digital technologies or mobile phones .As a side effect of increasingly popular social media, cyberbullying has emerged as a serious problem afflicting children, adolescents and young adults. In this paper, we propose a new representation learning method to tackle this problem. Our method named Semantic-Enhanced Marginalized Denoising Auto-Encoder (smSDA) is developed via semantic extension of the popular deep learning model stacked denoising autoencoder. The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. Our proposed method is able to exploit the hidden feature structure of bullying information and learn a robust and discriminative representation of text.

2. NETSPAM: A NETWORK-BASED SPAM DETECTION FRAMEWORK FOR REVIEWS IN ONLINE SOCIALMEDIA

Nowadays, a big part of people rely on available con-tent in social media in their decisions (e.g. reviews and feedback on a topic or product). The possibility that anybody can leave a review provide a golden opportunity for spammers to write spam reviews about products and services for different interests. Identifying these spammers and the spam content is a hot topic of research and although a considerable number of studies have been done recently toward this end, but so far the methodologies put forth still barely detect spam reviews, and none of them show the importance of each extracted feature type. In this study, we propose a novel framework, named NetSpam, which utilizes spam features for modeling review datasets as

heterogeneous information networks to map spam detection procedure into a classification problem in such networks. Using the importance of spam features help us to obtain better results in terms of different metrics experimented on real-world review datasets from Yelp and Amazon websites.

3. SOCIALQ&A: AN ONLINE SOCIAL NETWORK BASED QUESTION AND ANSWER SYSTEM

Question and Answer (Q&A) systems play a vital role in our daily life for information and knowledge sharing. Users post questions and pick questions to answer in the system. Due to the rapidly growing user population and the number of questions, it is unlikely for a user to stumble upon a question by chance that (s) he can answer. Also, altruism does not encourage all users to provide answers, not to mention high quality answers with a short answer wait time. The primary objective of this paper is to improve the performance of Q&A systems by actively forwarding questions to users who are capable and willing to answer the questions. To this end, we have designed and implemented SocialQ&A, an online social network based Q&A system. SocialQ&A leverages the social network properties of common-interest and mutual-trust friend relationship to identify an asker through friendship who are most likely to answer the question, and enhance the user security. We also improve SocialQ&A with security and efficiency enhancements by protecting user privacy and identifies, and retrieving answers automatically for recurrent questions. We describe the architecture and algorithms, and conducted comprehensive large-scale simulation to evaluate SocialQ&A in comparison with other methods. Our results suggest that social networks can be leveraged to improve the answer quality and asker's waiting time. We also implemented a real prototype of SocialQ&A, and analyze the Q&A behavior of real users and questions from a small-scale real-world SocialQ&A system.

4. SENTIMENT ANALYSIS OF TOP COLLEGES IN INDIA USING TWITTER DATA

In today's world, opinions and reviews accessible to us are one of the most critical factors in formulating our views and influencing the success of a brand, product or service. With the advent and growth of social media in the world, stakeholders often take to expressing their opinions on popular social media, namely twitter. While Twitter data is extremely informative, it presents a challenge for analysis because of its humongous and disorganized nature. This paper is a thorough effort to dive into the novel domain of performing sentiment analysis of people's opinions regarding top colleges in India. Besides taking additional preprocessing measures like the expansion of net lingo and removal of duplicate tweets.

5. FRAPPE: DETECTING MALICIOUS FACEBOOK APPLICATIONS

With 20 million installs a day, third-party apps are a major reason for the popularity and addictiveness of OSNs. Unfortunately, hackers have realized the potential of using apps for spreading malware and spam. The problem is already significant, as we find that at least 13% of apps in our dataset are malicious. So far, the research community has focused on detecting malicious posts, campaigns and applications.

6. MODELING URBAN BEHAVIOR BY MINING GEOTAGGED SOCIAL DATA

Data generated on location-based social networks provide rich information on the whereabouts of urban dwellers. Specifically, such data reveal who spends time where, when, and on what type of activity (e.g., shopping at a mall, or dining at a restaurant). That information can, in turn, be used to describe city regions in terms of activity that takes place therein. For example, the data might reveal that citizens visit one region mainly for shopping in the morning, while another for dining in the evening. Furthermore, once such a description is available, one can ask more elaborate questions. For example, one might ask what features distinguish one region from another—some regions might be different in terms of the type of venues they host and others in terms of the visitors they attract. As another example, one might ask which regions are similar across cities. In this paper, we present a method to answer such questions using publicly shared Foursquare data. Our analysis makes use of a probabilistic model, the features of which include the exact location of activity, the users who participate in the activity, as well as the time of the day and day of week the activity takes place. Compared to previous

approaches to similar tasks, our probabilistic modeling approach allows us to make minimal assumptions about the data—which relieves us from having to set arbitrary parameters in our analysis (e.g., regarding the granularity of discovered regions or the importance of different features). We demonstrate how the model learned with our method can be used to identify the most likely and distinctive features of a geographical area, quantify the importance features used in the model, and discover similar regions across different cities. Finally, we perform an empirical comparison with previous work and discuss insights obtained through our findings.

7. A WORKFLOW MANAGEMENT SYSTEM FOR SCALABLE DATA MINING ON CLOUDS

The extraction of useful information from data is often a complex process that can be conveniently modeled as a data analysis workflow. When very large data sets must be analyzed and/or complex data mining algorithms must be executed, data analysis workflows may take very long times to complete their execution. Therefore, efficient systems are required for the scalable execution of data analysis workflows, by exploiting the computing services of the Cloud platforms where data is increasingly being stored. The objective of the paper is to demonstrate how Cloud software technologies can be integrated to implement an effective environment for designing and executing scalable data analysis workflows. We describe the design and implementation of the Data Mining Cloud Framework (DMCF), a data analysis system that integrates a visual workflow language and a parallel runtime with the Software-as-a-Service (SaaS) model. DMCF was designed taking into account the needs of real data mining applications, with the goal of simplifying the development of data mining applications compared to generic workflow management systems that are not specifically designed for this domain. The result is a high-level environment that, through an integrated visual workflow language, minimizes the programming effort, making easier to domain experts the use of common patterns specifically designed for the development and the parallel execution of data mining applications. The DMCF's visual workflow language, system architecture and runtime mechanisms are presented. We also discuss several data mining workflows developed with DMCF and the scalability obtained executing such workflows on a public Cloud.

8. FIDOOP: PARALLEL MINING OF FREQUENT ITEMSETS USING MAPREDUCE

Existing parallel mining algorithms for frequent itemsets lack a mechanism that enables automatic parallelization, load balancing, data distribution, and fault tolerance on large

clusters. As a solution to this problem, we design a parallel frequent itemsets mining algorithm called FiDoop using the MapReduce programming model. To achieve compressed storage and avoid building conditional pattern bases, FiDoop incorporates the frequent items ultrametric tree, rather than conventional FP trees. In FiDoop, three MapReduce jobs are implemented to complete the mining task. In the crucial third MapReduce job, the mappers independently decompose itemsets, the reducers perform combination operations by constructing small ultrametric trees, and the actual mining of these trees separately. We implement FiDoop on our in-house Hadoop cluster. We show that FiDoop on the cluster is sensitive to data distribution and dimensions, because itemsets with different lengths have different decomposition and construction costs. To improve FiDoop's performance, we develop a workload balance metric to measure load balance across the cluster's computing nodes. We develop FiDoop-HD, an extension of FiDoop, to speed up the mining performance for high-dimensional data analysis. Extensive experiments using real-world celestial spectral data demonstrate that our proposed solution is efficient and scalable.

9. INVERTED LINEAR QUADTREE: EFFICIENT TOP K SPATIAL KEYWORD SEARCH

In this paper, With advances in geo-positioning technologies and geo-location services, there are a rapidly growing amount of spatiotextual objects collected in many applications such as location based services and social networks, in which an object is described by its spatial location and a set of keywords (terms). Consequently, the study of spatial keyword search which explores both location and textual description of the objects has attracted great attention from the commercial organizations and research communities. In the paper, we study two fundamental problems in the spatial keyword queries: top k spatial keyword search (TOPK-SK), and batch top k spatial keyword search (BTOPK-SK). Given a set of spatio-textual objects, a query location and a set of query keywords, the TOPK-SK retrieves the closest k objects each of which contains all keywords in the query. BTOPK-SK is the batch processing of sets of TOPK-SK queries. Based on the inverted index and the linear quadtree, we propose a novel index structure, called inverted linear quadtree (IL- Quadtree), which is carefully designed to exploit both spatial and keyword based pruning techniques to effectively reduce the search space. An efficient algorithm is then developed to tackle top k spatial keyword search. To further enhance the filtering capability of the signature of linear quadtree, we propose a partition based method. In addition, to deal with BTOPK-SK, we design a new computing paradigm which

partition the queries into groups based on both spatial proximity and the textual relevance between queries. We show that the IL-Quadtree technique can also efficiently support BTOPK-SK. Comprehensive experiments on real and synthetic data clearly demonstrate the efficiency of our methods.

10. TRUTH DISCOVERY IN CROWDSOURCED DETECTION OF SPATIAL EVENTS

The ubiquity of smartphones has led to the emergence of mobile crowdsourcing tasks such as the detection of spatial events when smartphone users move around in their daily lives. However, the credibility of those detected events can be negatively impacted by unreliable participants with low-quality data. Consequently, a major challenge in quality control is to discover true events from diverse and noisy participants' reports. This truth discovery problem is uniquely distinct from its online counterpart in that it involves uncertainties in both participants' mobility and reliability. Decoupling these two types of uncertainties through location tracking will raise severe privacy and energy issues, whereas simply ignoring missing reports or treating them as negative reports will significantly degrade the accuracy of the discovered truth. In this paper, we propose a new method to tackle this truth discovery problem through principled probabilistic modeling. In particular, we integrate the modeling of location popularity, location visit indicators, truth of events and three-way participant reliability in a unified framework. The proposed model is thus capable of efficiently handling various types of uncertainties and automatically discovering truth without any supervision or the need of location tracking. Experimental results demonstrate that our proposed method outperforms existing state-of-the-art truth discovery approaches in the mobile crowdsourcing environment.

11. SPORE: A SEQUENTIAL PERSONALIZED SPATIAL ITEM RECOMMENDER SYSTEM

With the rapid development of location-based social networks (LBSNs), spatial item recommendation has become an important way of helping users discover interesting locations to increase their engagement with location-based services. Although human movement exhibits sequential patterns in LBSNs, most current studies on spatial item recommendations do not consider the sequential influence of locations. Leveraging sequential patterns in spatial item recommendation is, however, very challenging, considering 1) users' check-in data in LBSNs has a low sampling rate in both space and time, which renders existing prediction techniques on GPS trajectories ineffective; 2) the prediction space is extremely large, with

millions of distinct locations as the next prediction target, which impedes the application of classical Markov chain models; and 3) there is no existing framework that unifies users' personal interests and the sequential influence in a principled manner. In light of the above challenges, we propose a sequential personalized spatial item recommendation framework (SPORE) which introduces a novel latent variable topic-region to model and fuse sequential influence with personal interests in the latent and exponential space. The advantages of modeling the sequential effect at the topic-region level include a significantly reduced prediction space, an effective alleviation of data sparsity and a direct expression of the semantic meaning of users' spatial activities. Furthermore, we design an asymmetric Locality Sensitive Hashing (ALSH) technique to speed up the online top-k recommendation process by extending the traditional LSH. We evaluate the performance of SPORE on two real datasets and one large-scale synthetic dataset. The results demonstrate a significant improvement in SPORE's ability to recommend spatial items, in terms of both effectiveness and efficiency, compared with the state-of-the-art methods.

12. A NOVEL RECOMMENDATION MODEL REGULARIZED WITH USER TRUST AND ITEM RATINGS

We propose TrustSVD, a trust-based matrix factorization technique for recommendations. TrustSVD integrates multiple information sources into the recommendation model in order to reduce the data sparsity and cold start problems and their degradation of recommendation performance. An analysis of social trust data from four real-world data sets suggests that not only the explicit but also the implicit influence of both ratings and trust should be taken into consideration in a recommendation model. TrustSVD therefore builds on top of a state-of-the-art recommendation algorithm, SVD++ (which uses the explicit and implicit influence of rated items), by further incorporating both the explicit and implicit influence of trusted and trusting users on the prediction of items for an active user. The proposed technique is the first to extend SVD++ with social trust information. Experimental results on the four data sets demonstrate that TrustSVD achieves better accuracy than other ten counterparts recommendation techniques.

13. AUTOMATICALLY MINING FACETS FOR QUERIES FROM THEIR SEARCH RESULTS

We address the problem of finding query facets which are multiple groups of words or phrases that explain and summarize the content covered by a query. We assume that the important aspects of a query are usually presented and repeated in the query's top retrieved documents

in the style of lists, and query facets can be mined out by aggregating these significant lists. We propose a systematic solution, which we refer to as QDMiner, to automatically mine query facets by extracting and grouping frequent lists from free text, HTML tags, and repeat regions within top search results. Experimental results show that a large number of lists do exist and useful query facets can be mined by QDMiner. We further analyze the problem of list duplication, and find better query facets can be mined by modeling fine-grained similarities between lists and penalizing the duplicated lists.

14. BUILDING AN INTRUSION DETECTION SYSTEM USING A FILTER-BASED FEATURE SELECTION ALGORITHM

Redundant and irrelevant features in data have caused a long-term problem in network traffic classification. These features not only slow down the process of classification but also prevent a classifier from making accurate decisions, especially when coping with big data. In this paper, we propose a mutual information based algorithm that analytically selects the optimal feature for classification. This mutual information based feature selection algorithm can handle linearly and nonlinearly dependent data features. Its effectiveness is evaluated in the cases of network intrusion detection. An Intrusion Detection System (IDS), named Least Square Support Vector Machine based IDS (LSSVM-IDS), is built using the features selected by our proposed feature selection algorithm. The performance of LSSVM-IDS is evaluated using three intrusion detection evaluation datasets, namely KDD Cup 99, NSL-KDD and Kyoto 2006+ dataset. The evaluation results show that our feature selection algorithm contributes more critical features for LSSVM-IDS to achieve better accuracy and lower computational cost compared with the state-of-the-art methods.

15. CONNECTING SOCIAL MEDIA TO E-COMMERCE: COLD-START PRODUCT RECOMMENDATION USING MICROBLOGGING INFORMATION

In recent years, the boundaries between e-commerce and social networking have become increasingly blurred. Many e-commerce websites support the mechanism of social login where users can sign on the websites using their social network identities such as their Facebook or Twitter accounts. Users can also post their newly purchased products on microblogs with links

to the e-commerce product web pages. In this paper we propose a novel solution for cross-site cold-start product recommendation which aims to recommend products from e-commerce websites to users at social networking sites in “cold-start” situations, a problem which has rarely been explored before. A major challenge is how to leverage knowledge extracted from social networking sites for cross-site cold-start product recommendation. We propose to use the linked users across social networking sites and e-commerce websites (users who have social networking accounts and have made purchases on e-commerce websites) as a bridge to map users’ social networking features to another feature representation for product recommendation. In specific, we propose learning both users’ and products’ feature representations (called user embeddings and product embeddings, respectively) from data collected from e-commerce websites using recurrent neural networks and then apply a modified gradient boosting trees method to transform users’ social networking features into user embeddings. We then develop a feature-based matrix factorization approach which can leverage the learnt user embeddings for cold-start product recommendation. Experimental results on a large dataset constructed from the largest Chinese microblogging service SINA WEIBO and the largest Chinese B2C e-commerce website JINGDONG have shown the effectiveness of our proposed framework.

16. CROSS-DOMAIN SENTIMENT CLASSIFICATION USING SENTIMENT SENSITIVE EMBEDDINGS

Unsupervised Cross-domain Sentiment Classification is the task of adapting a sentiment classifier trained on a particular domain (source domain), to a different domain (target domain), without requiring any labeled data for the target domain. By adapting an existing sentiment classifier to previously unseen target domains, we can avoid the cost for manual data annotation for the target domain. We model this problem as embedding learning, and construct three objective functions that capture: (a) distributional properties of pivots (i.e., common features that appear in both source and target domains), (b) label constraints in the source domain documents, and (c) geometric properties in the unlabeled documents in both source and target domains. Unlike prior proposals that first learn a lower-dimensional embedding independent of the source domain sentiment labels, and next a sentiment classifier in this embedding, our joint optimisation method learns embeddings that are sensitive to sentiment classification. Experimental results on a benchmark dataset show that by jointly optimising the three objectives we can obtain better performances in comparison to optimizing each objective function separately, thereby demonstrating the importance of task-specific

embedding learning for cross-domain sentiment classification. Among the individual objective functions, the best performance is obtained by (c). Moreover, the proposed method reports cross-domain sentiment classification accuracies that are statistically comparable to the current state-of-the-art embedding learning methods for cross-domain sentiment classification.

17. POINT-OF-INTEREST RECOMMENDATION FOR LOCATION PROMOTION IN LOCATIONBASED SOCIAL NETWORKS

Point-of-interest (POI) recommendation that suggests new places for users to visit arises with the popularity of location-based social networks (LBSNs). Due to the importance of POI recommendation in LBSNs, it has attracted much academic and industrial interest. In this paper, we offer a systematic review of this field, summarizing the contributions of individual efforts and exploring their relations. We discuss the new properties and challenges in POI recommendation, compared with traditional recommendation problems, e.g., movie recommendation. Then, we present a comprehensive review in three aspects: influential factors for POI recommendation, methodologies employed for POI recommendation, and different tasks in POI recommendation. Specifically, we propose three taxonomies to classify POI recommendation systems. First, we categorize the systems by the influential factors check-in characteristics, including the geographical information, social relationship, temporal influence, and content indications. Second, we categorize the systems by the methodology, including systems modeled by fused methods and joint methods. Third, we categorize the systems as general POI recommendation and successive POI recommendation by subtle differences in the recommendation task whether to be bias to the recent check-in. For each category, we summarize the contributions and system features, and highlight the representative work. Moreover, we discuss the available data sets and the popular metrics. Finally, we point out the possible future directions in this area and conclude this survey.