Master-Seminar 1: Data Analytics

Instructor : Ahmed Rashed

Tuesday 14:00 - 16:00
Room: B 026

Online Social Networks are a popular way for people to interact, communicate, express themselves, and share contents. These platforms provide a rich ground for various machine learning applications that can analyze user behavior and interactions for recommender systems, link prediction, location prediction, event detection, and sentiment analysis. The aim of this seminar is to expose the students to the applications of machine learning in the domain of social networks. It will enhance the students' abilities to comprehend, explain and criticize state-of-the-art research. On the other hand, it serves as a playground for developing analytical thinking.

Reading List

Category : Recommender Systems

#	Authors	Title & Abstract	Venue	Year	Pres. Date
1	Rendle et al.,	**Title:** BPR: Bayesian Personalized Ranking from Implicit Feedback **Abstract:** Item recommendation is the task of predicting a personalized ranking on a set of items (e.g. websites, movies, products). In this paper, we investigate the most common scenario with implicit feedback (e.g. clicks, purchases). There are many methods for item recommendation from implicit feedback like matrix factorization (MF) or adaptive k-nearest-neighbor (kNN). Even though these methods are designed for the item prediction task of personalized ranking, none of them is directly optimized for ranking. In this paper we present a generic optimization criterion BPR-Opt for personalized ranking that is the maximum posterior estimator derived from a Bayesian analysis of the problem. We also provide a generic learning algorithm for optimizing models with respect to BPR-Opt. The learning method is based on stochastic gradient descent with bootstrap sampling. We show how to apply our method to two state-of-the-art recommender models: matrix factorization and adaptive kNN. Our experiments indicate that for the task of personalized ranking our optimization method outperforms the standard learning techniques for MF and kNN. The results show the importance of optimizing models for the right criterion.	UAI	2009	20.11.2018
2	Krohn-Grimberghe et al.,	**Title:** Multi-relational matrix factorization using bayesian personalized ranking for social network data **Abstract:** A key element of the social networks on the internet such as Facebook and Flickr is that they encourage users to create connections between themselves, other users and objects. One important task that has been approached in the literature that deals with such data is to use social graphs to predict user behavior (e.g. joining a group of interest). More specifically, we study the cold-start problem, where users only participate in some relations, which we will call social relations, but not in the relation on which the predictions are made, which we will refer to as target relations. We propose a formalization of the problem and a principled approach to it based on multi-relational factorization techniques. Furthermore, we derive a principled feature extraction scheme from the social data to extract predictors for a classifier on the target relation. Experiments conducted on real world datasets show that our approach outperforms current methods.	WSDM	2012	20.11.2018

3	Tang et al.,	**Title:** Leveraging social media networks for classification **Abstract:** Social media has reshaped the way in which people interact with each other. The rapid development of participatory web and social networking sites like YouTube, Twitter, and Facebook, also brings about many data mining opportunities and novel challenges. In particular, we focus on classification tasks with user interaction information in a social network. Networks in social media are heterogeneous, consisting of various relations. Since the relation-type information may not be available in social media, most existing approaches treat these inhomogeneous connections homogeneously, leading to an unsatisfactory classification performance. In order to handle the network heterogeneity, we propose the concept of social dimension to represent actors' latent affiliations, and develop a classification framework based on that. The proposed framework, SocioDim, first extracts social dimensions based on the network structure to accurately capture prominent interaction patterns between actors, then learns a discriminative classifier to select relevant social dimensions. SocioDim, by differentiating different types of network connections, outperforms existing representative methods of classification in social media, and offers a simple yet effective approach to integrating two types of seemingly orthogonal information: the network of actors and their attributes.	DMKD	2011	27.11.2018
4	Perozzi et al.,	**Title:** DeepWalk: Online Learning of Social Representations **Abstract:**We present DeepWalk, a novel approach for learning latent representations of vertices in a network. These latent representations encode social relations in a continuous vector space, which is easily exploited by statistical models. DeepWalk generalizes recent advancements in language modeling and unsupervised feature learning (or deep learning) from sequences of words to graphs. DeepWalk uses local information obtained from truncated random walks to learn latent representations by treating walks as the equivalent of sentences. We demonstrate DeepWalk's latent representations on several multi-label network classification tasks for social networks such as BlogCatalog, Flickr, and YouTube. Our results show that DeepWalk outperforms challenging baselines which are allowed a global view of the network, especially in the presence of missing information. DeepWalk's representations can provide F1 scores up to 10% higher than competing methods when labeled data is sparse. In some experiments, DeepWalk's representations are able to outperform all baseline methods while using 60% less training data. DeepWalk is also scalable. It is an online learning algorithm which builds useful incremental results, and is trivially parallelizable. These qualities make it suitable for a broad class of real world applications such as network classification, and anomaly detection.	SIGKDD	2014	27.11.2018
5	Wang et al.,	**Title:** Collaborative Deep Learning for Recommender Systems **Abstract:** Collaborative filtering (CF) is a successful approach commonly used by many recommender systems. Conventional CF-based methods use the ratings given to items by users as the sole source of information for learning to make recommendation. However, the ratings are often very sparse in many applications, causing CF-based methods to degrade significantly in their recommendation performance. To address this sparsity problem, auxiliary information such as item content information may be utilized. Collaborative topic regression (CTR) is an appealing recent method taking this approach which tightly couples the two components that learn from two different sources of information. Nevertheless, the latent representation learned by CTR may not be very effective when the auxiliary information is very sparse. To address this problem, we generalize recently advances in deep learning from i.i.d. input to non-i.i.d. (CF-based) input and propose in this paper a hierarchical Bayesian model called collaborative deep learning (CDL), which jointly performs deep representation learning for the content information and collaborative filtering for the ratings (feedback) matrix. Extensive experiments on three real-world datasets from different domains show that CDL can significantly advance the state of the art.	SIGKDD	2015	4.12.2018

6	Chen et al.,	**Title:** On sampling strategies for neural network-based collaborative filtering **Abstract:** Recent advances in neural networks have inspired people to design hybrid recommendation algorithms that can incorporate both (1) user-item interaction information and (2) content information including image, audio, and text. Despite their promising results, neural network-based recommendation algorithms pose extensive computational costs, making it challenging to scale and improve upon. In this paper, we propose a general neural network-based recommendation framework, which subsumes several existing state-of-the-art recommendation algorithms, and address the efficiency issue by investigating sampling strategies in the stochastic gradient descent training for the framework. We tackle this issue by first establishing a connection between the loss functions and the user-item interaction bipartite graph, where the loss function terms are defined on links while major computation burdens are located at nodes. We call this type of loss functions "graph-based" loss functions, for which varied mini-batch sampling strategies can have different computational costs. Based on the insight, three novel sampling strategies are proposed, which can significantly improve the training efficiency of the proposed framework (up to $\times 30$ times speedup in our experiments), as well as improving the recommendation performance. Theoretical analysis is also provided for both the computational cost and the convergence. We believe the study of sampling strategies have further implications on general graph-based loss functions, and would also enable more research under the neural network-based recommendation framework.	SIGKDD	2017	4.12.2018

Category : Link Prediction

#	Authors	Title & Abstract	Venue	Year	Pres. Date
7	Tang et al.,	**Title:** Negative Link Prediction in Social Media **Abstract:** Signed network analysis has attracted increasing attention in recent years. This is in part because research on signed network analysis suggests that negative links have added value in the analytical process. A major impediment in their effective use is that most social media sites do not enable users to specify them explicitly. In other words, a gap exists between the importance of negative links and their availability in real data sets. Therefore, it is natural to explore whether one can predict negative links automatically from the commonly available social network data. In this paper, we investigate the novel problem of negative link prediction with only positive links and content-centric interactions in social media. We make a number of important observations about negative links, and propose a principled framework NeLP, which can exploit positive links and content-centric interactions to predict negative links. Our experimental results on real-world social networks demonstrate that the proposed NeLP framework can accurately predict negative links with positive links and content-centric interactions. Our detailed experiments also illustrate the relative importance of various factors to the effectiveness of the proposed framework.	WSDM	2015	11.12.2018

8	Dong et al.,	**Title:** Link Prediction and Recommendation across Heterogeneous Social Networks **Abstract:** Link prediction and recommendation is a fundamental problem in social network analysis. The key challenge of link prediction comes from the sparsity of networks due to the strong disproportion of links that they have potential to form to links that do form. Most previous work tries to solve the problem in single network, few research focus on capturing the general principles of link formation across heterogeneous networks. In this work, we give a formal definition of link recommendation across heterogeneous networks. Then we propose a ranking factor graph model (RFG) for predicting links in social networks, which effectively improves the predictive performance. Motivated by the intuition that people make friends in different networks with similar principles, we find several social patterns that are general across heterogeneous networks. With the general social patterns, we develop a transfer-based RFG model that combines them with network structure information. This model provides us insight into fundamental principles that drive the link formation and network evolution. Finally, we verify the predictive performance of the presented transfer model on 12 pairs of transfer cases. Our experimental results demonstrate that the transfer of general social patterns indeed help the prediction of links.	ICDM	2012	11.12.2018
9	Valverde-Rebaza et al.,	**Title:** Exploiting behaviors of communities of twitter users for link prediction **Abstract:** Currently, online social networks and social media have become increasingly popular showing an exponential growth. This fact have attracted increasing research interest and, in turn, facilitating the emergence of new interdisciplinary research directions, such as social network analysis. In this scenario, link prediction is one of the most important tasks since it deals with the problem of the existence of a future relation among members in a social network. Previous techniques for link prediction were based on structural (or topological) information. Nevertheless, structural information is not enough to achieve a good performance in the link prediction task on large-scale social networks. Thus, the use of additional information, such as interests or behaviors that nodes have into their communities, may improve the link prediction performance. In this paper, we analyze the viability of using a set of simple and non-expensive techniques that combine structural with community information for predicting the existence of future links in a large-scale online social network, such as Twitter. Twitter, a microblogging service, has emerged as a useful source of informative data shared by millions of users whose relationships require no reciprocation. Twitter network was chosen because it is not well understood, mainly due to the occurrence of directed and asymmetric links yet. Experiments show that our proposals can be used efficiently to improve unsupervised and supervised link prediction task in a directed and asymmetric large-scale network.	SNAM	2013	18.12.2018
10	Wang et al.,	**Title:** Signed Network Embedding in Social Media **Abstract:** Network embedding is to learn low-dimensional vector representations for nodes of a given social network, facilitating many tasks in social network analysis such as link prediction. The vast majority of existing embedding algorithms are designed for unsigned social networks or social networks with only positive links. However, networks in social media could have both positive and negative links, and little work exists for signed social networks. From recent findings of signed network analysis, it is evident that negative links have distinct properties and added value besides positive links, which brings about both challenges and opportunities for signed network embedding. In this paper, we propose a deep learning framework SiNE for signed network embedding. The framework optimizes an objective function guided by social theories that provide a fundamental understanding of signed social networks. Experimental results on two real-world datasets of social media demonstrate the effectiveness of the proposed framework SiNE.	SIAM	2017	18.12.2018

Category : Location Prediction

#	Authors	Title & Abstract	Venue	Year	Pres. Date
11	McGee et al.,	**Title:** Location prediction in social media based on tie strength **Abstract:** We propose a novel network-based approach for location estimation in social media that integrates evidence of the social tie strength between users for improved location estimation. Concretely, we propose a location estimator – FriendlyLocation – that leverages the relationship between the strength of the tie between a pair of users, and the distance between the pair. Based on an examination of over 100 million geo-encoded tweets and 73 million Twitter user profiles, we identify several factors such as the number of followers and how the users interact that can strongly reveal the distance between a pair of users. We use these factors to train a decision tree to distinguish between pairs of users who are likely to live nearby and pairs of users who are likely to live in different areas. We use the results of this decision tree as the input to a maximum likelihood estimator to predict a user's location. We find that this proposed method significantly improves the results of location estimation relative to a state-of-the-art technique. Our system reduces the average error distance for 80% of Twitter users from 40 miles to 21 miles using only information from the user's friends and friends-of-friends, which has great significance for augmenting traditional social media and enriching location-based services with more refined and accurate location estimates.	CIKM	2013	8.1.2019
12	Rout et al.,	**Title:** Where's @wally?: a classification approach to geolocating users based on their social ties **Abstract:** This paper presents an approach to geolocating users of online social networks, based solely on their 'friendship' connections. We observe that users interact more regularly with those closer to themselves and hypothesise that, in many cases, a person's social network is sufficient to reveal their location. The geolocation problem is formulated as a classification task, where the most likely city for a user without an explicit location is chosen amongst the known locations of their social ties. Our method uses an SVM classifier and a number of features that reflect different aspects and characteristics of Twitter user networks. The SVM classifier is trained and evaluated on a dataset of Twitter users with known locations. Our method outperforms a state-of-the-art method for geolocating users based on their social ties.	HT	2013	8.1.2019
13	Cho et al.,	**Title:** Friendship and mobility: user movement in location-based social networks **Abstract:**Even though human movement and mobility patterns have a high degree of freedom and variation, they also exhibit structural patterns due to geographic and social constraints. Using cell phone location data, as well as data from two online location-based social networks, we aim to understand what basic laws govern human motion and dynamics. We find that humans experience a combination of periodic movement that is geographically limited and seemingly random jumps correlated with their social networks. Short-ranged travel is periodic both spatially and temporally and not effected by the social network structure, while long-distance travel is more influenced by social network ties. We show that social relationships can explain about 10% to 30% of all human movement, while periodic behavior explains 50% to 70%. Based on our findings, we develop a model of human mobility that combines periodic short range movements with travel due to the social network structure. We show that our model reliably predicts the locations and dynamics of future human movement and gives an order of magnitude better performance than present models of human mobility.	SIGKDD	2011	15.1.2019

14	Cheng et al.,	**Title:** You Are Where You Tweet: A Content-Based Approach to Geo-locating Twitter Users **Abstract:**We propose and evaluate a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets, even in the absence of any other geospatial cues. By augmenting the massive human-powered sensing capabilities of Twitter and related microblogging services with content-derived location information, this framework can overcome the sparsity of geo-enabled features in these services and enable new location-based personalized information services, the targeting of regional advertisements, and so on. Three of the key features of the proposed approach are: (i) its reliance purely on tweet content, meaning no need for user IP information, private login information, or external knowledge bases; (ii) a classification component for automatically identifying words in tweets with a strong local geo-scope; and (iii) a lattice-based neighborhood smoothing model for refining a user's location estimate. The system estimates k possible locations for each user in descending order of confidence. On average we find that the location estimates converge quickly (needing just 100s of tweets), placing 51% of Twitter users within 100 miles of their actual location.	CIKM	2010	15.1.2019

Category : Event Detection

#	Authors	Title & Abstract	Venue	Year	Pres. Date
15	Ritter et al.,	**Title:** Open domain event extraction from twitter **Abstract:** Tweets are the most up-to-date and inclusive stream of in- formation and commentary on current events, but they are also fragmented and noisy, motivating the need for systems that can extract, aggregate and categorize important events. Previous work on extracting structured representations of events has focused largely on newswire text; Twitter's unique characteristics present new challenges and opportunities for open-domain event extraction. This paper describes TwiCal– the first open-domain event-extraction and categorization system for Twitter. We demonstrate that accurately extracting an open-domain calendar of significant events from Twitter is indeed feasible. In addition, we present a novel approach for discovering important event categories and classifying extracted events based on latent variable models. By leveraging large volumes of unlabeled data, our approach achieves a 14% increase in maximum F1 over a supervised baseline.	SIGKDD	2012	22.1.2019
16	Zhang et al.,	**Title:** Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams **Abstract:** Detecting local events (e.g., protest, disaster) at their onsets is an important task for a wide spectrum of applications, ranging from disaster control to crime monitoring and place recommendation. Recent years have witnessed growing interest in leveraging geo-tagged tweet streams for online local event detection. Nevertheless, the accuracies of existing methods still remain unsatisfactory for building reliable local event detection systems. We propose TrioVecEvent, a method that leverages multimodal embeddings to achieve accurate online local event detection. The effectiveness of TrioVecEvent is underpinned by its two-step detection scheme. First, it ensures a high coverage of the underlying local events by dividing the tweets in the query window into coherent geo-topic clusters. To generate quality geo-topic clusters, we capture short-text semantics by learning multimodal embeddings of the location, time, and text, and then perform online clustering with a novel Bayesian mixture model. Second, TrioVecEvent considers the geo-topic clusters as candidate events and extracts a set of features for classifying the candidates. Leveraging the multimodal embeddings as background knowledge, we introduce discriminative features that can well characterize local events, which enables pinpointing true local events from the candidate pool with a small amount of training data. We have used crowdsourcing to evaluate TrioVecEvent, and found that it improves the performance of the state-of-the-art method by a large margin.	SIGKDD	2017	22.1.2019

17	Zhang et al.,	**Title:** GeoBurst: Real-Time Local Event Detection in Geo-Tagged Tweet Streams **Abstract:**The real-time discovery of local events (e.g., protests, crimes, disasters) is of great importance to various applications, such as crime monitoring, disaster alarming, and activity recommendation. While this task was nearly impossible years ago due to the lack of timely and reliable data sources, the recent explosive growth in geo-tagged tweet data brings new opportunities to it. That said, how to extract quality local events from geo-tagged tweet streams in real time remains largely unsolved so far. We propose GeoBurst, a method that enables effective and real-time local event detection from geo-tagged tweet streams. With a novel authority measure that captures the geo-topic correlations among tweets, GeoBurst first identifies several pivots in the query window. Such pivots serve as representative tweets for potential local events and naturally attract similar tweets to form candidate events. To select truly interesting local events from the candidate list, GeoBurst further summarizes continuous tweet streams and compares the candidates against historical activities to obtain spatiotemporally bursty ones. Finally, GeoBurst also features an updating module that finds new pivots with little time cost when the query window shifts. As such, GeoBurst is capable of monitoring continuous streams in real time. We used crowdsourcing to evaluate GeoBurst on two real-life data sets that contain millions of geo-tagged tweets. The results demonstrate that GeoBurst significantly outperforms state-of-the-art methods in precision, and is orders of magnitude faster.	SIGIR	2016	29.1.2019
18	Abdelhaq et al.,	**Title:** EvenTweet: Online Localized Event Detection from Twitter **Abstract:** Microblogging services such as Twitter, Facebook, and Foursquare have become major sources for information about real-world events. Most approaches that aim at extracting event information from such sources typically use the temporal context of messages. However, exploiting the location information of georeferenced messages, too, is important to detect localized events, such as public events or emergency situations. Users posting messages that are close to the location of an event serve as human sensors to describe an event. In this demonstration, we present a novel framework to detect localized events in real-time from a Twitter stream and to track the evolution of such events over time. For this, spatio-temporal characteristics of keywords are continuously extracted to identify meaningful candidates for event descriptions. Then, localized event information is extracted by clustering keywords according to their spatial similarity. To determine the most important events in a (recent) time frame, we introduce a scoring scheme for events. We demonstrate the functionality of our system, called Even-Tweet, using a stream of tweets from Europe during the 2012 UEFA European Football Championship.	VLDB	2013	29.1.2019

Category : Sentiment Analysis

#	Authors	Title & Abstract	Venue	Year	Pres. Date
19	Tang et al.,	**Title:** Coooolll: A Deep Learning System for Twitter Sentiment Classification **Abstract:**In this paper, we develop a deep learning system for message-level Twitter sentiment classification. Among the 45 submitted systems including the SemEval 2013 participants, our system (Coooolll) is ranked 2nd on the Twitter2014 test set of SemEval 2014 Task 9. Coooolll is built in a supervised learning framework by concatenating the sentiment-specific word embedding (SSWE) features with the state-of-the-art hand-crafted features. We develop a neural network with hybrid loss function 1 to learn SSWE, which encodes the sentiment information of tweets in the continuous representation of words. To obtain large-scale training corpora, we train SSWE from 10M tweets collected by positive and negative emoticons, without any manual annotation. Our system can be easily re-implemented with the publicly available sentiment-specific word embedding.	SemEval	2014	5.2.2019

20	Tang et al.,	**Title:** Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification **Abstract:** We present a method that learns word embedding for Twitter sentiment classification in this paper. Most existing algorithms for learning continuous word representations typically only model the syntactic context of words but ignore the sentiment of text. This is problematic for sentiment analysis as they usually map words with similar syntactic context but opposite sentiment polarity, such as good and bad, to neighboring word vectors. We address this issue by learning sentimentspecific word embedding (SSWE), which encodes sentiment information in the continuous representation of words. Specifically, we develop three neural networks to effectively incorporate the supervision from sentiment polarity of text (e.g. sentences or tweets) in their loss functions. To obtain large scale training corpora, we learn the sentiment-specific word embedding from massive distant-supervised tweets collected by positive and negative emoticons. Experiments on applying SSWE to a benchmark Twitter sentiment classification dataset in SemEval 2013 show that (1) the SSWE feature performs comparably with hand-crafted features in the top-performed system; (2) the performance is further improved by concatenating SSWE with existing feature set.	VLDB	2014	5.2.2019