FINAL YEAR PROJECTS

DATA MINING 2018 - 2019

TITLES WITH ABSTRACTS

CALL US

(+91) 9944 7933 98
(+91) 452-439 0702, 439 2702

19 Years of Experience  |  Automated Services  |  24/7 Help Desk Support
Advanced Technologies and tools  |  Legitimate Members of all Journals

Elysium PRO

Titles with Abstracts 2018-19
Due to the flexibility in modelling data heterogeneity, heterogeneous information network (HIN) has been adopted to characterize complex and heterogeneous auxiliary data in recommender systems, called HIN based recommendation. It is challenging to develop effective methods for HIN based recommendation in both extraction and exploitation of the information from HINs. Most of HIN based recommendation methods rely on path based similarity, which cannot fully mine latent structure features of users and items. In this paper, we propose a novel heterogeneous network embedding based approach for HIN based recommendation, called HERec. To embed HINs, we design a meta-path based random walk strategy to generate meaningful node sequences for network embedding. The learned node embeddings are first transformed by a set of fusion functions, and subsequently integrated into an extended matrix factorization (MF) model. The extended MF model together with fusion functions are jointly optimized for the rating prediction task. Extensive experiments on three real-world datasets demonstrate the effectiveness of the HERec model. Moreover, we show the capability of the HERec model for the cold-start problem, and reveal that the transformed embedding information from HINs can improve the recommendation performance.

Mining High Average-Utility Itemsets (HAUIs) in a quantitative database is an extension of the traditional problem of frequent itemset mining, having several practical applications. Discovering HAUIs is more challenging than mining frequent itemsets using the traditional support model since the average-utilities of itemsets do not satisfy the downward-closure property. To design algorithms for mining HAUIs that reduce the search space of itemsets, prior studies have proposed various upper-bounds on the average-utilities of itemsets. However, these algorithms can generate a huge amount of unpromising HAUI candidates, which result in high memory consumption and long runtimes. To address this problem, this paper proposes four tight average-utility upper-bounds, based on a vertical database representation, and three efficient pruning strategies. Furthermore, a novel generic framework for comparing average-utility upper-bounds is presented. Based on these theoretical results, an efficient algorithm named dHAUIM is introduced for mining the complete set of HAUIs. dHAUIM represents the search space and quickly compute upper-bounds using a novel IDUL structure. Extensive experiments show that dHAUIM outperforms three state-of-the-art algorithms for mining HAUIs in terms of runtime on both real-life and synthetic databases. Moreover, results show that the proposed pruning strategies dramatically reduce the number of candidate HAUIs.
The increasing interest in collecting and publishing large amounts of individuals' data to public for purposes such as medical research, market analysis and economical measures has created major privacy concerns about individual's sensitive information. To deal with these concerns, many Privacy-Preserving Data Publishing (PPDP) techniques have been proposed in literature. However, they lack a proper privacy characterization and measurement. In this paper, we first present a novel multi-variable privacy characterization and quantification model. Based on this model, we are able to analyze the prior and posterior adversarial belief about attribute values of individuals. Then we show that privacy should not be measured based on one metric. We demonstrate how this could result in privacy misjudgment. We propose two different metrics for quantification of privacy leakage, distribution leakage and entropy leakage. Using these metrics, we analyzed some of the most well-known PPDP techniques such as $k$-anonymity, $l$-diversity and $t$-closeness. Based on our framework and the proposed metrics, we can determine that all the existing PPDP schemes have limitations in privacy characterization. Our proposed privacy characterization and measurement framework contributes to better understanding and evaluation of these techniques. Thus this paper provides a foundation for design and analysis of PPDP schemes.

Uncertainty is intrinsic to a wide spectrum of real-life applications, which inevitably applies to graph data. Representative uncertain graphs are seen in bio-informatics, social networks, etc. This paper motivates the problem of frequent subgraph mining on single uncertain graphs, and investigates two different - probabilistic and expected - semantics in terms of support definitions. First, we present an enumeration-evaluation algorithm to solve the problem under probabilistic semantics. By showing the support computation under probabilistic semantics is #P-complete, we develop an approximation algorithm with accuracy guarantee for efficient problem-solving. To enhance the solution, we devise computation sharing techniques to achieve better mining performance. Afterwards, the algorithm is extended in a similar flavor to handle the problem under expected semantics, where checkpoint-based pruning and validation techniques are integrated. Experiment results on real-life datasets confirm the practical usability of the mining algorithms.
The intelligence of Smart Cities (SC) is represented by its ability in collecting, managing, integrating, analyzing and mining multi-source data for valuable insights. In order to harness multi-source data for an informed place design, this paper presents "Public Sentiments and Activities in Places" multi-source data analysis flow (PSAP) in an Informed Design Platform (IDP). In terms of key contributions, PSAP implements 1) an Interconnected Data Model (IDM) to manage multi-source data independently and integrally, 2) an efficient and effective data mining mechanism based on multi-dimension and multi-measure queries (MMQs), and 3) concurrent data processing cascades with Sentiments in Places Analysis Mechanism (SPAM) and Activities in Places Analysis Mechanism (APAM), to fuse social network data with other data on public sentiment and activity comprehensively. As proved by a holistic evaluation, both SPAM and APAM outperform compared methods. Specifically, SPAM improves its classification accuracy gradually and significantly from 72.37% to about 85% within 9 crowd-calibration cycles, and APAM with an ensemble classifier achieves the highest precision of 92.13%, which is approximately 13% higher than the second best method. Finally, by applying MMQs on "Sentiment&Activity Linked Data", various place design insights of our testbed are mined to improve its livability.

A phrase is a natural, meaningful, and essential semantic unit. In topic modeling, visualizing phrases for individual topics is an effective way to explore and understand unstructured text corpora. However, from phrase quality and topical cohesion perspectives, the outcomes of existing approaches remain to be improved. Usually, the process of topical phrase mining is twofold: phrase mining and topic modeling. For phrase mining, existing approaches often suffer from order sensitive and inappropriate segmentation problems, which make them often extract inferior quality phrases. For topic modeling, traditional topic models do not fully consider the constraints induced by phrases, which may weaken the cohesion. Moreover, existing approaches often suffer from losing domain terminologies since they neglect the impact of domain-level topical distribution. In this paper, we propose an efficient method for high quality and cohesive topical phrase mining. In our framework, we integrate quality guaranteed phrase mining method, a novel topic model incorporating the constraint of phrases, and a novel document clustering method into an iterative framework to improve both phrase quality and topical cohesion. We also describe efficient algorithmic designs to execute these methods efficiently. The empirical verification demonstrates that our method outperforms the state-of-the-art methods from the aspects of both interpretability and efficiency.
Traditional cluster ensemble approaches have several limitations. (1) Few make use of prior knowledge provided by experts. (2) It is difficult to achieve good performance in high-dimensional datasets. (3) All of the weight values of the ensemble members are equal, which ignores different contributions from different ensemble members. (4) Not all pairwise constraints contribute to the final result. In the face of this situation, we propose double weighting semi-supervised ensemble clustering based on selected constraint projection (DCECP) to address these limitations. Specifically, DCECP first adopts the random subspace technique in combination with the constraint projection procedure to handle high-dimensional datasets. Second, it treats prior knowledge of experts as pairwise constraints, and assigns different subsets of pairwise constraints to different ensemble members. An adaptive ensemble member weighting process is designed to associate different weight values with different ensemble members. Third, the weighted normalized cut algorithm is adopted to summarize clustering solutions and generate the final result. Finally, nonparametric statistical tests are used to compare multiple algorithms on real-world datasets. Our experiments on 15 high-dimensional datasets show that DCECP performs better than most clustering algorithms.

The k Nearest Neighbor (k-NN) query has been gaining more importance in extensive applications involving information retrieval, data mining and databases. Specifically, in order to trade off accuracy for efficiency, approximate solutions for the k-NN query are extensively explored. However, the precision is usually order-insensitive, which is defined on the result set instead of the result sequence. In many situations, it cannot reasonably reflect the query result quality. In this paper, we focus on the approximate k-NN query problem with the order-sensitive precision requirement and propose a novel scheme based on the projection-filter-refinement framework. Basically, we adopt PCA to project the high-dimensional data objects into the low-dimensional space. Then, a filter condition is inferred to execute efficient pruning over the projected data. In addition, an index strategy named OR-tree is proposed to reduce the I/O cost. The extensive experiments based on several real-world data sets and a synthetic data set are conducted to verify the effectiveness and efficiency of the proposed solution. Compared to the state-of-the-art methods, our method can support order-sensitive k-NN queries with higher result precision while retaining satisfactory CPU and I/O efficiency.
Mining Summaries for Knowledge Graph Search

Querying heterogeneous and large-scale knowledge graphs is expensive. This paper studies a graph summarization framework to facilitate knowledge graph search. (1) We introduce a class of reduced summaries. Characterized by approximate graph pattern matching, these summaries are capable of summarizing entities in terms of their neighborhood similarity up to a certain hop, using small and informative graph patterns. (2) We study a diversified graph summarization problem. Given a knowledge graph, it is to discover top-$k$ summaries that maximize a bi-criteria function, characterized by both informativeness and diversity. We show that diversified summarization is feasible for large graphs, by developing both sequential and parallel summarization algorithms. (a) We show that there exists a 2-approximation algorithm to discover diversified summaries. We further develop an anytime sequential algorithm which discovers summaries under resource constraints. (b) We present a new parallel algorithm with quality guarantees. The algorithm is parallel scalable, which ensures its feasibility in distributed graphs. (3) We also develop a summary-based query evaluation scheme, which only refers to a small number of summaries. Using real-world knowledge graphs, we experimentally verify the effectiveness and efficiency of our summarization algorithms, and query processing using summaries.

CRAFTER: a Tree-ensemble Clustering Algorithm for Static Datasets with Mixed Attributes and High Dimensionality

Clustering is an important aspect of data mining, while clustering high-dimensional mixed-attribute data in a scalable fashion still remains a challenging problem. In this paper, we propose a tree-ensemble clustering algorithm for static datasets, CRAFTER, to tackle this problem. CRAFTER is able to handle categorical and numeric attributes simultaneously, and scales well with the dimensionality and the size of datasets. CRAFTER leverages the advantages of a tree-ensemble to handle mixed attributes and high dimensionality. The concept of the class probability estimates is utilized to identify the representative data points for clustering. Through a series of experiments on both synthetic and real datasets, we have demonstrated that CRAFTER is superior than Random Forest Clustering (RFC), an existing tree-based clustering method, in terms of both the clustering quality and the computational cost.
Given a large transactional database, correlation computing/association analysis aims at efficiently finding strongly correlated items. For traditional association analysis, relationships among variables are usually measured at a global level. In this study, we investigate confounding factors that can help to capture abnormal correlation behaviors at a local level. Indeed, many real-world phenomena are localized to specific markets or subpopulations. Such local relationships may not be visible or may be miscalculated when collectively analyzing the entire data. In particular, confounding effects that change the direction of correlation are a most severe problem because the global correlations alone leads to errant conclusions. To this end, we propose CONFOUND, an efficient algorithm to identify paradoxical correlation patterns (i.e., where controlling for a third item changes the direction of association for strongly correlated pairs) using effective pruning strategies. Moreover, we also provide an enhanced version of this algorithm, called CONFOUND+, which substantially speeds up the confounder search step. Finally, experimental results showed that our proposed CONFOUND and CONFOUND+ algorithms can effectively identify confounders and the computational performance is orders of magnitude faster than benchmark methods.

At the core of graph mining lies independent expansion where a substructure (also referred to as a subgraph) independently grows into a number of larger substructures in each iteration. Such an independent expansion, invariably, leads to the generation of duplicates. In the presence of graph partitions, duplicates are generated both within and across partitions. Eliminating these duplicates (for correctness) not only incurs generation and storage cost but also additional computation for its elimination. Our primary aim is to design techniques to reduce generating duplicate substructures as we show that they cannot be eliminated. This paper introduces three constraint-based optimization techniques, each significantly improving the overall mining cost by reducing the number of duplicates generated. These alternatives provide flexibility to choose the right technique based on graph properties. We establish theoretical correctness of each technique as well as its analysis with respect to graph characteristics such as degree, number of unique labels, and label distribution. We also investigate the applicability of their combination for improvements in duplicate reduction. Finally, we discuss the effects of the constraints with respect to the partitioning schemes used in graph mining. Our experiments demonstrate benefits of these constraints in terms of storage, computation, and communication cost (specific to partitioned approaches) across graphs with varied characteristics.
### EPRO DM - 013
**A Topic Modeling Approach for Traditional Chinese Medicine Prescriptions**

In traditional Chinese medicine (TCM), prescriptions are the daughters of doctors' clinical experiences, which have been the main way to cure diseases in China for several thousand years. In the long Chinese history, a large number of prescriptions have been invented based on TCM theories. Regularities in the prescriptions are important for both clinical practice and novel prescription development. Previous works used many methods to discover regularities in prescriptions, but rarely described how a prescription is generated using TCM theories. In this work, we propose a topic model which characterizes the generative process of prescriptions in TCM theories and further incorporate domain knowledge into the topic model. Using 33,765 prescriptions in TCM prescription books, the model can reflect the prescribing patterns in TCM. Our method can outperform several previous topic models and group recommendation methods on generalization performance, herbs recommendation, symptoms suggestion, and prescribing patterns discovery.

---

### EPRO DM - 014
**Mining Precise-Positioning Episode Rules from Event Sequences**

Episode Rule Mining is a popular framework for discovering sequential rules from event sequential data. However, traditional episode rule mining methods only tell that the consequent event is likely to happen within a given time interval after the occurrence of the antecedent events. As a result, they cannot satisfy the requirement of many time sensitive applications, such as program security trading and intelligent transportation management due to the lack of fine-grained response time. In this study, we come up with the concept of fixed-gap episode to address this problem. A fixed-gap episode consists of an ordered set of events where the elapsed time between any two consecutive events is a constant. Based on this concept, we formulate the problem of mining precise-positioning episode rules in which the occurrence time of each event in the consequent is clearly specified. In addition, we develop a trie-based data structure to mine such precise-positioning episode rules with several pruning strategies incorporated for improving the performance as well as reducing memory consumption. Experimental results on real datasets show the superiority of our proposed algorithms.
In this paper, we propose a prototype-based classification model for evolving data streams, called SyncStream, which allows dynamically modeling time-changing concepts, making predictions in a local fashion. Instead of learning a single model on a fixed or adaptive sliding window of historical data or ensemble learning a set of weighted base classifiers, SyncStream captures evolving concepts by dynamically maintaining a set of prototypes in a proposed P-Tree, which are obtained based on the error-driven representativeness learning and synchronization-inspired constrained clustering. To identify abrupt concept drifts in data streams, PCA and statistical analysis based heuristic approaches have been introduced. To further learn the associations among distributed data streams, the extended P-Tree structure and KNN-style strategy are introduced. We demonstrate that our new data stream classification approach has several attractive benefits: (a) SyncStream is capable of dynamically modeling the evolving concepts from even a small set of prototypes. (b) Owing to synchronization-based constrained clustering and P-Tree, SyncStream supports efficient and effective data representation and maintenance. (c) SyncStream is also tolerant of inappropriate or noisy examples via error-driven representativeness learning. (d) SyncStream allows learning relationship among distributed data streams at the instance level.

Spatial co-location pattern mining is an interesting and important task in spatial data mining which discovers the subsets of spatial features frequently observed together in nearby geographic space. However, the traditional framework of mining prevalent colocation patterns produces numerous redundant co-location patterns, which makes it hard for users to understand or apply. To address this issue, in this paper, we study the problem of reducing redundancy in a collection of prevalent co-location patterns by utilizing the spatial distribution information of co-location instances. We first introduce the concept of semantic distance between a co-location pattern and its super-patterns, and then define redundant co-locations by introducing the concept of δ-covered, where \( 0 \leq \delta \leq 1 \) is a coverage measure. We develop two algorithms RRclosed and RRnull to perform the redundancy reduction for prevalent co-location patterns. The former adopts the post-mining framework that is commonly used by existing redundancy reduction techniques, while the latter employs the mine-and-reduce framework that pushes redundancy reduction into the co-location mining process. Our performance studies on the synthetic and real-world data sets demonstrate that our method effectively reduces the size of the original collection of closed co-location patterns by about 50 percent. Furthermore, the RRnull method runs much faster than the related closed co-location pattern mining algorithm.
String similarity search is a fundamental query that has been widely used for DNA sequencing, error-tolerant query autocompletion, and data cleaning needed in database, data warehouse, and data mining. In this paper, we study string similarity search based on edit distance that is supported by many database management systems such as Oracle and PostgreSQL. Given the edit distance, ed(s, t), between two strings, s and t, the string similarity search is to find every string t in a string database D which is similar to a query string s such that ed(s, t) ≤ τ for a given threshold τ. In the literature, most existing work takes a filter-and-verify approach, where the filter step is introduced to reduce the high verification cost of two strings by utilizing an index built offline for D. The two up-to-date approaches are prefix filtering and local filtering. In this paper, we study string similarity search where strings can be either short or long. Our approach can support long strings, which are not well supported by the existing approaches due to the size of the index built and the time to build such index. We propose two new hash-based labeling techniques, named OX label and XX label, for string similarity search. We assign a hash-label, Hs, to a string s, and prune the dissimilar strings by comparing two hash-labels, Hs and Ht, for two strings s and t in the filter step. The key idea is to take the dissimilar bit-patterns between two hash-labels. We discuss our hash-based approaches, address their pruning power, and give the algorithms.

We formulate a document summarization method to extract passage-level answers for non-factoid queries, referred to as answer-biased summaries. We propose to use external information from related Community Question Answering (CQA) content to better identify answer bearing sentences. Three optimization-based methods are proposed: (i) query-biased, (ii) CQA-answer-biased, and (iii) expanded-query-biased, where expansion terms were derived from related CQA content. A learning-to-rank-based method is also proposed that incorporates a feature extracted from related CQA content. Our results show that even if a CQA answer does not contain a perfect answer to a query, their content can be exploited to improve the extraction of answer-biased summaries from other corpora. The quality of CQA content is found to impact on the accuracy of optimization-based summaries, though medium quality answers enable the system to achieve a comparable (and in some cases superior) accuracy to state-of-the-art techniques. The learning-to-rank-based summaries, on the other hand, are not significantly influenced by CQA quality. We provide a recommendation of the best use of our proposed approaches in regard to the availability of different quality levels of related CQA content. As a further investigation, the reliability of our approaches was tested on another publicly available dataset.
Modern networks are very large in size and also evolve with time. As their sizes grow, the complexity of performing network analysis grows as well. Getting a smaller representation of a temporal network with similar properties will help in various data mining tasks. In this paper, we study the novel problem of getting a smaller diffusion-equivalent representation of a set of time-evolving networks. We first formulate a well-founded and general temporal-network condensation problem based on the so-called systemmatrix of the network. We then propose NETCONDENSE, a scalable and effective algorithm which solves this problem using careful transformations in sub-quadratic running time, and linear space complexities. Our extensive experiments show that we can reduce the size of large real temporal networks (from multiple domains such as social, co-authorship, and email) significantly without much loss of information. We also show the wide-applicability of NETCONDENSE by leveraging it for several tasks: for example, we use it to understand, explore, and visualize the original datasets and to also speed-up algorithms for the influence-maximization and event detection problems on temporal networks.

Utilizing large-scale GPS data to improve taxi services has become a popular research problem in the areas of data mining, intelligent transportation, geographical information systems, and the Internet of Things. In this paper, we utilize a large-scale GPS data set generated by over 7,000 taxis in a period of one month in Nanjing, China, and propose TaxiRec: a framework for evaluating and discovering the passenger-finding potentials of road clusters, which is incorporated into a recommender system for taxi drivers to seek passengers. In TaxiRec, the underlying road network is first segmented into a number of road clusters, a set of features for each road cluster is extracted from real-life data sets, and then a ranking-based extreme learning machine (ELM) model is proposed to evaluate the passenger-finding potential of each road cluster. In addition, TaxiRec can use this model with a training cluster selection algorithm to provide road cluster recommendations when taxi trajectory data is incomplete or unavailable. Experimental results demonstrate the feasibility and effectiveness of TaxiRec.
<table>
<thead>
<tr>
<th>EPRO DM - 021</th>
<th>Applying Simulated Annealing and Parallel Computing to the Mobile Sequential Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>We speed up the solution of the mobile sequential recommendation (MSR) problem that requires searching optimal routes for empty taxi cabs through mining massive taxi GPS data. We develop new methods that combine parallel computing and the simulated annealing with novel global and local searches. While existing approaches usually involve costly offline algorithms and methodical pruning of the search space, our new methods provide direct real-time search for the optimal route without the offline preprocessing. Our methods significantly reduce computational time for the high dimensional MSR problems from days to seconds based on the real-world data as well as the synthetic ones. We efficiently provide solutions to MSR problems with thousands of pick-up points without offline training, compared to the published record of 25 pick-up points.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EPRO DM - 022</th>
<th>Privacy-Preserving Collaborative Model Learning: The Case of Word Vector Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nowadays machine learning is becoming a new paradigm for mining hidden knowledge in big data. The collection and manipulation of big data not only create considerable values, but also raise serious privacy concerns. To protect the huge amount of potentially sensitive data, a straightforward approach is to encrypt data with specialized cryptographic tools. However, it is challenging to utilize or operate on encrypted data, especially to perform machine learning algorithms. In this paper, we investigate the problem of training high quality word vectors over large-scale encrypted data (from distributed data owners) with the privacy-preserving collaborative neural network learning algorithms. We leverage and also design a suite of arithmetic primitives (e.g., multiplication, fixed-point representation and sigmoid function computation etc.) on encrypted data, served as components of our construction. We theoretically analyze the security and efficiency of our proposed construction, and conduct extensive experiments on representative real-world datasets to verify its practicality and effectiveness.</td>
<td></td>
</tr>
</tbody>
</table>

Elysium PRO Titles with Abstracts 2018-19
Automated Phrase Mining from Massive Text Corpora

As one of the fundamental tasks in text analysis, phrase mining aims at extracting quality phrases from a text corpus and has various downstream applications including information extraction/retrieval, taxonomy construction, and topic modeling. Most existing methods rely on complex, trained linguistic analyzers, and thus likely have unsatisfactory performance on text corpora of new domains and genres without extra but expensive adaption. None of the state-of-the-art models, even data-driven models, is fully automated because they require human experts for designing rules or labeling phrases. In this paper, we propose a novel framework for automated phrase mining, AutoPhrase, which supports any language as long as a general knowledge base (e.g., Wikipedia) in that language is available, while benefiting from, but not requiring, a POS tagger. Compared to the state-of-the-art methods, AutoPhrase has shown significant improvements in both effectiveness and efficiency on five real-world datasets across different domains and languages. Besides, AutoPhrase can be extend to model single-word quality phrases.

A Two-Phase Algorithm for Differentially Private Frequent Subgraph Mining

Mining frequent subgraphs from a collection of input graphs is an important task for exploratory data analysis on graph data. However, if the input graphs contain sensitive information, releasing discovered frequent subgraphs may pose considerable threats to individual privacy. In this paper, we study the problem of frequent subgraph mining (FSM) under the rigorous differential privacy model. We present a two-phase differentially private FSM algorithm, which is referred to as DFG. In DFG, frequent subgraphs are privately identified in the first phase, and the noisy support of each identified frequent subgraph is calculated in the second phase. In particular, to privately identity frequent subgraphs, we propose a frequent subgraph identification approach, which can improve the accuracy of discovered frequent subgraphs through candidate pruning. Moreover, to compute the noisy support of each identified frequent subgraph, we devise a lattice-based noisy support computation approach, which leverages the inclusion relations between the discovered frequent subgraphs to improve the accuracy of the noisy supports. Through formal privacy analysis, we prove that DFG satisfies $\varepsilon$-differential privacy. Extensive experimental results on real datasets show that DFG can privately find frequent subgraphs while achieving high data utility.
Graph Pattern Matching (GPM) has been used in lots of areas, like biology, medical science, and physics. With the advent of Online Social Networks (OSNs), recently, GPM has been playing a significant role in social network analysis, which has been widely used in, for example, finding experts, social community mining, and social position detection. Given a query which contains a pattern graph $G_Q$ and a data graph $G_D$, a GPM algorithm finds those subgraphs, $G_M$, that match $G_Q$ in $G_D$. However, the existing GPM methods do not consider the multiple end-to-end constraints of the social contexts, like social relationships, social trust, and social positions on edges in $G_Q$, which are commonly found in various applications, such as crowdsourcing travel, social network based ecommerce, and study group selection, etc. In this paper, we first conceptually extend Bounded Simulation to Multi-Constrained Simulation (MCS), and propose a novel NP-Complete Multi-Constrained Graph Pattern Matching (MC-GPM) problem. Then, to address the efficiency issue in large-scale MC-GPM, we propose a new concept called Strong Social Component (SSC), consisting of participants with strong social connections. We also propose an approach to identifying SSCs, and propose a novel index method and a graph compression method for SSC. Moreover, we devise a multithreading heuristic algorithm, called M-HAMC, to bidirectionally search the MC-GPM results in parallel without decompressing graphs.

Reliable propagation of information through large networks, e.g., communication networks, social networks, or sensor networks is very important in many applications concerning marketing, social networks, and wireless sensor networks. However, social ties of friendship may be obsolete, and communication links may fail, inducing the notion of uncertainty in such networks. In this paper, we address the problem of optimizing information propagation in uncertain networks given a constrained budget of edges. We show that this problem requires to solve two NP-hard subproblems: the computation of expected information flow, and the optimal choice of edges. To compute the expected information flow to a source vertex, we propose the F-tree as a specialized data structure that identifies independent components of the graph for which the information flow can either be computed analytically and efficiently, or for which traditional Monte-Carlo sampling can be applied independently of the remaining network. For the problem of finding the optimal edges, we propose a series of heuristics that exploit properties of this data structure. Our evaluation shows that these heuristics lead to high quality solutions, thus yielding high information flow, while maintaining low running time.
Stock market volatility is influenced by information release, dissemination, and public acceptance. With the increasing volume and speed of social media, the effects of Web information on stock markets are becoming increasingly salient. However, studies of the effects of Web media on stock markets lack both depth and breadth due to the challenges in automatically acquiring and analyzing massive amounts of relevant information. In this study, we systematically reviewed 229 research articles on quantifying the interplay between Web media and stock markets from the fields of Finance, Management Information Systems, and Computer Science. In particular, we first categorized the representative works in terms of media type and then summarized the core techniques for converting textual information into machine-friendly forms. Finally, we compared the analysis models used to capture the hidden relationships between Web media and stock movements. Our goal is to clarify current cutting-edge research and its possible future directions to fully understand the mechanisms of Web information percolation and its impact on stock markets from the perspectives of investors cognitive behaviors, corporate governance, and stock market regulation.

<table>
<thead>
<tr>
<th>EPRO DM - 027</th>
<th>Web Media and Stock Markets: A Survey and Future Directions from a Big Data Perspective</th>
</tr>
</thead>
</table>

Item-to-item collaborative filtering (aka. item-based CF) has been long used for building recommender systems in industrial settings, owing to its interpretability and efficiency in real-time personalization. In recent years, several works attempt to learn item similarities from data, by expressing the similarity as an underlying model and estimating model parameters by optimizing a recommendation-aware objective function. While extensive efforts have been made to use shallow linear models for learning item similarities, there has been relatively less work exploring nonlinear neural network models for item-based CF. In this work, we propose a neural network model named Neural Attentive Item Similarity model (NAIS) for item-based CF. The key to our design of NAIS is an attention network, which is capable of distinguishing which historical items in a user profile are more important for a prediction. Compared to the state-of-the-art item-based CF method Factored Item Similarity Model (FISM), our NAIS has stronger representation power with only a few additional parameters brought by the attention network. Extensive experiments on two public benchmarks demonstrate the effectiveness of NAIS. This work is the first attempt that designs neural network models for item-based CF, opening up new research possibilities for future developments of neural recommender systems.

<table>
<thead>
<tr>
<th>EPRO DM - 028</th>
<th>NAIS: Neural Attentive Item Similarity Model for Recommendation</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>EPRO DM - 029</th>
<th>Nonintrusive Smartphone User Verification Using Anonymized Multimodal Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smartphone user verification is important as personal daily activities are increasingly conducted on the phone and sensitive information is constantly logged. The commonly adopted user verification methods are typically active, i.e., they require a user's cooperative input of a security token to gain access permission. Though popular, these methods impose heavy burden to smartphone users to memorize, maintain and input the token at a high frequency. To alleviate this imposition onto the users and to provide additional security, we propose a new nonintrusive and continuous mobile user verification framework that can reduce the frequency required for a user to input his/her security token. Using tailored Hidden Markov Models and sequential likelihood ratio test, our verification is built on low-cost, readily available, anonymized, and multimodal smartphone data without additional effort of data collection and risk of privacy leakage. With extensive evaluation, we achieve a high rate of about 94% for detecting illegitimate smartphone uses and a rate of 74% for confirming legitimate uses. In a practical setting, this can translate into 74% of frequency reduction of inputting a security token using an active authentication method with only about 6% risk of miss detection of a random intruder, which is highly desirable.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EPRO DM - 030</th>
<th>Heuristic and Cost-based Optimization for Diverse Provenance Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>A well-established technique for capturing database provenance as annotations on data is to instrument queries to propagate such annotations. However, even sophisticated query optimizers often fail to produce efficient execution plans for instrumented queries. We develop provenance-aware optimization techniques to address this problem. Specifically, we study algebraic equivalences targeted at instrumented queries and alternative ways of instrumenting queries for provenance capture. Furthermore, we present an extensible heuristic and cost-based optimization framework utilizing these optimizations. Our experiments confirm that these optimizations are highly effective, improving performance by several orders of magnitude for diverse provenance tasks.</td>
<td></td>
</tr>
</tbody>
</table>
Differential Privacy (DP) has received increasing attention as a rigorous privacy framework. Many existing studies employ traditional DP mechanisms (e.g., the Laplace mechanism) as primitives to continuously release private data for protecting privacy at each time point (i.e., event-level privacy), which assume that the data at different time points are independent, or that adversaries do not have knowledge of correlation between data. However, continuously generated data tend to be temporally correlated, and such correlations can be acquired by adversaries. In this paper, we investigate the potential privacy loss of a traditional DP mechanism under temporal correlations. First, we analyze the privacy leakage of a DP mechanism under temporal correlation that can be modeled using Markov Chain. Our analysis reveals that, the event-level privacy loss of a DP mechanism may increase over time. We call the unexpected privacy loss temporal privacy leakage (TPL). Although TPL may increase over time, we find that its supremum may exist in some cases. Second, we design efficient algorithms for calculating TPL. Third, we propose data releasing mechanisms that convert any existing DP mechanism into one against TPL. Experiments confirm that our approach is efficient and effective.

The structured case representation improves case-based reasoning (CBR) by exploring structures in the case base and the relevance of case structures. Recent CBR classifiers have mostly been built upon the attribute-value case representation rather than structured case representation, in which the structural relations embodied in their representation structure are accordingly overlooked in improving the similarity measure. This results in retrieval inefficiency and limitations on the performance of CBR classifiers. This paper proposes a hierarchical case-based classifier, HCBC, which introduces a concept lattice to hierarchically organize cases. By exploiting structural case relations in the concept lattice, a novel dynamic weighting model is proposed to enhance the concept similarity measure. Based on this similarity measure, HCBC retrieves the top-K concepts that are most similar to a new case by using a bottom-up pruning-based recursive retrieval (PRR) algorithm. The concepts extracted in this way are applied to suggest a class label for the case by a weighted majority voting. Experimental results show that HCBC outperforms other classifiers in terms of classification performance and robustness on categorical data, and also works confidently well on numeric datasets. In addition, PRR effectively reduces the search space and greatly improves the retrieval efficiency of HCBC.
As a significant business paradigm, many online information platforms have emerged to satisfy society's needs for person-specific data, where a service provider collects raw data from data contributors, and then offers value-added data services to data consumers. However, in the data trading layer, the data consumers face a pressing problem, i.e., how to verify whether the service provider has truthfully collected and processed data? Furthermore, the data contributors are usually unwilling to reveal their sensitive personal data and real identities to the data consumers. In this paper, we propose TPDM, which efficiently integrates Truthfulness and Privacy preservation in Data Markets. TPDM is structured internally in an Encrypt-then-Sign fashion, using somewhat homomorphic encryption and identity-based signature. It simultaneously facilitates batch verification, data processing, and outcome verification, while maintaining identity preservation and data confidentiality. We also instantiate TPDM with a profile-matching service, and extensively evaluate its performance on Yahoo! Music ratings dataset. Our analysis and evaluation results reveal that TPDM achieves several desirable properties, while incurring low computation and communication overheads when supporting a large-scale data market.

Research of social recommendation aims at exploiting social information to improve the quality of a recommender system. It can be further divided into two classes. Explicit social recommendation assumes the existence of not only the users' ratings on items, but also the explicit social connections between users. Implicit social recommendation assumes the availability of only the ratings but not the social connections between users, and attempts to infer implicit social connections between users with the goal to boost recommendation accuracy. This paper proposes a unified framework that is applicable to both explicit and implicit social recommendation. We propose an optimization framework to learn the degree of social correlation and rating prediction jointly, so these two tasks can mutually boost the performance of each other. Furthermore, a well-known challenge for implicit social recommendation is that it takes quadratic time to learn the strength of pairwise connections. This paper further proposes several practical tricks to reduce the complexity of our model to be linear to the observed ratings. The experiments show that the proposed model, with only two parameters, can significantly outperform the state-of-the-art solutions for both explicit and implicit social recommender systems.
Given a set of multidimensional data points, skyline queries retrieve those points that are not dominated by any other points in the set. Due to the ubiquitous use of skyline queries, there are several outstanding challenges that have not been well addressed. More specifically, in this paper, we are tackling the data straggler and data skew challenges introduced by distributed skyline query processing as well as the ensuing high computing cost of merging skyline candidates. We thus introduce a new efficient three-phase approach for large scale processing of skyline queries. In the first preprocessing phase, the data is partitioned along the Z-order curve. We utilize a novel data partitioning approach that formulates data partitioning as an optimization problem to minimize the size of intermediate data. In the second phase, each computation node partitions the input data points into separate sets, and then performs the skyline computation on each set to produce skyline candidates in parallel. In the final phase, we build an index and employ an efficient algorithm to merge the generated skyline candidates. Extensive experiments demonstrate that the proposed skyline algorithm achieves more than one order of magnitude enhancement in performance compared to existing state-of-the-art approaches.

Shortest path query is one of the most fundamental and classic problems in graph analytics, which returns the complete shortest path between any two vertices. However, in many real-life scenarios, only critical vertices on the shortest path are desirable and it is unnecessary to search for the complete path. This paper investigates the shortest path sketch by defining a top-k critical vertices (kCV) query on the shortest path. Given a source vertex s and target vertex t in a graph, kCV query can return the top-k significant vertices on the shortest path SP(s,t). The significance of the vertices can be predefined. The key strategy for seeking the sketch is to apply off-line preprocessed distance oracle to accelerate on-line real-time queries. This allows to omit unnecessary vertices and obtain the most representative sketch of the shortest path directly. We further explore a series of methods and optimizations to answer kCV query on both centralized and distributed platforms, using exact and approximate approaches, respectively. We evaluate our methods in terms of time, space complexity and approximation quality. Experiments on large-scale real networks validate that our algorithms are of high efficiency and accuracy.
Locations, e.g. countries, states, cities, and point-of-interests, are central to news, emergency events, and people’s daily lives. Automatic identification of locations associated with or mentioned in documents has been explored for decades. As one of the most popular online social network platforms, Twitter has attracted a large number of users who send millions of tweets on daily basis. Due to the world-wide coverage of its users and real-time freshness of tweets, location prediction on Twitter has gained significant attention in recent years. Research efforts are spent on dealing with new challenges and opportunities brought by the noisy, short, and context-rich nature of tweets. In this survey, we aim at offering an overall picture of location prediction on Twitter. Specifically, we concentrate on the prediction of user home locations, tweet locations, and mentioned locations. We first define the three tasks and review the evaluation metrics. By summarizing Twitter network, tweet content, and tweet context as potential inputs, we then structurally highlight how the problems depend on these inputs. Each dependency is illustrated by a comprehensive review of the corresponding strategies adopted in state-of-the-art approaches. In addition, we also briefly review two related problems, i.e. semantic location prediction and point-of-interest recommendation. Finally, we make a conclusion of the survey.

Influence Maximization (IM), which selects a set of k users (called seed set) from a social network to maximize the expected number of influenced users (called influence spread), is a key algorithmic problem in social influence analysis. Due to its immense application potential and enormous technical challenges, IM has been extensively studied in the past decade. In this paper, we survey and synthesize a wide spectrum of existing studies on IM from an algorithmic perspective, with a special focus on the following key aspects (1) a review of well-accepted diffusion models that capture information diffusion process and build the foundation of the IM problem, (2) a fine-grained taxonomy to classify existing IM algorithms based on their design objectives, (3) a rigorous theoretical comparison of existing IM algorithms, and (4) a comprehensive study on the applications of IM techniques in combining with novel context features of social networks such as topic, location, and time. Based on this analysis, we then outline the key challenges and research directions to expand the boundary of IM research.
With the proliferation of spatial-textual data such as location-based services and geo-tagged websites, spatial keyword queries are ubiquitous in real life. One example of spatial-keyword query is the so-called collective spatial keyword query (CoSKQ) which is to find for a given query consisting a query location and several query keywords a set of objects which covers the query keywords collectively and has the smallest cost wrt the query location. In the literature, many different functions were proposed for defining the cost and correspondingly, many different approaches were developed for the CoSKQ problem. In this paper, we study the CoSKQ problem systematically by proposing a unified cost function and a unified approach for the CoSKQ problem (with the unified cost function). The unified cost function includes all existing cost functions as special cases and the unified approach solves the CoSKQ problem with the unified cost function in a unified way. Experiments were conducted on both real and synthetic datasets which verified our proposed approach.

In the big data era, the generation of data presents some new characteristics, including wide distribution, high velocity, high dimensionality and privacy concern. To address these challenges for big data analytics, we develop a privacy-preserving distributed online learning framework on the data collected from distributed data sources. Specifically, each node (i.e., data source) has the capacity of learning a model from its local dataset, and exchanges intermediate parameters with a random part of their own neighboring (logically connected) nodes. Hence, the topology of the communications in our distributed computing framework is unfixed in practice. As online learning always performs on the sensitive data, we introduce the notion of differential privacy (DP) into our distributed online learning algorithm (DOLA) to protect the data privacy during the learning, which prevents an adversary from inferring any significant sensitive information. Our model is of general value for big data analytics in the distributed setting, because it can provide rigorous and scalable privacy proof and have much less computational complexity when compared to classic schemes, e.g., secure multiparty computation (SMC). To tackle high-dimensional incoming data entries, we study a sparse version of the DOLA. Numerical experiment results validate the feasibility of our private DOLAs.
Information can be disseminated widely and rapidly through Online Social Networks (OSNs) with “word-of-mouth” effects. Viral marketing is such a typical application in which new products or commercial activities are advertised by some seed users in OSNs to other users in a cascading manner. The selection of initial seed users yields a tradeoff between the expense and reward of viral marketing. In this paper, we define a general profit metric that naturally combines the benefit of influence spread with the cost of seed selection in viral marketing. We carry out a comprehensive study on finding a set of seed nodes to maximize the profit of viral marketing. We show that the profit metric is significantly different from the influence metric in that it is no longer monotone. This characteristic differentiates the profit maximization problem from the traditional influence maximization problem. We develop new seed selection algorithms for profit maximization with strong approximation guarantees. We also derive several upper bounds to benchmark the practical performance of an algorithm on any specific problem instance. Experimental evaluations with real OSN datasets demonstrate the effectiveness of our algorithms and techniques.

A new density-based clustering algorithm, RNN-DBSCAN, is presented which uses reverse nearest neighbor counts as an estimate of observation density. Clustering is performed using a DBSCAN-like approach based on k nearest neighbor graph traversals through dense observations. RNN-DBSCAN is preferable to the popular density-based clustering algorithm DBSCAN in two aspects. First, problem complexity is reduced to the use of a single parameter (choice of k nearest neighbors), and second, an improved ability for handling large variations in cluster density (heterogeneous density). The superiority of RNN-DBSCAN is demonstrated on several artificial and real-world datasets with respect to prior work on reverse nearest neighbor based clustering approaches (RECORD, IS-DBSCAN, and ISB-DBSCAN) along with DBSCAN and OPTICS. Each of these clustering approaches is described by a common graph-based interpretation wherein clusters of dense observations are defined as connected components, along with a discussion on their computational complexity. Heuristics for RNN-DBSCAN parameter selection are presented, and the effects of k on RNN-DBSCAN clusterings discussed. Additionally, with respect to scalability, an approximate version of RNN-DBSCAN is presented leveraging an existing approximate k nearest neighbor technique.
GPS enables mobile devices to continuously provide new opportunities to improve our daily lives. For example, the data collected in applications created by Uber or Public Transport Authorities can be used to plan transportation routes, estimate capacities, and proactively identify low coverage areas. In this paper, we study a new kind of query-Reverse k Nearest Neighbor Search over Trajectories (RkNNT), which can be used for route planning and capacity estimation. Given a set of existing routes $D_R$, a set of passenger transitions $D_T$, and a query route $Q$, an RkNNT query returns all transitions that take $Q$ as one of its $k$ nearest travel routes. To solve the problem, we first develop an index to handle dynamic trajectory updates, so that the most up-to-date transition data are available for answering an RkNNT query. Then we introduce a filter refinement framework for processing RkNNT queries using the proposed indexes. Next, we show how to use RkNNT to solve the optimal route planning problem MaxRkNNT (MinRkNNT), which is to search for the optimal route from a start location to an end location that could attract the maximum (or minimum) number of passengers based on a predefined travel distance threshold. Experiments on real datasets demonstrate the efficiency and scalability of our approaches. To the best of our knowledge, this is the first work to study the RkNNT problem for route planning.

The classical K Shortest Paths (KSP) problem, which identifies the $k$ shortest paths in a directed graph, plays an important role in many application domains, such as providing alternative paths for vehicle routing services. However, the returned $k$ shortest paths may be highly similar, i.e., sharing significant amounts of edges, thus adversely affecting service qualities. In this paper, we formalize the K Shortest Paths with Diversity (KSPD) problem that identifies top-$k$ shortest paths such that the paths are dissimilar with each other and the total length of the paths is minimized. We first prove that the KSPD problem is NP-hard and then propose a generic greedy framework to solve the KSPD problem in the sense that (1) it supports a wide variety of path similarity metrics which are widely adopted in the literature and (2) it is also able to efficiently solve the traditional KSP problem if no path similarity metric is specified. The core of the framework includes the use of two judiciously designed lower bounds, where one is dependent on and the other one is independent on the chosen path similarity metric, which effectively reduces the search space and significantly improves efficiency. Empirical studies on five real-world and synthetic graphs and five different path similarity metrics offer insight into the design properties of the proposed general framework and offer evidence that the proposed lower bounds are effective.
### ComClus: A Self-Grouping Framework for Multi-Network Clustering

Joint clustering of multiple networks has been shown to be more accurate than performing clustering on individual networks separately. This is because multi-network clustering algorithms typically assume there is a common clustering structure shared by all networks, and different networks can provide compatible and complementary information for uncovering this underlying clustering structure. However, this assumption is too strict to hold in many emerging applications, where multiple networks usually have diverse data distributions. More popularly, the networks in consideration belong to different underlying groups. Only networks in the same underlying group share similar clustering structures. Better clustering performance can be achieved by considering such groups differently. As a result, an ideal method should be able to automatically detect network groups so that networks in the same group share a common clustering structure. To address this problem, we propose a new method, COMCLUS, to simultaneously group and cluster multiple networks. COMCLUS is novel in combining the clustering approach of non-negative matrix factorization (NMF) and the feature subspace learning approach of metric learning. Specifically, it treats node clusters as features of networks and learns proper subspaces from such features to differentiate different network groups. During the learning process, the two procedures of network grouping and clustering are coupled and mutually enhanced.

### Efficient Recommendation of Aggregate Data Visualizations

Efficient recommendation of aggregate data visualizations is crucial in many applications. The goal is to provide users with intuitive and informative visualizations that highlight important patterns and insights. This is achieved by automatically selecting and aggregating relevant data from multiple sources, and visualizing the results in a way that is accessible and meaningful. The process involves several steps, including data selection, feature extraction, and visualization design. Each step requires careful consideration to ensure that the resulting visualizations are effective and user-friendly. The proposed method uses a combination of machine learning techniques and user feedback to optimize the visualization process, making it more efficient and accurate.
Given a set of facilities and a set of users, a reverse nearest neighbors (RNN) query retrieves every user \( u \) for which the query facility \( q \) is its closest facility. Since \( q \) is the closest facility to \( u \), the user \( u \) is said to be influenced by \( q \). In this paper, we propose a relaxed definition of influence where a user \( u \) is said to be influenced by not only its closest facility but also every other facility that is almost as close to \( u \) as its closest facility is. Based on this definition of influence, we propose reverse approximate nearest neighbors (RANN) queries. Formally, given a value \( x > 1 \), an RANN query \( q \) returns every user \( u \) for which \( \text{dist}(u,q) \leq x \times \text{NNDist}(u) \) where \( \text{NNDist}(u) \) denotes the distance between a user \( u \) and its nearest facility, i.e., \( q \) is an approximate nearest neighbor of \( u \). In this paper, we study both snapshot and continuous versions of RANN queries. In a snapshot RANN query, the underlying data sets do not change and the results of a query are to be computed only once. In the continuous version, the users continuously change their locations and the results of RANN queries are to be continuously monitored. Based on effective pruning techniques and several non-trivial observations, we propose efficient RANN query processing algorithms for both the snapshot and continuous RANN queries. We conduct extensive experiments on both real and synthetic data sets and demonstrate that our algorithm for both snapshot and continuous queries are significantly better than the competitors.

Data mining technique has attracted attention in the information industry and society as a whole, because of the big amount of data and the imminent need to transform that data into useful information and knowledge. Recently conducted studies with successfully demarcated results using this technique, to estimate several parameters in a variety of domains. However, the effective use of data in some areas is still developing, as is the case of sports, which has shown moderate growth. In this context, the objective of this article is to present a systematic review of the literature about research involving sports data mining. As systematic searches were made out in five databases, resulting in 21 articles that answered a question that grounded this article.
Automatic text summarization is a fundamental NLP application that aims to condense a source text into a shorter version. The rapid increase in multimedia data transmission over the Internet necessitates multi-modal summarization (MMS) from asynchronous collections of text, image, audio and video. In this work, we propose an extractive MMS method that unites the techniques of NLP, speech processing and computer vision to explore the rich information contained in multi-modal data and to improve the quality of MMS. The key idea is to bridge the semantic gaps between multi-modal content. For audio information, we design an approach to selectively use its transcription and to infer the salience of the transcription with audio signals. For visual information, we learn the joint representations of text and images using a neural network. Then, we capture the coverage of the generated summary for important visual information through text-image matching or multi-modal topic modeling. Finally, all the multi-modal aspects are considered to generate a textual summary by maximizing the salience, non-redundancy, readability and coverage through the budgeted optimization of submodular functions. We further introduce a publicly available MMS corpus in English and Chinese. The experimental results demonstrate that our methods outperform other competitive baseline methods.

Networked data are common in domains where instances are characterized by both feature values and inter-dependency relationships. Finding cluster structures for networked instances and discovering representative features for each cluster represent a special co-clustering task usefully for many real-world applications, such as automatic categorization of scientific publications and finding representative key-words for each cluster. To date, although co-clustering has been commonly used for finding clusters for both instances and features, all existing methods are focused on instance-feature values, without leveraging valuable topology relationships between instances to help boost co-clustering performance. In this paper, we propose CFOND, a consensus factorization based framework for co-clustering networked data. We argue that feature values and linkages provide useful information from different perspectives, yet they are not always consistent and therefore need to be carefully aligned for best clustering results. In the paper, we advocate a consensus factorization principle, which simultaneously factorizes information from three aspects: network topology structures, instance-feature content relationships, and feature-feature correlations.
Thank you!