Data Mining for java/ dot net

## 1. A Supermodularity-Based Differential Privacy Preserving Algorithm for Data Anonymization.

Maximizing data usage and minimizing privacy risk are two conflicting goals. Organizations always apply a set of transformations on their data before releasing it. While determining the best set of transformations has been the focus of extensive work in the database community, most of this work suffered from one or both of the following major problems: scalability and privacy guarantee. Differential Privacy provides a theoretical formulation for privacy that ensures that the system essentially behaves the same way regardless of whether any individual is included in the database. In this paper, we address both scalability and privacy risk of data anonymization. We propose a scalable algorithm that meets differential privacy when applying a specific random sampling. The contribution of the paper is two-fold: 1) we propose a personalized anonymization technique based on an aggregate formulation and prove that it can be implemented in polynomial time; and 2) we show that combining the proposed aggregate formulation with specific sampling gives an anonymization algorithm that satisfies differential privacy. Our results rely heavily on exploring the supermodularity properties of the risk function, which allow us to employ techniques from convex optimization. Through experimental studies we compare our proposed algorithm with other anonymization schemes in terms of both time and privacy risk.

## 2. A Data-Mining Model for Protection of FACTS-Based Transmission Line.

### Synopsis:

This paper presents a data-mining model for fault-zone identification of a flexible ac transmission systems (FACTS)-based transmission line including a thyristor-controlled series compensator (TCSC) and unified power-flow controller (UPFC), using ensemble decision trees. Given the randomness in the ensemble of decision trees stacked inside the random forests model, it provides effective decision on fault-zone identification. Half-cycle postfault current and voltage samples from the fault inception are used as an input vector against target output “1” for the fault after TCSC/UPFC and “ - 1” for the fault before TCSC/UPFC for fault-zone identification. The algorithm is tested on simulated fault data with wide variations in operating parameters of the power system network, including noisy environment providing a reliability measure of 99% with faster response time (3/4th cycle from fault inception). The results of the presented approach using the RF model indicate reliable identification of the fault zone in FACTS-based transmission lines.

## 3. A Temporal Pattern Search Algorithm for Personal History Event Visualization.

### Synopsis:

We present Temporal Pattern Search (TPS), a novel algorithm for searching for temporal patterns of events in historical personal histories. The traditional method of searching for such patterns uses an automaton-based approach over a single array of events, sorted by time stamps. Instead, TPS operates on a set of arrays, where each array contains all events of the same type, sorted by time stamps. TPS searches for a particular item in the pattern using a binary search over the appropriate arrays. Although binary search is considerably more expensive per item, it allows TPS to skip many unnecessary events in personal histories. We show that TPS's running time is bounded by O(m2n lg(n)), where m is the length of (number of events) a search pattern, and n is the number of events in a record (history). Although the asymptotic running time of TPS is inferior to that of a nondeterministic finite automaton (NFA) approach (O(mn)), TPS performs better than NFA under our experimental conditions. We also show TPS is very competitive with Shift-And, a bit-parallel approach, with real data. Since the experimental conditions we describe here subsume the conditions under which analysts would typically use TPS (i.e., within an interactive visualization program), we argue that TPS is an appropriate design choice for us.

## 4. Adaptive Cluster Distance Bounding for High Dimensional Indexing.

### Synopsis:

We consider approaches for similarity search in correlated, high-dimensional data sets, which are derived within a clustering framework. We note that indexing by “vector approximation” (VA-File), which was proposed as a technique to combat the “Curse of Dimensionality,” employs scalar quantization, and hence necessarily ignores dependencies across dimensions, which represents a source of suboptimality. Clustering, on the other hand, exploits interdimensional correlations and is thus a more compact representation of the data set. However, existing methods to prune irrelevant clusters are based on bounding hyperspheres and/or bounding rectangles, whose lack of tightness compromises their efficiency in exact nearest neighbor search. We propose a new cluster-adaptive distance bound based on separating hyperplane boundaries of Voronoi clusters to complement our cluster based index. This bound enables efficient spatial filtering, with a relatively small preprocessing storage overhead and is applicable to euclidean and Mahalanobis similarity measures. Experiments in exact nearest-neighbor set retrieval, conducted on real data sets, show that our indexing method is scalable with data set size and data dimensionality and outperforms several recently proposed indexes. Relative to the VA-File, over a wide range of quantization resolutions, it is able to reduce random IO accesses, given (roughly) the same amount of sequential IO operations, by factors reaching 100X and more.

## 5. Approximate Shortest Distance Computing: A Query-Dependent Local Landmark Scheme.

### Synopsis:

Shortest distance query is a fundamental operation in large-scale networks. Many existing methods in the literature take a landmark embedding approach, which selects a set of graph nodes as landmarks and computes the shortest distances from each landmark to all nodes as an embedding. To answer a shortest distance query, the precomputed distances from the landmarks to the two query nodes are used to compute an approximate shortest distance based on the triangle inequality. In this paper, we analyze the factors that affect the accuracy of distance estimation in landmark embedding. In particular, we find that a globally selected, query-independent landmark set may introduce a large relative error, especially for nearby query nodes. To address this issue, we propose a query-dependent local landmark scheme, which identifies a local landmark close to both query nodes and provides more accurate distance estimation than the traditional global landmark approach. We propose efficient local landmark indexing and retrieval techniques, which achieve low offline indexing complexity and online query complexity. Two optimization techniques on graph compression and graph online search are also proposed, with the goal of further reducing index size and improving query accuracy. Furthermore, the challenge of immense graphs whose index may not fit in the memory leads us to store the embedding in relational database, so that a query of the local landmark scheme can be expressed with relational operators. Effective indexing and query optimization mechanisms are designed in this context. Our experimental results on large-scale social networks and road networks demonstrate that the local landmark scheme reduces the shortest distance estimation error significantly when compared with global landmark embedding and the state-of-the-art sketch-based embedding.

## 6. A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data.

### Synopsis:

Feature selection involves identifying a subset of the most useful features that produces compatible results as the original entire set of features. A feature selection algorithm may be evaluated from both the efficiency and effectiveness points of view. While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. Based on these criteria, a fast clustering-based feature selection algorithm (FAST) is proposed and experimentally evaluated in this paper. The FAST algorithm works in two steps. In the first step, features are divided into clusters by using graph-theoretic clustering methods. In the second step, the most representative feature that is strongly related to target classes is selected from each cluster to form a subset of features. Features in different clusters are relatively independent, the clustering-based strategy of FAST has a high probability of producing a subset of useful and independent features. To ensure the efficiency of FAST, we adopt the efficient minimum-spanning tree (MST) clustering method. The efficiency and effectiveness of the FAST algorithm are evaluated through an empirical study. Extensive experiments are carried out to compare FAST and several representative feature selection algorithms, namely, FCBF, ReliefF, CFS, Consist, and FOCUS-SF, with respect to four types of well-known classifiers, namely, the probability-based Naive Bayes, the tree-based C4.5, the instance-based IB1, and the rule-based RIPPER before and after feature selection. The results, on 35 publicly available real-world high-dimensional image, microarray, and text data, demonstrate that the FAST not only produces smaller subsets of features but also improves the performances of the four types of classifiers.

## 7. Advance Mining of Temporal High Utility Itemset.

### Synopsis:

The stock market domain is a dynamic and unpredictable environment. Traditional techniques, such as fundamental and technical analysis can provide investors with some tools for managing their stocks and predicting their prices. However, these techniques cannot discover all the possible relations between stocks and thus there is a need for a different approach that will provide a deeper kind of analysis. Data mining can be used extensively in the financial markets and help in stock-price forecasting. Therefore, we propose in this paper a portfolio management solution with business intelligence characteristics. We know that the temporal high utility itemsets are the itemsets with support larger than a pre-specified threshold in current time window of data stream. Discovery of temporal high utility itemsets is an important process for mining interesting patterns like association rules from data streams. We proposed the novel algorithm for temporal association mining with utility approach. This make us to find the temporal high utility itemset which can generate less candidate itemsets.

## 8. Data Leakage Detection.

### Synopsis:

We study the following problem: A data distributor has given sensitive data to a set of supposedly trusted agents (third parties). Some of the data are leaked and found in an unauthorized place (e.g., on the web or somebody's laptop). The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. We propose data allocation strategies (across the agents) that improve the probability of identifying leakages. These methods do not rely on alterations of the released data (e.g., watermarks). In some cases, we can also inject “realistic but fake” data records to further improve our chances of detecting leakage and identifying the guilty party.

## 9. Best Keyword Cover Search.

### Synopsis:

It is common that the objects in a spatial database (e.g., restaurants/hotels) are associated with keyword(s) to indicate their businesses/services/features. An interesting problem known as Closest Keywords search is to query objects, called keyword cover, which together cover a set of query keywords and have the minimum inter-objects distance. In recent years, we observe the increasing availability and importance of keyword rating in object evaluation for the better decision making. This motivates us to investigate a generic version of Closest Keywords search called Best Keyword Cover which considers inter-objects distance as well as the keyword rating of objects. The baseline algorithm is inspired by the methods of Closest Keywords search which is based on exhaustively combining objects from different query keywords to generate candidate keyword covers. When the number of query keywords increases, the performance of the baseline algorithm drops dramatically as a result of massive candidate keyword covers generated. To attack this drawback, this work proposes a much more scalable algorithm called keyword nearest neighbor expansion (keyword-NNE). Compared to the baseline algorithm, keyword-NNE algorithm significantly reduces the number of candidate keyword covers generated. The in-depth analysis and extensive experiments on real data sets have justified the superiority of our keyword-NNE algorithm.

## 10. A Generalized Flow-Based Method for Analysis of Implicit Relationships on Wikipedia.

### Synopsis:

We focus on measuring relationships between pairs of objects in Wikipedia whose pages can be regarded as individual objects. Two kinds of relationships between two objects exist: in Wikipedia, an explicit relationship is represented by a single link between the two pages for the objects, and an implicit relationship is represented by a link structure containing the two pages. Some of the previously proposed methods for measuring relationships are cohesion-based methods, which underestimate objects having high degrees, although such objects could be important in constituting relationships in Wikipedia. The other methods are inadequate for measuring implicit relationships because they use only one or two of the following three important factors: distance, connectivity, and cocitation. We propose a new method using a generalized maximum flow which reflects all the three factors and does not underestimate objects having high degree. We confirm through experiments that our method can measure the strength of a relationship more appropriately than these previously proposed methods do. Another remarkable aspect of our method is mining elucidatory objects, that is, objects constituting a relationship. We explain that mining elucidatory objects would open a novel way to deeply understand a relationship.

## 11. An Exploration of Improving Collaborative Recommender Systems via User-Item Subgroups.

### Synopsis:

Collaborative filtering (CF) is one of the most successful recommendation approaches. It typically associates a user with a group of like-minded users based on their preferences over all the items, and recommends to the user those items enjoyed by others in the group. However we find that two users with similar tastes on one item subset may have totally different tastes on another set. In other words, there exist many user-item subgroups each consisting of a subset of items and a group of like-minded users on these items. It is more natural to make preference predictions for a user via the correlated subgroups than the entire user-item matrix. In this paper, to find meaningful subgroups, we formulate the Multiclass Co-Clustering (MCoC) problem and propose an effective solution to it. Then we propose an unified framework to extend the traditional CF algorithms by utilizing the subgroups information for improving their top-N recommendation performance. Our approach can be seen as an extension of traditional clustering CF models. Systematic experiments on three real world data sets have demonstrated the effectiveness of our proposed approach.

## 12. Decision Trees for Uncertain Data.

### Synopsis:

Traditional decision tree classifiers work with data whose values are known and precise. We extend such classifiers to handle data with uncertain information. Value uncertainty arises in many applications during the data collection process. Example sources of uncertainty include measurement/quantization errors, data staleness, and multiple repeated measurements. With uncertainty, the value of a data item is often represented not by one single value, but by multiple values forming a probability distribution. Rather than abstracting uncertain data by statistical derivatives (such as mean and median), we discover that the accuracy of a decision tree classifier can be much improved if the "complete information" of a data item (taking into account the probability density function (pdf)) is utilized. We extend classical decision tree building algorithms to handle data tuples with uncertain values. Extensive experiments have been conducted which show that the resulting classifiers are more accurate than those using value averages. Since processing pdfs is computationally more costly than processing single values (e.g., averages), decision tree construction on uncertain data is more CPU demanding than that for certain data. To tackle this problem, we propose a series of pruning techniques that can greatly improve construction efficiency.

# 13. Building Confidential and Efficient Query Services in the Cloud with RASP Data Perturbation..

## Synopsis:

With the wide deployment of public cloud computing infrastructures, using clouds to host data query services has become an appealing solution for the advantages on scalability and cost-saving. However, some data might be sensitive that the data owner does not want to move to the cloud unless the data confidentiality and query privacy are guaranteed. On the other hand, a secured query service should still provide efficient query processing and significantly reduce the in-house workload to fully realize the benefits of cloud computing. We propose the random space perturbation (RASP) data perturbation method to provide secure and efficient range query and kNN query services for protected data in the cloud. The RASP data perturbation method combines order preserving encryption, dimensionality expansion, random noise injection, and random projection, to provide strong resilience to attacks on the perturbed data and queries. It also preserves multidimensional ranges, which allows existing indexing techniques to be applied to speedup range query processing. The kNN-R algorithm is designed to work with the RASP range query algorithm to process the kNN queries. We have carefully analyzed the attacks on data and queries under a precisely defined threat model and realistic security assumptions. Extensive experiments have been conducted to show the advantages of this approach on efficiency and security.

# 14. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining.

## Synopsis:

Data mining is an increasingly important technology for extracting useful knowledge hidden in large collections of data. There are, however, negative social perceptions about data mining, among which potential privacy invasion and potential discrimination. The latter consists of unfairly treating people on the basis of their belonging to a specific group. Automated data collection and data mining techniques such as classification rule mining have paved the way to making automated decisions, like loan granting/denial, insurance premium computation, etc. If the training data sets are biased in what regards discriminatory (sensitive) attributes like gender, race, religion, etc., discriminatory decisions may ensue. For this reason, anti-discrimination techniques including discrimination discovery and prevention have been introduced in data mining. Discrimination can be either direct or indirect. Direct discrimination occurs when decisions are made based on sensitive attributes. Indirect discrimination occurs when decisions are made based on nonsensitive attributes which are strongly correlated with biased sensitive ones. In this paper, we tackle discrimination prevention in data mining and propose new techniques applicable for direct or indirect discrimination prevention individually or both at the same time. We discuss how to clean training data sets and outsourced data sets in such a way that direct and/or indirect discriminatory decision rules are converted to legitimate (nondiscriminatory) classification rules. We also propose new metrics to evaluate the utility of the proposed approaches and we compare these approaches. The experimental evaluations demonstrate that the proposed techniques are effective at removing direct and/or indirect discrimination biases in the original data set while preserving data quality.

## 15. Anomaly Detection for Discrete Sequences: A Survey.

### Synopsis:

This survey attempts to provide a comprehensive and structured overview of the existing research for the problem of detecting anomalies in discrete/symbolic sequences. The objective is to provide a global understanding of the sequence anomaly detection problem and how existing techniques relate to each other. The key contribution of this survey is the classification of the existing research into three distinct categories, based on the problem formulation that they are trying to solve. These problem formulations are: 1) identifying anomalous sequences with respect to a database of normal sequences; 2) identifying an anomalous subsequence within a long sequence; and 3) identifying a pattern in a sequence whose frequency of occurrence is anomalous. We show how each of these problem formulations is characteristically distinct from each other and discuss their relevance in various application domains. We review techniques from many disparate and disconnected application domains that address each of these formulations. Within each problem formulation, we group techniques into categories based on the nature of the underlying algorithm. For each category, we provide a basic anomaly detection technique, and show how the existing techniques are variants of the basic technique. This approach shows how different techniques within a category are related or different from each other. Our categorization reveals new variants and combinations that have not been investigated before for anomaly detection. We also provide a discussion of relative strengths and weaknesses of different techniques. We show how techniques developed for one problem formulation can be adapted to solve a different formulation, thereby providing several novel adaptations to solve the different problem formulations. We also highlight the applicability of the techniques that handle discrete sequences to other related areas such as online anomaly detection and time series anomaly detection.

## 16. Discovering Conditional Functional Dependencies.

Synopsis:

This paper investigates the discovery of conditional functional dependencies (CFDs). CFDs are a recent extension of functional dependencies (FDs) by supporting patterns of semantically related constants, and can be used as rules for cleaning relational data. However, finding quality CFDs is an expensive process that involves intensive manual effort. To effectively identify data cleaning rules, we develop techniques for discovering CFDs from relations. Already hard for traditional FDs, the discovery problem is more difficult for CFDs. Indeed, mining patterns in CFDs introduces new challenges. We provide three methods for CFD discovery. The first, referred to as CFDMiner, is based on techniques for mining closed item sets, and is used to discover constant CFDs, namely, CFDs with constant patterns only. Constant CFDs are particularly important for object identification, which is essential to data cleaning and data integration. The other two algorithms are developed for discovering general CFDs. One algorithm, referred to as CTANE, is a levelwise algorithm that extends TANE, a well-known algorithm for mining FDs. The other, referred to as FastCFD, is based on the depth-first approach used in FastFD, a method for discovering FDs. It leverages closed-item-set mining to reduce the search space. As verified by our experimental study, CFDMiner can be multiple orders of magnitude faster than CTANE and FastCFD for constant CFD discovery. CTANE works well when a given relation is large, but it does not scale well with the arity of the relation. FastCFD is far more efficient than CTANE when the arity of the relation is large; better still, leveraging optimization based on closed-item-set mining, FastCFD also scales well with the size of the relation. These algorithms provide a set of cleaning-rule discovery tools for users to choose for different applications.

## 17. Capturing Telic/Atelic Temporal Data Semantics: Generalizing Conventional Conceptual Models.

Synopsis:

Time provides context for all our experiences, cognition, and coordinated collective action. Prior research in linguistics, artificial intelligence, and temporal databases suggests the need to differentiate between temporal facts with goal-related semantics (i.e., telic) from those are intrinsically devoid of culmination (i.e., atelic). To differentiate between telic and atelic data semantics in conceptual database design, we propose an annotation-based temporal conceptual model that generalizes the semantics of a conventional conceptual model. Our temporal conceptual design approach involves: 1) capturing "what" semantics using a conventional conceptual model; 2) employing annotations to differentiate between telic and atelic data semantics that help capture "when" semantics; 3) specifying temporal constraints, specifically nonsequenced semantics, in the temporal data dictionary as metadata. Our proposed approach provides a mechanism to represent telic/atelic temporal semantics using temporal annotations. We also show how these semantics can be formally defined using constructs of the conventional conceptual model and axioms in first-order logic. Via what we refer to as the "semantics of composition," i.e., semantics implied by the interaction of annotations, we illustrate the logical consequences of representing telic/atelic data semantics during temporal conceptual design.

## 18. A New Algorithm for Inferring User Search Goals with Feedback Sessions.

### Synopsis:

For a broad-topic and ambiguous query, different users may have different search goals when they submit it to a search engine. The inference and analysis of user search goals can be very useful in improving search engine relevance and user experience. In this paper, we propose a novel approach to infer user search goals by analyzing search engine query logs. First, we propose a framework to discover different user search goals for a query by clustering the proposed feedback sessions. Feedback sessions are constructed from user click-through logs and can efficiently reflect the information needs of users. Second, we propose a novel approach to generate pseudo-documents to better represent the feedback sessions for clustering. Finally, we propose a new criterion )“Classified Average Precision (CAP)” to evaluate the performance of inferring user search goals. Experimental results are presented using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

## 19. Automatic Discovery of Association Orders between Name and Aliases from the Web using Anchor Texts-based Co- occurrences.

### Synopsis:

Many celebrities and experts from various fields may have been referred by not only their personal names but also by their aliases on web. Aliases are very important in information retrieval to retrieve complete information about a personal name from the web, as some of the web pages of the person may also be referred by his aliases. The aliases for a personal name are extracted by previously proposed alias extraction method. In information retrieval, the web search engine automatically expands the search query on a person name by tagging his aliases for complete information retrieval thereby improving recall in relation detection task and achieving a significant mean reciprocal rank (MRR) of search engine. For the further substantial improvement on recall and MRR from the previously proposed methods, our proposed method will order the aliases based on their associations with the name using the definition of anchor texts-based co-occurrences between name and aliases in order to help the search engine tag the aliases according to the order of associations. The association orders will automatically be discovered by creating an anchor texts-based co-occurrence graph between name and aliases. Ranking support vector machine (SVM) will be used to create connections between name and aliases in the graph by performing ranking on anchor texts-based co-occurrence measures. The hop distances between nodes in the graph will lead to have the associations between name and aliases. The hop distances will be found by mining the graph. The proposed method will outperform previously proposed methods, achieving substantial growth on recall and MRR.

## 20. Effective Navigation of Query Results Based on Concept Hierarchies.

### Synopsis:

Search queries on biomedical databases, such as PubMed, often return a large number of results, only a small subset of which is relevant to the user. Ranking and categorization, which can also be combined, have been proposed to alleviate this information overload problem. Results categorization for biomedical databases is the focus of this work. A natural way to organize biomedical citations is according to their MeSH annotations. MeSH is a comprehensive concept hierarchy used by PubMed. In this paper, we present the BioNav system, a novel search interface that enables the user to navigate large number of query results by organizing them using the MeSH concept hierarchy. First, the query results are organized into a navigation tree. At each node expansion step, BioNav reveals only a small subset of the concept nodes, selected such that the expected user navigation cost is minimized. In contrast, previous works expand the hierarchy in a predefined static manner, without navigation cost modeling. We show that the problem of selecting the best concepts to reveal at each node expansion is NP-complete and propose an efficient heuristic as well as a feasible optimal algorithm for relatively small trees. We show experimentally that BioNav outperforms state-of-the-art categorization systems by up to an order of magnitude, with respect to the user navigation cost. BioNav for the MEDLINE database is available at http://db.cse.buffalo.edu/bionav.

## 21. Dealing With Concept Drifts in Process MiningServices.

Synopsis:

Although most business processes change over time, contemporary process mining techniques tend to analyze these processes as if they are in a steady state. Processes may change suddenly or gradually. The drift may be periodic (e.g., because of seasonal influences) or one-of-a-kind (e.g., the effects of new legislation). For the process management, it is crucial to discover and understand such concept drifts in processes. This paper presents a generic framework and specific techniques to detect when a process changes and to localize the parts of the process that have changed. Different features are proposed to characterize relationships among activities. These features are used to discover differences between successive populations. The approach has been implemented as a plug-in of the ProM process mining framework and has been evaluated using both simulated event data exhibiting controlled concept drifts and real-life event data from a Dutch municipality.

## 22. A Probabilistic Approach to String Transformation.

Synopsis:

Many problems in natural language processing, data mining, information retrieval, and bioinformatics can be formalized as string transformation, which is a task as follows. Given an input string, the system generates the k most likely output strings corresponding to the input string. This paper proposes a novel and probabilistic approach to string transformation, which is both accurate and efficient. The approach includes the use of a log linear model, a method for training the model, and an algorithm for generating the top k candidates, whether there is or is not a predefined dictionary. The log linear model is defined as a conditional probability distribution of an output string and a rule set for the transformation conditioned on an input string. The learning method employs maximum likelihood estimation for parameter estimation. The string generation algorithm based on pruning is guaranteed to generate the optimal top k candidates. The proposed method is applied to correction of spelling errors in queries as well as reformulation of queries in web search. Experimental results on large scale data show that the proposed approach is very accurate and efficient improving upon existing methods in terms of accuracy and efficiency in different settings.

## 23. Confucius A Tool Supporting Collaborative Scientific Workflow Composition.

Synopsis:

Modern scientific data management and analysis usually rely on multiple scientists with diverse expertise. In recent years, such a collaborative effort is often structured and automated by a data flow-oriented process called scientific workflow. However, such workflows may have to be designed and revised among multiple scientists over a long time period. Existing workbenches are single user-oriented and do not support scientific workflow application development in a "collaborative fashion". In this paper, we report our research on the enabling techniques in the aspects of collaboration provenance management and reproduciability. Based on a scientific collaboration ontology, we propose a service-oriented collaboration model supported by a set of composable collaboration primitives and patterns. The collaboration protocols are then applied to support effective concurrency control in the process of collaborative workflow composition. We also report the design and development of Confucius, a service-oriented collaborative scientific workflow composition tool that extends an open-source, single-user development environment.

## 24Extended XML Tree Pattern Matching: Theories and Algorithms.

### Synopsis:

As business and enterprises generate and exchange XML data more often, there is an increasing need for efficient processing of queries on XML data. Searching for the occurrences of a tree pattern query in an XML database is a core operation in XML query processing. Prior works demonstrate that holistic twig pattern matching algorithm is an efficient technique to answer an XML tree pattern with parent-child (P-C) and ancestor-descendant (A-D) relationships, as it can effectively control the size of intermediate results during query processing. However, XML query languages (e.g., XPath and XQuery) define more axes and functions such as negation function, order-based axis, and wildcards. In this paper, we research a large set of XML tree pattern, called extended XML tree pattern, which may include P-C, A-D relationships, negation functions, wildcards, and order restriction. We establish a theoretical framework about “matching cross” which demonstrates the intrinsic reason in the proof of optimality on holistic algorithms. Based on our theorems, we propose a set of novel algorithms to efficiently process three categories of extended XML tree patterns. A set of experimental results on both real-life and synthetic data sets demonstrate the effectiveness and efficiency of our proposed theories and algorithms.

## 25. Efficient Ranking on Entity Graphs with Personalized Relationships.

### Synopsis:

Authority flow techniques like PageRank and ObjectRank can provide personalized ranking of typed entity-relationship graphs. There are two main ways to personalize authority flow ranking: Node-based personalization, where authority originates from a set of user-specific nodes; edge-based personalization, where the importance of different edge types is user-specific. We propose the first approach to achieve efficient edge-based personalization using a combination of precomputation and runtime algorithms. In particular, we apply our method to ObjectRank, where a personalized weight assignment vector (WAV) assigns different weights to each edge type or relationship type. Our approach includes a repository of rankings for various WAVs. We consider the following two classes of approximation: (a) SchemaApprox is formulated as a distance minimization problem at the schema level; (b) DataApprox is a distance minimization problem at the data graph level. SchemaApprox is not robust since it does not distinguish between important and trivial edge types based on the edge distribution in the data graph. In contrast, DataApprox has a provable error bound. Both SchemaApprox and DataApprox are expensive so we develop efficient heuristic implementations, ScaleRank and PickOne respectively. Extensive experiments on the DBLP data graph show that ScaleRank provides a fast and accurate personalized authority flow ranking.

## 26. A Survey of XML Tree Patterns.

### Synopsis:

With XML becoming a ubiquitous language for data interoperability purposes in various domains, efficiently querying XML data is a critical issue. This has lead to the design of algebraic frameworks based on tree-shaped patterns akin to the tree-structured data model of XML. Tree patterns are graphic representations of queries over data trees. They are actually matched against an input data tree to answer a query. Since the turn of the 21st century, an astounding research effort has been focusing on tree pattern models and matching optimization (a primordial issue). This paper is a comprehensive survey of these topics, in which we outline and compare the various features of tree patterns. We also review and discuss the two main families of approaches for optimizing tree pattern matching, namely pattern tree minimization and holistic matching. We finally present actual tree pattern-based developments, to provide a global overview of this significant research topic.

## 27. Coupled Behavior Analysis for Capturing Coupling Relationships in Group-based Market Manipulations.

## Synopsis:

Coupled behaviors, which refer to behaviors having some relationships between them, are usually seen in many real-world scenarios, especially in stock markets. Recently, the coupled hidden Markov model (CHMM)-based coupled behavior analysis has been proposed to consider the coupled relationships in a hidden state space. However, it requires aggregation of the behavioral data to cater for the CHMM modeling, which may overlook the couplings within the aggregated behaviors to some extent. In addition, the Markov assumption limits its capability to capturing temporal couplings. Thus, this paper proposes a novel graph-based framework for detecting abnormal coupled behaviors. The proposed framework represents the coupled behaviors in a graph view without aggregating the behavioral data and is flexible to capture richer coupling information of the behaviors (not necessarily temporal relations). On top of that, the couplings are learned via relational learning methods and an efficient anomaly detection algorithm is proposed as well. Experimental results on a real-world data set in stock markets show that the proposed framework outperforms the CHMM-based one in both technical and business measures.

# 28. Group Enclosing Queries..

## Synopsis:

Given a set of points P and a query set Q, a group enclosing query (Geq) fetches the point p* ∊ P such that the maximum distance of p* to all points in Q is minimized. This problem is equivalent to the Min-Max case (minimizing the maximum distance) of aggregate nearest neighbor queries for spatial databases. This work first designs a new exact solution by exploring new geometric insights, such as the minimum enclosing ball, the convex hull, and the furthest voronoi diagram of the query group. To further reduce the query cost, especially when the dimensionality increases, we turn to approximation algorithms. Our main approximation algorithm has a worst case √2-approximation ratio if one can find the exact nearest neighbor of a point. In practice, its approximation ratio never exceeds 1.05 for a large number of data sets up to six dimensions. We also discuss how to extend it to higher dimensions (up to 74 in our experiment) and show that it still maintains a very good approximation quality (still close to 1) and low query cost. In fixed dimensions, we extend the √2-approximation algorithm to get a (1 + ε)-approximate solution for the Geq problem. Both approximation algorithms have O(log N + M) query cost in any fixed dimension, where N and M are the sizes of the data set P and query group Q. Extensive experiments on both synthetic and real data sets, up to 10 million points and 74 dimensions, confirm the efficiency, effectiveness, and scalability of the proposed algorithms, especially their significant improvement over the state-of-the-art method.

## 29. Facilitating Document Annotation using Content and Querying Value.

### Synopsis:

A large number of organizations today generate and share textual descriptions of their products, services, and actions. Such collections of textual data contain significant amount of structured information, which remains buried in the unstructured text. While information extraction algorithms facilitate the extraction of structured relations, they are often expensive and inaccurate, especially when operating on top of text that does not contain any instances of the targeted structured information. We present a novel alternative approach that facilitates the generation of the structured metadata by identifying documents that are likely to contain information of interest and this information is going to be subsequently useful for querying the database. Our approach relies on the idea that humans are more likely to add the necessary metadata during creation time, if prompted by the interface; or that it is much easier for humans (and/or algorithms) to identify the metadata when such information actually exists in the document, instead of naively prompting users to fill in forms with information that is not available in the document. As a major contribution of this paper, we present algorithms that identify structured attributes that are likely to appear within the document, by jointly utilizing the content of the text and the query workload. Our experimental evaluation shows that our approach generates superior results compared to approaches that rely only on the textual content or only on the query workload, to identify attributes of interest.

## 30. A System to Filter Unwanted Messages from OSN User Walls.

### Synopsis:

One fundamental issue in today's Online Social Networks (OSNs) is to give users the ability to control the messages posted on their own private space to avoid that unwanted content is displayed. Up to now, OSNs provide little support to this requirement. To fill the gap, in this paper, we propose a system allowing OSN users to have a direct control on the messages posted on their walls. This is achieved through a flexible rule-based system, that allows users to customize the filtering criteria to be applied to their walls, and a Machine Learning-based soft classifier automatically labeling messages in support of content-based filtering.

## 31. Creating Evolving User Behaviour Profiles Automatically.

## Synopsis:

Knowledge about computer users is very beneficial for assisting them, predicting their future actions or detecting masqueraders. In this paper, a new approach for creating and recognizing automatically the behavior profile of a computer user is presented. In this case, a computer user behavior is represented as the sequence of the commands she/he types during her/his work. This sequence is transformed into a distribution of relevant subsequences of commands in order to find out a profile that defines its behavior. Also, because a user profile is not necessarily fixed but rather it evolves/changes, we propose an evolving method to keep up to date the created profiles using an Evolving Systems approach. In this paper, we combine the evolving classifier with a trie-based user profiling to obtain a powerful self-learning online scheme. We also develop further the recursive formula of the potential of a data point to become a cluster center using cosine distance, which is provided in the Appendix. The novel approach proposed in this paper can be applicable to any problem of dynamic/evolving user behavior modeling where it can be represented as a sequence of actions or events. It has been evaluated on several real data streams.

# 32. Load Shedding in Mobile Systems with MobiQual.

## Synopsis:

In location-based, mobile continual query (CQ) systems, two key measures of quality-of-service (QoS) are: freshness and accuracy. To achieve freshness, the CQ server must perform frequent query reevaluations. To attain accuracy, the CQ server must receive and process frequent position updates from the mobile nodes. However, it is often difficult to obtain fresh and accurate CQ results simultaneously, due to 1) limited resources in computing and communication and 2) fast-changing load conditions caused by continuous mobile node movement. Hence, a key challenge for a mobile CQ system is: How do we achieve the highest possible quality of the CQ results, in both freshness and accuracy, with currently available resources? In this paper, we formulate this problem as a load shedding one, and develop MobiQual-a QoS-aware approach to performing both update load shedding and query load shedding. The design of MobiQual highlights three important features. 1) Differentiated load shedding: We apply different amounts of query load shedding and update load shedding to different groups of queries and mobile nodes, respectively. 2) Per-query QoS specification: Individualized QoS specifications are used to maximize the overall freshness and accuracy of the query results. 3) Low-cost adaptation: MobiQual dynamically adapts, with a minimal overhead, to changing load conditions and available resources. We conduct a set of comprehensive experiments to evaluate the effectiveness of MobiQual. The results show that, through a careful combination of update and query load shedding, the MobiQual approach leads to much higher freshness and accuracy in the query results in all cases, compared to existing approaches that lack the

QoS-awareness properties of MobiQual, as well as the solutions that perform query-only or update-only load shedding.

## 33. Fast Nearest Neighbor Search with Keywords.

### Synopsis:

Conventional spatial queries, such as range search and nearest neighbor retrieval, involve only conditions on objects' geometric properties. Today, many modern applications call for novel forms of queries that aim to find objects satisfying both a spatial predicate, and a predicate on their associated texts. For example, instead of considering all the restaurants, a nearest neighbor query would instead ask for the restaurant that is the closest among those whose menus contain “steak, spaghetti, brandy” all at the same time. Currently, the best solution to such queries is based on the IR 2-tree, which, as shown in this paper, has a few deficiencies that seriously impact its efficiency. Motivated by this, we develop a new access method called the spatial inverted index that extends the conventional inverted index to cope with multidimensional data, and comes with algorithms that can answer nearest neighbor queries with keywords in real time. As verified by experiments, the proposed techniques outperform the IR 2-tree in query response time significantly, often by a factor of orders of magnitude.

## 34. Annotating Search Results from Web Databases..

### Synopsis:

An increasing number of databases have become web accessible through HTML form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine processable, which is essential for many applications such as deep web data collection and Internet comparison shopping, they need to be extracted out and assigned meaningful labels. In this paper, we present an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantic. Then, for each group we annotate it from different aspects and aggregate the different annotations to predict a final annotation label for it. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database. Our experiments indicate that the proposed approach is highly effective.

## 35. Credibility Ranking of Tweets during High Impact Events.

### Synopsis:

Twitter has evolved from being a conversation or opinion sharing medium among friends into a platform to share and disseminate information about current events. Events in the real world create a corresponding spur of posts (tweets) on Twitter. Not all content posted on Twitter is trustworthy or useful in providing information about the event. In this paper, we analyzed the credibility of information in tweets corresponding to fourteen high impact news events of 2011 around the globe. From the data we analyzed, on average 30% of total tweets posted about an event contained situational information about the event while 14% was spam. Only 17% of the total tweets posted about the event contained situational awareness information that was credible. Using regression analysis, we identified the important content and sourced based features, which can predict the credibility of information in a tweet. Prominent content based features were number of unique characters, swear words, pronouns, and emoticons in a tweet, and user based features like the number of followers and length of username. We adopted a supervised machine learning and relevance feedback approach using the above features, to rank tweets according to their credibility score. The performance of our ranking algorithm significantly enhanced when we applied re-ranking strategy. Results show that extraction of credible information from Twitter can be automated with high confidence.

## 36. Making Aggregation Work in Uncertain and Probabilistic Databases.

### Synopsis:

We describe how aggregation is handled in the Trio system for uncertain and probabilistic data. Because “exact” aggregation in uncertain databases can produce exponentially sized results, we provide three alternatives: a low bound on the aggregate value, a high bound on the value, and the expected value. These variants return a single result instead of a set of possible results, and they are generally efficient to compute for both full-table and grouped aggregation queries. We provide formal definitions and semantics and a description of our open source implementation for single-table aggregation queries. We study the performance and scalability of our algorithms through experiments over a large synthetic data set. We also provide some preliminary results on aggregations over joins.

## 37. Incremental Affinity Propagation Clustering Based on Message Passing.

## Synopsis:

Affinity Propagation (AP) clustering has been successfully used in a lot of clustering problems. However, most of the applications deal with static data. This paper considers how to apply AP in incremental clustering problems. First, we point out the difficulties in Incremental Affinity Propagation (IAP) clustering, and then propose two strategies to solve them. Correspondingly, two IAP clustering algorithms are proposed. They are IAP clustering based on K-Medoids (IAPKM) and IAP clustering based on Nearest Neighbor Assignment (IAPNA). Five popular labeled data sets, real world time series and a video are used to test the performance of IAPKM and IAPNA. Traditional AP clustering is also implemented to provide benchmark performance. Experimental results show that IAPKM and IAPNA can achieve comparable clustering performance with traditional AP clustering on all the data sets. Meanwhile, the time cost is dramatically reduced in IAPKM and IAPNA. Both the effectiveness and the efficiency make IAPKM and IAPNA able to be well used in incremental clustering tasks.

# 38. Anomaly Detection Approach Using Hidden Markov Model.

## Synopsis:

Anomaly detection is an important problem that has been researched within diverse research areas. Numerous methods and approaches based on Hidden Markov Model regarding the anomaly detection have been proposed and reported in the literature. However, the potential applications using Hidden Markov Model classification based anomaly detection technique have not yet been fully explored and still in its infancy. This paper investigates the capabilities the use of Hidden Markov Model in anomaly detection for discrete sequences.

# 39. DDD: A New Ensemble Approach for Dealing with Concept Drift.

## Synopsis:

Online learning algorithms often have to operate in the presence of concept drifts. A recent study revealed that different diversity levels in an ensemble of learning machines are required in order to maintain high generalization on both old and new concepts. Inspired by this study and based on a further study of diversity with different strategies to deal with drifts, we propose a new online ensemble learning approach called Diversity for Dealing with Drifts (DDD). DDD maintains ensembles with different diversity levels and is able to attain better accuracy than other approaches. Furthermore, it is very robust, outperforming other drift handling approaches in terms of accuracy when there are false positive drift detections. In all the experimental comparisons we have carried out, DDD always performed at least as well as other drift handling approaches under various conditions, with very few exceptions.

## 40. Ranking Spatial Data by Quality Preferences.

### Synopsis:

A spatial preference query ranks objects based on the qualities of features in their spatial neighborhood. For example, using a real estate agency database of flats for lease, a customer may want to rank the flats with respect to the appropriateness of their location, defined after aggregating the qualities of other features (e.g., restaurants, cafes, hospital, market, etc.) within their spatial neighborhood. Such a neighborhood concept can be specified by the user via different functions. It can be an explicit circular region within a given distance from the flat. Another intuitive definition is to assign higher weights to the features based on their proximity to the flat. In this paper, we formally define spatial preference queries and propose appropriate indexing techniques and search algorithms for them. Extensive evaluation of our methods on both real and synthetic data reveals that an optimized branch-and-bound solution is efficient and robust with respect to different parameters.

## 41. Infrequent Weighted Itemset Mining Using Frequent Pattern Growth.

### Synopsis:

Frequent weighted itemsets represent correlations frequently holding in data in which items may weight differently. However, in some contexts, e.g., when the need is to minimize a certain cost function, discovering rare data correlations is more interesting than mining frequent ones. This paper tackles the issue of discovering rare and weighted itemsets, i.e., the infrequent weighted itemset (IWI) mining problem. Two novel quality measures are proposed to drive the IWI mining process. Furthermore, two algorithms that perform IWI and Minimal IWI mining efficiently, driven by the proposed measures, are presented. Experimental results show efficiency and effectiveness of the proposed approach.

## 42. Anomaly Detection via Online Oversampling Principal Component Analysis.

### Synopsis:

Anomaly detection has been an important research topic in data mining and machine learning. Many real-world applications such as intrusion or credit card fraud detection require an effective and efficient framework to identify deviated data instances. However, most anomaly detection methods are typically implemented in batch mode, and thus cannot be easily extended to large-scale problems without sacrificing computation and memory requirements. In this paper, we propose an online oversampling principal component analysis (osPCA) algorithm to address this problem, and we aim at detecting the presence of outliers from a large amount of data via an online updating technique. Unlike prior principal component analysis (PCA)-based approaches, we do not store the entire data matrix or covariance matrix, and thus our approach is especially of interest in online or large-scale problems. By oversampling the target instance and extracting the principal direction of the data, the proposed osPCA allows us to determine the anomaly of the target instance according to the variation of the resulting dominant eigenvector. Since our osPCA need not perform eigen analysis explicitly, the proposed framework is favored for online applications which have computation or memory limitations. Compared with the well-known power method for PCA and other popular anomaly detection algorithms, our experimental results verify the feasibility of our proposed method in terms of both accuracy and efficiency.

## 43. Effective Pattern Discovery for Text Mining.

### Synopsis:

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance.

## 44. Relevance-based Retrieval on Hidden-Web Text Databases without Ranking Support.

## Synopsis:

Many online or local data sources provide powerful querying mechanisms but limited ranking capabilities. For instance, PubMed allows users to submit highly expressive Boolean keyword queries, but ranks the query results by date only. However, a user would typically prefer a ranking by relevance, measured by an information retrieval (IR) ranking function. A naive approach would be to submit a disjunctive query with all query keywords, retrieve all the returned matching documents, and then rerank them. Unfortunately, such an operation would be very expensive due to the large number of results returned by disjunctive queries. In this paper, we present algorithms that return the top results for a query, ranked according to an IR-style ranking function, while operating on top of a source with a Boolean query interface with no ranking capabilities (or a ranking capability of no interest to the end user). The algorithms generate a series of conjunctive queries that return only documents that are candidates for being highly ranked according to a relevance metric. Our approach can also be applied to other settings where the ranking is monotonic on a set of factors (query keywords in IR) and the source query interface is a Boolean expression of these factors. Our comprehensive experimental evaluation on the PubMed database and a TREC data set show that we achieve order of magnitude improvement compared to the current baseline approaches.

# 45. A Meta-Top-Down Method for Large-Scale Hierarchical Classification.

## Synopsis:

Recent large-scale hierarchical classification tasks typically have tens of thousands of classes on which the most widely used approach to multiclass classification--one-versus-rest--becomes intractable due to computational complexity. The top-down methods are usually adopted instead, but they are less accurate because of the so-called error-propagation problem in their classifying phase. To address this problem, this paper proposes a meta-top-down method that employs metaclassification to enhance the normal top-down classifying procedure. The proposed method is first analyzed theoretically on complexity and accuracy, and then applied to five real-world large-scale data sets. The experimental results indicate that the classification accuracy is largely improved, while the increased time costs are smaller than most of the existing approaches.

# 46. CloudMoV: Cloud-based Mobile Social TV.

Synopsis:

The rapidly increasing power of personal mobile devices (smartphones, tablets, etc.) is providing much richer contents and social interactions to users on the move. This trend however is throttled by the limited battery lifetime of mobile devices and unstable wireless connectivity, making the highest possible quality of service experienced by mobile users not feasible. The recent cloud computing technology, with its rich resources to compensate for the limitations of mobile devices and connections, can potentially provide an ideal platform to support the desired mobile services. Tough challenges arise on how to effectively exploit cloud resources to facilitate mobile services, especially those with stringent interaction delay requirements. In this paper, we propose the design of a Cloud-based, novel Mobile sOcial tV system (CloudMoV). The system effectively utilizes both PaaS (Platform-as-a-Service) and IaaS (Infrastructure-as-a-Service) cloud services to offer the living-room experience of video watching to a group of disparate mobile users who can interact socially while sharing the video. To guarantee good streaming quality as experienced by the mobile users with time-varying wireless connectivity, we employ a surrogate for each user in the IaaS cloud for video downloading and social exchanges on behalf of the user. The surrogate performs efficient stream transcoding that matches the current connectivity quality of the mobile user. Given the battery life as a key performance bottleneck, we advocate the use of burst transmission from the surrogates to the mobile users, and carefully decide the burst size which can lead to high energy efficiency and streaming quality. Social interactions among the users, in terms of spontaneous textual exchanges, are effectively achieved by efficient designs of data storage with BigTable and dynamic handling of large volumes of concurrent messages in a typical PaaS cloud. These various designs for flexible transcoding capabilities- battery efficiency of mobile devices and spontaneous social interactivity together provide an ideal platform for mobile social TV services. We have implemented CloudMoV on Amazon EC2 and Google App Engine and verified its superior performance based on real-world experiments.

## 47.A  Efficient Anonymous Message Submission.

Synopsis:

In online surveys, many people are not willing to provide true answers due to privacy concerns. Thus, anonymity is important for online message collection. Existing solutions let each member blindly shuffle the submitted messages by using the IND-CCA2 secure cryptosystem. In the end, all messages are randomly shuffled and no one knows the message order. However, the heavy computational overhead and linear communication rounds make it only useful for small groups. In this paper, we propose an efficient anonymous message submission protocol aimed at a practical group size. Our protocol is based on a simplified secret sharing scheme and a symmetric key cryptosystem. We propose a novel method to let all members secretly aggregate their messages into a message vector such that a member knows nothing about other members' message positions.We provide a theoretical proof showing that our protocol is anonymous under malicious attacks. We then conduct a thorough analysis of our protocol, showing that our protocol is computationally more efficient than existing solutions and results in a constant communication rounds with a high probability.

## 48. The CoQUOS Approach to Continuous Queries in Unstructured Overlays.

### Synopsis:

The current peer-to-peer (P2P) content distribution systems are constricted by their simple on-demand content discovery mechanism. The utility of these systems can be greatly enhanced by incorporating two capabilities, namely a mechanism through which peers can register their long term interests with the network so that they can be continuously notified of new data items, and a means for the peers to advertise their contents. Although researchers have proposed a few unstructured overlay-based publish-subscribe systems that provide the above capabilities, most of these systems require intricate indexing and routing schemes, which not only make them highly complex but also render the overlay network less flexible toward transient peers. This paper argues that for many P2P applications, implementing full-fledged publish-subscribe systems is an overkill. For these applications, we study the alternate continuous query paradigm, which is a best-effort service providing the above two capabilities. We present a scalable and effective middleware, called CoQUOS, for supporting continuous queries in unstructured overlay networks. Besides being independent of the overlay topology, CoQUOS preserves the simplicity and flexibility of the unstructured P2P network. Our design of the CoQUOS system is characterized by two novel techniques, namely cluster-resilient random walk algorithm for propagating the queries to various regions of the network and dynamic probability-based query registration scheme to ensure that the registrations are well distributed in the overlay. Further, we also develop effective and efficient schemes for providing resilience to the churn of the P2P network and for ensuring a fair distribution of the notification load among the peers. This paper studies the properties of our algorithms through theoretical analysis. We also report series of experiments evaluating the effectiveness and the costs of the proposed schemes.

## 49. Mining Weakly Labeled Web Facial Images for Search-Based Face Annotation.

### Synopsis:

This paper investigates a framework of search-based face annotation (SBFA) by mining weakly labeled facial images that are freely available on the World Wide Web (WWW). One challenging problem for search-based face annotation scheme is how to effectively perform annotation by exploiting the list of most similar facial images and their weak labels that are often noisy and incomplete. To tackle this problem, we propose an effective unsupervised label refinement (ULR) approach for refining the labels of web facial images using machine learning techniques. We formulate the learning problem as a convex optimization and develop effective optimization algorithms to solve the large-scale learning task efficiently. To further speed up the proposed scheme, we also propose a clustering-based approximation algorithm which can improve the scalability considerably. We have conducted an extensive set of empirical studies on a large-scale web facial image testbed, in which encouraging results showed that the proposed ULR algorithms can significantly boost the performance of the promising SBFA scheme.

## 50. Comparable Entity Mining from Comparative Questions.

### Synopsis:

Comparing one thing with another is a typical part of human decision making process. However, it is not always easy to know what to compare and what are the alternatives. In this paper, we present a novel way to automatically mine comparable entities from comparative questions that users posted online to address this difficulty. To ensure high precision and high recall, we develop a weakly supervised bootstrapping approach for comparative question identification and comparable entity extraction by leveraging a large collection of online question archive. The experimental results show our method achieves F1-measure of 82.5 percent in comparative question identification and 83.3 percent in comparable entity extraction. Both significantly outperform an existing state-of-the-art method. Additionally, our ranking results show highly relevance to user's comparison intents in web.

## 51. Efficient Computation of Range Aggregates against Uncertain Location Based Queries.

### Synopsis:

In many applications, including location-based services, queries may not be precise. In this paper, we study the problem of efficiently computing range aggregates in a multidimensional space when the query location is uncertain. Specifically, for a query point Q whose location is uncertain and a set S of points in a multidimensional space, we want to calculate the aggregate (e.g., count, average and sum) over the subset S' of S such that for each p ϵ S', Q has at least probability θ within the distance γ to p. We propose novel, efficient techniques to solve the problem following the filtering-and-verification paradigm. In particular, two novel filtering techniques are proposed to effectively and efficiently remove data points from verification. Our comprehensive experiments based on both real and synthetic data demonstrate the efficiency and scalability of our techniques.

## 52. On Skyline Groups.

### Synopsis:

We formulate and investigate the novel problem of finding the skyline k-tuple groups from an n-tuple data set-i.e., groups of k tuples which are not dominated by any other group of equal size, based on aggregate-based group dominance relationship. The major technical challenge is to identify effective anti-monotonic properties for pruning the search space of skyline groups. To this end, we first show that the anti-monotonic property in the well-known Apriori algorithm does not hold for skyline group pruning. Then, we identify two anti-monotonic properties with varying degrees of applicability: order-specific property which applies to SUM, MIN, and MAX as well as weak candidate-generation property which applies to MIN and MAX only. Experimental results on both real and synthetic data sets verify that the proposed algorithms achieve orders of magnitude performance gain over the baseline method.

## 53. Co-Occurrence-Based Diffusion for Expert Search on the Web.

### Synopsis:

Expert search has been studied in different contexts, e.g., enterprises, academic communities. We examine a general expert search problem: searching experts on the web, where millions of webpages and thousands of names are considered. It has mainly two challenging issues: 1) webpages could be of varying quality and full of noises; 2) The expertise evidences scattered in webpages are usually vague and ambiguous. We propose to leverage the large amount of co-occurrence information to assess relevance and reputation of a person name for a query topic. The co-occurrence structure is modeled using a hypergraph, on which a heat diffusion based ranking algorithm is proposed. Query keywords are regarded as heat sources, and a person name which has strong connection with the query (i.e., frequently co-occur with query keywords and co-occur with other names related to query keywords) will receive most of the heat, thus being ranked high.

Experiments on the ClueWeb09 web collection show that our algorithm is effective for retrieving experts and outperforms baseline algorithms significantly. This work would be regarded as one step toward addressing the more general entity search problem without sophisticated NLP techniques.

## 54. Efficient Extended Boolean Retrieval.

### Synopsis:

Extended Boolean retrieval (EBR) models were proposed nearly three decades ago, but have had little practical impact, despite their significant advantages compared to either ranked keyword or pure Boolean retrieval. In particular, EBR models produce meaningful rankings; their query model allows the representation of complex concepts in an and-or format; and they are scrutable, in that the score assigned to a document depends solely on the content of that document, unaffected by any collection statistics or other external factors. These characteristics make EBR models attractive in domains typified by medical and legal searching, where the emphasis is on iterative development of reproducible complex queries of dozens or even hundreds of terms. However, EBR is much more computationally expensive than the alternatives. We consider the implementation of the p-norm approach to EBR, and demonstrate that ideas used in the max-score and wand exact optimization techniques for ranked keyword retrieval can be adapted to allow selective bypass of documents via a low-cost screening process for this and similar retrieval models. We also propose term-independent bounds that are able to further reduce the number of score calculations for short, simple queries under the extended Boolean retrieval model. Together, these methods yield an overall saving from 50 to 80 percent of the evaluation cost on test queries drawn from biomedical search.

## 55. On the use of Side Information for Mining Text Data.

### Synopsis:

In many text mining applications, side-information is available along with the text documents. Such side-information may be of different kinds, such as document provenance information, the links in the document, user-access behavior from web logs, or other non-textual attributes which are embedded into the text document. Such attributes may contain a tremendous amount of information for clustering purposes. However, the relative importance of this side-information may be difficult to estimate, especially when some of the information is noisy. In such cases, it can be risky to incorporate side-information into the mining process, because it can either improve the quality of the representation for the mining process, or can add noise to the process. Therefore, we need a principled way to perform the mining process, so as to maximize the advantages from using this side information. In this paper, we design an algorithm which combines classical partitioning algorithms with probabilistic models in order to create an effective clustering approach. We then show how to extend the approach to the classification problem. We present experimental results on a number of real data sets in order to illustrate the advantages of using such an approach.

## 56. Crowdsourcing Predictors of Behavioral Outcomes.

### Synopsis:

Generating models from large data sets-and determining which subsets of data to mine-is becoming increasingly automated. However, choosing what data to collect in the first place requires human intuition or experience, usually supplied by a domain expert. This paper describes a new approach to machine science which demonstrates for the first time that nondomain experts can collectively formulate features and provide values for those features such that they are predictive of some behavioral outcome of interest. This was accomplished by building a Web platform in which human groups interact to both respond to questions likely to help predict a behavioral outcome and pose new questions to their peers. This results in a dynamically growing online survey, but the result of this cooperative behavior also leads to models that can predict the user's outcomes based on their responses to the user-generated survey questions. Here, we describe two Web-based experiments that instantiate this approach: The first site led to models that can predict users' monthly electric energy consumption, and the other led to models that can predict users' body mass index. As exponential increases in content are often observed in successful online collaborative communities, the proposed methodology may, in the future, lead to similar exponential rises in discovery and insight into the causal factors of behavioral outcomes.

## 57. Efficient Similarity Search over Encrypted Data.

### Synopsis:

In recent years, due to the appealing features of cloud computing, large amount of data have been stored in the cloud. Although cloud based services offer many advantages, privacy and security of the sensitive data is a big concern. To mitigate the concerns, it is desirable to outsource sensitive data in encrypted form. Encrypted storage protects the data against illegal access, but it complicates some basic, yet important functionality such as the search on the data. To achieve search over encrypted data without compromising the privacy, considerable amount of searchable encryption schemes have been proposed in the literature. However, almost all of them handle exact query matching but not similarity matching, a crucial requirement for real world applications. Although some sophisticated secure multi-party computation based cryptographic techniques are available for similarity tests, they are computationally intensive and do not scale for large data sources. In this paper, we propose an efficient scheme for similarity search over encrypted data. To do so, we utilize a state-of-the-art algorithm for fast near neighbor search in high dimensional spaces called locality sensitive hashing. To ensure the confidentiality of the sensitive data, we provide a rigorous security definition and prove the security of the proposed scheme under the provided definition. In addition, we provide a real world application of the proposed scheme and verify the theoretical results with empirical observations on a real dataset.

## 58. Online Feature Selection and Its Applications.

### Synopsis:

Feature selection is an important technique for data mining. Despite its importance, most studies of feature selection are restricted to batch learning. Unlike traditional batch learning methods, online learning represents a promising family of efficient and scalable machine learning algorithms for large-scale applications. Most existing studies of online learning require accessing all the attributes/features of training instances. Such a classical setting is not always appropriate for real-world applications when data instances are of high dimensionality or it is expensive to acquire the full set of attributes/features. To address this limitation, we investigate the problem of online feature selection (OFS) in which an online learner is only allowed to maintain a classifier involved only a small and fixed number of features. The key challenge of online feature selection is how to make accurate prediction for an instance using a small number of active features. This is in contrast to the classical setup of online learning where all the features can be used for prediction. We attempt to tackle this challenge by studying sparsity regularization and truncation techniques. Specifically, this article addresses two different tasks of online feature selection: 1) learning with full input, where an learner is allowed to access all the features to decide the subset of active features, and 2) learning with partial input, where only a limited number of features is allowed to be accessed for each instance by the learner. We present novel algorithms to solve each of the two problems and give their performance analysis. We evaluate the performance of the proposed algorithms for online feature selection on several public data sets, and demonstrate their applications to real-world problems including image classification in computer vision and microarray gene expression analysis in bioinformatics. The encouraging results of our experiments validate the efficacy and efficiency of th-proposed techniques.

## 59. Dynamic Personalized Recommendation on Sparse Data.

### Synopsis:

Recommendation techniques are very important in the fields of E-commerce and other web-based services. One of the main difficulties is dynamically providing high-quality recommendation on sparse data. In this paper, a novel dynamic personalized recommendation algorithm is proposed, in which information contained in both ratings and profile contents are utilized by exploring latent relations between ratings, a set of dynamic features are designed to describe user preferences in multiple phases, and finally, a recommendation is made by adaptively weighting the features. Experimental results on public data sets show that the proposed algorithm has satisfying performance.

## 60. Enabling cross-site interactions in social networks.

### Synopsis:

Social Networks is one of the major technological phenomena on the Web 2.0. Hundreds of millions of people are posting articles, photos, and videos on their profiles and interacting with other people, but the sharing and interaction are limited within a same social network site. Although users can share some contents in a social network site with people outside of the social network site using a public link of content, appropriate access control mechanisms are still not supported. To overcome those limitations, we propose a cross-site content sharing framework named x-mngr, allowing users to interact with others in other social network sites, with a cross-site access control policy, which enables users to specify policies that allow/deny access to their shared contents across social network sites. We implemented our proposed framework through a photo sharing application that shares user's photos between Face book and My Space based on the cross-site access control policy. To evaluate our approach, we conducted a user study for the x-mngr framework.

## 61. Personalized Recommendation Combining User Interest and Social Circle.

### Synopsis:

With the advent and popularity of social network, more and more users like to share their experiences, such as ratings, reviews, and blogs. The new factors of social network like interpersonal influence and interest based on circles of friends bring opportunities and challenges for recommender system (RS) to solve the cold start and sparsity problem of datasets. Some of the social factors have been used in RS, but have not been fully considered. In this paper, three social factors, personal interest, interpersonal interest similarity, and interpersonal influence, fuse into a unified personalized recommendation model based on probabilistic matrix factorization. The factor of personal interest can make the RS recommend items to meet users' individualities, especially for experienced users. Moreover, for cold start users, the interpersonal interest similarity and interpersonal influence can enhance the intrinsic link among features in the latent space. We conduct a series of experiments on three rating datasets: Yelp, MovieLens, and Douban Movie. Experimental results show the proposed approach outperforms the existing RS approaches.

## 62. Dynamic Query Forms for Database Queries.

### Synopsis:

Modern scientific databases and web databases maintain large and heterogeneous data. These real-world databases contain hundreds or even thousands of relations and attributes. Traditional predefined query forms are not able to satisfy various ad-hoc queries from users on those databases. This paper proposes DQF, a novel database query form interface, which is able to dynamically generate query forms. The essence of DQF is to capture a user's preference and rank query form components, assisting him/her in making decisions. The generation of a query form is an iterative process and is guided by the user. At each iteration, the system automatically generates ranking lists of form components and the user then adds the desired form components into the query form. The ranking of form components is based on the captured user preference. A user can also fill the query form and submit queries to view the query result at each iteration. In this way, a query form could be dynamically refined until the user is satisfied with the query results. We utilize the expected F-measure for measuring the goodness of a query form. A probabilistic model is developed for estimating the goodness of a query form in DQF. Our experimental evaluation and user study demonstrate the effectiveness and efficiency of the system.

## 63. Enabling Multilevel Trust in Privacy Preserving Data Mining.

### Synopsis:

Privacy Preserving Data Mining (PPDM) addresses the problem of developing accurate models about aggregated data without access to precise information in individual data record. A widely studied perturbation-based PPDM approach introduces random perturbation to individual values to preserve privacy before data are published. Previous solutions of this approach are limited in their tacit assumption of single-level trust on data miners. In this work, we relax this assumption and expand the scope of perturbation-based PPDM to Multilevel Trust (MLT-PPDM). In our setting, the more trusted a data miner is, the less perturbed copy of the data it can access. Under this setting, a malicious data miner may have access to differently perturbed copies of the same data through various means, and may combine these diverse copies to jointly infer additional information about the original data that the data owner does not intend to release. Preventing such diversity attacks is the key challenge of providing MLT-PPDM services. We address this challenge by properly correlating perturbation across copies at different trust levels. We prove that our solution is robust against diversity attacks with respect to our privacy goal. That is, for data miners who have access to an arbitrary collection of the perturbed copies, our solution prevent them from jointly reconstructing the original data more accurately than the best effort using any individual copy in the collection. Our solution allows a data owner to generate perturbed copies of its data for arbitrary trust levels on-demand. This feature offers data owners maximum flexibility.

## 64. Privacy-Preserving and Content-Protecting Location Based Queries.

### Synopsis:

In this paper we present a solution to one of the location-based query problems. This problem is defined as follows: (i) a user wants to query a database of location data, known as Points Of Interest (POIs), and does not want to reveal his/her location to the server due to privacy concerns; (ii) the owner of the location data, that is, the location server, does not want to simply distribute its data to all users. The location server desires to have some control over its data, since the data is its asset. We propose a major enhancement upon previous solutions by introducing a two stage approach, where the first step is based on Oblivious Transfer and the second step is based on Private Information Retrieval, to achieve a secure solution for both parties. The solution we present is efficient and practical in many scenarios. We implement our solution on a desktop machine and a mobile device to assess the efficiency of our protocol. We also introduce a security model and analyse the security in the context of our protocol. Finally, we highlight a security weakness of our previous work and present a solution to overcome it.

## 65. Efficient Algorithms for Mining High Utility Itemsets from Transactional Databases.

### Synopsis:

Mining high utility itemsets from a transactional database refers to the discovery of itemsets with high utility like profits. Although a number of relevant algorithms have been proposed in recent years, they incur the problem of producing a large number of candidate itemsets for high utility itemsets. Such a large number of candidate itemsets degrades the mining performance in terms of execution time and space requirement. The situation may become worse when the database contains lots of long transactions or long high utility itemsets. In this paper, we propose two algorithms, namely utility pattern growth (UP-Growth) and UP-Growth+, for mining high utility itemsets with a set of effective strategies for pruning candidate itemsets. The information of high utility itemsets is maintained in a tree-based data structure named utility pattern tree (UP-Tree) such that candidate itemsets can be generated efficiently with only two scans of database. The performance of UP-Growth and UP-Growth+ is compared with the state-of-the-art algorithms on many types of both real and synthetic data sets. Experimental results show that the proposed algorithms, especially UP-Growth+, not only reduce the number of candidates effectively but also outperform other algorithms substantially in terms of runtime, especially when databases contain lots of long transactions.

## 66. A Framework for Personal Mobile Commerce Pattern Mining and Prediction.

### Synopsis:

Due to a wide range of potential applications, research on mobile commerce has received a lot of interests from both of the industry and academia. Among them, one of the active topic areas is the mining and prediction of users' mobile commerce behaviors such as their movements and purchase transactions. In this paper, we propose a novel framework, called Mobile Commerce Explorer (MCE), for mining and prediction of mobile users' movements and purchase transactions under the context of mobile commerce. The MCE framework consists of three major components: 1) Similarity Inference Model (SIM) for measuring the similarities among stores and items, which are two basic mobile commerce entities considered in this paper; 2) Personal Mobile Commerce Pattern Mine (PMCP-Mine) algorithm for efficient discovery of mobile users' Personal Mobile Commerce Patterns (PMCPs); and 3) Mobile Commerce Behavior Predictor (MCBP) for prediction of possible mobile user behaviors. To our best knowledge, this is the first work that facilitates mining and prediction of mobile users' commerce behaviors in order to recommend stores and items previously unknown to a user. We perform an extensive experimental evaluation by simulation and show that our proposals produce excellent results.

## 67. Privacy-Preserving Enhanced Collaborative Tagging.

### Synopsis:

Collaborative tagging is one of the most popular services available online, and it allows end user to loosely classify either online or offline resources based on their feedback, expressed in the form of free-text labels (i.e., tags). Although tags may not be per se sensitive information, the wide use of collaborative tagging services increases the risk of cross referencing, thereby seriously compromising user privacy. In this paper, we make a first contribution toward the development of a privacy-preserving collaborative tagging service, by showing how a specific privacy-enhancing technology, namely tag suppression, can be used to protect end-user privacy. Moreover, we analyze how our approach can affect the effectiveness of a policy-based collaborative tagging system that supports enhanced web access functionalities, like content filtering and discovery, based on preferences specified by end users.

## 68. Efficient Evaluation of SUM Queries over Probabilistic Data.

### Synopsis:

SUM queries are crucial for many applications that need to deal with uncertain data. In this paper, we are interested in the queries, called ALL_SUM, that return all possible sum values and their probabilities. In general, there is no efficient solution for the problem of evaluating ALL_SUM queries. But, for many practical applications, where aggregate values are small integers or real numbers with small precision, it is possible to develop efficient solutions. In this paper, based on a recursive approach, we propose a new solution for those applications. We implemented our solution and conducted an extensive experimental evaluation over synthetic and real-world data sets; the results show its effectiveness.

## 69. Fuzzy Order-of-Magnitude Based Link Analysis for Qualitative Alias Detection.

### Synopsis:

Alias detection has been the significant subject being extensively studied for several domain applications, especially intelligence data analysis. Many preliminary methods rely on text-based measures, which are ineffective with false descriptions of terrorists' name, date-of-birth, and address. This barrier may be overcome through link information presented in relationships among objects of interests. Several numerical link-based similarity techniques have proven effective for identifying similar objects in the Internet and publication domains. However, as a result of exceptional cases with unduly high measure, these methods usually generate inaccurate similarity descriptions. Yet, they are either computationally inefficient or ineffective for alias detection with a single-property based model. This paper presents a novel orders-of-magnitude based similarity measure that integrates multiple link properties to refine the estimation process and derive semantic-rich similarity descriptions. The approach is based on order-of-magnitude reasoning with which the theory of fuzzy set is blended to provide quantitative semantics of descriptors and their unambiguous mathematical manipulation. With such explanatory formalism, analysts can validate the generated results and partly resolve the problem of false positives. It also allows coherent interpretation and communication within a decision-making group, using this computing-with-word capability. Its performance is evaluated over a terrorism-related data set, with further generalization over publication and email data collections.

## 70. Secure KNN Query Processing in Untrusted Cloud Environments.

### Synopsis:

Mobile devices with geo-positioning capabilities (e.g., GPS) enable users to access information that is relevant to their present location. Users are interested in querying about points of interest (POI) in their physical proximity, such as restaurants, cafes, ongoing events, etc. Entities specialized in various areas of interest (e.g., certain niche directions in arts, entertainment, travel) gather large amounts of geo-tagged data that appeal to subscribed users. Such data may be sensitive due to their contents. Furthermore, keeping such information up-to-date and relevant to the users is not an easy task, so the owners of such data sets will make the data accessible only to paying customers. Users send their current location as the query parameter, and wish to receive as result the nearest POIs, i.e., nearest-neighbors (NNs). But typical data owners do not have the technical means to support processing queries on a large scale, so they outsource data storage and querying to a cloud service provider. Many such cloud providers exist who offer powerful storage and computational infrastructures at low cost. However, cloud providers are not fully trusted, and typically behave in an honest-but-curious fashion. Specifically, they follow the protocol to answer queries correctly, but they also collect the locations of the POIs and the subscribers for other purposes. Leakage of POI locations can lead to privacy breaches as well as financial losses to the data owners, for whom the POI data set is an important source of revenue. Disclosure of user locations leads to privacy violations and may deter subscribers from using the service altogether. In this paper, we propose a family of techniques that allow processing of NN queries in an untrusted outsourced environment, while at the same time protecting both the POI and querying users' positions. Our techniques rely on mutable order preserving encoding (mOPE), the only secure order-preserving encryption method known to-date. W- also provide performance optimizations to decrease the computational cost inherent to processing on encrypted data, and we consider the case of incrementally updating data sets. We present an extensive performance evaluation of our techniques to illustrate their viability in practice.

## 71. Facilitating Effective User Navigation through Website Structure Improvement.

### Synopsis:

Designing well-structured websites to facilitate effective user navigation has long been a challenge. A primary reason is that the web developers' understanding of how a website should be structured can be considerably different from that of the users. While various methods have been proposed to relink webpages to improve navigability using user navigation data, the completely reorganized new structure can be highly unpredictable, and the cost of disorienting users after the changes remains unanalyzed. This paper addresses how to improve a website without introducing substantial changes. Specifically, we propose a mathematical programming model to improve the user navigation on a website while minimizing alterations to its current structure. Results from extensive tests conducted on a publicly available real data set indicate that our model not only significantly improves the user navigation with very few changes, but also can be effectively solved. We have also tested the model on large synthetic data sets to demonstrate that it scales up very well. In addition, we define two evaluation metrics and use them to assess the performance of the improved website using the real data set. Evaluation results confirm that the user navigation on the improved structure is indeed greatly enhanced. More interestingly, we find that heavily disoriented users are more likely to benefit from the improved structure than the less disoriented users.

## 72. Geometric Data Perturbation for Privacy Preserving Outsourced Data Mining.

### Synopsis:

Data perturbation is a popular technique in privacy-preserving data mining. A major challenge in data perturbation is to balance privacy protection and data utility, which are normally considered as a pair of conflicting factors. We argue that selectively preserving the task/model specific information in perturbation will help achieve better privacy guarantee and better data utility. One type of such information is the multidimensional geometric information, which is implicitly utilized by many data-mining models. To preserve this information in data perturbation, we propose the Geometric Data Perturbation (GDP) method. In this paper, we describe several aspects of the GDP method. First, we show that several types of well-known data-mining models will deliver a comparable level of model quality over the geometrically perturbed data set as over the original data set. Second, we discuss the intuition behind the GDP method and compare it with other multidimensional perturbation methods such as random projection perturbation. Third, we propose a multi-column privacy evaluation framework for evaluating the effectiveness of geometric data perturbation with respect to different level of attacks. Finally, we use this evaluation framework to study a few attacks to geometrically perturbed data sets. Our experimental study also shows that geometric data perturbation can not only provide satisfactory privacy guarantee but also preserve modeling accuracy well.

## 73. Secure Outsourced Attribute-Based Signatures.

Synopsis:

Attribute-based signature (ABS) enables users to sign messages over attributes without revealing any information other than the fact that they have attested to the messages. However, heavy computational cost is required during signing in existing work of ABS, which grows linearly with the size of the predicate formula. As a result, this presents a significant challenge for resource-constrained devices (such as mobile devices or RFID tags) to perform such heavy computations independently. Aiming at tackling the challenge above, we first propose and formalize a new paradigm called Outsourced ABS, i.e., OABS, in which the computational overhead at user side is greatly reduced through outsourcing intensive computations to an untrusted signing-cloud service provider (S-CSP). Furthermore, we apply this novel paradigm to existing ABS schemes to reduce the complexity. As a result, we present two concrete OABS schemes: i) in the first OABS scheme, the number of exponentiations involving in signing is reduced from O(d) to O(1) (nearly three), where d is the upper bound of threshold value defined in the predicate; ii) our second scheme is built on Herranz et al.'s construction with constant-size signatures. The number of exponentiations in signing is reduced from O(d2) to O(d) and the communication overhead is O(1). Security analysis demonstrates that both OABS schemes are secure in terms of the unforgeability and attribute-signer privacy definitions specified in the proposed security model. Finally, to allow for high efficiency and flexibility, we discuss extensions of OABS and show how to achieve accountability as well.

## 74. Fairness-aware and Privacy-Preserving Friend Matching Protocol in Mobile Social Networks.

Synopsis:

Mobile social networks represent a promising cyber-physical system, which connects mobile

nodes within a local physical proximity using mobile smart phones as well as wireless communication.

In mobile social networks, the mobile users may, however, face the risk of leaking their personal information

and location privacy. In this paper, we first model the secure friend discovery process as a generalized privacypreserving

interest and profile matching problem. We identify a new security threat arising from existing

secure friend discovery protocols, coined as runaway attack, which can introduce a serious unfairness issue.

To thwart this new threat, we introduce a novel blind vector transformation technique, which could hide

the correlation between the original vector and transformed results. Based on this, we propose our privacy preserving

and fairness-aware interest and profile matching protocol, which allows one party to match its

interest with the profile of another, without revealing its real interest and profile and vice versa. The detailed

security analysis as well as real-world implementations demonstrate the effectiveness and efficiency of the

proposed protocol.

## 75. How do Facebookers use Friendlists.

### Synopsis:

Facebook friend lists are used to classify friends into groups and assist users in controlling access to their information. In this paper, we study the effectiveness of Facebook friend lists from two aspects: Friend Management and Policy Patterns by examining how users build friend lists and to what extent they use them in their policy templates. We have collected real Facebook profile information and photo privacy policies of 222 participants, through their consent in our Facebook survey application posted on Mechanical Turk. Our data analysis shows that users' customized friend lists are less frequently created and have fewer overlaps as compared to Facebook created friend lists. Also, users do not place all of their friends into lists. Moreover, friends in more than one friend lists have higher values of node betweenness and outgoing to incoming edge ratio values among all the friends of a particular user. Last but not the least, friend list and user based exceptions are less frequently used in policies as compared to allowing all friends, friends of friends and everyone to view photos.

## 76. Security Evaluation of Pattern Classifiers under Attack.

### Synopsis:

Pattern classification systems are commonly used in adversarial applications, like biometric authentication, network intrusion detection, and spam filtering, in which data can be purposely manipulated by humans to undermine their operation. As this adversarial scenario is not taken into account by classical design methods, pattern classification systems may exhibit vulnerabilities, whose exploitation may severely affect their performance, and consequently limit their practical utility. Extending pattern classification theory and design methods to adversarial settings is thus a novel and very relevant research direction, which has not yet been pursued in a systematic way. In this paper, we address one of the main open issues: evaluating at design phase the security of pattern classifiers, namely, the performance degradation under potential attacks they may incur during operation. We propose a framework for empirical evaluation of classifier security that formalizes and generalizes the main ideas proposed in the literature, and give examples of its use in three real applications. Reported results show that security evaluation can provide a more complete understanding of the classifier's behavior in adversarial environments, and lead to better design choices.

## 78. Investigation and Analysis of New Approach of Intelligent Semantic Web Search Engines

### Synopsis:

As we know that www is allowing peoples to share the huge information from big database repositories. The amount of information grows billions of databases. Hence to search particular information from these huge databases we need specialized mechanism which helps to retrive that information efficiently. now days various types of search engines are available which makes information retrieving is difficult. but to provide the better solution to this proplem ,semantic web search engines are playing vital role.basically main aim of this kind of search engines is providing the required information is small time with maximum accuracy.

## 79. Self-Adaptive Semantic Focused Crawler for Mining Services Information Discovery.

### Synopsis:

It is well recognized that the Internet has become the largest marketplace in the world, and online advertising is very popular with numerous industries, including the traditional mining service industry where mining service advertisements are effective carriers of mining service information. However, service users may encounter three major issues - heterogeneity, ubiquity, and ambiguity, when searching for mining service information over the Internet. In this paper, we present the framework of a novel self-adaptive semantic focused crawler - SASF crawler, with the purpose of precisely and efficiently discovering, formatting, and indexing mining service information over the Internet, by taking into account the three major issues. This framework incorporates the technologies of semantic focused crawling and ontology learning, in order to maintain the performance of this crawler, regardless of the variety in the Web environment. The innovations of this research lie in the design of an unsupervised framework for vocabulary-based ontology learning, and a hybrid algorithm for matching semantically relevant concepts and metadata. A series of experiments are conducted in order to evaluate the performance of this crawler. The conclusion and the direction of future work are given in the final section.

## 80. FoCUS: Learning to Crawl Web Forums.

### Synopsis:

In this paper, we present Forum Crawler Under Supervision (FoCUS), a supervised web-scale forum crawler. The goal of FoCUS is to crawl relevant forum content from the web with minimal overhead. Forum threads contain information content that is the target of forum crawlers. Although forums have different layouts or styles and are powered by different forum software packages, they always have similar implicit navigation paths connected by specific URL types to lead users from entry pages to thread pages. Based on this observation, we reduce the web forum crawling problem to a URL-type recognition problem. And we show how to learn accurate and effective regular expression patterns of implicit navigation paths from automatically created training sets using aggregated results from weak page type classifiers. Robust page type classifiers can be trained from as few as five annotated forums and applied to a large set of unseen forums. Our test results show that FoCUS achieved over 98 percent effectiveness and 97 percent coverage on a large set of test forums powered by over 150 different forum software packages. In addition, the results of applying FoCUS on more than 100 community Question and Answer sites and Blog sites demonstrated that the concept of implicit navigation path could apply to other social media sites.

## 81. Efficient Multi-dimensional Fuzzy Search for Personal Information Management Systems.

### Synopsis:

With the explosion in the amount of semistructured data users access and store in personal information management systems, there is a critical need for powerful search tools to retrieve often very heterogeneous data in a simple and efficient way. Existing tools typically support some IR-style ranking on the textual part of the query, but only consider structure (e.g., file directory) and metadata (e.g., date, file type) as filtering conditions. We propose a novel multidimensional search approach that allows users to perform fuzzy searches for structure and metadata conditions in addition to keyword conditions. Our techniques individually score each dimension and integrate the three dimension scores into a meaningful unified score. We also design indexes and algorithms to efficiently identify the most relevant files that match multidimensional queries. We perform a thorough experimental evaluation of our approach and show that our relaxation and scoring framework for fuzzy query conditions in noncontent dimensions can significantly improve ranking accuracy. We also show that our query processing strategies perform and scale well, making our fuzzy search approach practical for every day usage.

## 82. Supporting Privacy Protection in Personalized Web Search.

### Synopsis:

Personalized web search (PWS) has demonstrated its effectiveness in improving the quality of various search services on the Internet. However, evidences show that users' reluctance to disclose their private information during search has become a major barrier for the wide proliferation of PWS. We study privacy protection in PWS applications that model user preferences as hierarchical user profiles. We propose a PWS framework called UPS that can adaptively generalize profiles by queries while respecting user-specified privacy requirements. Our runtime generalization aims at striking a balance between two predictive metrics that evaluate the utility of personalization and the privacy risk of exposing the generalized profile. We present two greedy algorithms, namely GreedyDP and GreedyIL, for runtime generalization. We also provide an online prediction mechanism for deciding whether personalizing a query is beneficial. Extensive experiments demonstrate the effectiveness of our framework. The experimental results also reveal that GreedyIL significantly outperforms GreedyDP in terms of efficiency.

## 83. Improving Security and Efficiency in Attribute-Based Data Sharing.

### Synopsis:

With the recent adoption and diffusion of the data sharing paradigm in distributed systems such as online social networks or cloud computing, there have been increasing demands and concerns for distributed data security. One of the most challenging issues in data sharing systems is the enforcement of access policies and the support of policies updates. Ciphertext policy attribute-based encryption (CP-ABE) is becoming a promising cryptographic solution to this issue. It enables data owners to define their own access policies over user attributes and enforce the policies on the data to be distributed. However, the advantage comes with a major drawback which is known as a key escrow problem. The key generation center could decrypt any messages addressed to specific users by generating their private keys. This is not suitable for data sharing scenarios where the data owner would like to make their private data only accessible to designated users. In addition, applying CP-ABE in the data sharing system introduces another challenge with regard to the user revocation since the access policies are defined only over the attribute universe. Therefore, in this study, we propose a novel CP-ABE scheme for a data sharing system by exploiting the characteristic of the system architecture. The proposed scheme features the following achievements: 1) the key escrow problem could be solved by escrow-free key issuing protocol, which is constructed using the secure two-party computation between the key generation center and the data-storing center, and 2) fine-grained user revocation per each attribute could be done by proxy encryption which takes advantage of the selective attribute group key distribution on top of the ABE. The performance and security analyses indicate that the proposed scheme is efficient to securely manage the data distributed in the data sharing system.

## 84. Multiparty Access Control for Online Social Networks: Model and Mechanisms..

### Synopsis:

Online social networks (OSNs) have experienced tremendous growth in recent years and become a de facto portal for hundreds of millions of Internet users. These OSNs offer attractive means for digital social interactions and information sharing, but also raise a number of security and privacy issues. While OSNs allow users to restrict access to shared data, they currently do not provide any mechanism to enforce privacy concerns over data associated with multiple users. To this end, we propose an approach to enable the protection of shared data associated with multiple users in OSNs. We formulate an access control model to capture the essence of multiparty authorization requirements, along with a multiparty policy specification scheme and a policy enforcement mechanism. Besides, we present a logical representation of our access control model that allows us to leverage the features of existing logic solvers to perform various analysis tasks on our model. We also discuss a proof-of-concept prototype of our approach as part of an application in Facebook and provide usability study and system evaluation of our method.

## 85. Task Trail An Effective Segmentation of User Search Behavior.

### Synopsis:

In this paper, we introduce “task trail” to understand user search behaviors. We define a task to be an atomic user information need, whereas a task trail represents all user activities within that particular task, such as query reformulations, URL clicks. Previously, web search logs have been studied mainly at session or query level where users may submit several queries within one task and handle several tasks within one session. Although previous studies have addressed the problem of task identification, little is known about the advantage of using task over session or query for search applications. In this paper, we conduct extensive analyses and comparisons to evaluate the effectiveness of task trails in several search applications: determining user satisfaction, predicting user search interests, and suggesting related queries. Experiments on large scale data sets of a commercial search engine show that: (1) Task trail performs better than session and query trails in determining user satisfaction; (2) Task trail increases webpage utilities of end users comparing to session and query trails; (3) Task trails are comparable to query trails but more sensitive than session trails in measuring different ranking functions; (4) Query terms from the same task are more topically consistent to each other than query terms from different tasks; (5) Query suggestion based on task trail is a good complement of query suggestions based on session trail and click-through bipartite. The findings in this paper verify the need of extracting task trails from web search logs and enhance applications in search and recommendation systems.

## 86. Incentive Compatible Privacy-Preserving Data Analysis.

### Synopsis:

In many cases, competing parties who have private data may collaboratively conduct privacy-preserving distributed data analysis (PPDA) tasks to learn beneficial data models or analysis results. Most often, the competing parties have different incentives. Although certain PPDA techniques guarantee that nothing other than the final analysis result is revealed, it is impossible to verify whether participating parties are truthful about their private input data. Unless proper incentives are set, current PPDA techniques cannot prevent participating parties from modifying their private inputs.incentive compatible privacy-preserving data analysis techniques This raises the question of how to design incentive compatible privacy-preserving data analysis techniques that motivate participating parties to provide truthful inputs. In this paper, we first develop key theorems, then base on these theorems, we analyze certain important privacy-preserving data analysis tasks that could be conducted in a way that telling the truth is the best choice for any participating party.

## 87. On the Spectral Characterization and Scalable Mining of Network Communities.

### Synopsis:

Network communities refer to groups of vertices within which their connecting links are dense but between which they are sparse. A network community mining problem (or NCMP for short) is concerned with the problem of finding all such communities from a given network. A wide variety of applications can be formulated as NCMPs, ranging from social and/or biological network analysis to web mining and searching. So far, many algorithms addressing NCMPs have been developed and most of them fall into the categories of either optimization based or heuristic methods. Distinct from the existing studies, the work presented in this paper explores the notion of network communities and their properties based on the dynamics of a stochastic model naturally introduced. In the paper, a relationship between the hierarchical community structure of a network and the local mixing properties of such a stochastic model has been established with the large-deviation theory. Topological information regarding to the community structures hidden in networks can be inferred from their spectral signatures. Based on the above-mentioned relationship, this work proposes a general framework for characterizing, analyzing, and mining network communities. Utilizing the two basic properties of metastability, i.e., being locally uniform and temporarily fixed, an efficient implementation of the framework, called the LM algorithm, has been developed that can scalably mine communities hidden in large-scale networks. The effectiveness and efficiency of the LM algorithm have been theoretically analyzed as well as experimentally validated.

## 88. Towards Differential Query Services in Cost-Efficient Clouds.

### Synopsis:

Cloud computing as an emerging technology trend is expected to reshape the advances in information technology. In a cost-efficient cloud environment, a user can tolerate a certain degree of delay while retrieving information from the cloud to reduce costs. In this paper, we address two fundamental issues in such an environment: privacy and efficiency. We first review a private keyword-based file retrieval scheme that was originally proposed by Ostrovsky. Their scheme allows a user to retrieve files of interest from an untrusted server without leaking any information. The main drawback is that it will cause a heavy querying overhead incurred on the cloud and thus goes against the original intention of cost efficiency. In this paper, we present three efficient information retrieval for ranked query (EIRQ) schemes to reduce querying overhead incurred on the cloud. In EIRQ, queries are classified into multiple ranks, where a higher ranked query can retrieve a higher percentage of matched files. A user can retrieve files on demand by choosing queries of different ranks. This feature is useful when there are a large number of matched files, but the user only needs a small subset of them. Under different parameter settings, extensive evaluations have been conducted on both analytical models and on a real cloud environment, in order to examine the effectiveness of our schemes.

## 89. Learning Regularized, Query-dependent Bilinear Similarities for Large Scale Image Retrieval.

### Synopsis:

An effective way to improve the quality of image retrieval is by employing a query-dependent similarity measure. However, implementing this in a large scale system is non-trivial because we want neither hurting the efficiency nor relying on too many training samples. In this paper, we introduce a query-dependent bilinear similarity measure to address the first issue. Based on our bilinear similarity model, query adaptation can be achieved by simply applying any existing efficient indexing/retrieval method to a transformed version (surrogate) of a query. To address the issue of limited training samples, we further propose a novel angular regularization constraint for learning the similarity measure. The learning is formulated as a Quadratic Programming (QP) problem and can be solved efficiently by a SMO-type algorithm. Experiments on two public datasets and our 1-million web-image dataset validate that our proposed method can consistently bring improvements and the whole solution is practical in large scale applications.

## 90. Organizing User Search Histories.

### Synopsis:

Users are increasingly pursuing complex task-oriented goals on the web, such as making travel arrangements, managing finances, or planning purchases. To this end, they usually break down the tasks into a few codependent steps and issue multiple queries around these steps repeatedly over long periods of time. To better support users in their long-term information quests on the web, search engines keep track of their queries and clicks while searching online. In this paper, we study the problem of organizing a user's historical queries into groups in a dynamic and automated fashion. Automatically identifying query groups is helpful for a number of different search engine components and applications, such as query suggestions, result ranking, query alterations, sessionization, and collaborative search. In our approach, we go beyond approaches that rely on textual similarity or time thresholds, and we propose a more robust approach that leverages search query logs. We experimentally study the performance of different techniques, and showcase their potential, especially when combined together.

## 91. XSPath: Navigation on XML Schemas Made Easy.

### Synopsis:

Schemas are often used to constrain the content and structure of XML documents. They can be quite big and complex and, thus, difficult to be accessed manually. The ability to query a single schema, a collection of schemas or to retrieve schema components that meet certain structural constraints significantly eases schema management and is, thus, useful in many contexts. In this paper, we propose a query language, named XSPath, specifically tailored for XML schema that works on logical graph-based representations of schemas, on which it enables the navigation, and allows the selection of nodes. We also propose XPath/XQuery-based translations that can be exploited for the evaluation of XSPath queries. An extensive evaluation of the usability and efficiency of the proposed approach is finally presented within the EXup system.

## 92. Mining User Queries with Markov Chains: Application to Online Image Retrieval..

### Synopsis:

We propose a novel method for automatic annotation, indexing and annotation-based retrieval of images. The new method, that we call Markovian Semantic Indexing (MSI), is presented in the context of an online image retrieval system. Assuming such a system, the users' queries are used to construct an Aggregate Markov Chain (AMC) through which the relevance between the keywords seen by the system is defined. The users' queries are also used to automatically annotate the images. A stochastic distance between images, based on their annotation and the keyword relevance captured in the AMC, is then introduced. Geometric interpretations of the proposed distance are provided and its relation to a clustering in the keyword space is investigated. By means of a new measure of Markovian state similarity, the mean first cross passage time (CPT), optimality properties of the proposed distance are proved. Images are modeled as points in a vector space and their similarity is measured with MSI. The new method is shown to possess certain theoretical advantages and also to achieve better Precision versus Recall results when compared to

Latent Semantic Indexing (LSI) and probabilistic Latent Semantic Indexing (pLSI) methods in Annotation-Based Image Retrieval (ABIR) tasks.

## 93. Ranking Model Adaptation for Domain-Specific Search.

### Synopsis:

With the explosive emergence of vertical search domains, applying the broad-based ranking model directly to different domains is no longer desirable due to domain differences, while building a unique ranking model for each domain is both laborious for labeling data and time consuming for training models. In this paper, we address these difficulties by proposing a regularization-based algorithm called ranking adaptation SVM (RA-SVM), through which we can adapt an existing ranking model to a new domain, so that the amount of labeled data and the training cost is reduced while the performance is still guaranteed. Our algorithm only requires the prediction from the existing ranking models, rather than their internal representations or the data from auxiliary domains. In addition, we assume that documents similar in the domain-specific feature space should have consistent rankings, and add some constraints to control the margin and slack variables of RA-SVM adaptively. Finally, ranking adaptability measurement is proposed to quantitatively estimate if an existing ranking model can be adapted to a new domain. Experiments performed over Letor and two large scale data sets crawled from a commercial search engine demonstrate the applicabilities of the proposed ranking adaptation algorithms and the ranking adaptability measurement.

## 94. m-Privacy for Collaborative Data Publishing.

### Synopsis:

We propose a novel method for automatic annotation, indexing and annotation-based retrieval of images. The new method, that we call Markovian Semantic Indexing (MSI), is presented in the context of an online image retrieval system. Assuming such a system, the users' queries are used to construct an Aggregate Markov Chain (AMC) through which the relevance between the keywords seen by the system is defined. The users' queries are also used to automatically annotate the images. A stochastic distance between images, based on their annotation and the keyword relevance captured in the AMC, is then introduced. Geometric interpretations of the proposed distance are provided and its relation to a clustering in the keyword space is investigated. By means of a new measure of Markovian state similarity, the mean first cross passage time (CPT), optimality properties of the proposed distance are proved. Images are modeled as points in a vector space and their similarity is measured with MSI. The new method is shown to possess certain theoretical advantages and also to achieve better Precision versus Recall results when compared to Latent Semantic Indexing (LSI) and probabilistic Latent Semantic Indexing (pLSI) methods in Annotation-Based Image Retrieval (ABIR) tasks.

## 94. Secure Outsourced Attribute-Based Signatures.

### Synopsis:

We propose a novel method for automatic annotation, indexing and annotation-based retrieval of images. The new method, that we call Markovian Semantic Indexing (MSI), is presented in the context of an online image retrieval system. Assuming such a system, the users' queries are used to construct an Aggregate Markov Chain (AMC) through which the relevance between the keywords seen by the system is defined. The users' queries are also used to automatically annotate the images. A stochastic distance between images, based on their annotation and the keyword relevance captured in the AMC, is then introduced. Geometric interpretations of the proposed distance are provided and its relation to a clustering in the keyword space is investigated. By means of a new measure of Markovian state similarity, the mean first cross passage time (CPT), optimality properties of the proposed distance are proved. Images are modeled as points in a vector space and their similarity is measured with MSI. The new method is shown to possess certain theoretical advantages and also to achieve better Precision versus Recall results when compared to Latent Semantic Indexing (LSI) and probabilistic Latent Semantic Indexing (pLSI) methods in Annotation-Based Image Retrieval (ABIR) tasks.

## 95. Resilient Identity Crime Detection.

### Synopsis:

Identity crime is well known, prevalent, and costly; and credit application fraud is a specific case of identity crime. The existing nondata mining detection system of business rules and scorecards, and known fraud matching have limitations. To address these limitations and combat identity crime in real time, this paper proposes a new multilayered detection system complemented with two additional layers: communal detection (CD) and spike detection (SD). CD finds real social relationships to reduce the suspicion score, and is tamper resistant to synthetic social relationships. It is the whitelist-oriented approach on a fixed set of attributes. SD finds spikes in duplicates to increase the suspicion score, and is probe-resistant for attributes. It is the attribute-oriented approach on a variable-size set of attributes. Together, CD and SD can detect more types of attacks, better account for changing legal behavior, and remove the redundant attributes. Experiments were carried out on CD and SD with several million real credit applications. Results on the data support the hypothesis that successful credit application fraud patterns are sudden and exhibit sharp spikes in duplicates. Although this research is specific to credit application fraud detection, the concept of resilience, together with adaptivity and quality data discussed in the paper, are general to the design, implementation, and evaluation of all detection systems.

## 96. Online Search and Buying Behaviour in Consumer Markets.

### Synopsis:

Online search behaviour is analysed using a novel methodology based on an international panel of two million users. Consumer search is measured by the size and distribution of online consideration sets and the use of price comparison engines in a range of US and UK consumer markets. It is shown that most online researchers who are considering competing suppliers only view two or three competitor websites, which results in an average online consideration set of between 2.1 and 2.8, regardless of the use of price comparison websites. Consumer perceived risk is negatively correlated with the size of online consideration sets and online price competition intensity. Using international data from fifteen countries it is shown that online research and online purchasing are negatively correlated with shop density. The implications for managers are outlined, in particular the importance of branding and advertising to improve the likelihood of inclusion in online consideration sets.

## 97. Sequential Anomaly Detection in the Presence of Noise and Limited Feedback.

### Synopsis:

This paper describes a methodology for detecting anomalies from sequentially observed and potentially noisy data. The proposed approach consists of two main elements: 1) filtering, or assigning a belief or likelihood to each successive measurement based upon our ability to predict it from previous noisy observations and 2) hedging, or flagging potential anomalies by comparing the current belief against a time-varying and data-adaptive threshold. The threshold is adjusted based on the available feedback from an end user. Our algorithms, which combine universal prediction with recent work on online convex programming, do not require computing posterior distributions given all current observations and involve simple primal-dual parameter updates. At the heart of the proposed approach lie exponential-family models which can be used in a wide variety of contexts and applications, and which yield methods that achieve sublinear per-round regret against both static and slowly varying product distributions with marginals drawn from the same exponential family. Moreover, the regret against static distributions coincides with the minimax value of the corresponding online strongly convex game. We also prove bounds on the number of mistakes made during the hedging step relative to the best offline choice of the threshold with access to all estimated beliefs and feedback signals. We validate the theory on synthetic data drawn from a time-varying distribution over binary vectors of high dimensionality, as well as on the Enron email dataset.

## 98. Optimal Client-Server Assignment for Internet Distributed Systems.

### Synopsis:

We investigate an underlying mathematical model and algorithms for optimizing the performance of a class of distributed systems over the Internet. Such a system consists of a large number of clients who communicate with each other indirectly via a number of intermediate servers. Optimizing the overall performance of such a system then can be formulated as a client-server assignment problem whose aim is to assign the clients to the servers in such a way to satisfy some prespecified requirements on the communication cost and load balancing. We show that 1) the total communication load and load balancing are two opposing metrics, and consequently, their tradeoff is inherent in this class of distributed systems; 2) in general, finding the optimal client-server assignment for some prespecified requirements on the total load and load balancing is NP-hard, and therefore; 3) we propose a heuristic via relaxed convex optimization for finding the approximate solution. Our simulation results indicate that the proposed algorithm produces superior performance than other heuristics, including the popular Normalized Cuts algorithm.

## 99. Slicing: A New Approach to Privacy Preserving Data Publishing.

### Synopsis:

Several anonymization techniques, such as generalization and bucketization, have been designed for privacy preserving microdata publishing. Recent work has shown that generalization loses considerable amount of information, especially for high-dimensional data. Bucketization, on the other hand, does not prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. In this paper, we present a novel technique called slicing, which partitions the data both horizontally and vertically. We show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. We show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ℓ-diversity requirement. Our workload experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. Our experiments also demonstrate that slicing can be used to prevent membership disclosure.

## 100. Personalized QoS-Aware Web Service Recommendation and Visualization.

### Synopsis:

With the proliferation of web services, effective QoS-based approach to service recommendation is becoming more and more important. Although service recommendation has been studied in the recent literature, the performance of existing ones is not satisfactory, since (1) previous approaches fail to consider the QoS variance according to users' locations; and (2) previous recommender systems are all black boxes providing limited information on the performance of the service candidates. In this paper, we propose a novel collaborative filtering algorithm designed for large-scale web service recommendation. Different from previous work, our approach employs the characteristic of QoS and achieves considerable improvement on the recommendation accuracy. To help service users better understand the rationale of the recommendation and remove some of the mystery, we use a recommendation visualization technique to show how a recommendation is grouped with other choices. Comprehensive experiments are conducted using more than 1.5 million QoS records of real-world web service invocations. The experimental results show the efficiency and effectiveness of our approach.

## 101. Statistical Entity Extraction from Web .

### Synopsis:

There are various kinds of valuable semantic information about real-world entities embedded in webpages and databases. Extracting and integrating these entity information from the Web is of great significance. Comparing to traditional information extraction problems, web entity extraction needs to solve several new challenges to fully take advantage of the unique characteristic of the Web. In this paper, we introduce our recent work on statistical extraction of structured entities, named entities, entity facts and relations from Web. We also briefly introduce iKnoweb, an interactive knowledge mining framework for entity information integration. We will use two novel web applications, Microsoft Academic Search (aka Libra) and EntityCube, as working examples.

## 102. PMSE -A Personalized Mobile Search Engine.

### Synopsis:

We propose a personalized mobile search engine (PMSE) that captures the users' preferences in the form of concepts by mining their clickthrough data. Due to the importance of location information in mobile search, PMSE classifies these concepts into content concepts and location concepts. In addition, users' locations (positioned by GPS) are used to supplement the location concepts in PMSE. The user preferences are organized in an ontology-based, multifacet user profile, which are used to adapt a personalized ranking function for rank adaptation of future search results. To characterize the diversity of the concepts associated with a query and their relevances to the user's need, four entropies are introduced to balance the weights between the content and location facets. Based on the client-server model, we also present a detailed architecture and design for implementation of PMSE. In our design, the client collects and stores locally the clickthrough data to protect privacy, whereas heavy tasks such as concept extraction, training, and reranking are performed at the PMSE server. Moreover, we address the privacy issue by restricting the information in the user profile exposed to the PMSE server with two privacy parameters. We prototype PMSE on the Google Android platform. Experimental results show that PMSE significantly improves the precision comparing to the baseline.

## 103. Toward Private Joins on Outsourced Data..

### Synopsis:

In an outsourced database framework, clients place data management responsibilities with specialized service providers. Of essential concern in such frameworks is data privacy. Potential clients are reluctant to outsource sensitive data to a foreign party without strong privacy assurances beyond policy &#x201C;fine prints.&#x201D; In this paper, we introduce a mechanism for executing general binary JOIN operations (for predicates that satisfy certain properties) in an outsourced relational database framework with computational privacy and low overhead&#x2014;the first, to the best of our knowledge. We illustrate via a set of relevant instances of JOIN predicates, including: range and equality (e.g., for geographical data), Hamming distance (e.g., for DNA matching), and semantics (i.e., in health-care scenarios&#x2014;mapping antibiotics to bacteria). We experimentally evaluate the main overhead components and show they are reasonable. The initial client computation overhead for 100,000 data items is around 5 minutes and our privacy mechanisms can sustain theoretical throughputs of several million predicate evaluations per second, even for an unoptimized OpenSSL-based implementation.

## 104. Preventing Private Information Inference Attacks on Social Networks.

### Synopsis:

Online social networks, such as Facebook, are increasingly utilized by many people. These networks allow users to publish details about themselves and to connect to their friends. Some of the information revealed inside these networks is meant to be private. Yet it is possible to use learning algorithms on released data to predict private information. In this paper, we explore how to launch inference attacks using released social networking data to predict private information. We then devise three possible sanitization techniques that could be used in various situations. Then, we explore the effectiveness of these techniques and attempt to use methods of collective inference to discover sensitive attributes of the data set. We show that we can decrease the effectiveness of both local and relational classification algorithms by using the sanitization methods we described.

## 105. Privacy against Aggregate Knowledge Attacks.

### Synopsis:

This paper focuses on protecting the privacy of individuals in publication scenarios where the attacker is expected to have only abstract or aggregate knowledge about each record. Whereas, data privacy research usually focuses on defining stricter privacy guarantees that assume increasingly more sophisticated attack scenarios, it is also important to have anonymization methods and guarantees that will address any attack scenario. Enforcing a stricter guarantee than required increases unnecessarily the information loss. Consider for example the publication of tax records, where attackers might only know the total income, and not its constituent parts. Traditional anonymization methods would protect user privacy by creating equivalence classes of identical records. Alternatively, in this work we propose an anonymization technique that generalizes attributes, only as much as needed to guarantee that aggregate values over the complete record, will create equivalence classes of at size k. The experimental evaluation on real data shows that the proposed method produces anonymized data that lie closer to the original ones, with respect to traditional anonymization algorithms.

## 106. Privacy-preserving Mining of Association Rules from Outsourced Transaction Databases.

### Synopsis:

Spurred by developments such as cloud computing, there has been considerable recent interest in the paradigm of data mining-as-a-service. A company (data owner) lacking in expertise or computational resources can outsource its mining needs to a third party service provider (server). However, both the items and the association rules of the outsourced database are considered private property of the corporation (data owner). To protect corporate privacy, the data owner transforms its data and ships it to the server, sends mining queries to the server, and recovers the true patterns from the extracted patterns received from the server. In this paper, we study the problem of outsourcing the association rule mining task within a corporate privacy-preserving framework. We propose an attack model based on background knowledge and devise a scheme for privacy preserving outsourced mining. Our scheme ensures that each transformed item is indistinguishable with respect to the attacker's background knowledge, from at least k-1 other transformed items. Our comprehensive experiments on a very large and real transaction database demonstrate that our techniques are effective, scalable, and protect privacy.

## 107. Ranking on Data Manifold with Sink Points.

### Synopsis:

Ranking is an important problem in various applications, such as Information Retrieval (IR), natural language processing, computational biology, and social sciences. Many ranking approaches have been proposed to rank objects according to their degrees of relevance or importance. Beyond these two goals, diversity has also been recognized as a crucial criterion in ranking. Top ranked results are expected to convey as little redundant information as possible, and cover as many aspects as possible. However, existing ranking approaches either take no account of diversity, or handle it separately with some heuristics. In this paper, we introduce a novel approach, Manifold Ranking with Sink Points (MRSPs), to address diversity as well as relevance and importance in ranking. Specifically, our approach uses a manifold ranking process over the data manifold, which can naturally find the most relevant and important data objects. Meanwhile, by turning ranked objects into sink points on data manifold, we can effectively prevent redundant objects from receiving a high rank. MRSP not only shows a nice convergence property, but also has an interesting and satisfying optimization explanation. We applied MRSP on two application tasks, update summarization and query recommendation, where diversity is of great concern in ranking. Experimental results on both tasks present a strong empirical performance of MRSP as compared to existing ranking approaches.

## 108. Robust Module-based Data Management.

### Synopsis:

The current trend for building an ontology-based data management system (DMS) is to capitalize on efforts made to design a preexisting well-established DMS (a reference system). The method amounts to extracting from the reference DMS a piece of schema relevant to the new application needs-a module-, possibly personalizing it with extra constraints w.r.t. the application under construction, and then managing a data set using the resulting schema. In this paper, we extend the existing definitions of modules and we introduce novel properties of robustness that provide means for checking easily that a robust module-based DMS evolves safely w.r.t. both the schema and the data of the reference DMS. We carry out our investigations in the setting of description logics which underlie modern ontology languages, like RDFS, OWL, and OWL2 from W3C. Notably, we focus on the DL-liteA dialect of the DL-lite family, which encompasses the foundations of the QL profile of OWL2 (i.e., DL-liteR): the W3C recommendation for efficiently managing large data sets.

## 109. Secure Mining of Association Rules in Horizontally Distributed Databases.

### Synopsis:

We propose a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Kantarcioglu and Clifton . Our protocol, like theirs, is based on the Fast Distributed Mining (FDM)algorithm of Cheung et al. , which is an unsecured distributed version of the Apriori algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms-one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy with respect to the protocol in . In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

## 110. Sensitive Label Privacy Protection on Social Network Data.

### Synopsis:

Privacy is one of the major concerns when publishing or sharing social network data for social science research and business analysis. Recently, researchers have developed privacy models similar to k-anonymity to prevent node reidentification through structure information. However, even when these privacy models are enforced, an attacker may still be able to infer one's private information if a group of nodes largely share the same sensitive labels (i.e., attributes). In other words, the label-node relationship is not well protected by pure structure anonymization methods. Furthermore, existing approaches, which rely on edge editing or node clustering, may significantly alter key graph properties. In this paper, we define a k-degree-l-diversity anonymity model that considers the protection of structural information as well as sensitive labels of individuals. We further propose a novel anonymization methodology based on adding noise nodes. We develop a new algorithm by adding noise nodes into the original graph with the consideration of introducing the least distortion to graph properties. Most importantly, we provide a rigorous analysis of the theoretical bounds on the number of noise nodes added and their impacts on an important graph property. We conduct extensive experiments to evaluate the effectiveness of the proposed technique.

## 111. Spatial Approximate String Search.

### Synopsis:

This work deals with the approximate string search in large spatial databases. Specifically, we investigate range queries augmented with a string similarity search predicate in both euclidean space and road networks. We dub this query the spatial approximate string (SAS) query. In euclidean space, we propose an approximate solution, the MHR-tree, which embeds min-wise signatures into an R-tree. The min-wise signature for an index node u keeps a concise representation of the union of q-grams from strings under the subtree of u. We analyze the pruning functionality of such signatures based on the set resemblance between the query string and the q-grams from the subtrees of index nodes. We also discuss how to estimate the selectivity of a SAS query in euclidean space, for which we present a novel adaptive algorithm to find balanced partitions using both the spatial and string information stored in the tree. For queries on road networks, we propose a novel exact method, RSASSOL, which significantly outperforms the baseline algorithm in practice. The RSASSOL combines the q-gram-based inverted lists and the reference nodes based pruning. Extensive experiments on large real data sets demonstrate the efficiency and effectiveness of our approaches.

## 112. Spatial Query Integrity with Voronoi Neighbors.

### Synopsis:

With the popularity of location-based services and the abundant usage of smart phones and GPS-enabled devices, the necessity of outsourcing spatial data has grown rapidly over the past few years. Meanwhile, the fast arising trend of cloud storage and cloud computing services has provided a flexible and cost-effective platform for hosting data from businesses and individuals, further enabling many location-based applications. Nevertheless, in this database outsourcing paradigm, the authentication of the query results at the client remains a challenging problem. In this paper, we focus on the Outsourced Spatial Database (OSDB) model and propose an efficient scheme, called VN-Auth, which allows a client to verify the correctness and completeness of the result set. Our approach is based on neighborhood information derived from the Voronoi diagram of the underlying spatial data set and can handle fundamental spatial query types, such as k nearest neighbor and range queries, as well as more advanced query types like reverse k nearest neighbor, aggregate nearest neighbor, and spatial skyline. We evaluated VN-Auth based on real-world data sets using mobile devices (Google Droid smart phones with Android OS) as query clients. Compared to the current state-of-the-art approaches (i.e., methods based on Merkle Hash Trees), our experiments show that VN-Auth produces significantly smaller verification objects and is more computationally efficient, especially for queries with low selectivity.

## 113. SybilDefender Defend Against Sybil Attacks in Large Social Networks.

### Synopsis:

Distributed systems without trusted identities are particularly vulnerable to sybil attacks, where an adversary creates multiple bogus identities to compromise the running of the system. This paper presents SybilDefender, a sybil defense mechanism that leverages the network topologies to defend against sybil attacks in social networks. Based on performing a limited number of random walks within the social graphs, SybilDefender is efficient and scalable to large social networks. Our experiments on two 3,000,000 node real-world social topologies show that SybilDefender outperforms the state of the art by one to two orders of magnitude in both accuracy and running time. SybilDefender can effectively identify the sybil nodes and detect the sybil community around a sybil node, even when the number of sybil nodes introduced by each attack edge is close to the theoretically detectable lower bound. Besides, we propose two approaches to limiting the number of attack edges in online social networks. The survey results of our Facebook application show that the assumption made by previous work that all the relationships in social networks are trusted does not apply to online social networks, and it is feasible to limit the number of attack edges in online social networks by relationship rating.

## 114. User Action Interpretation for Online Content Optimization.

### Synopsis:

Web portal services have become an important medium to deliver digital content and service, such as news, advertisements, and so on, to Web users in a timely fashion. To attract more users to various content modules on the Web portal, it is necessary to design a recommender system that can effectively achieve online content optimization by automatically estimating content items' attractiveness and relevance to users' interests. User interaction plays a vital role in building effective content optimization, as both implicit user feedbacks and explicit user ratings on the recommended items form the basis for designing and learning recommendation models. However, user actions on real-world Web portal services are likely to represent many implicit signals about users' interests and content attractiveness, which need more accurate interpretation to be fully leveraged in the recommendation models. To address this challenge, we investigate a couple of critical aspects of the online learning framework for personalized content optimization on Web portal services, and, in this paper, we propose deeper user action interpretation to enhance those critical aspects. In particular, we first propose an approach to leverage historical user activity to build behavior-driven user segmentation; then, we introduce an approach for interpreting users' actions from the factors of both user engagement and position bias to achieve unbiased estimation of content attractiveness. Our experiments on the large-scale data from a commercial Web recommender system demonstrate that recommendation models with our user action interpretation can reach significant improvement in terms of online content optimization over the baseline method. The effectiveness of our user action interpretation is also proved by the online test results on real user traffic.

## 115. A Cocktail Approach for Travel Package Recommendation.

### Synopsis:

Recent years have witnessed an increased interest in recommender systems. Despite significant progress in this field, there still remain numerous avenues to explore. Indeed, this paper provides a study of exploiting online travel information for personalized travel package recommendation. A critical challenge along this line is to address the unique characteristics of travel data, which distinguish travel packages from traditional items for recommendation. To that end, in this paper, we first analyze the characteristics of the existing travel packages and develop a tourist-area-season topic (TAST) model. This TAST model can represent travel packages and tourists by different topic distributions, where the topic extraction is conditioned on both the tourists and the intrinsic features (i.e., locations, travel seasons) of the landscapes. Then, based on this topic model representation, we propose a cocktail approach to generate the lists for personalized travel package recommendation. Furthermore, we extend the TAST model to the tourist-relation-area-season topic (TRAST) model for capturing the latent relationships among the tourists in each travel group. Finally, we evaluate the TAST model, the TRAST model, and the cocktail recommendation approach on the real-world travel package data. Experimental results show that the TAST model can effectively capture the unique characteristics of the travel data and the cocktail approach is, thus, much more effective than traditional recommendation techniques for travel package recommendation. Also, by considering tourist relationships, the TRAST model can be used as an effective assessment for travel group formation.

## 116. A Decentralized Privacy Preserving Reputation Protocol for the Malicious Adversarial Model.

### Synopsis:

Users hesitate to submit negative feedback in reputation systems due to the fear of retaliation from the recipient user. A privacy preserving reputation protocol protects users by hiding their individual feedback and revealing only the reputation score. We present a privacy preserving reputation protocol for the malicious adversarial model. The malicious users in this model actively attempt to learn the private feedback values of honest users as well as to disrupt the protocol. Our protocol does not require centralized entities, trusted third parties, or specialized platforms, such as anonymous networks and trusted hardware. Moreover, our protocol is efficient. It requires an exchange of messages, where and are the number of users in the protocol and the environment, respectively.

## 117. A Query Formulation Language for the data web.

### Synopsis:

We present a query formulation language (called MashQL) in order to easily query and fuse structured data on the web. The main novelty of MashQL is that it allows people with limited IT skills to explore and query one (or multiple) data sources without prior knowledge about the schema, structure, vocabulary, or any technical details of these sources. More importantly, to be robust and cover most cases in practice, we do not assume that a data source should have - an offline or inline - schema. This poses several language-design and performance complexities that we fundamentally tackle. To illustrate the query formulation power of MashQL, and without loss of generality, we chose the Data web scenario. We also chose querying RDF, as it is the most primitive data model; hence, MashQL can be similarly used for querying relational databases and XML. We present two implementations of MashQL, an online mashup editor, and a Firefox add on. The former illustrates how MashQL can be used to query and mash up the Data web as simple as filtering and piping web feeds; and the Firefox add on illustrates using the browser as a web composer rather than only a navigator. To end, we evaluate MashQL on querying two data sets, DBLP and DBPedia, and show that our indexing techniques allow instant user interaction.

## 118. A Dual Framework and Algorithms for Targeted Data Delivery.

### Synopsis:

A variety of emerging online data delivery applications challenge existing techniques for data delivery to human users, applications, or middleware that are accessing data from multiple autonomous servers. In this paper, we develop a framework for formalizing and comparing pull-based solutions and present dual optimization approaches. The first approach, most commonly used nowadays, maximizes user utility under the strict setting of meeting a priori constraints on the usage of system resources. We present an alternative and more flexible approach that maximizes user utility by satisfying all users. It does this while minimizing the usage of system resources. We discuss the benefits of this latter approach and develop an adaptive monitoring solution Satisfy User Profiles (SUPs).

Through formal analysis, we identify sufficient optimality conditions for SUP. Using real (RSS feeds) and synthetic traces, we empirically analyze the behavior of SUP under varying conditions. Our experiments show that we can achieve a high degree of satisfaction of user utility when the estimations of SUP closely estimate the real event stream, and has the potential to save a significant amount of system resources. We further show that SUP can exploit feedback to improve user utility with only a moderate increase in resource utilization.

## 119. An Efficient Certificateless Encryption for Secure Data Sharing in Public Clouds..

### Synopsis:

We propose a mediated certificateless encryption scheme without pairing operations for securely sharing sensitive information in public clouds. Mediated certificateless public key encryption (mCL-PKE) solves the key escrow problem in identity based encryption and certificate revocation problem in public key cryptography. However, existing mCL-PKE schemes are either inefficient because of the use of expensive pairing operations or vulnerable against partial decryption attacks. In order to address the performance and security issues, in this paper, we first propose a mCL-PKE scheme without using pairing operations. We apply our mCL-PKE scheme to construct a practical solution to the problem of sharing sensitive information in public clouds. The cloud is employed as a secure storage as well as a key generation center. In our system, the data owner encrypts the sensitive data using the cloud generated users' public keys based on its access control policies and uploads the encrypted data to the cloud. Upon successful authorization, the cloud partially decrypts the encrypted data for the users. The users subsequently fully decrypt the partially decrypted data using their private keys. The confidentiality of the content and the keys is preserved with respect to the cloud, because the cloud cannot fully decrypt the information. We also propose an extension to the above approach to improve the efficiency of encryption at the data owner. We implement our mCL-PKE scheme and the overall cloud based system, and evaluate its security and performance. Our results show that our schemes are efficient and practical.

## 120. Achieving Data Privacy through Secrecy Views and Null-Based Virtual Updates.

### Synopsis:

We may want to keep sensitive information in a relational database hidden from a user or group thereof. We characterize sensitive data as the extensions of secrecy views. The database, before returning the answers to a query posed by a restricted user, is updated to make the secrecy views empty or a single tuple with null values. Then, a query about any of those views returns no meaningful information. Since the database is not supposed to be physically changed for this purpose, the updates are only virtual, and also minimal. Minimality makes sure that query answers, while being privacy preserving, are also maximally informative. The virtual updates are based on null values as used in the SQL standard. We provide the semantics of secrecy views, virtual updates, and secret answers (SAs) to queries. The different instances resulting from the virtually updates are specified as the models of a logic program with stable model semantics, which becomes the basis for computation of the SAs.

## 121. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication.

### Synopsis:

Record linkage is the process of matching records from several databases that refer to the same entities. When applied on a single database, this process is known as deduplication. Increasingly, matched data are becoming important in many application areas, because they can contain information that is not available otherwise, or that is too costly to acquire. Removing duplicate records in a single database is a crucial step in the data cleaning process, because duplicates can severely influence the outcomes of any subsequent data processing or data mining. With the increasing size of today's databases, the complexity of the matching process becomes one of the major challenges for record linkage and deduplication. In recent years, various indexing techniques have been developed for record linkage and deduplication. They are aimed at reducing the number of record pairs to be compared in the matching process by removing obvious nonmatching pairs, while at the same time maintaining high matching quality. This paper presents a survey of 12 variations of 6 indexing techniques. Their complexity is analyzed, and their performance and scalability is evaluated within an experimental framework using both synthetic and real data sets. No such detailed survey has so far been published.

## 122. A Link Analysis Extension of Correspondence Analysis for Mining Relational Databases.

### Synopsis:

This work introduces a link analysis procedure for discovering relationships in a relational database or a graph, generalizing both simple and multiple correspondence analysis. It is based on a random walk model through the database defining a Markov chain having as many states as elements in the database. Suppose we are interested in analyzing the relationships between some elements (or records) contained in two different tables of the relational database. To this end, in a first step, a reduced, much smaller, Markov chain containing only the elements of interest and preserving the main characteristics of the initial chain, is extracted by stochastic complementation. This reduced chain is then analyzed by projecting jointly the elements of interest in the diffusion map subspace and visualizing the results. This two-step procedure reduces to simple correspondence analysis when only two tables are defined, and to multiple correspondence analysis when the database takes the form of a simple star-schema. On the other hand, a kernel version of the diffusion map distance, generalizing the basic diffusion map distance to directed graphs, is also introduced and the links with spectral clustering are discussed. Several data sets are analyzed by using the proposed methodology, showing the usefulness of the technique for extracting relationships in relational databases or graphs.

## 123. An Empirical Performance Evaluation of Relational Keyword Search Systems.

### Synopsis:

Extending the keyword search paradigm to relational data has been an active area of research within the database and IR community during the past decade. Many approaches have been proposed, but despite numerous publications, there remains a severe lack of standardization for the evaluation of proposed search techniques. Lack of standardization has resulted in contradictory results from different evaluations, and the numerous discrepancies muddle what advantages are proffered by different approaches. In this paper, we present the most extensive empirical performance evaluation of relational keyword search techniques to appear to date in the literature. Our results indicate that many existing search techniques do not provide acceptable performance for realistic retrieval tasks. In particular, memory consumption precludes many search techniques from scaling beyond small data sets with tens of thousands of vertices. We also explore the relationship between execution time and factors varied in previous evaluations; our analysis indicates that most of these factors have relatively little impact on performance. In summary, our work confirms previous claims regarding the unacceptable performance of these search techniques and underscores the need for standardization in evaluations--standardization exemplified by the IR community.

## 124. Anonymization of Centralized and Distributed Social Networks by Sequential Clustering.

### Synopsis:

We study the problem of privacy-preservation in social networks. We consider the distributed setting in which the network data is split between several data holders. The goal is to arrive at an anonymized view of the unified network without revealing to any of the data holders information about links between nodes that are controlled by other data holders. To that end, we start with the centralized setting and offer two variants of an anonymization algorithm which is based on sequential clustering (Sq). Our algorithms significantly outperform the SaNGreeA algorithm due to Campan and Truta which is the leading algorithm for achieving anonymity in networks by means of clustering. We then devise secure distributed versions of our algorithms. To the best of our knowledge, this is the first study of privacy preservation in distributed social networks. We conclude by outlining future research proposals in that direction.

## 125. Answering General Time-Sensitive Queries..

## Synopsis:

Time is an important dimension of relevance for a large number of searches, such as over blogs and news archives. So far, research on searching over such collections has largely focused on locating topically similar documents for a query. Unfortunately, topic similarity alone is not always sufficient for document ranking. In this paper, we observe that, for an important class of queries that we call time-sensitive queries, the publication time of the documents in a news archive is important and should be considered in conjunction with the topic similarity to derive the final document ranking. Earlier work has focused on improving retrieval for “recency” queries that target recent documents. We propose a more general framework for handling time-sensitive queries and we automatically identify the important time intervals that are likely to be of interest for a query. Then, we build scoring techniques that seamlessly integrate the temporal aspect into the overall ranking mechanism. We present an extensive experimental evaluation using a variety of news article data sets, including TREC data as well as real web data analyzed using the Amazon Mechanical Turk. We examine several techniques for detecting the important time intervals for a query over a news archive and for incorporating this information in the retrieval process. We show that our techniques are robust and significantly improve result quality for time-sensitive queries compared to state-of-the-art retrieval techniques.

## 126. A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts.

## Synopsis:

The Machine Learning (ML) field has gained its momentum in almost any domain of research and just recently has become a reliable tool in the medical domain. The empirical domain of automatic learning is used in tasks such as medical decision support, medical imaging, protein-protein interaction, extraction of medical knowledge, and for overall patient management care. ML is envisioned as a tool by which computer-based systems can be integrated in the healthcare field in order to get a better, more efficient medical care. This paper describes a ML-based methodology for building an application that is capable of identifying and disseminating healthcare information. It extracts sentences from published medical papers that mention diseases and treatments, and identifies semantic relations that exist between diseases and treatments. Our evaluation results for these tasks show that the proposed methodology obtains reliable outcomes that could be integrated in an application to be used in the medical care domain. The potential value of this paper stands in the ML settings that we propose and in the fact that we outperform previous results on the same data set.

## 127. BestPeer++: A Peer-to-Peer Based Large-Scale Data Processing Platform.

### Synopsis:

The corporate network is often used for sharing information among the participating companies and facilitating collaboration in a certain industry sector where companies share a common interest. It can effectively help the companies to reduce their operational costs and increase the revenues. However, the inter-company data sharing and processing poses unique challenges to such a data management system including scalability, performance, throughput, and security. In this paper, we present BestPeer++, a system which delivers elastic data sharing services for corporate network applications in the cloud based on BestPeer - a peer-to-peer (P2P) based data management platform. By integrating cloud computing, database, and P2P technologies into one system, BestPeer++ provides an economical, flexible and scalable platform for corporate network applications and delivers data sharing services to participants based on the widely accepted pay-as-you-go business model. We evaluate BestPeer++ on Amazon EC2 Cloud platform. The benchmarking results show that BestPeer++ outperforms HadoopDB, a recently proposed large-scale data processing system, in performance when both systems are employed to handle typical corporate network workloads. The benchmarking results also demonstrate that BestPeer++ achieves near linear scalability for throughput with respect to the number of peer nodes.

## 128. Automatic Extraction of Top-k Lists from the Web.

### Synopsis:

This paper is concerned with information extraction from top-k web pages, which are web pages that describe top k instances of a topic which is of general interest. Examples include “the 10 tallest buildings in the world”, “the 50 hits of 2010 you don't want to miss”, etc. Compared to other structured information on the web (including web tables), information in top-k lists is larger and richer, of higher quality, and generally more interesting. Therefore top-k lists are highly valuable. For example, it can help enrich open-domain knowledge bases (to support applications such as search or fact answering). In this paper, we present an efficient method that extracts top-k lists from web pages with high performance. Specifically, we extract more than 1.7 million top-k lists from a web corpus of 1.6 billion pages with 92.0% precision and 72.3% recall.

## 129. Bridging Social and Data Networks.

## Synopsis:

Social networking applications have emerged as the platform of choice for carrying out a number of different activities online. In addition to their primary target of social interaction, we now also employ such applications to search for information online or to share multimedia content with our friends and families. For instance, according to recent statistics, each of us spends on average 15 min on YouTube every day.

## 130. A Privacy-Preserving Remote Data Integrity Checking Protocol with Data Dynamics and Public Verifiability.

## Synopsis:

Remote data integrity checking is a crucial technology in cloud computing. Recently, many works focus on providing data dynamics and/or public verifiability to this type of protocols. Existing protocols can support both features with the help of a third-party auditor. In a previous work, Sebé et al. propose a remote data integrity checking protocol that supports data dynamics. In this paper, we adapt Sebé et al.'s protocol to support public verifiability. The proposed protocol supports public verifiability without help of a third-party auditor. In addition, the proposed protocol does not leak any private information to third-party verifiers. Through a formal analysis, we show the correctness and security of the protocol. After that, through theoretical analysis and experimental results, we demonstrate that the proposed protocol has a good performance.

## 131. Demand Bidding Program and Its Application in Hotel Energy Management..

## Synopsis:

Demand bidding program (DBP) is recently adopted in practice by some energy operators. DBP is a risk-free demand response program targeting large energy consumers. In this paper, we consider DBP with the application in hotel energy management. For DBP, optimization problem is formulated with the objective of maximizing expected reward, which is received when the the amount of energy saving satisfies the contract. For a general distribution of energy consumption, we give a general condition for the optimal bid and outline an algorithm to find the solution without numerical integration. Furthermore, for Gaussian distribution, we derive closed-form expressions of the optimal bid and the corresponding expected reward. Regarding hotel energy, we characterize loads in the hotel and introduce several energy consumption models that capture major energy use. With the proposed models and DBP, simulation results show that DBP provides economics benefits to the hotel and encourages load scheduling. Furthermore, when only mean and variance of energy consumption are known, the validity of Gaussian approximation for computing optimal load and expected reward is also discussed.

## 132. Constructing a Global Social Service Network for Better Quality of Web Service Discovery.

### Synopsis:

Web services have had a tremendous impact on the Web for supporting a distributed service-based economy on a global scale. However, despite the outstanding progress, their uptake on a Web scale has been significantly less than initially anticipated. The isolation of services and the lack of social relationships among related services have been identified as reasons for the poor uptake. In this paper, we propose connecting the isolated service islands into a global social service network to enhance the services' sociability on a global scale. First, we propose linked social service-specific principles based on linked data principles for publishing services on the open Web as linked social services. Then, we suggest a new framework for constructing the global social service network following linked social service-specific principles based on complex network theories. Next, an approach is proposed to enable the exploitation of the global social service network, providing Linked Social Services as a Service. Finally, experimental results show that our approach can solve the quality of service discovery problem, improving both the service discovering time and the success rate by exploring service-to-service based on the global social service network.

## 133. Computing Structural Statistics by Keywords in Databases.

### Synopsis:

Keyword search in RDBs has been extensively studied in recent years. The existing studies focused on finding all or top-k interconnected tuple-structures that contain keywords. In reality, the number of such interconnected tuple-structures for a keyword query can be large. It becomes very difficult for users to obtain any valuable information more than individual interconnected tuple-structures. Also, it becomes challenging to provide a similar mechanism like group-&-aggregate for those interconnected tuple-structures. In this paper, we study computing structural statistics keyword queries by extending the group-&-aggregate framework. We consider an RDB as a large directed graph where nodes represent tuples, and edges represent the links among tuples. Instead of using tuples as a member in a group to be grouped, we consider rooted subgraphs. Such a rooted subgraph represents an interconnected tuple-structure among tuples and some of the tuples contain keywords. The dimensions of the rooted subgraphs are determined by dimensional-keywords in a data driven fashion. Two rooted subgraphs are grouped into the same group if they are isomorphic based on the dimensions or in other words the dimensional-keywords. The scores of the rooted subgraphs are computed by a user-given score function if the rooted subgraphs contain some of general keywords. Here, the general keywords are used to compute scores rather than determining dimensions. The aggregates are computed using an SQL aggregate function for every group based on the scores computed. We give our motivation using a real dataset. We propose new approaches to compute structural statistics keyword queries, perform extensive performance studies using two large real datasets and a large synthetic dataset, and confirm the effectiveness and efficiency of our approach.

## 134. A Query Formulation Language for the Data Web.

### Synopsis:

We present a query formulation language (called MashQL) in order to easily query and fuse structured data on the web. The main novelty of MashQL is that it allows people with limited IT skills to explore and query one (or multiple) data sources without prior knowledge about the schema, structure, vocabulary, or any technical details of these sources. More importantly, to be robust and cover most cases in practice, we do not assume that a data source should have - an offline or inline - schema. This poses several language-design and performance complexities that we fundamentally tackle. To illustrate the query formulation power of MashQL, and without loss of generality, we chose the Data web scenario. We also chose querying RDF, as it is the most primitive data model; hence, MashQL can be similarly used for querying relational databases and XML. We present two implementations of MashQL, an online mashup editor, and a Firefox add on. The former illustrates how MashQL can be used to query and mash up the Data web as simple as filtering and piping web feeds; and the Firefox add on illustrates using the browser as a web composer rather than only a navigator. To end, we evaluate MashQL on querying two data sets, DBLP and DBPedia, and show that our indexing techniques allow instant user interaction.

## 135. Dynamic Query Forms for Database Queries.

Synopsis:

Modern scientific databases and web databases maintain large and heterogeneous data. These real-world databases contain hundreds or even thousands of relations and attributes. Traditional predefined query forms are not able to satisfy various ad-hoc queries from users on those databases. This paper proposes DQF, a novel database query form interface, which is able to dynamically generate query forms. The essence of DQF is to capture a user's preference and rank query form components, assisting him/her in making decisions. The generation of a query form is an iterative process and is guided by the user. At each iteration, the system automatically generates ranking lists of form components and the user then adds the desired form components into the query form. The ranking of form components is based on the captured user preference. A user can also fill the query form and submit queries to view the query result at each iteration. In this way, a query form could be dynamically refined until the user is satisfied with the query results. We utilize the expected F-measure for measuring the goodness of a query form. A probabilistic model is developed for estimating the goodness of a query form in DQF. Our experimental evaluation and user study demonstrate the effectiveness and efficiency of the system.

## 136. Constructing E-Tourism Platform Based on Service Value Broker: A Knowledge Management Perspective..

Synopsis:

In our previous work, we have introduced various service value broker (SVB) patterns which integrate business modeling, knowledge management and economic analysis. In this paper, working towards the target of maximizing the potential usage of available resource to achieve the optimization of the satisfaction on both the service provider side and the service consumer side under the guidance of the public administrative, we propose to build the E-Tourism platform based on SVB. This paper demonstrates the mechanism for SVB based E-Tourism framework. The advantages of employing SVB include that the SVB can help to increase the value added in a realtime and balanced manner which conforms to the economical goal of both long run and short run. An experiment is shown using a personnel recommendation system.

## 137. Decentralized Probabilistic Text Clustering.

Synopsis:

Text clustering is an established technique for improving quality in information retrieval, for both centralized and distributed environments. However, traditional text clustering algorithms fail to scale on highly distributed environments, such as peer-to-peer networks.

Our algorithm for peer-to-peer clustering achieves high scalability by using a probabilistic approach for assigning documents to clusters. It enables a peer to compare each of its documents only with very few selected clusters, without significant loss of clustering quality. The algorithm offers probabilistic guarantees for the correctness of each document assignment to a cluster. Extensive experimental evaluation with up to 1 million peers and 1 million documents demonstrates the scalability and effectiveness of the algorithm.

## 138. Adaptive Fault Tolerant QoS Control Algorithms for Maximizing.

### Synopsis:

Data sensing and retrieval in wireless sensor systems have a widespread application in areas such as security and surveillance monitoring, and command and control in battlefields. In query-based wireless sensor systems, a user would issue a query and expect a response to be returned within the deadline. While the use of fault tolerance mechanisms through redundancy improves query reliability in the presence of unreliable wireless communication and sensor faults, it could cause the energy of the system to be quickly depleted. Therefore, there is an inherent trade-off between query reliability versus energy consumption in query-based wireless sensor systems. In this paper, we develop adaptive fault-tolerant quality of service (QoS) control algorithms based on hop-by-hop data delivery utilizing “source” and “path” redundancy, with the goal to satisfy application QoS requirements while prolonging the lifetime of the sensor system. We develop a mathematical model for the lifetime of the sensor system as a function of system parameters including the “source” and “path” redundancy levels utilized. We discover that there exists optimal “source” and “path” redundancy under which the lifetime of the system is maximized while satisfying application QoS requirements. Numerical data are presented and validated through extensive simulation, with physical interpretations given, to demonstrate the feasibility of our algorithm design.

## 139. Incremental Detection of Inconsistencies in Distributed Data.

### Synopsis:

This paper investigates incremental detection of errors in distributed data. Given a distributed database D, a set Σ of conditional functional dependencies (CFDs), the set V of violations of the CFDs in D, and updates ΔD to D, it is to find, with minimum data shipment, changes ΔV to V in response to ΔD. The need for the study is evident since real-life data is often dirty, distributed and frequently updated. It is often prohibitively expensive to recompute the entire set of violations when D is updated. We show that the incremental detection problem is NP-complete for database D that is partitioned either vertically or horizontally, even when Σ and D are fixed. Nevertheless, we show that it is bounded: there exist algorithms to detect errors such that their computational cost and data shipment are both linear in the size of ΔD and ΔV, independent of the size of the database D. We provide such incremental algorithms for vertically partitioned data and horizontally partitioned data, and show that the algorithms are optimal. We further propose optimization techniques for the incremental algorithm over vertical partitions to reduce data shipment. We verify experimentally, using real-life data on Amazon Elastic Compute Cloud (EC2), that our algorithms substantially outperform their batch counterparts.

## 140. Cost-Based Optimization of Service Compositions.

### Synopsis:

For providers of composite services, preventing cases of SLA violations is crucial. Previous work has established runtime adaptation of compositions as a promising tool to achieve SLA conformance. However, to get a realistic and complete view of the decision process of service providers, the costs of adaptation need to be taken into account. In this paper, we formalize the problem of finding the optimal set of adaptations, which minimizes the total costs arising from SLA violations and the adaptations to prevent them. We present possible algorithms to solve this complex optimization problem, and detail an end-to-end system based on our earlier work on the PREvent (prediction and prevention based on event monitoring) framework, which clearly indicates the usefulness of our model. We discuss experimental results that show how the application of our approach leads to reduced costs for the service provider, and explain the circumstances in which different algorithms lead to more or less satisfactory results.

## 141. Combined Mining: Discovering Informative Knowledge in Complex Data.

### Synopsis:

Enterprise data mining applications often involve complex data such as multiple large heterogeneous data sources, user preferences, and business impact. In such situations, a single method or one-step mining is often limited in discovering informative knowledge. It would also be very time and space consuming, if not impossible, to join relevant large data sources for mining patterns consisting of multiple aspects of information. It is crucial to develop effective approaches for mining patterns combining necessary information from multiple relevant business lines, catering for real business settings and decision-making actions rather than just providing a single line of patterns. The recent years have seen increasing efforts on mining more informative patterns, e.g., integrating frequent pattern mining with classifications to generate frequent pattern-based classifiers. Rather than presenting a specific algorithm, this paper builds on our existing works and proposes combined mining as a general approach to mining for informative patterns combining components from either multiple data sets or multiple features or by multiple methods on demand. We summarize general frameworks, paradigms, and basic processes for multifeature combined mining, multisource combined mining, and multimethod combined mining. Novel types of combined patterns, such as incremental cluster patterns, can result from such frameworks, which cannot be directly produced by the existing methods. A set of real-world case studies has been conducted to test the frameworks, with some of them briefed in this paper. They identify combined patterns for informing government debt prevention and improving government service objectives, which show the flexibility and instantiation capability of combined mining in discovering informative knowledge in complex data.

## 142. Adaptive Provisioning of Human Expertise in Service-oriented Systems.

### Synopsis:

Web-based collaborations have become essential in today's business environments. Due to the availability of various SOA frameworks, Web services emerged as the de facto technology to realize flexible compositions of services. While most existing work focuses on the discovery and composition of software based services, we highlight concepts for a people-centric Web. Knowledge-intensive environments clearly demand for provisioning of human expertise along with sharing of computing resources or business data through software-based services. To address these challenges, we introduce an adaptive approach allowing humans to provide their expertise through services using SOA standards, such as WSDL and SOAP. The seamless integration of humans in the SOA loop triggers numerous social implications, such as evolving expertise and drifting interests of human service providers. Here we propose a framework that is based on interaction monitoring techniques enabling adaptations in SOA-based socio-technical systems.

## 143. Keyword Query Routing.

### Synopsis:

Keyword search is an intuitive paradigm for searching linked data sources on the web. We propose to route keywords only to relevant sources to reduce the high cost of processing keyword search queries over all sources. We propose a novel method for computing top-k routing plans based on their potentials to contain results for a given keyword query. We employ a keyword-element relationship summary that compactly represents relationships between keywords and the data elements mentioning them. A multilevel scoring mechanism is proposed for computing the relevance of routing plans based on scores at the level of keywords, data elements, element sets, and subgraphs that connect these elements. Experiments carried out using 150 publicly available sources on the web showed that valid plans (precision@1 of 0.92) that are highly relevant (mean reciprocal rank of 0.89) can be computed in 1 second on average on a single PC. Further, we show routing greatly helps to improve the performance of keyword search, without compromising its result quality.

# 144. Dynamic Query Forms for Database Queries..

## Synopsis:

Modern scientific databases and web databases maintain large and heterogeneous data. These real-world databases contain hundreds or even thousands of relations and attributes. Traditional predefined query forms are not able to satisfy various ad-hoc queries from users on those databases. This paper proposes DQF, a novel database query form interface, which is able to dynamically generate query forms. The essence of DQF is to capture a user's preference and rank query form components, assisting him/her in making decisions. The generation of a query form is an iterative process and is guided by the user. At each iteration, the system automatically generates ranking lists of form components and the user then adds the desired form components into the query form. The ranking of form components is based on the captured user preference. A user can also fill the query form and submit queries to view the query result at each iteration. In this way, a query form could be dynamically refined until the user is satisfied with the query results. We utilize the expected F-measure for measuring the goodness of a query form. A probabilistic model is developed for estimating the goodness of a query form in DQF. Our experimental evaluation and user study demonstrate the effectiveness and efficiency of the system.

# 145. Effective Pattern Discovery for Text Mining.

## Synopsis:

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance.

## 146. A Personalized Ontology Model for Web Information Gathering by Domain Specific Search.

Synopsis:

As a model for knowledge description and formalization, ontologies are widely used to represent user profiles in personalized web information gathering. However, when representing user profiles, many models have utilized only knowledge from either a global knowledge base or a user local information. In this paper, a personalized ontology model is proposed for knowledge representation and reasoning over user profiles. This model learns ontological user profiles from both a world knowledge base and user local instance repositories. The ontology model is evaluated by comparing it against benchmark models in web information gathering. The results show that this ontology model is successful.

## 147. Secure Mining of Association Rules in Horizontally Distributed Databases.

Synopsis:

We propose a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Kantarcioglu and Clifton . Our protocol, like theirs, is based on the Fast Distributed Mining (FDM)algorithm of Cheung et al. , which is an unsecured distributed version of the Apriori algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms-one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy with respect to the protocol in . In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

## 148. ELCA Evaluation for Keyword Search on Probabilistic XML Data.

Synopsis:

As probabilistic data management is becoming one of the main research focuses and keyword search is turning into a more popular query means, it is natural to think how to support keyword queries on probabilistic XML data. With regards to keyword query on deterministic XML documents, ELCA (Exclusive Lowest Common Ancestor) semantics allows more relevant fragments rooted at the ELCAs to appear as results and is more popular compared with other keyword query result semantics (such as SLCAs).

In this paper, we investigate how to evaluate ELCA results for keyword queries on probabilistic XML documents. After defining probabilistic ELCA semantics in terms of possible world semantics, we propose an approach to compute ELCA probabilities without generating possible worlds. Then we develop an efficient stack-based algorithm that can find all probabilistic ELCA results and their ELCA probabilities for a given keyword query on a probabilistic XML document. Finally, we experimentally evaluate the proposed ELCA algorithm and compare it with its SLCA counterpart in aspects of result effectiveness, time and space efficiency, and scalability.

## 149. Secure Efficient and Accurate Discovery of Patterns in Sequence Data Sets.

### Synopsis:

Existing sequence mining algorithms mostly focus on mining for subsequences. However, a large class of applications, such as biological DNA and protein motif mining, require efficient mining of “approximate” patterns that are contiguous. The few existing algorithms that can be applied to find such contiguous approximate pattern mining have drawbacks like poor scalability, lack of guarantees in finding the pattern, and difficulty in adapting to other applications. In this paper, we present a new algorithm called FLexible and Accurate Motif DEtector (FLAME). FLAME is a flexible suffix-tree-based algorithm that can be used to find frequent patterns with a variety of definitions of motif (pattern) models. It is also accurate, as it always finds the pattern if it exists. Using both real and synthetic data sets, we demonstrate that FLAME is fast, scalable, and outperforms existing algorithms on a variety of performance metrics. In addition, based on FLAME, we also address a more general problem, named extended structured motif extraction, which allows mining frequent combinations of motifs under relaxed constraints.

## 150. Automatic Discovery of Personal Name Aliases from the Web.

### Synopsis:

An individual is typically referred by numerous name aliases on the web. Accurate identification of aliases of a given person name is useful in various web related tasks such as information retrieval, sentiment analysis, personal name disambiguation, and relation extraction. We propose a method to extract aliases of a given personal name from the web.

Given a personal name, the proposed method first extracts a set of candidate aliases. Second, we rank the extracted candidates according to the likelihood of a candidate being a correct alias of the given name. We propose a novel, automatically extracted lexical pattern-based approach to efficiently extract a large set of candidate aliases from snippets retrieved from a web search engine. We define numerous ranking scores to evaluate candidate aliases using three approaches: lexical pattern frequency, word co-occurrences in an anchor text graph, and page counts on the web. To construct a robust alias detection system, we integrate the different ranking scores into a single ranking function using ranking support vector machines. We evaluate the proposed method on three data sets: an English personal names data set, an English place names data set, and a Japanese personal names data set. The proposed method outperforms numerous baselines and previously proposed name alias extraction methods, achieving a statistically significant mean reciprocal rank (MRR) of 0.67. Experiments carried out using location names and Japanese personal names suggest the possibility of extending the proposed method to extract aliases for different types of named entities, and for different languages. Moreover, the aliases extracted using the proposed method are successfully utilized in an information retrieval task and improve recall by 20 percent in a relation-detection task.

## 151. Secure KNN Query Processing in Untrusted Cloud Environments..

### Synopsis:

Mobile devices with geo-positioning capabilities (e.g., GPS) enable users to access information that is relevant to their present location. Users are interested in querying about points of interest (POI) in their physical proximity, such as restaurants, cafes, ongoing events, etc. Entities specialized in various areas of interest (e.g., certain niche directions in arts, entertainment, travel) gather large amounts of geo-tagged data that appeal to subscribed users. Such data may be sensitive due to their contents. Furthermore, keeping such information up-to-date and relevant to the users is not an easy task, so the owners of such data sets will make the data accessible only to paying customers. Users send their current location as the query parameter, and wish to receive as result the nearest POIs, i.e., nearest-neighbors (NNs). But typical data owners do not have the technical means to support processing queries on a large scale, so they outsource data storage and querying to a cloud service provider. Many such cloud providers exist who offer powerful storage and computational infrastructures at low cost. However, cloud providers are not fully trusted, and typically behave in an honest-but-curious fashion. Specifically, they follow the protocol to answer queries correctly, but they also collect the locations of the POIs and the subscribers for other purposes. Leakage of POI locations can lead to privacy breaches as well as financial losses to the data owners, for whom the POI data set is an important source of revenue. Disclosure of user locations leads to privacy violations and may deter subscribers from using the service altogether. In this paper, we propose a family of techniques that allow processing of NN queries in an untrusted outsourced environment, while at the same time protecting both the POI and querying users' positions. Our techniques rely on mutable order preserving encoding (mOPE), the only secure order-preserving encryption method known to-date. W- also provide performance optimizations to decrease the computational cost inherent to processing on encrypted data, and we consider the case of incrementally updating data sets. We present an extensive performance evaluation of our techniques to illustrate their viability in practice.

## 152. Evaluating the Vulnerability of Network Mechanisms to Sophisticated DDoS Attacks.

### Synopsis:

In recent years, we have experienced a wave of DDoS attacks threatening the welfare of the internet. These are launched by malicious users whose only incentive is to degrade the performance of other, innocent, users. The traditional systems turn out to be quite vulnerable to these attacks. The objective of this work is to take a first step to close this fundamental gap, aiming at laying a foundation that can be used in future computer/network designs taking into account the malicious users. Our approach is based on proposing a metric that evaluates the vulnerability of a system. We then use our vulnerability metric to evaluate a data structure which is commonly used in network mechanisms-the Hash table data structure. We show that Closed Hash is much more vulnerable to DDoS attacks than Open Hash, even though the two systems are considered to be equivalent by traditional performance evaluation. We also apply the metric to queuing mechanisms common to computer and communications systems. Furthermore, we apply it to the practical case of a hash table whose requests are controlled by a queue, showing that even after the attack has ended, the regular users still suffer from performance degradation or even a total denial of service.

## 153. Efficient audit service outsourcing for data integrity in clouds.

### Synopsis:

Cloud-based outsourced storage relieves the client's burden for storage management and maintenance by providing a comparably low-cost, scalable, location-independent platform. However, the fact that clients no longer have physical possession of data indicates that they are facing a potentially formidable risk for missing or corrupted data. To avoid the security risks, audit services are critical to ensure the integrity and availability of outsourced data and to achieve digital forensics and credibility on cloud computing. Provable data possession (PDP), which is a cryptographic technique for verifying the integrity of data without retrieving it at an untrusted server, can be used to realize audit services.

In this paper, profiting from the interactive zero-knowledge proof system, we address the construction of an interactive PDP protocol to prevent the fraudulence of prover (soundness property) and the leakage of verified data (zero-knowledge property). We prove that our construction holds these properties based on the computation Diffie–Hellman assumption and the rewindable black-box knowledge extractor. We also propose an efficient mechanism with respect to probabilistic queries and periodic verification to reduce the audit costs per verification and implement abnormal detection timely. In addition, we present an efficient method for selecting an optimal parameter value to minimize computational overheads of cloud audit services. Our experimental results demonstrate the effectiveness of our approach.

## 154. Bridging Socially-Enhanced Virtual Communities.

### Synopsis:

Interactions spanning multiple organizations have become an important aspect in today's collaboration landscape. Organizations create alliances to fulfill strategic objectives. The dynamic nature of collaborations increasingly demands for automated techniques and algorithms to support the creation of such alliances. Our approach bases on the recommendation of potential alliances by discovery of currently relevant competence sources and the support of semi-automatic formation. The environment is service-oriented comprising humans and software services with distinct capabilities. To mediate between previously separated groups and organizations, we introduce the broker concept that bridges disconnected networks. We present a dynamic broker discovery approach based on interaction mining techniques and trust metrics.

## 155. The Role of Hubness in Clustering High-Dimensional Data.

### Synopsis:

High-dimensional data arise naturally in many domains, and have regularly presented a great challenge for traditional data mining techniques, both in terms of effectiveness and efficiency. Clustering becomes difficult due to the increasing sparsity of such data, as well as the increasing difficulty in distinguishing distances between data points. In this paper, we take a novel perspective on the problem of clustering high-dimensional data. Instead of attempting to avoid the curse of dimensionality by observing a lower dimensional feature subspace, we embrace dimensionality by taking advantage of inherently high-dimensional phenomena. More specifically, we show that hubness, i.e., the tendency of high-dimensional data to contain points (hubs) that frequently occur in k-nearest-neighbor lists of other points, can be successfully exploited in clustering. We validate our hypothesis by demonstrating that hubness is a good measure of point centrality within a high-dimensional data cluster, and by proposing several hubness-based clustering algorithms, showing that major hubs can be used effectively as cluster prototypes or as guides during the search for centroid-based cluster configurations. Experimental results demonstrate good performance of our algorithms in multiple settings, particularly in the presence of large quantities of noise. The proposed methods are tailored mostly for detecting approximately hyperspherical clusters and need to be extended to properly handle clusters of arbitrary shapes.

## 156. Exploring the Impact of Information System Introduction: The Case of an Australian Hospital Emergency Department .

### Synopsis:

In recent years, a large number of healthcare organisations have adopted information systems to improve their operations. An analysis of the existing literature shows that there is little concrete understanding about information systems&#x00E2; impact within the healthcare context. This study aims to improve the current understanding in the field by conducting an in-depth, exploratory study of the impact of IS in the healthcare industry. A longitudinal single case study was carried out in a major emergency and trauma centre in Australia, which just recently introduced a large scale IS. By focusing on a number of key work routines involved in the case organisation, this study gathered rich and deep insights into how the introduction of an advanced IS affects what healthcare professionals do as well as how they work and think. The findings of this study offer some important implications to both research and practice.

## 157. Efficient Data Mining for XML Queries – Answering Support..

### Synopsis:

Extracting information from semistructured documents is a very hard task, and is going to become more and more critical as the amount of digital information available on the Internet grows. Indeed, documents are often so large that the data set returned as answer to a query may be too big to convey interpretable knowledge. In this paper, we describe an approach based on Tree-Based Association Rules (TARs): mined rules, which provide approximate, intensional information on both the structure and the contents of Extensible Markup Language (XML) documents, and can be stored in XML format as well. This mined knowledge is later used to provide: 1) a concise idea—the gist—of both the structure and the content of the XML document and 2) quick, approximate answers to queries. In this paper, we focus on the second feature. A prototype system and experimental results demonstrate the effectiveness of the approach.

## 158. Collaborative Filtering with Personalized Skylines.

### Synopsis:

Collaborative filtering (CF) systems exploit previous ratings and similarity in user behavior to recommend the top-k objects/records which are potentially most interesting to the user assuming a single score per object. However, in various applications, a record (e.g., hotel) maybe rated on several attributes (value, service, etc.), in which case simply returning the ones with the highest overall scores fails to capture the individual attribute characteristics and to accommodate different selection criteria. In order to enhance the flexibility of CF, we propose Collaborative Filtering Skyline (CFS), a general framework that combines the advantages of CF with those of the skyline operator. CFS generates a personalized skyline for each user based on scores of other users with similar behavior. The personalized skyline includes objects that are good on certain aspects, and eliminates the ones that are not interesting on any attribute combination. Although the integration of skylines and CF has several attractive properties, it also involves rather expensive computations. We face this challenge through a comprehensive set of algorithms and optimizations that reduce the cost of generating personalized skylines. In addition to exact skyline processing, we develop an approximate method that provides error guarantees. Finally, we propose the top-k personalized skyline, where the user specifies the required output cardinality.

## 159. Web Image Re-Ranking UsingQuery-Specific Semantic Signatures.

### Synopsis:

Image re-ranking, as an effective way to improve the results of web-based image search, has been adopted by current commercial search engines such as Bing and Google. Given a query keyword, a pool of images are first retrieved based on textual information. By asking the user to select a query image from the pool, the remaining images are re-ranked based on their visual similarities with the query image. A major challenge is that the similarities of visual features do not well correlate with images' semantic meanings which interpret users' search intention. Recently people proposed to match images in a semantic space which used attributes or reference classes closely related to the semantic meanings of images as basis. However, learning a universal visual semantic space to characterize highly diverse images from the web is difficult and inefficient. In this paper, we propose a novel image re-ranking framework, which automatically offline learns different semantic spaces for different query keywords. The visual features of images are projected into their related semantic spaces to get semantic signatures. At the online stage, images are re-ranked by comparing their semantic signatures obtained from the semantic space specified by the query keyword. The proposed query-specific semantic signatures significantly improve both the accuracy and efficiency of image re-ranking. The original visual features of thousands of dimensions can be projected to the semantic signatures as short as 25 dimensions. Experimental results show that 25-40 percent relative improvement has been achieved on re-ranking precisions compared with the state-of-the-art methods.

## 160. Ginix Generalized Inverted Index for Keyword Search.

### Synopsis:

Keyword search has become a ubiquitous method for users to access text data in the face of information explosion. Inverted lists are usually used to index underlying documents to retrieve documents according to a set of keywords efficiently. Since inverted lists are usually large, many compression techniques have been proposed to reduce the storage space and disk I/O time. However, these techniques usually perform decompression operations on the fly, which increases the CPU time. This paper presents a more efficient index structure, the Generalized INverted IndeX (Ginix), which merges consecutive IDs in inverted lists into intervals to save storage space. With this index structure, more efficient algorithms can be devised to perform basic keyword search operations, i.e., the union and the intersection operations, by taking the advantage of intervals. Specifically, these algorithms do not require conversions from interval lists back to ID lists. As a result, keyword search using Ginix can be more efficient than those using traditional inverted indices. The performance of Ginix is also improved by reordering the documents in datasets using two scalable algorithms. Experiments on the performance and scalability of Ginix on real datasets show that Ginix not only requires less storage space, but also improves the keyword search performance, compared with traditional inverted indexes.

## 161. Generative Models for Item Adoptions Using Social Correlation.

### Synopsis:

Users face many choices on the web when it comes to choosing which product to buy, which video to watch, and so on. In making adoption decisions, users rely not only on their own preferences, but also on friends. We call the latter social correlation, which may be caused by the homophily and social influence effects. In this paper, we focus on modeling social correlation on users item adoptions. Given a user-user social graph and an item-user adoption graph, our research seeks to answer the following questions: Whether the items adopted by a user correlate with items adopted by her friends, and how to model item adoptions using social correlation. We propose a social correlation framework that considers a social correlation matrix representing the degrees of correlation from every user to the users friends, in addition to a set of latent factors representing topics of interests of individual users. Based on the framework, we develop two generative models, namely sequential and unified, and the corresponding parameter estimation approaches. From each model, we devise the social correlation only and hybrid methods for predicting missing adoption links. Experiments on LiveJournal and Epinions data sets show that our proposed models outperform the approach based on latent factors only (LDA).

## 162. Cost-aware rank join with random and sorted access.

### Synopsis:

In this paper, we address the problem of joining ranked results produced by two or more services on the web. We consider services endowed with two kinds of access that are often available: 1) sorted access, which returns tuples sorted by score; 2) random access, which returns tuples matching a given join attribute value. Rank join operators combine objects of two or more relations and output the k combinations with the highest aggregate score. While the past literature has studied suitable bounding schemes for this setting, in this paper we focus on the definition of a pulling strategy, which determines the order of invocation of the joined services. We propose the Cost-Aware with Random and Sorted access (CARS) pulling strategy, which is derived at compile-time and is oblivious of the query-dependent score distributions. We cast CARS as the solution of an optimization problem based on a small set of parameters characterizing the joined services. We validate the proposed strategy with experiments on both real and synthetic data sets. We show that CARS outperforms prior proposals and that its overall access cost is always within a very short margin from that of an oracle-based optimal strategy. In addition, CARS is shown to be robust w.r.t. the uncertainty that may characterize the estimated parameters.

## 163. One Size Does Not Fit All: Towards User- and Query-Dependent Ranking For Web Databases.

### Synopsis:

With the emergence of the deep web, searching web databases in domains such as vehicles, real estate, etc., has become a routine task. One of the problems in this context is ranking the results of a user query. Earlier approaches for addressing this problem have used frequencies of database values, query logs, and user profiles. A common thread in most of these approaches is that ranking is done in a user- and/or query-independent manner. This paper proposes a novel query- and user-dependent approach for ranking query results in web databases. We present a ranking model, based on two complementary notions of user and query similarity, to derive a ranking function for a given user query. This function is acquired from a sparse workload comprising of several such ranking functions derived for various user-query pairs. The model is based on the intuition that similar users display comparable ranking preferences over the results of similar queries. We define these similarities formally in alternative ways and discuss their effectiveness analytically and experimentally over two distinct web databases.

## 164. Gmatch: Secure and Privacy-Preserving Group Matching in Social Networks..

### Synopsis:

Groups are becoming one of the most compelling features in both online social networks and Twitter-like micro-blogging services. A stranger outside of an existing group may have the need to find out more information about attributes of current members in the group, in order to make a decision to join. However, in many cases, attributes of both group members and the stranger need to be kept private and should not be revealed to others, as they may contain sensitive and personal information. How can we find out matching information exists between the stranger and members of the group, based on their attributes that are not to be disclosed? In this paper, we present a new group matching mechanism, by taking advantage private set intersection and ring signatures. With our scheme, a stranger is able to collect correct group matching information while sensitive information of the stranger and group members are not disclosed. Finally, we propose to use batch verification to significantly improve the performance of the matching process.

## 165. Heuristics Based Query Processing for Large RDF Graphs Using Cloud Computing.

### Synopsis:

Semantic web is an emerging area to augment human reasoning. Various technologies are being developed in this arena which have been standardized by the World Wide Web Consortium (W3C). One such standard is the Resource Description Framework (RDF). Semantic web technologies can be utilized to build efficient and scalable systems for Cloud Computing. With the explosion of semantic web technologies, large RDF graphs are common place. This poses significant challenges for the storage and retrieval of RDF graphs. Current frameworks do not scale for large RDF graphs and as a result do not address these challenges. In this paper, we describe a framework that we built using Hadoop to store and retrieve large numbers of RDF triples by exploiting the cloud computing paradigm. We describe a scheme to store RDF data in Hadoop Distributed File System. More than one Hadoop job (the smallest unit of execution in Hadoop) may be needed to answer a query because a single triple pattern in a query cannot simultaneously take part in more than one join in a single Hadoop job. To determine the jobs, we present an algorithm to generate query plan, whose worst case cost is bounded, based on a greedy approach to answer a SPARQL Protocol and RDF Query Language (SPARQL) query. We use Hadoop's MapReduce framework to answer the queries. Our results show that we can store large RDF graphs in Hadoop clusters built with cheap commodity class hardware. Furthermore, we show that our framework is scalable and efficient and can handle large amounts of RDF data, unlike traditional approaches.

## 166. Data Leakage Detection.

## Synopsis:

We study the following problem: A data distributor has given sensitive data to a set of supposedly trusted agents (third parties). Some of the data are leaked and found in an unauthorized place (e.g., on the web or somebody's laptop). The distributor must assess the likelihood that the leaked data came from one or more agents, as opposed to having been independently gathered by other means. We propose data allocation strategies (across the agents) that improve the probability of identifying leakages. These methods do not rely on alterations of the released data (e.g., watermarks). In some cases, we can also inject “realistic but fake” data records to further improve our chances of detecting leakage and identifying the guilty party.

## 167. Optimal Service Pricing for a Cloud Cache.

## Synopsis:

Cloud applications that offer data management services are emerging. Such clouds support caching of data in order to provide quality query services. The users can query the cloud data, paying the price for the infrastructure they use. Cloud management necessitates an economy that manages the service of multiple users in an efficient, but also, resource-economic way that allows for cloud profit. Naturally, the maximization of cloud profit given some guarantees for user satisfaction presumes an appropriate price-demand model that enables optimal pricing of query services. The model should be plausible in that it reflects the correlation of cache structures involved in the queries. Optimal pricing is achieved based on a dynamic pricing scheme that adapts to time changes. This paper proposes a novel price-demand model designed for a cloud cache and a dynamic pricing scheme for queries executed in the cloud cache. The pricing solution employs a novel method that estimates the correlations of the cache services in an time-efficient manner. The experimental study shows the efficiency of the solution.

## 168. Mining Order-Preserving Submatrices from Data with Repeated Measurements.

### Synopsis:

Order-preserving submatrices (OPSM's) have been shown useful in capturing concurrent patterns in data when the relative magnitudes of data items are more important than their exact values. For instance, in analyzing gene expression profiles obtained from microarray experiments, the relative magnitudes are important both because they represent the change of gene activities across the experiments, and because there is typically a high level of noise in data that makes the exact values untrustable. To cope with data noise, repeated experiments are often conducted to collect multiple measurements. We propose and study a more robust version of OPSM, where each data item is represented by a set of values obtained from replicated experiments. We call the new problem OPSM-RM (OPSM with repeated measurements). We define OPSM-RM based on a number of practical requirements. We discuss the computational challenges of OPSM-RM and propose a generic mining algorithm. We further propose a series of techniques to speed up two time dominating components of the algorithm. We show the effectiveness and efficiency of our methods through a series of experiments conducted on real microarray data.

## 169. Identifying Content for Planned Events Across Social Media Sites.

### Synopsis:

User-contributed Web data contains rich and diverse information about a variety of events in the physical world, such as shows, festivals, conferences and more. This information ranges from known event features (e.g., title, time, location) posted on event aggregation platforms (e.g., Last.fm events, EventBrite, Facebook events) to discussions and reactions related to events shared on different social media sites (e.g., Twitter, YouTube, Flickr). In this paper, we focus on the challenge of automatically identifying user-contributed content for events that are planned and, therefore, known in advance, across different social media sites. We mine event aggregation platforms to extract event features, which are often noisy or missing. We use these features to develop query formulation strategies for retrieving content associated with an event on different social media sites. Further, we explore ways in which event content identified on one social media site can be used to retrieve additional relevant event content on other social media sites. We apply our strategies to a large set of user-contributed events, and analyze their effectiveness in retrieving relevant event content from Twitter, YouTube, and Flickr.

## 170. Efficient and Accurate Discovery of Patterns in Sequence Data Sets.

### Synopsis:

Existing sequence mining algorithms mostly focus on mining for subsequences. However, a large class of applications, such as biological DNA and protein motif mining, require efficient mining of “approximate” patterns that are contiguous. The few existing algorithms that can be applied to find such contiguous approximate pattern mining have drawbacks like poor scalability, lack of guarantees in finding the pattern, and difficulty in adapting to other applications. In this paper, we present a new algorithm called FLexible and Accurate Motif DEtector (FLAME). FLAME is a flexible suffix-tree-based algorithm that can be used to find frequent patterns with a variety of definitions of motif (pattern) models. It is also accurate, as it always finds the pattern if it exists. Using both real and synthetic data sets, we demonstrate that FLAME is fast, scalable, and outperforms existing algorithms on a variety of performance metrics. In addition, based on FLAME, we also address a more general problem, named extended structured motif extraction, which allows mining frequent combinations of motifs under relaxed constraints.

## 171. Optimal Stochastic Location Updates In Mobile Ad Hoc Networks.

### Synopsis:

We consider the location service in a mobile ad-hoc network (MANET), where each node needs to maintain its location information by 1) frequently updating its location information within its neighboring region, which is called neighborhood update (NU), and 2) occasionally updating its location information to certain distributed location server in the network, which is called location server update (LSU). The trade off between the operation costs in location updates and the performance losses of the target application due to location inaccuracies (i.e., application costs) imposes a crucial question for nodes to decide the optimal strategy to update their location information, where the optimality is in the sense of minimizing the overall costs. In this paper, we develop a stochastic sequential decision framework to analyze this problem. Under a Markovian mobility model, the location update decision problem is modeled as a Markov Decision Process (MDP). We first investigate the monotonicity properties of optimal NU and LSU operations with respect to location inaccuracies under a general cost setting. Then, given a separable cost structure, we show that the location update decisions of NU and LSU can be independently carried out without loss of optimality, i.e., a separation property. From the discovered separation property of the problem structure and the monotonicity properties of optimal actions, we find that 1) there always exists a simple optimal threshold-based update rule for LSU operations; 2) for NU operations, an optimal threshold-based update rule exists in a low-mobility scenario. In the case that no a priori knowledge of the MDP model is available, we also introduce a practical model-free learning approach to find a near-optimal solution for the problem.

# 172. SAMPLING ONLINE SOCIAL NETWORK.

## Synopsis:

As online social networking emerges, there has been increased interest to utilize the underlying network structure as well as the available information on social peers to improve the information needs of a user. In this paper, we focus on improving the performance of information collection from the neighborhood of a user in a dynamic social network. We introduce sampling-based algorithms to efficiently explore a user's social network respecting its structure and to quickly approximate quantities of interest. We introduce and analyze variants of the basic sampling scheme exploring correlations across our samples. Models of centralized and distributed social networks are considered. We show that our algorithms can be utilized to rank items in the neighborhood of a user, assuming that information for each user in the network is available. Using real and synthetic data sets, we validate the results of our analysis and demonstrate the efficiency of our algorithms in approximating quantities of interest. The methods we describe are general and can probably be easily adopted in a variety of strategies aiming to efficiently collect information from a social graph.

# 173. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques.

## Synopsis:

Recommender systems are becoming increasingly important to individual users and businesses for providing personalized recommendations. However, while the majority of algorithms proposed in recommender systems literature have focused on improving recommendation accuracy (as exemplified by the recent Netflix Prize competition), other important aspects of recommendation quality, such as the diversity of recommendations, have often been overlooked. In this paper, we introduce and explore a number of item ranking techniques that can generate substantially more diverse recommendations across all users while maintaining comparable levels of recommendation accuracy. Comprehensive empirical evaluation consistently shows the diversity gains of the proposed techniques using several real-world rating data sets and different rating prediction algorithms..

## 174. Exploring Application-Level Semantics for Data Compression.

### Synopsis:

Natural phenomena show that many creatures form large social groups and move in regular patterns. However, previous works focus on finding the movement patterns of each single object or all objects. In this paper, we first propose an efficient distributed mining algorithm to jointly identify a group of moving objects and discover their movement patterns in wireless sensor networks. Afterward, we propose a compression algorithm, called 2P2D, which exploits the obtained group movement patterns to reduce the amount of delivered data. The compression algorithm includes a sequence merge and an entropy reduction phases. In the sequence merge phase, we propose a Merge algorithm to merge and compress the location data of a group of moving objects. In the entropy reduction phase, we formulate a Hit Item Replacement (HIR) problem and propose a Replace algorithm that obtains the optimal solution. Moreover, we devise three replacement rules and derive the maximum compression ratio. The experimental results show that the proposed compression algorithm leverages the group movement patterns to reduce the amount of delivered data effectively and efficiently.

## 175. Publishing Search Logs A Comparative Study of Privacy Guarantees.

### Synopsis:

Search engine companies collect the “database of intentions,” the histories of their users' search queries. These search logs are a gold mine for researchers. Search engine companies, however, are wary of publishing search logs in order not to disclose sensitive information. In this paper, we analyze algorithms for publishing frequent keywords, queries, and clicks of a search log. We first show how methods that achieve variants of k-anonymity are vulnerable to active attacks. We then demonstrate that the stronger guarantee ensured by ε-differential privacy unfortunately does not provide any utility for this problem. We then propose an algorithm ZEALOUS and show how to set its parameters to achieve (ε, δ)-probabilistic privacy. We also contrast our analysis of ZEALOUS with an analysis by Korolova et al. [17] that achieves (ε',δ')-indistinguishability. Our paper concludes with a large experimental study using real applications where we compare ZEALOUS and previous work that achieves k-anonymity in search log publishing. Our results show that ZEALOUS yields comparable utility to k-anonymity while at the same time achieving much stronger privacy guarantees.

## 176. Secure Mining of Association Rules in Horizontally Distributed Databases..

Synopsis:

We propose a protocol for secure mining of association rules in horizontally distributed databases. The current leading protocol is that of Kantarcioglu and Clifton . Our protocol, like theirs, is based on the Fast Distributed Mining (FDM)algorithm of Cheung et al. , which is an unsecured distributed version of the Apriori algorithm. The main ingredients in our protocol are two novel secure multi-party algorithms-one that computes the union of private subsets that each of the interacting players hold, and another that tests the inclusion of an element held by one player in a subset held by another. Our protocol offers enhanced privacy with respect to the protocol in . In addition, it is simpler and is significantly more efficient in terms of communication rounds, communication cost and computational cost.

## 177. Joint Top-K Spatial Keyword Query Processing.

Synopsis:

Web users and content are increasingly being geopositioned, and increased focus is being given to serving local content in response to web queries. This development calls for spatial keyword queries that take into account both the locations and textual descriptions of content. We study the efficient, joint processing of multiple top-k spatial keyword queries. Such joint processing is attractive during high query loads and also occurs when multiple queries are used to obfuscate a user's true query. We propose a novel algorithm and index structure for the joint processing of top-k spatial keyword queries. Empirical studies show that the proposed solution is efficient on real data sets. We also offer analytical studies on synthetic data sets to demonstrate the efficiency of the proposed solution.

## 178. Heuristics Based Query Processing for Large RDF Graphs Using Cloud Computing.

Synopsis:

Semantic web is an emerging area to augment human reasoning. Various technologies are being developed in this arena which have been standardized by the World Wide Web Consortium (W3C). One such standard is the Resource Description Framework (RDF). Semantic web technologies can be utilized to build efficient and scalable systems for Cloud Computing. With the explosion of semantic web technologies, large RDF graphs are common place. This poses significant challenges for the storage and retrieval of RDF graphs. Current frameworks do not scale for large RDF graphs and as a result do not address these challenges. In this paper, we describe a framework that we built using Hadoop to store and retrieve large numbers of RDF triples by exploiting the cloud computing paradigm. We describe a scheme to store RDF data in Hadoop Distributed File System. More than one Hadoop job (the smallest unit of execution in Hadoop) may be needed to answer a query because a single triple pattern in a query cannot simultaneously take part in more than one join in a single Hadoop job. To determine the jobs, we present an algorithm to generate query plan, whose worst case cost is bounded, based on a greedy approach to answer a SPARQL Protocol and RDF Query Language (SPARQL) query. We use Hadoop's MapReduce framework to answer the queries. Our results show that we can store large RDF graphs in Hadoop clusters built with cheap commodity class hardware. Furthermore, we show that our framework is scalable and efficient and can handle large amounts of RDF data, unlike traditional approaches.

## 179. Query Planning for Continuous Aggregation Queries over a Network of Data Aggregators.

### Synopsis:

Continuous queries are used to monitor changes to time varying data and to provide results useful for online decision making. Typically a user desires to obtain the value of some aggregation function over distributed data items, for example, to know value of portfolio for a client; or the AVG of temperatures sensed by a set of sensors. In these queries a client specifies a coherency requirement as part of the query. We present a low-cost, scalable technique to answer continuous aggregation queries using a network of aggregators of dynamic data items. In such a network of data aggregators, each data aggregator serves a set of data items at specific coherencies. Just as various fragments of a dynamic webpage are served by one or more nodes of a content distribution network, our technique involves decomposing a client query into subqueries and executing subqueries on judiciously chosen data aggregators with their individual subquery incoherency bounds. We provide a technique for getting the optimal set of subqueries with their incoherency bounds which satisfies client query's coherency requirement with least number of refresh messages sent from aggregators to the client. For estimating the number of refresh messages, we build a query cost model which can be used to estimate the number of messages required to satisfy the client specified incoherency bound. Performance results using real-world traces show that our cost-based query planning leads to queries being executed using less than one third the number of messages required by existing schemes.

## 180. SeDas: A Self-Destructing Data System Based on Active Storage Framework.

Synopsis:

Personal data stored in the Cloud may contain account numbers, passwords, notes, and other important information that could be used and misused by a miscreant, a competitor, or a court of law. These data are cached, copied, and archived by Cloud Service Providers (CSPs), often without users' authorization and control. Self-destructing data mainly aims at protecting the user data's privacy. All the data and their copies become destructed or unreadable after a user-specified time, without any user intervention. In addition, the decryption key is destructed after the user-specified time. In this paper, we present SeDas, a system that meets this challenge through a novel integration of cryptographic techniques with active storage techniques based on T10 OSD standard. We implemented a proof-of-concept SeDas prototype. Through functionality and security properties evaluations of the SeDas prototype, the results demonstrate that SeDas is practical to use and meets all the privacy-preserving goals described. Compared to the system without self-destructing data mechanism, throughput for uploading and downloading with the proposed SeDas acceptably decreases by less than 72%, while latency for upload/download operations with self-destructing data mechanism increases by less than 60%.

## 181. Mining Web Graphs for Recommendations.

Synopsis:

As the exponential explosion of various contents generated on the Web, Recommendation techniques have become increasingly indispensable. Innumerable different kinds of recommendations are made on the Web every day, including movies, music, images, books recommendations, query suggestions, tags recommendations, etc. No matter what types of data sources are used for the recommendations, essentially these data sources can be modeled in the form of various types of graphs. In this paper, aiming at providing a general framework on mining Web graphs for recommendations, (1) we first propose a novel diffusion method which propagates similarities between different nodes and generates recommendations; (2) then we illustrate how to generalize different recommendation problems into our graph diffusion framework. The proposed framework can be utilized in many recommendation tasks on the World Wide Web, including query suggestions, tag recommendations, expert finding, image recommendations, image annotations, etc. The experimental analysis on large data sets shows the promising future of our work.

## 182. Horizontal Aggregations in SQL to Prepare Data Sets for Data Mining Analysis..

Synopsis:

Preparing a data set for analysis is generally the most time consuming task in a data mining project, requiring many complex SQL queries, joining tables, and aggregating columns. Existing SQL aggregations have limitations to prepare data sets because they return one column per aggregated group. In general, a significant manual effort is required to build data sets, where a horizontal layout is required. We propose simple, yet powerful, methods to generate SQL code to return aggregated columns in a horizontal tabular layout, returning a set of numbers instead of one number per row. This new class of functions is called horizontal aggregations. Horizontal aggregations build data sets with a horizontal denormalized layout (e.g., point-dimension, observation-variable, instance-feature), which is the standard layout required by most data mining algorithms. We propose three fundamental methods to evaluate horizontal aggregations: CASE: Exploiting the programming CASE construct; SPJ: Based on standard relational algebra operators (SPJ queries); PIVOT: Using the PIVOT operator, which is offered by some DBMSs. Experiments with large tables compare the proposed query evaluation methods. Our CASE method has similar speed to the PIVOT operator and it is much faster than the SPJ method. In general, the CASE and PIVOT methods exhibit linear scalability, whereas the SPJ method does not.

## 183. Scalable Learning of Collective Behavior.

## Synopsis:

This study of collective behavior is to understand how individuals behave in a social networking environment. Oceans of data generated by social media like Facebook, Twitter, Flickr, and YouTube present opportunities and challenges to study collective behavior on a large scale. In this work, we aim to learn to predict collective behavior in social media. In particular, given information about some individuals, how can we infer the behavior of unobserved individuals in the same network? A social-dimension-based approach has been shown effective in addressing the heterogeneity of connections presented in social media. However, the networks in social media are normally of colossal size, involving hundreds of thousands of actors. The scale of these networks entails scalable learning of models for collective behavior prediction. To address the scalability issue, we propose an edge-centric clustering scheme to extract sparse social dimensions. With sparse social dimensions, the proposed approach can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other nonscalable methods.

## 184. SUSIE: Search Using Services and Information Extraction.

## Synopsis:

The API of a Web service restricts the types of queries that the service can answer. For example, a Web service might provide a method that returns the songs of a given singer, but it might not provide a method that returns the singers of a given song. If the user asks for the singer of some specific song, then the Web service cannot be called - even though the underlying database might have the desired piece of information. This asymmetry is particularly problematic if the service is used in a Web service orchestration system. In this paper, we propose to use on-the-fly information extraction to collect values that can be used as parameter bindings for the Web service. We show how this idea can be integrated into a Web service orchestration system. Our approach is fully implemented in a prototype called SUSIE. We present experiments with real-life data and services to demonstrate the practical viability and good performance of our approach.

## 185. Optimization of Horizontal Aggregation in SQL by Using K-Means Clustering.

### Synopsis:

Data mining systems use datasets with columns in a horizontal tabular layout in order to analyze data efficiently. In data mining project, preparation of data set is more complex process and it requires many SQL queries, joining tables and aggregating columns. So this is an important problem in data mining. Horizontal aggregation solves this problem by preparing the data set in a horizontal tabular layout and returns a set of numbers instead of single number per row. Integrating data mining algorithms with a relational database management system is an important problem for database programmers. K means algorithms using SQL is a best clustering algorithm[10]. When K means algorithm use with horizontal aggregation, it partitions the large set of data get from the horizontal aggregation into k cluster in order to reduce effort in data preparation phase of data mining. Here describing three SQL implementations of K means algorithm to integrate it with a relational database management systems: 1) Standard K Means, a direct translation of K means into SQL, 2) Optimized K means, an optimized version based on improved data organization, efficient indexing and sufficient statistics and 3) Incremental K means which is an incremental version that uses the optimized version as a building block with fast convergence and automated reseeding. Horizontal aggregation solves problem of data mining in preparing summary data set and K means clustering algorithm integrated with a relational DBMS using SQL optimize the data set generated by horizontal aggregation.

## 186. Integration of Sound Signature Authentication System.

### Synopsis:

This document provides guidelines for implementing anauthentication system which works on graphical password and includes sound signature. Click based graphical password provides security from brute force and dictionary attacks and they are not predictive thus it’s not easy to breachthem and a sound signature is integrated along with which enhances the security as this sound signature also under goes the password verification, and once the graphical password along with the sound signature is verified the user is allowed to log into the system.

## 187. Scalable Scheduling of Updates in Streaming Data Warehouses.

### Synopsis:

We discuss update scheduling in streaming data warehouses, which combine the features of traditional data warehouses and data stream systems. In our setting, external sources push append-only data streams into the warehouse with a wide range of interarrival times. While traditional data warehouses are typically refreshed during downtimes, streaming warehouses are updated as new data arrive. We model the streaming warehouse update problem as a scheduling problem, where jobs correspond to processes that load new data into tables, and whose objective is to minimize data staleness over time (at time t, if a table has been updated with information up to some earlier time r, its staleness is t minus r). We then propose a scheduling framework that handles the complications encountered by a stream warehouse: view hierarchies and priorities, data consistency, inability to preempt updates, heterogeneity of update jobs caused by different interarrival times and data volumes among different sources, and transient overload. A novel feature of our framework is that scheduling decisions do not depend on properties of update jobs (such as deadlines), but rather on the effect of update jobs on data staleness. Finally, we present a suite of update scheduling algorithms and extensive simulation experiments to map out factors which affect their performance.

## 188. Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development.

### Synopsis:

Twitter has received much attention recently. An important characteristic of Twitter is its real-time nature. We investigate the real-time interaction of events such as earthquakes in Twitter and propose an algorithm to monitor tweets and to detect a target event. To detect a target event, we devise a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. Subsequently, we produce a probabilistic spatiotemporal model for the target event that can find the center of the event location. We regard each Twitter user as a sensor and apply particle filtering, which are widely used for location estimation. The particle filter works better than other comparable methods for estimating the locations of target events. As an application, we develop an earthquake reporting system for use in Japan. Because of the numerous earthquakes and the large number of Twitter users throughout the country, we can detect an earthquake with high probability (93 percent of earthquakes of Japan Meteorological Agency (JMA) seismic intensity scale 3 or more are detected) merely by monitoring tweets. Our system detects earthquakes promptly and notification is delivered much faster than JMA broadcast announcements.

## 189. Organizing User Search Histories.

### Synopsis:

Users are increasingly pursuing complex task-oriented goals on the web, such as making travel arrangements, managing finances, or planning purchases. To this end, they usually break down the tasks into a few codependent steps and issue multiple queries around these steps repeatedly over long periods of time. To better support users in their long-term information quests on the web, search engines keep track of their queries and clicks while searching online. In this paper, we study the problem of organizing a user's historical queries into groups in a dynamic and automated fashion. Automatically identifying query groups is helpful for a number of different search engine components and applications, such as query suggestions, result ranking, query alterations, sessionization, and collaborative search. In our approach, we go beyond approaches that rely on textual similarity or time thresholds, and we propose a more robust approach that leverages search query logs. We experimentally study the performance of different techniques, and showcase their potential, especially when combined together.

## 190. Knowledge-Based Interactive Postmining of Association Rules Using Ontologies.

### Synopsis:

In Data Mining, the usefulness of association rules is strongly limited by the huge amount of delivered rules. To overcome this drawback, several methods were proposed in the literature such as itemset concise representations, redundancy reduction, and postprocessing. However, being generally based on statistical information, most of these methods do not guarantee that the extracted rules are interesting for the user. Thus, it is crucial to help the decision-maker with an efficient postprocessing step in order to reduce the number of rules. This paper proposes a new interactive approach to prune and filter discovered rules. First, we propose to use ontologies in order to improve the integration of user knowledge in the postprocessing task. Second, we propose the Rule Schema formalism extending the specification language proposed by Liu et al. for user expectations. Furthermore, an interactive framework is designed to assist the user throughout the analyzing task. Applying our new approach over voluminous sets of rules, we were able, by integrating domain expert knowledge in the postprocessing step, to reduce the number of rules to several dozens or less. Moreover, the quality of the filtered rules was validated by the domain expert at various points in the interactive process.

## 191. Selecting Attributes for Sentiment Classification Using Feature Relation Networks.

## Synopsis:

A major concern when incorporating large sets of diverse n-gram features for sentiment classification is the presence of noisy, irrelevant, and redundant attributes. These concerns can often make it difficult to harness the augmented discriminatory potential of extended feature sets. We propose a rule-based multivariate text feature selection method called Feature Relation Network (FRN) that considers semantic information and also leverages the syntactic relationships between n-gram features. FRN is intended to efficiently enable the inclusion of extended sets of heterogeneous n-gram features for enhanced sentiment classification. Experiments were conducted on three online review testbeds in comparison with methods used in prior sentiment classification research. FRN outperformed the comparison univariate, multivariate, and hybrid feature selection methods; it was able to select attributes resulting in significantly better classification accuracy irrespective of the feature subset sizes. Furthermore, by incorporating syntactic information about n-gram relations, FRN is able to select features in a more computationally efficient manner than many multivariate and hybrid techniques.

## 192. Outsourced Similarity Search on Metric Data Assets.

## Synopsis:

This paper considers a cloud computing setting in which similarity querying of metric data is outsourced to a service provider. The data is to be revealed only to trusted users, not to the service provider or anyone else. Users query the server for the most similar data objects to a query example. Outsourcing offers the data owner scalability and a low-initial investment. The need for privacy may be due to the data being sensitive (e.g., in medicine), valuable (e.g., in astronomy), or otherwise confidential. Given this setting, the paper presents techniques that transform the data prior to supplying it to the service provider for similarity queries on the transformed data. Our techniques provide interesting trade-offs between query cost and accuracy. They are then further extended to offer an intuitive privacy guarantee. Empirical studies with real data demonstrate that the techniques are capable of offering privacy while enabling efficient and accurate processing of similarity queries.

## 193. USHER: Improving Data Quality with Dynamic Forms.

### Synopsis:

Data quality is a critical problem in modern databases. data-entry forms present the first and arguably best opportunity for detecting and mitigating errors, but there has been little research into automatic methods for improving data quality at entry time. In this paper, we propose Usher, an end-to-end system for form design, entry, and data quality assurance. Using previous form submissions, Usher learns a probabilistic model over the questions of the form. Usher then applies this model at every step of the data-entry process to improve data quality. Before entry, it induces a form layout that captures the most important data values of a form instance as quickly as possible and reduces the complexity of error-prone questions. During entry, it dynamically adapts the form to the values being entered by providing real-time interface feedback, reasking questions with dubious responses, and simplifying questions by reformulating them. After entry, it revisits question responses that it deems likely to have been entered incorrectly by reasking the question or a reformulation thereof. We evaluate these components of Usher using two real-world data sets. Our results demonstrate that Usher can improve data quality considerably at a reduced cost when compared to current practice.

## 194. Mining Web Graphs for Recommendations.

### Synopsis:

As the exponential explosion of various contents generated on the Web, Recommendation techniques have become increasingly indispensable. Innumerable different kinds of recommendations are made on the Web every day, including movies, music, images, books recommendations, query suggestions, tags recommendations, etc. No matter what types of data sources are used for the recommendations, essentially these data sources can be modeled in the form of various types of graphs. In this paper, aiming at providing a general framework on mining Web graphs for recommendations, (1) we first propose a novel diffusion method which propagates similarities between different nodes and generates recommendations; (2) then we illustrate how to generalize different recommendation problems into our graph diffusion framework. The proposed framework can be utilized in many recommendation tasks on the World Wide Web, including query suggestions, tag recommendations, expert finding, image recommendations, image annotations, etc. The experimental analysis on large data sets shows the promising future of our work.

## 195. The World in a Nutshell Concise Range Queries.

## Synopsis:

With the advance of wireless communication technology, it is quite common for people to view maps or get related services from the handheld devices, such as mobile phones and PDAs. Range queries, as one of the most commonly used tools, are often posed by the users to retrieve needful information from a spatial database. However, due to the limits of communication bandwidth and hardware power of handheld devices, displaying all the results of a range query on a handheld device is neither communication-efficient nor informative to the users. This is simply because that there are often too many results returned from a range query. In view of this problem, we present a novel idea that a concise representation of a specified size for the range query results, while incurring minimal information loss, shall be computed and returned to the user. Such a concise range query not only reduces communication costs, but also offers better usability to the users, providing an opportunity for interactive exploration. The usefulness of the concise range queries is confirmed by comparing it with other possible alternatives, such as sampling and clustering. Unfortunately, we prove that finding the optimal representation with minimum information loss is an NP-hard problem. Therefore, we propose several effective and nontrivial algorithms to find a good approximate result. Extensive experiments on real-world data have demonstrated the effectiveness and efficiency of the proposed techniques.

# 196. Network Coding Based Privacy Preservation against Traffic Analysis in Multi-hop Wireless Networks.

## Synopsis:

Privacy threat is one of the critical issues in multi-hop wireless networks, where attacks such as traffic analysis and flow tracing can be easily launched by a malicious adversary due to the open wireless medium. Network coding has the potential to thwart these attacks since the coding/mixing operation is encouraged at intermediate nodes. However, the simple deployment of network coding cannot achieve the goal once enough packets are collected by the adversaries. On the other hand, the coding/mixing nature precludes the feasibility of employing the existing privacy-preserving techniques, such as Onion Routing. In this paper, we propose a novel network coding based privacy-preserving scheme against traffic analysis in multi-hop wireless networks. With homomorphic encryption on Global Encoding Vectors (GEVs), the proposed scheme offers two significant privacy-preserving features, packet flow untraceability and message content confidentiality, for efficiently thwarting the traffic analysis attacks. Moreover, the proposed scheme keeps the random coding feature, and each sink can recover the source packets by inverting the GEVs with a very high probability. Theoretical analysis and simulative evaluation demonstrate the validity and efficiency of the proposed scheme.

## 197. Query Planning for Continuous Aggregation Queries over a Network of Data Aggregators.

### Synopsis:

Continuous queries are used to monitor changes to time varying data and to provide results useful for online decision making. Typically a user desires to obtain the value of some aggregation function over distributed data items, for example, to know value of portfolio for a client; or the AVG of temperatures sensed by a set of sensors. In these queries a client specifies a coherency requirement as part of the query. We present a low-cost, scalable technique to answer continuous aggregation queries using a network of aggregators of dynamic data items. In such a network of data aggregators, each data aggregator serves a set of data items at specific coherencies. Just as various fragments of a dynamic webpage are served by one or more nodes of a content distribution network, our technique involves decomposing a client query into subqueries and executing subqueries on judiciously chosen data aggregators with their individual subquery incoherency bounds. We provide a technique for getting the optimal set of subqueries with their incoherency bounds which satisfies client query's coherency requirement with least number of refresh messages sent from aggregators to the client. For estimating the number of refresh messages, we build a query cost model which can be used to estimate the number of messages required to satisfy the client specified incoherency bound. Performance results using real-world traces show that our cost-based query planning leads to queries being executed using less than one third the number of messages required by existing schemes.

## 198. Ranking Model Adaptation for Domain-Specific Search.

### Synopsis:

With the explosive emergence of vertical search domains, applying the broad-based ranking model directly to different domains is no longer desirable due to domain differences, while building a unique ranking model for each domain is both laborious for labeling data and time consuming for training models. In this paper, we address these difficulties by proposing a regularization-based algorithm called ranking adaptation SVM (RA-SVM), through which we can adapt an existing ranking model to a new domain, so that the amount of labeled data and the training cost is reduced while the performance is still guaranteed. Our algorithm only requires the prediction from the existing ranking models, rather than their internal representations or the data from auxiliary domains. In addition, we assume that documents similar in the domain-specific feature space should have consistent rankings, and add some constraints to control the margin and slack variables of RA-SVM adaptively. Finally, ranking adaptability measurement is proposed to quantitatively estimate if an existing ranking model can be adapted to a new domain. Experiments performed over Letor and two large scale data sets crawled from a commercial search engine demonstrate the applicabilities of the proposed ranking adaptation algorithms and the ranking adaptability measurement.

## 199. SCALABLE LEARNING OF COLLECTIVE BEHAVIOUR .

### Synopsis:

This study of collective behavior is to understand how individuals behave in a social networking environment. Oceans of data generated by social media like Facebook, Twitter, Flickr, and YouTube present opportunities and challenges to study collective behavior on a large scale. In this work, we aim to learn to predict collective behavior in social media. In particular, given information about some individuals, how can we infer the behavior of unobserved individuals in the same network? A social-dimension-based approach has been shown effective in addressing the heterogeneity of connections presented in social media. However, the networks in social media are normally of colossal size, involving hundreds of thousands of actors. The scale of these networks entails scalable learning of models for collective behavior prediction. To address the scalability issue, we propose an edge-centric clustering scheme to extract sparse social dimensions. With sparse social dimensions, the proposed approach can efficiently handle networks of millions of actors while demonstrating a comparable prediction performance to other nonscalable methods.

## 200. Scalable Scheduling of Updates in Streaming Data Warehouses.

### Synopsis:

We discuss update scheduling in streaming data warehouses, which combine the features of traditional data warehouses and data stream systems. In our setting, external sources push append-only data streams into the warehouse with a wide range of interarrival times. While traditional data warehouses are typically refreshed during downtimes, streaming warehouses are updated as new data arrive. We model the streaming warehouse update problem as a scheduling problem, where jobs correspond to processes that load new data into tables, and whose objective is to minimize data staleness over time (at time t, if a table has been updated with information up to some earlier time r, its staleness is t minus r). We then propose a scheduling framework that handles the complications encountered by a stream warehouse: view hierarchies and priorities, data consistency, inability to preempt updates, heterogeneity of update jobs caused by different interarrival times and data volumes among different sources, and transient overload. A novel feature of our framework is that scheduling decisions do not depend on properties of update jobs (such as deadlines), but rather on the effect of update jobs on data staleness. Finally, we present a suite of update scheduling algorithms and extensive simulation experiments to map out factors which affect their performance.