Latest IEEE 2017-18 Big Data project list

1) SOCIALQ&A: AN ONLINE SOCIAL NETWORK BASED QUESTION AND ANSWER SYSTEM

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

2) PRIVACY-PRESERVING DATA ENCRYPTION STRATEGY FOR BIG DATA IN MOBILE CLOUD COMPUTING

Privacy has become a considerable issue when the applications of big data are dramatically growing in cloud computing. The benefits of the implementation for these emerging technologies have improved or changed service models and improve application performances in various perspectives. However, the remarkably growing volume of data sizes has also resulted in many challenges in practice. The execution time of the data encryption is one of the serious issues during the data processing and transmissions. Many current applications abandon data encryptions in order to reach an adoptive performance level companioning with privacy concerns. In this paper, we concentrate on privacy and propose a novel data encryption approach, which is called Dynamic Data Encryption Strategy (D2ES). Our proposed approach aims to selectively encrypt data and use privacy classification methods under timing constraints. This approach is designed to maximize the privacy protection scope by using a selective encryption strategy within the required execution time requirements.

3) NETSPAM: A NETWORK-BASED SPAM DETECTION FRAMEWORK FOR REVIEWS IN ONLINE SOCIAL MEDIA

Nowadays, a big part of people rely on available con-tent in social media in their decisions (e.g. reviews and feedback on a topic or product). The possibility that anybody can leave a review provide a golden opportunity for spammers to write spam reviews about products and services for different interests. Identifying these spammers and the spam content is a hot topic of research and although a considerable number of studies have been done recently toward this end, but so far the methodologies put forth still barely detect spam reviews, and none of them show the importance of each extracted feature type. In this study, we propose a novel framework, named NetSpam, which utilizes spam features for modeling review datasets as heterogeneous information networks to map spam detection procedure into a classification problem in such networks. Using the importance of spam features help us to obtain better results in terms of different metrics experimented on real-world review datasets from Yelp and Amazon websites.

4) EFFICIENT PROCESSING OF SKYLINE QUERIES USING MAPREDUCE.

The skyline operator has attracted considerable attention recently due to its broad applications. However, computing a skyline is challenging today since we have to deal with big data. For data-intensive applications, the Map Reduce framework has been widely used recently. In this paper, we propose the efficient parallel algorithm SKY-MR+ for processing skyline queries using Map Reduce. We first build a quad tree-based histogram for space partitioning by deciding whether to split each leaf node judiciously based on the benefit of splitting in terms of the estimated execution time. In addition, we apply the dominance power filtering method to effectively prune non-skyline points in advance. We next partition data based on the regions divided by the quad tree and compute candidate skyline points for each partition using Map Reduce.

5) FIDOOP-DP: DATA PARTITIONING IN FREQUENT ITEMSET MINING ON HADOOP CLUSTERS.

Traditional parallel algorithms for mining frequent itemsets aim to balance load by equally partitioning data among a group of computing nodes. We start this study by discovering a serious performance problem of the existing parallel Frequent Itemset Mining algorithms. Given a large dataset, data partitioning strategies in the existing solutions suffer high communication and mining overhead induced by redundant transactions transmitted

among computing nodes. We address this problem by developing a data partitioning approach called FiDoop-DP using the MapReduce programming model. The overarching goal of FiDoop-DP is to boost the performance of parallel Frequent Itemset Mining on Hadoop clusters. At the heart of FiDoop-DP is the Voronoi diagram-based data partitioning technique, which exploits correlations among transactions. Incorporating the similarity metric and the Locality-Sensitive Hashing technique, FiDoop-DP places highly similar transactions into a data partition to improve locality without creating an excessive number of redundant transactions.

6) RESEARCH DIRECTIONS FOR ENGINEERING BIG DATA

Many software startups and research and development efforts are actively taking place to harness the power of big data and create software with potential to improve almost every aspect of human life. As these efforts continue to increase, full consideration needs to be given to engineering aspects of big data software. Since these systems exist to make predictions on complex and continuous massive datasets, they pose unique problems during specification, design, and verification of software that needs to be delivered on-time and within budget. But given the nature of big data software, can this be done? Does big data software engineering really work? This article explores details of big data software, discusses the main problems encountered when engineering big data software, and proposes avenues for future research.

7) SCALABLE DATA CHUNK SIMILARITY BASED COMPRESSION APPROACH FOR EFFICIENT BIG SENSING DATA PROCESSING.

Big sensing data is prevalent in both industry and scientific research applications where the data is generated with high volume and velocity. Cloud computing provides a promising platform for big sensing data processing and storage as it provides a flexible stack of massive computing, storage, and software services in a scalable manner. Current big sensing data processing on Cloud have adopted some data compression techniques. However, due to the high volume and velocity of big sensing data, traditional data compression techniques lack sufficient efficiency and scalability for data processing. Based on specific on-Cloud data compression requirements, we propose a novel scalable data compression approach based on calculating similarity among the partitioned data chunks. Instead of compressing basic data units, the compression will be conducted over partitioned data chunks. To restore original data sets, some restoration functions and predictions will be designed. Map Reduce is used for algorithm implementation to achieve extra scalability on Cloud. With real world meteorological big sensing data experiments

on U-Cloud platform, we demonstrate that the proposed scalable compression approach based on data chunk similarity can significantly improve data compression efficiency with affordable data accuracy loss.

8) USER-CENTRIC SIMILARITY SEARCH

User preferences play a significant role in market analysis. In the database literature there has been extensive work on query primitives, such as the well known top-k query that can be used for the ranking of products based on the preferences customers have expressed. Still, the fundamental operation that evaluates the similarity between products is typically done ignoring these preferences. Instead products are depicted in a feature space based on their attributes and similarity is computed via traditional distance metrics on that space. In this work we utilize the rankings of the products based on the opinions of their customers in order to map the products in a user-centric space where similarity calculations are performed. We identify important properties of this mapping that result in upper and lower similarity bounds, which in turn permit us to utilize conventional multidimensional indexes on the original product space in order to perform these user-centric similarity computations.

9) SECURE BIG DATA STORAGE AND SHARING SCHEME FOR CLOUD TENANTS.

The Cloud is increasingly being used to store and process big data for its tenants and classical security mechanisms using encryption are neither sufficiently efficient nor suited to the task of protecting big data in the Cloud. In this paper, we present an alternative approach which divides big data into sequenced parts and stores them among multiple Cloud storage service providers. Instead of protecting the big data itself, the proposed scheme protects the mapping of the various data elements to each provider using a trapdoor function.

10) EFFICIENT RECOMMENDATION OF DE-IDENTIFICATION POLICIES USING MAPREDUCE.

Many data owners are required to release the data in a variety of real world application, since it is of vital importance to discovery valuable information stay behind the data. However, existing re-identification attacks on the AOL and ADULTS datasets have shown that publish such data directly may cause tremendous threads to the individual privacy. Thus, it is urgent to resolve all kinds of re-identification risks by recommending effective de-identification policies to guarantee both privacy and utility of the data. De-identification policies is one of the models that can be used to achieve such requirements, however, the number of de-identification policies is exponentially large due to the broad domain of quasi-identifier attributes. To better control the trade off between data utility and data privacy, skyline computation can be used to select such policies, but it is yet challenging for efficient skyline processing over large number of policies. In this paper, we propose one parallel algorithm called SKY-FILTER-MR, which is based on MapReduce to overcome this challenge by computing skylines over large scale de-identification policies that is represented by bit-strings. To further improve the performance, a novel approximate skyline computation scheme was proposed to prune unqualified policies using the approximately domination relationship. With approximate skyline, the power of filtering in the policy space generation stage was greatly strengthened to effectively decrease the cost of skyline computation over alternative policies.

11) HIERARCHY-CUTTING MODEL BASED ASSOCIATION SEMANTIC FOR ANALYZING DOMAIN TOPIC ON THE WEB.

Association link network (ALN) can organize massive web information to provide many intelligent services in the era of Big Data. Effective semantic layered technology not only can provide theoretical support for knowledge discovery in Web resources, but also can improve the searching efficiency of the related information system such as Web information system and industrial information system. How to realize the layer division of association semantic by the hierarcy analysis of ALN is an important research topic. To solve this problem, this paper proposes a hierarchy-cutting model of association semantic. First, some experiments of four types of keywords with different linking roles are conducted to discover the possible distribution law. Experimental results show that these keywords with association role reveal previous power-law distribution. Then, based on the discovered power-law distribution, up-cutting and down-cutting points are presented to divide the association semantic into three layers. At the same time, the theories of hierarcy-cutting model are presented. Finally, the examples of the current core topic and permernent topics belonging to a domain are given. The experiments show that hierarcy-cutting points have high accuracy. The multilayer theory of association semantic can provide a theoretical support for knowledge recommendation with different particle sizes on ALNs.

12) LARGE-SCALE MULTI-MODALITY ATTRIBUTE REDUCTION WITH MULTI-KERNEL FUZZY ROUGH SETS.

In complex pattern recognition tasks, objects are typically characterized by means of multi-modality attributes, including categorical, numerical, text, image, audio and even videos. In these cases, data are usually high dimensional, structurally complex, and granular. Those attributes exhibit some redundancy and irrelevant information. The evaluation, selection, and combination of multi-modality attributes pose great challenges to traditional classification algorithms. Multi-kernel learning handles multi-modality attributes by using di_erent kernels to extract information coming from di_erent attributes. However, it cannot consider the aspects fuzziness in fuzzy classification. Fuzzy rough sets emerge as a powerful vehicle to handle fuzzy and uncertain attribute reduction. In this paper, we design a framework of multi-modality attribute reduction based on multi-kernel fuzzy rough sets. First, a combination of kernels based on set theory is defined to extract fuzzy similarity for fuzzy classification with multi-modality attributes. Then, a model of multi-kernel fuzzy rough sets is constructed.

13) A SECURE AND VERIFIABLE ACCESS CONTROL SCHEME FOR BIG DATA STORAGE IN CLOUDS.

Due to the complexity and volume, outsourcing ciphertexts to a cloud is deemed to be one of the most effective approaches for big data storage and access. Nevertheless, verifying the access legitimacy of a user and securely updating a ciphertext in the cloud based on a new access policy designated by the data owner are two critical challenges to make cloud-based big data storage practical and effective. Traditional approaches either completely ignore the issue of access policy update or delegate the update to a third party authority; but in practice, access policy update is important for enhancing security and dealing with the dynamism caused by user join and leave activities. In this paper, we propose a secure and verifiable access control scheme based on the NTRU cryptosystem for big data storage in clouds. We first propose a new NTRU decryption algorithm to overcome the decryption failures of the original NTRU, and then detail our scheme and analyze its correctness, security strengths, and computational efficiency.

14) QUESTION QUALITY ANALYSIS AND PREDICTION IN COMMUNITY QUESTION ANSWERING SERVICES WITH COUPLED MUTUAL REINFORCEMENT.

Community Question Answering Services (CQAS) (e.g., Yahoo! Answers) provides a platform where people post questions and answer questions posed by others. Previous works analyzed the Answer Quality (AQ) based on answer-related features, but neglect the question-related features on AQ. Previous work analyzed how asker- and question-related features affect the Question Quality (QQ) regarding the amount of attention from users, the number of answers and the question solving latency, but neglect the correlation between QQ and AQ (measured by the rating of the best answer), which is critical to Quality of Service (QoS). We handle this problem
from two aspects. First, we additionally use QQ in measuring AQ, and analyze the correlation between a comprehensive list of features (including answer-related features) and QQ. Second, we propose the first method that estimates the probability for a given question to obtain high AQ. Our analysis on the Yahoo! Answers trace confirmed that the list of our identified features exert influence on AQ, which determines QQ. For the correlation analysis, the previous classification algorithms cannot consider the mutual interactions between multiple (>2) classes of features. We then propose a novel Coupled Semi-Supervised Mutual Reinforcement-based Label Propagation (CSMRLP) algorithm for this purpose. algorithms in the accuracy of AQ classification, and the effectiveness of our proposed method in AQ prediction. Finally, we provide suggestions on how to create a question that will receive high AQ, which can be exploited to improve the QoS of CQAS.

15)FIDOOP: PARALLEL MINING OF FREQUENT ITEMSETS USING MAPREDUCE

Existing parallel mining algorithms for frequent item sets lack a mechanism that enables automatic parallelization, load balancing, data distribution, and fault tolerance on large clusters. As a solution to this problem, we design a parallel frequent item sets mining algorithm called FiDoop using the

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

16) SENTIMENT ANALYSIS OF TOP COLLEGES USING TWITTER DATA.

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

17) THE SP THEORY OF INTELLIGENCE: DISTINCTIVE FEATURES AND ADVANTAGE.

Community question answering services (CQAS) (e.g., Yahoo! Answers) provides a platform where people post questions and answer questions posed by others. Previous works analyzed the answer quality (AQ) based on answer-related features, but neglect the question-related features on AQ. Previous work analyzed how asker- and question-related features affect the question quality (QQ) regarding the amount of attention from users, the number of answers and the question solving latency, but neglect the correlation between QQ and AQ (measured by the rating of the best answer), which is critical to quality of service (QoS). We handle this problem from two aspects. First, we additionally use QQ in measuring AQ, and analyze the correlation between a comprehensive list of features (including answer-related features) and QQ. Second, we propose the first method that estimates the probability for a given question to obtain high AQ. Our analysis on the Yahoo! Answers trace confirmed that the list of our identified features exert influence on AQ, which determines QQ. For the correlation analysis, the previous classification algorithms cannot consider the mutual interactions between multiple (>2) classes of features. We then propose a novel Coupled Semi-Supervised Mutual Reinforcement-based Label Propagation (CSMRLP) algorithm for this purpose. Our extensive experiments show that CSMRLP outperforms the Mutual Reinforcement-based Label Propagation (MRLP) and five other traditional classification algorithms in the accuracy of AQ classification, and the effectiveness of our proposed method in AQ prediction. Finally, we provide suggestions on how to create a question that will receive high AQ, which can be exploited to improve the QoS of CQAS.

18) ON TRAFFIC-AWARE PARTITION AND AGGREGATION IN MAPREDUCE FOR BIG DATA APPLICATIONS.

The MapReduce programming model simplifies large-scale data processing on commodity cluster by exploiting parallel map tasks and reduce tasks. Although many efforts have been made to improve the performance of MapReduce jobs, they ignore the network traffic generated in the shuffle phase, which plays a critical role in performance enhancement. Traditionally, a hash function is used to partition intermediate data among reduce tasks, which, however, is not traffic-efficient because network topology and data size associated with each key are not taken into consideration. In this paper, we study to reduce network traffic cost for a MapReduce job by designing a novel intermediate data partition scheme. Furthermore, we jointly consider the aggregator placement problem, where each aggregator can reduce merged traffic from multiple map tasks. A decomposition-based distributed algorithm is proposed to deal with the large-scale optimization problem for big data application and an online algorithm is also designed to adjust data partition and aggregation in a dynamic manner. Finally, extensive simulation results demonstrate that our proposals can significantly reduce network traffic cost under both offline and online cases.

19) Protection of Big Data Privacy

In recent years, big data have become a hot research topic. The increasing amount of big data also increases the chance of breaching the privacy of individuals. Since big data require high computational power and large storage, distributed systems are used. As multiple parties are involved in these systems, the risk of privacy violation is increased. There have been a number of privacy-preserving mechanisms developed for privacy protection at different stages (e.g., data generation, data storage, and data processing) of a big data life cycle. The goal of this paper is to provide a comprehensive overview of the privacy preservation mechanisms in big data and present the challenges for existing mechanisms. In particular, in this paper, we illustrate the infrastructure of big data and the state-of-the-art privacy-preserving mechanisms in each stage of the big data life cycle. Furthermore, we discuss the challenges and future research directions related to privacy preservation in big data.