| IEEE 2018-19 PROJECT LIST | |
|---|---|
| DATA MINING | |
| CODE | TITLE AND ABSTRACT |
| 19ANSP-DM-01 | **Mining Online Discussion Data for Understanding Teachers’ Reflective Thinking**<br>***Abstract-***<br>Teachers’ online discussion text data shed light on their reflective thinking. With the growing scale of text data, the traditional way of manual coding, however, has been challenged. In order to process the large-scale unstructured text data, it is necessary to integrate the inductive content analysis method and educational data mining techniques. An inductive content analysis on samples taken from 17,624 posts was implemented and the categories of teachers’ reflective thinking were obtained. Based on the results of inductive content analysis, we implemented a single-label text classification algorithm to classify the sample data. Then, we applied the trained classification model on a large-scale and unexplored online discussion text data set and two types of visualizations of the results were provided. By using the categories gained from inductive content analysis to create a radar map, teachers’ reflection level was represented. In addition, a cumulative adjacency matrix was created to characterize the evolution of teachers’ reflective thinking. This study could partly explain how teachers reflected in online professional learning environments and brought awareness to educational policy makers, teacher training managers, and education researchers. |
| 19ANSP-DM -002 | **Privacy preserving big data mining: association rule hiding using fuzzy logic approach**<br>***Abstract—***<br>Recently, privacy preserving data mining has been studied widely. Association rule mining can cause potential threat toward privacy of data. So, association rule hiding techniques are employed to avoid the risk of sensitive knowledge leakage. Many researches have been done on association rule hiding, but most of them focus on proposing algorithms with least side effect for static databases (with no new data entrance), while now the authors confront with streaming data which are |

| | |
|---|---|
| | continuous data. Furthermore, in the age of big data, it is necessary to optimize existing methods to be executable for large volume of data. In this study, data anonymization is used to fit the proposed model for big data mining. Besides, special features of big data such as velocity make it necessary to consider each rule as a sensitive association rule with an appropriate membership degree. Furthermore, parallelization techniques which are embedded in the proposed model, can help to speed up data mining process. |
| 19ANSP-DM -003 | **Using Data Mining to Predict Hospital Admissions from the Emergency Department**<br>***Abstract***<br>Crowding within emergency departments (EDs) can have significant negative consequences for patients. EDs therefore need to explore the use of innovative methods to improve patient flow and prevent overcrowding. One potential method is the use of data mining using machine learning techniques to predict ED admissions. This paper uses routinely collected administrative data (120 600 records) from two major acute hospitals in Northern Ireland to compare contrasting machine learning algorithms in predicting the risk of admission from the ED. We use three algorithms to build the predictive models: 1) logistic regression; 2) decision trees; and 3) gradient boosted machines (GBM). The GBM performed better (accuracy D 80:31%, AUC-ROC D 0:859) than the decision tree (accuracy D 80:06%, AUC-ROC D 0:824) and the logistic regression model (accuracy D 79:94%, AUC-ROC D 0:849). Drawing on logistic regression, we identify several factors related to hospital admissions, including hospital site, age, arrival mode, triage category, care group, previous admission in the past month, and previous admission in the past year. This paper highlights the potential utility of three common machine learning algorithms in predicting patient admissions. Practical implementation of the models developed in this paper in decision support tools would provide a snapshot of predicted admissions from the ED at a given time, allowing for advance resource planning and the avoidance bottlenecks in patient _ow, as well as comparison of predicted and actual admission rates. When interpretability is a key consideration, EDs should consider adopting logistic regression models, although GBM's will be useful where accuracy is paramount. |
| 19ANSP-DM-004 | **Corporate Communication Network and Stock Price Movements: Insights from Data Mining**<br>***Abstract*** |

| | |
|---|---|
| | Grounded on communication theories, we propose to use a data-mining algorithm to detect communication patterns within a company to determine if such patterns may reveal the performance of the company. Specifically, we would like to find out whether or not there exist any association relationships between the frequency of e-mail exchange of the key employees in a company and the performance of the company as reflected in its stock prices. If such relationships do exist, we would also like to know whether or not the company's stock price could be accurately predicted based on the detected relationships. To detect the association relationships, a data-mining algorithm is proposed here to mine e-mail communication records and historical stock prices so that based on the detected relationship, rules that can predict changes in stock prices can be constructed. Using the data-mining algorithm and a set of publicly available Enron e-mail corpus and Enron's stock prices recorded during the same period, we discovered the existence of interesting, statistically significant, association relationships in the data. In addition, we also discovered that these relationships can predict stock price movements with an average accuracy of around 80%. The results confirm the belief that corporate communication has identifiable patterns and such patterns can reveal meaningful information of corporate performance as reflected by such indicators as stock market performance. Given the increasing popularity of social networks, the mining of interesting communication patterns could provide insights into the development of many useful applications in many areas. |
| 19ANSP-DM-005 | **Systematic Approach to Analyze Travel Time in Road-Based Mass Transit Systems Based on Data Mining**<br>***Abstract***<br>Road-based mass transit systems are an effective means to combat the negative impact of transport that is based on private vehicles. Providing quality of service in this type of transit system is a priority for transport authorities. In these systems, travel time (TT) is a basic factor in quality of service. This paper presents a methodology, based on data mining, for analyzing TT in a mass transit system that is planned by timetable. The objective of the methodology is to understand the behavior patterns of TTs on the different routes of the transport network, as well as the factors that influence these patterns. To achieve this objective, the methodology uses clustering techniques to process the GPS data provided by the vehicles of the public transport fleet. The results that were obtained when implementing this methodology in a public transport company are presented as a use case, demonstrating its validity. |

| | |
|---|---|
| 19ANSP-DM-006 | **KnowEdu: A System to Construct Knowledge Graph for Education**<br>***Abstract***<br>Motivated by the vast applications of knowledge graph and the increasing demand in education domain, we propose a system, called *KnowEdu*, to automatically construct knowledge graph for education. By leveraging on heterogeneous data (e.g., pedagogical data and learning assessment data) from the education domain, this system first extracts the concepts of subjects or courses and then identifies the educational relations between the concepts. More specifically, it adopts the neural sequence labeling algorithm on pedagogical data to extract instructional concepts and employs probabilistic association rule mining on learning assessment data to identify the relations with educational significance. We detail all the above-mentioned efforts through an exemplary case of constructing a demonstrative knowledge graph for mathematics, where the instructional concepts and their prerequisite relations are derived from curriculum standards and concept-based performance data of students. Evaluation results show that the F1 score for concept extraction exceeds 0.70, and for relation identification, the area under the curve and mean average precision achieve 0.95 and 0.87, respectively. |
| 19ANSP- DM -007 | **Frequent Itemsets Mining with Differential Privacy Over Large-Scale Data**<br>***Abstract:***<br>Frequent itemsets mining with differential privacy refers to the problem of mining all frequent itemsets whose supports are above a given threshold in a given transactional dataset, with the constraint that the mined results should not break the privacy of any single transaction. Current solutions for this problem cannot well balance efficiency, privacy, and data utility over large-scale data. Toward this end, we propose an efficient, differential private frequent itemsets mining algorithm over large-scale data. Based on the ideas of sampling and transaction truncation using length constraints, our algorithm reduces the computation intensity, reduces mining sensitivity, and thus improves data utility given a fixed privacy budget. Experimental results show that our algorithm achieves better performance than prior approaches on multiple datasets. |
| 19ANSP- DM -008 | **HPPQ: A Parallel Package Queries Processing Approach for Large-Scale Data**<br>***Abstract:*** |

| | |
|---|---|
| | A lot of scholars have focused on developing effective techniques for package queries, and a lot of excellent approaches have been proposed. Unfortunately, most of the existing methods focus on a small volume of data. The rapid increase in data volume means that traditional methods of package queries find it difficult to meet the increasing requirements. To solve this problem, a novel optimization method of package queries (HPPQ) is proposed in this paper. First, the data is preprocessed into regions. Data preprocessing segments the dataset into multiple subsets and the centroid of the subsets is used for package queries, this effectively reduces the volume of candidate results. Furthermore, an efficient heuristic algorithm is proposed (namely IPOL-HS) based on the preprocessing results. This improves the quality of the candidate results in the iterative stage and improves the convergence rate of the heuristic algorithm. Finally, a strategy called HPR is proposed, which relies on a greedy algorithm and parallel processing to accelerate the rate of query. The experimental results show that our method can significantly reduce time consumption compared with existing methods. |
| 19ANSP- DM -009 | **Natural Neighborhood-Based Classification Algorithm Without Parameter k**<br>***Abstract:***<br>Various kinds of k-Nearest Neighbor (KNN) based classification methods are the bases of many well established and high-performance pattern recognition techniques. However, such methods are vulnerable to parameter choice. Essentially, the challenge is to detect the neighborhood of various datasets while ignoring the data characteristics. This article introduces a new supervised classification algorithm, Natural Neighborhood Based Classification Algorithm (NNBCA). Findings indicate that this new algorithm provides a good classification result without artificially selecting the neighborhood parameter. Unlike the original KNN-based method, which needs a prior k, NNBCA predicts different k for different samples. Therefore, NNBCA is able to learn more from flexible neighbor information both in the training and testing stages. Thus, NNBCA provides a better classification result than other methods. |
| 19ANSP- DM -010 | **A Goal-oriented Requirement Analysis Approach for the Selection of Data Mining Techniques for Non-Expert Users**<br>***Abstract:***<br>The application of data mining techniques for obtaining knowledge has historically required the intervention of experts to obtain satisfactory results. This paper presents a solution proposal for the complex topic of |

| | |
|---|---|
| | identifying the requirements of non-expert users when trying to perform data mining techniques. The modeling language for objective-oriented requirements analysis i * (i star) has been used to facilitate the use of a taxonomy of requirements. As a result, it is intended that nonexpert users can represent their requirements without having relevant knowledge of data mining techniques. The application in a case study allows as proof of concept, validate the proposed model. |
| 19ANSP- DM -011 | **Rapid-Response Framework for Defensive Driving Based on Internet of Vehicles Using Message-Oriented Middleware**<br>***Abstract:***<br>With the rapid development of automatic driving and advanced driver assistance systems, vehicle safety has improved greatly. These systems mainly use sensors installed on the vehicle and help drivers deal with human operation errors. Traffic accidents are inherently unpredictable, and it is difficult to prevent mistakes made by others. Therefore, the concept of defensive driving has attracted much interest. Defensive driving aims to increase drivers' self-awareness to prevent accidents. Future self-driving vehicles should integrate defensive driving to improve driver safety. This paper proposes a framework based on the risk evaluation value of defensive driving that rapidly transmits information about high-accident-likelihood zones to drivers or vehicles by using Internet of Vehicles technology. This should enable drivers or self-driving vehicles to predict risks and operate vehicles safely. To send alert messages in a timely manner, it is essential to overcome the challenge of processing real-time data during driving. We design five kinds of services in this rapid response framework, including raw data receiver, warning area decision, accident pattern recognition, message generator, and user profile to analyze driver information using distributed system architecture. Message-oriented middleware is used for communication between services. This framework identifies high-accident-likelihood zones by using density-based spatial clustering of applications with noise, simplifying the process of association calculation. After the calculation, this framework uses the weighted severity index to weight and compare risk severities. According to our experimental results,<br>the service-oriented middleware design increases the speed and stability of information transmission. |
| 19ANSP- DM -012 | **From Latency, Through Outbreak, to Decline: Detecting Different States of Emergency Events Using Web** |

| | |
|---|---|
| | **Resources**<br>***Abstract:***<br>An emergency event is a sudden, urgent, usually unexpected incident or occurrence that requires an immediate reaction or assistance for emergency situations, which plays an increasingly important role in the global economy and in our daily lives. Recently, the web is becoming an important event information provider and repository due to its real-time, open, and dynamic features. In this paper, web resources-based states detecting algorithm of an event is developed in order to let the people know of an emergency event clearly and help the social group or government process the emergency events effectively. The relationship between web and emergency events is first introduced, which is the foundation of using web resources to detect the state of emergency events imaged on the web. Second, five temporal features of emergency events are developed to provide the basis for state detection. Moreover, the outbreak power and the fluctuation power are presented to integrate the above temporal features for measuring the different states of an emergency event. Using these two powers, an automatic state detecting algorithm for emergency events is proposed. In addition, heuristic rules for detecting the states of emergency event on the web are discussed. Our evaluations using real-world data sets demonstrate the utility of the proposed algorithm, in terms of performance and effectiveness in the analysis of emergency events. |
| 19ANSP- DM -013 | **Web Media and Stock Markets: A Survey and Future Directions from a Big Data Perspective**<br>***Abstract:***<br>Stock market volatility is influenced by information release, dissemination, and public acceptance. With the increasing volume and speed of social media, the effects of Web information on stock markets are becoming increasingly salient. However, studies of the effects of Web media on stock markets lack both depth and breadth due to the challenges in automatically acquiring and analyzing massive amounts of relevant information. In this study, we systematically reviewed 229 research articles on quantifying the interplay between Web media and stock markets from the fields of Finance, Management Information Systems, and Computer Science. In particular, we first categorized the representative works in terms of media type and then summarized the core techniques for converting textual information into machine-friendly forms. Finally, we compared the analysis models used to capture the hidden relationships between Web media and stock movements. Our goal |

| | |
|---|---|
| | is to clarify current cutting-edge research and its possible future directions to fully understand the mechanisms of Web information percolation and its impact on stock markets from the perspectives of investors cognitive behaviors, corporate governance, and stock market regulation. |
| 19ANSP- DM -014 | **Event Detection and Identification of Influential Spreaders in Social Media Data Streams**<br>***Abstract:***<br>Microblogging, a popular social media service platform, has become a new information channel for users to receive and exchange the most up-to-date information on current events. Consequently, it is a crucial platform for detecting newly emerging events and for identifying influential spreaders who have the potential to actively disseminate knowledge about events through microblogs. However, traditional event detection models require human intervention to detect the number of topics to be explored, which significantly reduces the efficiency and accuracy of event detection. In addition, most existing methods focus only on event detection and are unable to identify either influential spreaders or key event-related posts, thus making it challenging to track momentous events in a timely manner. To address these problems, we propose a Hypertext-Induced Topic Search (HITS) based Topic-Decision method (TD-HITS), and a Latent Dirichlet Allocation (LDA) based Three-Step model (TS-LDA). TDHITS can automatically detect the number of topics as well as identify associated key posts in a large number of posts. TS-LDA can identify influential spreaders of hot event topics based on both post and user information. The experimental results, using a Twitter dataset, demonstrate the effectiveness of our proposed methods for both detecting events and identifying influential spreaders. |