CHAPTER 19

COMPUTER SCIENCE

Doctoral Theses

01. GROVER (Sapna)
Approximation Algorithms for Facility Location Problem and its Variants.
Supervisor: Prof. Neelima Gupta
Th 28247

Abstract

In {\em facility location problem} (FL), we are given a set $\clientset$ of clients, each with a non-negative request $r: \clientset \rightarrow \mathcal{I}^+ \cup \{0\}$ denoted as r_j for $j \in \clientset$, a set $\facilityset$ of facilities each with a location-specific opening cost $f: \facilityset \rightarrow \mathcal{R}^+ \cup \{0\},$ called {\em facility opening cost}, denoted as f_i for $i \in \facilityset$ and a metric cost function $c: \facilityset \cup \clientset \times \facilityset \cup \clientset \rightarrow \mathcal{R}^+ \cup \{0\}$, a solution $S = (\facilityset`, \sigma)$ to the problem consists of a subset $\facilityset` \subseteq \facilityset$ of opened facilities and an assignment function $\sigma: \clientset \rightarrow \facilityset`$ that assigns the requests of clients to the facilities in $\facilityset`$. $\dist{j}{\sigma(j)}$ then denotes the cost of serving one unit of request of client j by $\sigma(j)$ and is called the {\em assignment cost} or the {\em service cost} of j. Each such solution is associated with a cost value, which is the sum of total facility opening cost and service cost paid by the clients. The objective is to obtain a solution with minimum total cost. The problem is NP-hard even when the cost function is metric (i.e., it satisfies triangle inequality). Various approximation algorithms have been developed for the problem. However, constraints occur naturally in real world which makes the problem harder. The most common occurring natural constraints are an {\em upper bound} and a {\em lower bound}. Upper bound (/capacity) $\capacity_i$ places an upper limit on the maximum amount of client request a facility i can serve. Lower bound $\B_i$ denotes the minimum amount of client request a facility i must serve. In this thesis, we study variants of facility location problem with lower and/or upper bounds using LP rounding, reduction and combination techniques. In particular, we develop approximation algorithms for the facility location problem, the k-median problem and their generalizations with lower and/or upper bounds. We first study facility location problem and present the following results: \begin{enumerate} \item An $O(1/\epsilon)$ factor approximation algorithm for FL with uniform upper bounds (UFL). The result, obtained using LP rounding, violates the upper bounds by a small factor of ($1+\epsilon$) for a fixed $\epsilon >0$. Though true approximation algorithms are known for the problem for general capacities, our result is interesting as the approach is simple and more time-efficient than the true approximation algorithms obtained using strengthened LP and local search. Our result shows that the LP is not too bad. \item an approximation algorithm for LUFL when one of the bounds is

uniform. The result violates the upper bounds by a factor of $(\beta+1)$ where β is the violation in upper bounds in the solution of UFL. \end{enumerate}.

Contents

1. Introduction 2. Relatedwork 3. Facility location problem with uniform upper bounds 4. Knapsack median problem with uniform upper bounds 5. Introducing lower bounds in UFL 6. Frame work for LUkFL 7. Conclusion.

02. MEENA (Sunil Kumar)
Advancing Influence Maximization in Social Networks through Diversity and Diffusion Dynamics.
Supervisors: Dr. Kuldeep Singh and Dr. S. S. Singh
Th 28248

Abstract

Influence maximization (IM) is a discrete optimization problem aiming to select the top k nodes that maximize the influence. Diversity in influence is important in various applications and removes the overlap influence as well. Along with diversity, inactive nodes contribute to the diffusion process in diffusion dynamics. Further, privacy also needs to be preserved in the diffusion process. The existing IM algorithm primarily focuses on the static social network and ignores the diversity. Also, do not consider the role of inactive nodes and do not preserve privacy in the diffusion process. To bridge the gap for real-world applications in IM, we have addressed the diversified. To address the diversified IM, this research introduces novel algorithms by utilizing the proposed objective functions and strategies in dynamic social networks. We have suggested objective functions to achieve diversity. First, in Dynamic Community Diversified Influence Maximization (DCDIM) algorithm, we propose the objective function that jointly considers the number of influenced nodes and the number of influenced communities. This work further shows the Monotone, Submodular, and NP-Hard properties of objective functions and validates the algorithm through experiments. Second, in the Dynamic Community Diversified Influence Maximization using Bridge nodes (DCDIMB) algorithm, this work utilizes and investigates the role of bridge nodes in enhancing diversity. The experimental results show that the proposed algorithm improves the reach and effectiveness of IM. Third, limitations in the DCDIM and DCDIMB algorithms include the consideration of equal cost of the nodes, high complexity in selecting the seed nodes, and imbalance results. To address these, this work proposed a Diversified Budgeted Influence Maximization (DBIM) algorithm that integrates the diversified and budgeted IM. To achieve the objectives, this work proposes an objective function and shows the Monotone, Submodular, and NP-Hard properties of the objective function. The proposed work validated the results through experimental analysis. Along with diversified IM, this work studies the diffusion model to make IM advance. To bridge the gap, this work proposes the Extended Linear Threshold Model (ELTM) and Extended Linear Threshold Model with Privacy (ELTMP) models. In the ELTM model, this work considers the role of inactive nodes with the control parameters. We further showed the properties of the proposed objective function under the ELTM models. The experimental results conclude that the ELTM model has a higher number of influenced nodes due to the weightage of inactive nodes. To consider privacy in the

ELTMP model, this work introduces a threshold-based revealing mechanism. This ensures that influence weights are revealed only when the probability of exposure surpasses a defined threshold. The experimental results show that the ELTMP has a lower number of influenced nodes as some weights are not revealed.

Contents

1. Introduction 2. Related work 3. Diversified influence maximization diversified influence maximization using bridge nodes 5. Budgeted diversified influence maximization 6. Diffusion models considering inactive nodes and privacy 7. Conclusion.

03. NISHA

Privacy Preserving Techniques for Educational Data.

Supervisors: Prof. Archana Singhal and Prof. Sunil K Muttoo

Th 28249

Abstract

Educational data is available in today's world in abundance. Valuable information can be extracted from educational data with the help of analytical tools that can benefit the stakeholder. It can be leveraged to improve students' performance based on their academic records to predict their future performances. For harvesting valuable information from educational data, it needs to be shared with third party for analysis purpose. Data sharing without intruding the privacy of individuals is a major concern. One of the most popular privacy-preserving technique is data publishing, in which, the data curator anonymizes the dataset and shares the anonymized data with the data analyst. First technique proposed in the current work is an improved privacy preserving k-anonymization Cluster-based technique for a multi-relational educational dataset. To overcome the limitations of k-Anonymization, anonymized data is further l-diversified to protect sensitive data from attacks. Text Steganography is applied to sensitive attribute before k-anonymization to avoid similarity attack. Since the utility of data is an important factor, it must be maintained along with privacy to get useful information from the analysis. A Loss Metric is used to find the distortion of k-anonymized data to evaluate the balance between privacy and utility. Earth's mover distance has been calculated to validate the results. Another privacy-preserving technique suggested in this work is ε-differential privacy for securing the data. The benefit of this technique is that background knowledge, homogeneity and similarity attacks cannot be applied on the data. In this thesis, a predictive model with privacy-preserving algorithms has been used which provides future insights as well as provide two-layers of privacy to the sensitive information of individuals. Educational dataset has been used in this work for the experimental purpose.

Contents

1. Introduction 2. Related work 3. Anonymization using multi-relational datase 4. Privacy preserving technique using steganography and ldiversity 5. An improved differential privacy preserving technique for educational dataset 6. A privacy-preserving framework for predictive modelling using

machine learning 7. Conclusion and future directions. References and list of publications.

04. RAJNI
Approximation Algorithms for Capacitated Facility Location with Outliers.
Supervisor: Prof. Neelima Gupta
Th 28250

Abstract

In the thesis, we study the Capacitated Facility Location with Outliers (CFLO) problem. In this problem, we are given a set of clients and a set of potential facilities. Each facility has an associated opening cost and a capacity that limits the number of clients it can serve. For every client and facility, the cost of serving the client by the facility is given by the distance between them (service cost). Additionally, we are provided with an upper bound on the number of clients that can remain unserved, referred to as "outliers." The objective is to determine which facilities to open and which clients to serve in order to minimize the total cost, which includes both the facility opening costs of the selected facilities and the service costs of the selected clients, while satisfying the capacity and outlier constraints. We assume that the distances form a metric. We first study the case when the facility opening costs are uniform, meaning all facilities have the same opening cost. For this special case, we present a (6.373 + \epsilon)-approximation algorithm using a 2-operation local search approach, where \epsilon > 0 is a fixed constant. To the best of our knowledge, this is the first constant-factor approximation for this problem. Furthermore, a simplified version of our local search algorithm and analysis leads to a (3.733 + \epsilon)-approximation algorithm for the Capacitated Facility Location with Uniform Facility Cost problem, improving the current best-known factor of 4 by Kao[ISAAC`24]. Next, we relax the uniformity assumption on facility costs and explore LP-based algorithms for the Uniform Capacitated Facility Location with Outliers problem, where all facilities have the same capacity. We present a tri-criteria approximation, where the solution approximates the cost up to a constant factor, while allowing small violations in both the capacity and outlier constraints. In particular, we give a O(1/\epsilon^2) factor approximation for the problem violating both capacities and outliers by (1 + \epsilon), for a fixed constant \epsilon>0. We then study Capacitated k-Median with Outliers (CkMO) problem, which is similar to the CFLO but instead of facility opening costs, it imposes a hard constraint on the number of facilities that can be opened. We obtain a (3 + \epsilon)-approximation algorithm for CkMO, which runs in FPT time with respect to k, the number of outliers, and \epsilon.

Contents

1. Introduction 2. Related work 3. (6.373+€)-approximation for cflo with uniform facility openingcosts 4. Tri-criteriaforcflowithuniformcapacities 5. (3+€)-fptapproximationforckmo 6. Conclusion.

05. SHIVANI
Explainable Multiple Disease Diagnosis System using Agentic AI on Multimodal Knowledge Graph.
Supervisor: Prof. Punam Bedi
Th 28788

Abstract

Healthcare is one of the fundamental needs of people. Providing access to quality healthcare services is crucial as it enhances the quality of life, reduces mortality rates, and increases life expectancy. Disease diagnosis is a process that determines the disease causing the unusual symptoms in a patient's body. It is the primary and most critical component of a healthcare system, as the treatment is only prescribed after disease diagnosis. Misdiagnosis (identifying the wrong disease) can worsen the patient's condition due to the adverse effects of wrong medication suggested based on a misdiagnosis. Hence, identifying the correct disease is vital for the patient's treatment. This work aims to assist medical practitioners in making accurate disease diagnosis using Artificial Intelligence (AI) techniques for timely clinical decisions. Despite many advances in AI-based disease diagnosis systems, challenges such as processing unstructured and multimodal clinical data, detecting multiple diseases in a patient, and providing explanations still require the attention of AI researchers. This work addresses these challenges by proposing Clinical Narratives to Knowledge Graph (N2K) Mapper, which maps unstructured patient clinical data to construct a Knowledge Graph (KG), which is used for single and multiple disease diagnosis. Further to improve the accuracy of disease diagnosis, multimodal patient clinical data, such as medical images, demographics, and vitals, have been added to the KG. A Modified-Bidirectional Encoder Representation from Transformers is applied to make multiple disease predictions based on the semantics in the KG. The proposed XLR-KGDD framework uses Large Language Models (LLMs) and Retrieval Augmented Generation to generate precise explanations in natural language. A web application has been developed based on the Agentic AI disease diagnosis framework with explanations. The application consists of multiple autonomous agents that are backed by the LLMs, can learn proactively, make decisions, and adapt to the environment with minimal human intervention.

Contents

1. Introduction 2. Background concepts 3. Knowledge graph enrichment from clinical information 4. Multiple disease diagnosis using heterogeneous ehr curated knowledge graph and machine learning models 5. Multiple disease diagnosis using multimodal ehr curated knowledge graph and a semantically modified bert model 6. Explainable agentic ai system for disease diagnosis using llm, rag and knowledge graph 7. Conclusion and directions for future work 8. References.

06. SINGH (Onkar)
Hybrid Techniques for Generating Cancelable Biometric Templates.
Supervisors: Prof. Naveen Kumar and Prof. Ajay Jaiswal
Th 28251

Abstract

Cancelable biometrics enhance the security and privacy of biometric authentication by converting biometric data into non-invertible templates. However, achieving non-invertibility often compromises discriminability. This thesis presents four research studies presenting hybrid techniques for generating cancelable biometric templates offering better recognition

accuracy while maintaining non-invertibility. The first technique is based on random permutation and linear regression. For each biometric image, a virtual image is generated using linear regression, which is then randomly permuted to produce the cancelable template. This virtual representation reduces intra-user variation, improving classification accuracy. The second work introduces two methods built using eigenfeature regularization. Initially, biometric images are transformed into cryptic patterns using random permutation. These are then processed through eigenfeature extraction based on total and between-class scatter matrices to create discriminative cancelable templates. In the third study, we employ a user-specific singular matrix alongside random permutation to generate intermediate templates. These intermediate templates are then transformed into cancelable templates using the Hadamard product and Boolean-XOR, ensuring security through user-specific transformation and randomness. The fourth work proposes two related methods involving bitwise transformations in the binary domain. Each image is transformed pixel by pixel using Boolean-XOR. The two methods differ in the way the transformation is applied and how randomness is introduced, offering flexibility in template generation. All proposed techniques outperform existing random permutation- and XOR-based baselines in terms of recognition accuracy, while strictly adhering to the principles of cancelable biometrics: non-invertibility, diversity, and revocability. These contributions provide a robust foundation for secure and accurate biometric authentication.

Contents

1. Introduction 2. Background 3. Random permutation-based linear regression for cancelable biometrics 4. Cancelable biometric template generation using eigen feature regularization 5. Real-numbered singular matrix transformation for non-invertibleand cancelable biometrictemplates 6. Boolean approach for cancel able biometric template generation. Conclusion.