**研 究 計 画 書**

**Research Plan**

|  |  |  |  |
| --- | --- | --- | --- |
| 志願者氏名Applicant's Name | XXXXX | 受験番号App. No. | ※1 |
| 予定指導教員Prospective Research Adviser | 氏名　Name | 押印または署名　Seal or Signature　※2 |
| **RAGE Uday Kiran** |  |
| 予定副指導教員Prospective Sub Research Adviser(該当者のみ if applicable) | 氏名　Name | 押印または署名　Seal or Signature　※2 |
|  |  |
| 英文で記述すること。枠を広げページを追加してよい。　Write in English. You can expand the input area to add pages. |
| **Introduction:**Many real-world applications generate uncertain temporal data naturally due to the veracity. Discovering knowledgeable patterns and correlations can help the users achieve socio-economic development. Periodic-Frequent Pattern Mining (PFPM) is an important class of regularities that exists in data mining. PFPM aims to find all the patterns in an uncertain temporal database that satisfy the user-specified constraints of expected support (expSup) and maximum periodicity(maxPer). The epSup controls the number of transactions that a pattern must appear in the database. The maxPer controls the maximum interval of a time that a pattern must reappear in the database. The successful application of PFPM is traffic congestion analytics. It helps to identify all the sets of roads where traffic congestion is faced by people regularly.Figure. 1 (a) shows the road sensor network in Kobe prefecture, Japan. For every 5 minutes these sensors generate the road identifiers, congestion length, and the probability of congestion shown in Figure. 1 (b). Without the loss of any generality, the generated data can be converted into the uncertain temporal database as shown in Figure. 1 (c).Periodic-frequent pattern mining on this uncertain temporal database identifies the complete sets of road segments on which people have regularly faced traffic congestion. An example of a periodic-frequent pattern is as follows:{RID5, RID6, RID7, RID9} [expected support=0.6, periodicity=2 hours].The above pattern indicates that 60% of congestions happened on road identifiers RID5, RID\_6, $RID\_7$, and $RID\_9$. Moreover, people have encountered these congestions every two hours. When this information is combined with other data sources, such as rainfall data as shown in Figure. 1 (e), the generated patterns may found to be useful in monitoring traffic and suggesting alternative routes during natural disasters.The existing study [1] is the first to work on the PFPM with uncertain data. However, this algorithm is computationally expensive and suffers from many limitations. We observed, that there exists a possibility to reduce the search space. The contribution of my thesis are as follows:* To reduce the search space, a new tighter-upper bound constraint called *cutoff expected support* (CES) was proposed. It helps in determining whether a superset of a pattern can be a periodic-frequent pattern or not by exploiting the anti-monotonic nature of the probability.
* A fast depth-first algorithm called Uncertain Periodic-Frequent Pattern-growth++ (UPFP-growth++) was also proposed, to discover the complete set of periodic-frequent patterns in a database effectively.
* Experimental results on real-world and synthetic databases demonstrate that CES significantly reduces the search space, and decreases the number of false patterns making UPFP-growth++ more efficient.

Please, refer to the experimental section in paper [2] to have a brief look at the conducted experiments. This work has been accepted and published in one of the top data mining conferences, ICONIP 2022. However, all the existing studies in periodic-frequent pattern mining with uncertain data implicitly assume that the geo-referenced (or spatial) information of the items will not impact the interestingness of the pattern. This application limits the applicability of PFPM to real-world applications. ”Neighbors" is one of the key concepts to develop efficient solutions to problems in the real world. The patterns in the coordinate system sound more interesting when the items are close to each other than the items that are far from each other. I argue that considering the geo-referenced information of items in the database will help users to find the more user-interesting based patterns. The usefulness of considering geo-reference information is shown in Figure 2.Figure 2: The usefulness of geo-referenced patterns in traffic congestion dataDuring my graduate course (Masters program), developed a novel algorithm to find the geo-referenced frequent patterns in geo-referenced uncertain databases (GFPMiner) [4] which got accepted in the world's top conference related to data mining **PAKDD 2023**(Core Rank A). During my doctoral course, I intend to solve this problem and find solution by exploring novel mathematical problems to find an efficient solution. To the best of my knowledge, this will be the first study that aims to explore novel spatiotemporal analytical techniques and energy-efficient algorithms to discover useful information in uncertain traffic congestion data. |
| **Objectives of the Research:*** Study the nature of geo-referenced (or spatiotemporal) uncertainty in big data.
* Introduce the big data analytical techniques to find useful information in the congestion data.
* Develop a novel mathematical model of geo-referenced patterns that may exist in uncertain congestion data.
* Develop a green-computing inspired practicable framework to get the desired results.
* Finally, demonstrate the usefulness of the proposed work with a case study performed on road sensor network data of Fukushima.

**Future Research Plan for the Doctoral Program:**First, I will employ fundamental big data analytics to find the useful information for the traffic congestion data. In big data analytics, we try to find the trends, patterns, and correlations in large amounts of raw data which are very useful in making data-driven decisions. The congestion data will be the test data. Our decision will depend on the generated results and findings.  Second, a key question that immediately raises after proving our observation is how to conduct analytics on uncertain congestion data? I aim to develop fundamental uncertain analytical theory to answer this question. I will also develop an open-source toolkit to promote further research in TCA. The activities and the time frame for each phase are shown in Figure 3. **Phase I (First year):** I will go through the literature and formulate a hypothesis to prove our claim. Necessary arguments will be investigated. We will develop multi-sensor information fusion theory and gather congestion data. **Phase II (Second year):** I will conduct experiments and evaluate using real-world databases.**Phase III (Third year)**: I will explore new analytical theory to uncover hidden patterns in uncertain spatiotemporal data. An open-source Python toolkit will be developed to promote further research in congestion analytics. References:[1] Uday Kiran, R., Likhitha, P., Dao, M.S., Zettsu, K., Zhang, J.: Discovering periodic-frequent patterns in uncertain temporal databases. In: Mantoro, T., Lee, M., Ayu, M.A., Wong, K.W., Hidayanto, A.N. (eds.) Neural Information Processing. pp. 710–718. Springer International Publishing, Cham (2021).[2] Likhitha, P., Veena, R., Kiran, R.U., Zettsu, K., Toyoda, M., Fournier- Viger, P.: Upfp-growth++: An efficient algorithm to find periodic-frequent patterns in uncertain temporal databases. In: Tanveer, M., Agarwal, S.,Ozawa, S., Ekbal, A., Jatowt, A. (eds.) Neural Information Processing. pp. 182–194. Springer Nature Singapore, Singapore (2023)[3] Likhitha, P., Veena, P., Rage, U.K., Zettsu, K.: Discovering geo-referenced frequent patterns in uncertain geo-referenced transactional databases. In: Kashima, H., Ide, T., Peng, W.C. (eds.) Advances in Knowledge Discovery and Data Mining. pp. 29–41. Springer Nature Switzerland, Cham (2023) |

※1 記入しないこと。　Do not write in the box.

※2 入学後の研究計画について、出願の前に必ず**指導を受けようとする教員に承認を得ること**。遠隔地に居住している等の理由により、押印や署名を得られない場合は、教員が了承していることを示すE-mailの写しを添付すること。
Prior to making the application, you **must obtain approval from your prospective research adviser** regarding this research plan. If you are in a remote place and not able to obtain a seal or a signature, a copy of an email that verifies their agreement should be attached.