Review of Exceptional Mobility Mining through Topic Modeling on Urban Social Media

Freddie Liu
x.liu1@student.tue.nl
Technical University of Eindhoven

ABSTRACT

This review evaluated if the paper *Exceptional Mobility Mining through Topic Modeling on Urban Social Media* should be accepted by ICDM. The paper proposed topic modeling using LDA to extract exceptional mobility patterns of a subgroup that have the potential to improve advertisement targeting, sale strategies, and urban planning, etc. The experiments provided valid evidence that the approach can find substantial phenomena in mobility dataset. However, the technical innovation is marginal since they just adjusted beam search and LDA according to the input. There is also no clear motivation for the chosen methods. Additionally, this work has unclear originality and they missed an important reference which also studied trajectory pattern mining. The authors did experiments on both synthetic and real-world social media mobility data which provide valid evidence that the proposed approach has the ability to find exceptional mobility patterns. But the exceptional mobility patterns are only based on KL divergence and these experiments lack other quantitative evaluations. There is also no comparison between different methods or to a baseline method. On the other hand, this paper is hard to follow. Some definitions and concepts are confusing and not explained well such as the location. They initially avoid the use of regions due to semantic ambiguities and sparsity of mobility patterns. However, in later sections and experiments, they do use local regions to represent GPS location, which contradicts the initial motivation and definitions. Some sentences contain too many sub clauses and could be simplified. To summarize, I agree with the reviewers that this paper should not be accepted by ICDM.

ACM Reference Format:
Freddie Liu. 2023. *Review of Exceptional Mobility Mining through Topic Modeling on Urban Social Media.* In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 KNOWLEDGE GAP

This research paper introduced the influence of spatial structure and semantic categories of features in city environments on human activities as motivation of the research topic. The way that a city is constructed and the features it includes can shape and influence people's interests of movement to some places.

I did not identify an explicit knowledge gap, but the need for a method to analyze the preferences and movements of individuals in a city is motivated by the growing amount of social media (such as Twitter and Instagram) data. The authors also mentioned that location is the most important factor in describing these human activities. Frequently visited places can reflect people's preferences for a certain semantic class of places which could have many potentials. The authors argued that the traditional methods for analyzing human mobility patterns based on social media data, such as frequent pattern mining and association rule mining, were limited in their ability to identify interesting patterns and subgroups of people with distinctive mobility patterns. Thus, they proposed a new model class for exceptional model mining based on topic modeling which applied Latent Dirichlet Allocation(LDA) to a metaphor of mobility patterns. The mobility data is mapped to LDA where the whole mobilities of people can be considered as corpus, mobilities of a subgroup can be considered as document, and mobility as word, distributions of mobility patterns as topic. So the "topic" distribution of each subgroup (i.e. mobility patterns) can be drawn from the mobility dataset. The mined exceptional mobility patterns can be used to improve advertisement targeting, sale strategies, and urban planning, etc.

The authors used the aforementioned methods to bridge the gap in finding exceptional subgroups of people with unusual mobility patterns. They determine if a subgroup is exceptional based on KL Divergence (ϕKL). They proved this to be robust in the synthetic dataset since ϕKL drops off significantly in lower-ranked subgroups. The experiments conducted on synthetic and real-world data show that the approach is sound and has the ability to generate substantial mobility patterns in the dataset. More details of the experiment will be discussed in Section 4. Based on these results, the authors believe that their approach has successfully addressed the knowledge gap. However, there are no evaluation metrics introduced, and the quality of exceptional mobility patterns is purely based on KL Divergence. So it is difficult to determine to what extent the knowledge gap has actually been addressed, whether the chosen method is effective and the generated results are useful.

The reviewers all mentioned that this paper lacks clear motivation for the chosen methods which I think stands a strong point. The paper did not discuss the reason to choose Latent Dirichlet Allocation over the others in exceptional mobility mining but just jumped to it. LDA is not the only option in this context. There are other methods that can infer latent patterns for all mobility. For example, clustering algorithms like Kim, Y., et.al.[1] used in the paper that the first reviewer mentioned. They also did research in mining latent mobility patterns. Pattern trees[2], random forests[3] and association rule like the authors mentioned[4] can also identify latent subgroups and are possible to address the problem. Additionally, non-negative matrix factorization (NMF) is also an effective

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

standard topic model especially for sparse, noisy and ambiguous texts[5]. But none of the other methods are discussed and compared in the paper. Thus, there is no evidence to support the choice of LDA as the most appropriate one.

2 EMBEDDING IN RELATED WORK

The paper does not have a related work section. It talked about other scientific literature mostly in introduction. Others were embedded in the paragraphs. It listed some references but did not discuss them adequately. For example, the advantages and disadvantages of the chosen methods of these literature, the ways in which they were influenced by previous research, and how they overcame the constraints of prior work.

The reviewers criticized the lack of technical innovation of this paper. The technical improvement is barely incremental. The authors just applied LDA and beam search and adjusted accordingly based on the input data. The motivation of using this method is unclear. The authors mentioned they chose LDA to avoid replacing GPS location with regions, but regions were eventually used in the experiments.

The first reviewer also pointed out that the paper did not include an important reference which studies a similar problem *Topical trajectory pattern mining*[1]. I think it is reasonable to say that since they also employed latent topic-based clustering to identify patterns in mobility and text data. The knowledge gap presented in their work is clearer. It was because the local context in text messages was nearly neglected while GPS data was widely used to discover mobility. Additionally, their technical innovation is also more prominent since they introduced a new algorithm.

On top of that, the paper could also include the following two references that both made significant contributions to semantic trajectory pattern mining. Alvares et al.[6] introduced the semantic trajectory data model and a preprocessing method to incorporate semantic information with raw trajectories. Zhang et al.[7] considered groups of similar points of interest instead of single POI to combine mining semantic category sequences with mining frequent spatial routes. These studies are close to the topic of this paper and should be discussed.

3 QUALITY OF THE PAPER WRITING

The paper is generally well written and the English used in the paper is proper, but some of the sentences are structured weird and could be more concise. For example, *"Those subgroups whose distribution is the most different from that of their complements are considered the most exceptional"* (p.1) has too many subordinate clauses and makes it less comprehensible. It could be written as *"The subgroups with the most distinct distribution from their complements are considered the most exceptional."* to increase readability. The similar problem lies in these sentences:

"Each subgroup is input into the trained LDA to predict the distribution of latent mobility patterns, the result of which will be input for the quality measure function along with the same value for the complement of the subgroup." (p.4)

They could be rewritten as *"The trained LDA analyzes each subgroup to predict its distribution of latent mobility patterns. These predictions, along with the complement of each subgroup, are then used in the quality measure function."*

The third reviewer also mentioned small grammar mistakes and typos, which did exist in the paper and should definitely be avoided. For instance, *"..., and then a location pair which indicates a mobility can have a semantic meaning that represents the pattern."* (p.2) should be *"...that indicates mobility..."*. Also *"more smaller parts"* (p.6) should be *"smaller parts"*.

One reviewer mentioned that this paper is hard to follow and I agree. The logic flow in this paper is not quite good and it causes a considerable amount of confusion. For example, in the beginning session of introduction, the paper wrote "semantic categories of features in city environment influence human activities". This sentence might want to say that types of features (such as buildings, parks, etc.) present in the city environment have an impact on human activities. Later it adds that "semantic interests" refers to the latent patterns that indicate the preferences or inclinations of people's mobility. But back to the introduction, it is the human activities that influence and indicate their semantic interests, not the other way around. So the concepts related to "semantics" mentioned in the paper are very confusing, and lack of clear explanation.

The paper requires a high level of prior knowledge from the readers. For example, the paper first mentioned TF-IDF in table 6, but it did not explain TF-IDF at all. And table 6 lacks a clear header below the regions to indicate what exactly are the data and how the importance is calculated. The same issue applies with table 5. These tables with many numbers not explained adequately are barely self-contained since one can get easily lost when just reading them. Additionally, the paper gave many concepts and definitions. Some of them are not very intuitive, such as the definition of the mobility $M(p, t) = (L\_i, L\_j)$ where L stand for different locations, and t stands for time.

The definition of location causes huge confusion in this paper. In page 2 the authors mentioned that regions would not be used because it causes ambiguities if they set the same semantic meaning for all the locations in one region. It also causes the sparsity of mobility patterns for groups. They claimed that this is the reason that they introduced LDA. However, in the end of 3.1 and the experiment, they partitioned the city space to local regions and used regions to represent GPS location. This leads to conflicts with the motivation of methods and all definitions related to location, which is a significant flaw since the paper studies mobility.

4 EXPERIMENTS

The authors conduct experiments on both synthetic and real-world social media mobility data to demonstrate the effectiveness of their approach. This is an advantage since synthetic data can be a good supplement to real-world data. For the synthetic data, the authors generated normal and exceptional behavior documents using the LDA process and set the number of topics to 100, vocabulary size to 2500, and parameter η to 0.01 for both types of documents. A single descriptive attribute was generated in two different ways, and an exhaustive search algorithm was run twice to evaluate disjunctive subgroup descriptions. For the real-world data, the authors collected social media data from Shenzhen and partitioned the city space into 60 local regions using OpenStreetMap. Noise in the dataset

was filtered by removing users spanning only one location or with a transition time longer than 24 hours. The authors trained an LDA model with 21,950 user records and ran beam search with search depth 2, using interval-valued conditions on numeric descriptors and set-valued ones on nominals. They are reproducible technically, but the authors did not provide any publicly accessible data or implementation so it is not possible to reproduce in practice.

The experiments on synthetic data provide evidence that the method can successfully find pure exceptional subgroups. And the experiments on real-world data from Shenzhen illustrate the kinds of exceptional findings that the Exceptional Mobility Mining approach can generate. The top-10 subgroups identified in the experiments display a distinctive difference in topic distribution, which was further explored through TF-IDF analysis and visually on a map of the city.

There is a huge disadvantage for the experiments conducted in the paper, lack of quantitative evaluation. The authors just presented all the exceptional mobility patterns of subgroups, and ended there. There is no other quantitative evaluation introduced than KL divergence. Exclusively relying on KL divergence would lead to a limited understanding of the results. Other evaluation could be introduced to determine whether the subgroup mobility is exceptional. The authors could then compare if the exceptional mobility based on other evaluation is similar, and what caused the discrepancies. The reviewers also think that different methods should be compared in the experiments. And there is also no comparison between the current method and a baseline method, so the effectiveness of the chosen approach cannot be measured. I think their concerns are appropriate since lack of those comparisons would lead to weak support of the chosen method. Therefore, from this perspective, the experiments should be expanded by introducing quantitative evaluation and comparison among methods.

5 CONCLUSIONS

The paper did wrap up in a good manner. The authors described the results of their experiments on both synthetic and real-world social media mobility data, and they reported and analyzed carefully the top-10 subgroups displaying a distinctive difference in topic distribution. They also provide evidence that their Exceptional Mobility Mining approach finds substantial phenomena in the dataset which is valid. Overall they did not overclaim in the conclusion and the knowledge gap they proposed was successfully addressed. The reviewers did not criticize the conclusion, but more in the lack of motivation, originality, technical innovation and comparison of the methods which were discussed before.

The first reviewer offered a previous paper with the similar idea of using LDA for mining latent mobility pattern as a strong argument. All three reviewers mentioned lack of comparison between methods which is quite solid. Based on that flaw, the second reviewer argued that this paper did not have a clear motivation for the chosen method. I think their arguments are all convincing and they all provided some examples and evidence to support that.

Additionally, the authors argue that their approach is applicable to any source of data that comes in textual form or can be mapped to the LDA terminology. They conclude that Exceptional Mobility Mining finds substantial phenomena of the mobility in the dataset, and that the underlying principle is applicable to data beyond mobility. But they did not give any examples or use cases to support or demonstrate this idea which could be added.

REFERENCES

[1] Younghoon Kim, Jiawei Han, and Cangzhou Yuan. Toptrac: Topical trajectory pattern mining. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 587–596, 2015.

[2] Florian Lemmerich, Martin Becker, and Martin Atzmueller. Generic pattern trees for exhaustive exceptional model mining. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 277–292. Springer, 2012.

[3] Jared C Foster, Jeremy MG Taylor, and Stephen J Ruberg. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

[4] Martin Atzmueller, Mark Kibanov, Naveed Hayat, Matthias Trojahn, and Dennis Kroll. Adaptive class association rule mining for human activity recognition. In *MUSE@ PKDD/ECML*, pages 19–34, 2015.

[5] Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1105–1114, 2018.

[6] Luis Otavio Alvares, Vania Bogorny, Bart Kuijpers, Jose Antonio Fernandes de Macedo, Bart Moelans, and Alejandro Vaisman. A model for enriching trajectories with semantic geographical information. In *Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems*, pages 1–8, 2007.

[7] Chao Zhang, Jiawei Han, Lidan Shou, Jiajun Lu, and Thomas La Porta. Splitter: Mining fine-grained sequential patterns in semantic trajectories. *Proceedings of the VLDB Endowment*, 7(9):769–780, 2014.