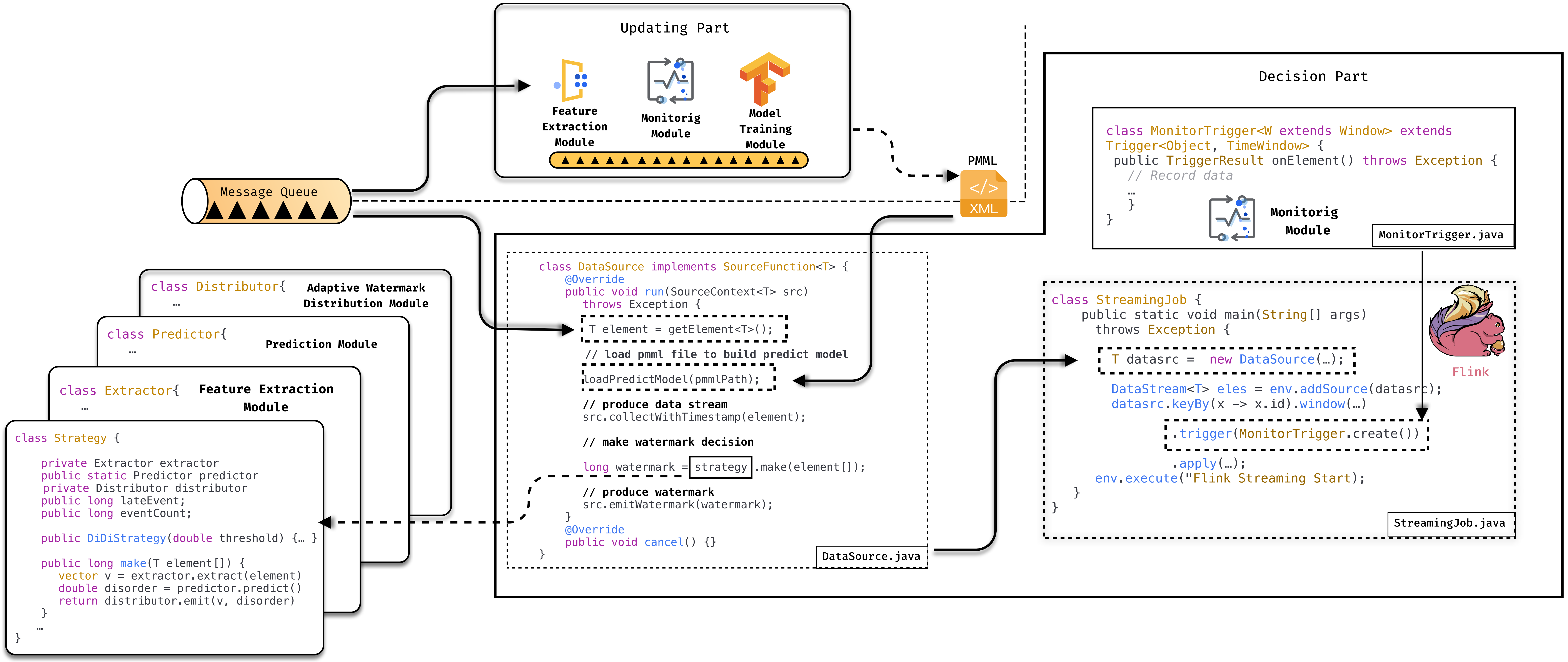
论文审稿意见总结

Adaptive Watermark Generation Mechanism based on
Time Series Prediction for Stream Processing 为例

意见类型

系统实现

1. There are missing details on how the proposed method has been implemented into Flink.



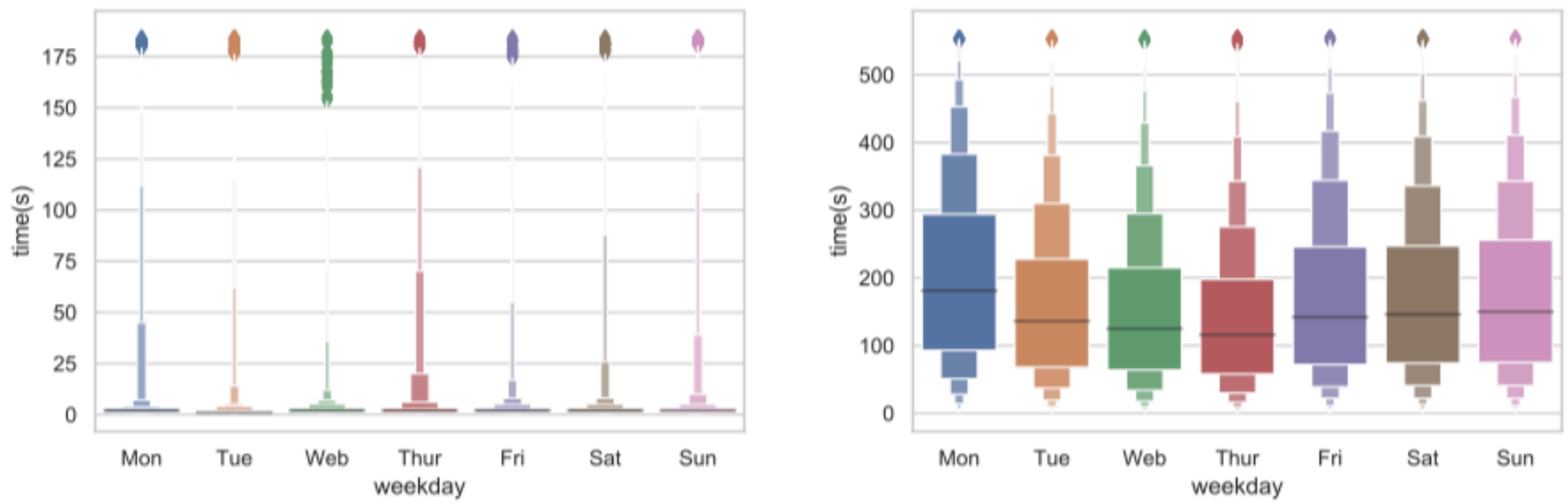
意见类型

参数设定

3. Please explain why the specific time windows have been selected in Section 5.1.3.

4. Please explain why the 1 minute time window has been selected for periodic watermark distribution strategy in Section 5.2.3. What would the results be if other time window is selected?

5. How the data lateness ratio threshold has been determined in Figure 7 and 8?



(a) 共享单车 (b) 滴滴出行

图 10 迟到时间分布

意见类型

相关工作对比

1. As adaptive watermark generation mechanism is a key contribution of this paper, what the difference between the proposed one and that in [6]?

Table 2 Comparison of existing watermark distribution mechanisms

	Our adaptive watermark	Concept drift watermark [6]	Periodic watermark [4,5]
Type of prediction model	Supervised prediction model (LSTM, XGBoost)	Unsupervised prediction model (Concept drift)	No machine learning model used
Input/Output of prediction model	Input: Take the LSTM model for example, it is a feature vector of disordered data ratio composed of multiple history windows Output: The disordered data ratio within a time interval in the future, used to describe the degree of data disorder in this period of time	Input: The difference between stream data **E** processing time and event time tp(**E**)-te **E**) Output: Determine whether a data is abnormal data, that is, the tp(**E**)-te(**E**) value drifts of data e will be considered as an abnormal value	Users need to adjust the watermark delay time according to their own experiences
Tuning parameters	One parameter, only need to set the expected data lateness ratio	A large number of parameters need to be tuned empirically, such as out-of-order threshold, sensitivity, sensitivity change rate, which is both error-prune and time-consuming to reach the optimal settings	No parameters

意见类型

实验相关

2. To objectively illustrate the advantages of the proposed scheme, the authors should compare the performance with some similar research work such as [4,5,6].

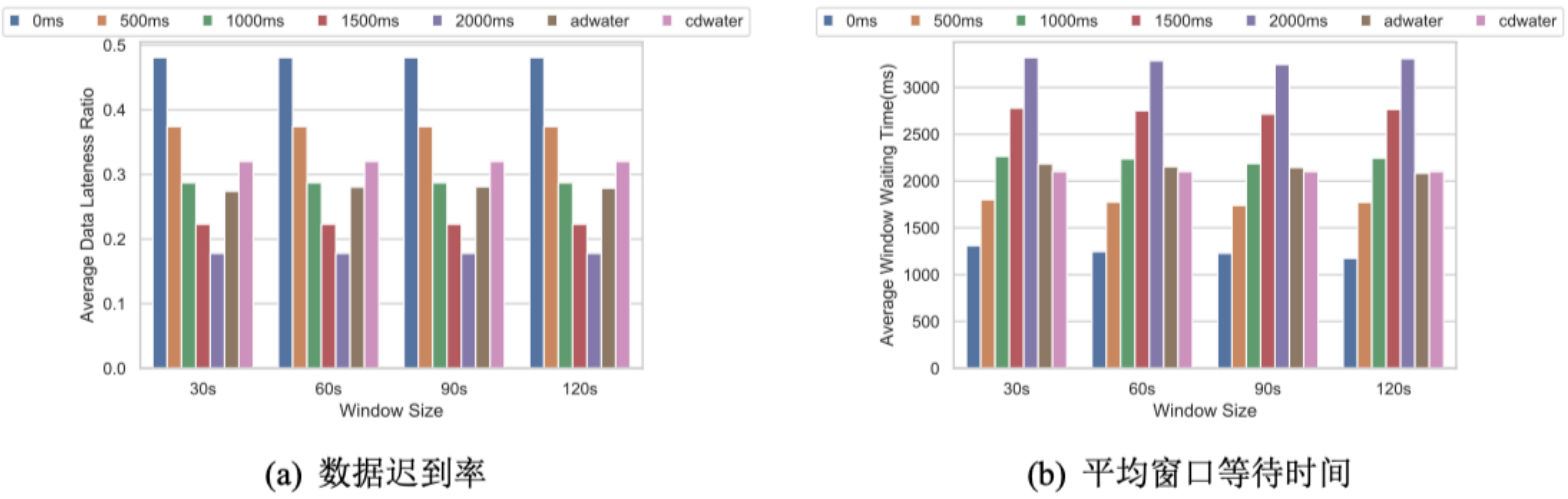


图 16　共享单车数据集实验结果对比图

意见类型

格式文笔相关

3. The organization and format of the paper need significant improvement.

Minors:

1.Line 45, page 11, “…predicted using the updated model is It is…

2.Line 30, page 3, “…the system remain”→ remains

3.Line 61, page 4, “… approaches and considering…”→ consider

4.Algorithm 1 has not been explicitly cited throughout the text

审稿意见回复格式

1．开头第一段需要表达对审稿人意见的感谢，一定要有礼貌，这个直接用现成的模板就行

We would like to thank all reviewers and editors for their valuable feedbacks. Your comments have helped us greatly in improving the manuscript. We have carefully addressed all your comments and suggestions in the revision. Below, we provide detailed responses to each reviewer. We color all new text in red to highlight the changes that we have made in the manuscript. The references are in the same order, as shown in the manuscript.

Editor:

This paper proposes a watermark generation method based one time-series prediction. It implement the proposed method on Flink and evaluate with real-world datasets such as New York City and Didi datasets. Results show that the proposed method can adapt to data stream changes and adjust the watermark generation strategies for better responsiveness and accuracy compared to existing approaches.

Positive:

1. It verifies that LSTM is effective for temporal prediction.
2. A nice implementation and experiments on real-world datasets.

审稿意见回复格式

2．一般有多个审稿人，每个审稿人的意见都要单独列出来，每个审稿人都要专门写一句感谢的话

Reviewer 1

This paper proposes a watermark generation method that can adjust the frequency and timing of watermark distribution. The adjustment is based on the disordered data ratio and lateness of the data stream, predicted using the LSTM model, that improves system responsiveness with acceptable accuracy. The authors implement the proposed method on Flink and evaluate with real-world datasets such as New York City and Didi datasets. The experiment results show that the proposed method can adapt to data stream changes and adjust the watermark generation strategies for better responsiveness and accuracy compared to existing approaches.

Strength:
1.The proposed time series based prediction model using LSTM seems quite effective.
2.The proposed method has been implemented on real system software.
3.The evaluation has been conducted on real-world datasets.

Response: We appreciate very much for your effort in reviewing our paper.

1. There are missing details on how the proposed method has been implemented into Flink.

Response: Thank you very much for these comments. The implementation of our proposed adaptive watermark

审稿意见回复格式

3．对每个审稿意见做回复的同时，需要把在论文中添加和修改的地方标红列出。

4. Please explain why the 1 minute time window has been selected for periodic watermark distribution strategy in Section 5.2.3. What would the results be if other time window is selected?

Response: Thank you very much for these comments. The purpose of the experiments in Section 5.2.3 (Section 6.2.3 in the revision) is to demonstrate that there is no need to use the same delay time to wait for the late data in all periods of the day. From Figures 8 and 9, we can see that the disordered data ratio fluctuates within a day. For the CityBike dataset, it is more obvious that the disordered data ratio peaks in the morning and afternoon periods. Thus, in the time period when the data lateness ratio is small, using lower watermark delay time is enough to ensure a low disordered data ratio. The Didi dataset has a low disordered data ratio in the morning period. To show the detailed changes of data late ratio in Figures 10 and 11, we chose to calculate the data lateness ratio within 1-minute window. If a smaller window is used, the calculated value of the data lateness ratio will be inaccurate due to insufficient data in the window, which in turn cannot clearly show the data changes during this time period. In contrast, if a larger window is selected, the changes of the data lateness ratio become insignificant to show in the figure. Based on the characteristics of our evaluation datasets, we choose 1 minute to present the changes of the data lateness ratio within a day.

The relevant text appears in Section 6.2.3 on page 13. We reproduce the relevant text below for your convenience.

"To show the detailed changes of data late ratio in Figure 10 and Figure 11, we chose to calculate the data lateness ratio within 1-minute window. If a smaller window is used, the calculated value of the data lateness ratio will be inaccurate due to insufficient data in the window, which in turn cannot clearly show the data changes during this time period. In contrast, if a larger window is selected, the data lateness ratio changes become insignificant to show in the figures. Based on our evaluation datasets' characteristics, we choose 1 minute to present the changes of the data lateness ratio within a day. "

"To demonstrate the advantage of our approach, we compare with periodic watermark [4,5], and concept drift watermark [6]. On the New York CityBike dataset, we select four different time windows of 0.5min, 1min, 1.5min, and 2min. In the taxi dataset, we select four different time windows of 1min, 2min, 3min, and 4min. The time window is selected based on the application scenario. Through the analysis of the evaluation datasets, we find that the length of delay of the late data is at the second and minute granularity. These late data have the greatest impact on the minute-level window. That is, the window may wait for N time windows of the late data.

Figure 6 and Figure 7 show the distribution of the data lateness of the two datasets in a week. Since the volume of daily data is relatively large, we use a box diagram to represent the portion of a certain value with box width. For the CityBike dataset, we can see that most of the data delay time is concentrated within the 0-3s range. For the Didi dataset, most of the delay time is concentrated within 1-4 minutes range. Comparing these two datasets, we can also see that the delay time of the Didi dataset is longer than the CityBike dataset. Therefore, in the experiments, we set a larger time window for the Didi dataset. "

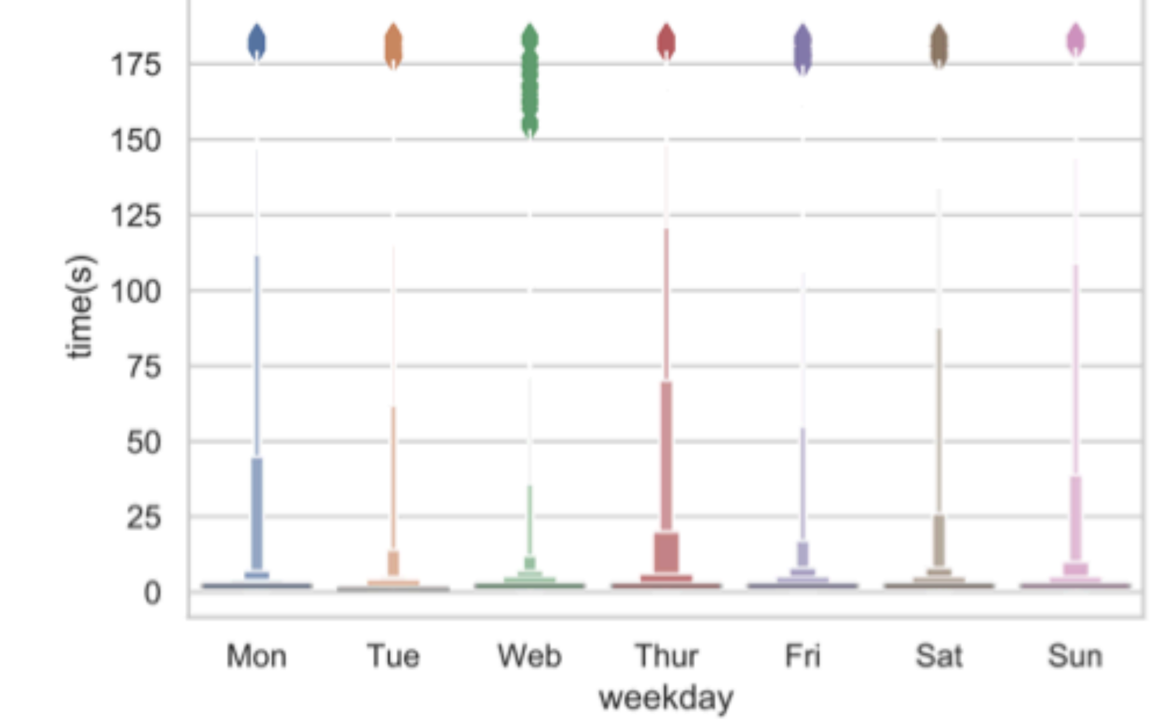


Figure 6 The distribution of the data lateness of the CityBike dataset.

审稿意见回复格式

4. 格式类型问题的回复

3. The organization and format of the paper need significant improvement.

Response: Thank you very much for your suggestions. We have carefully gone through the manuscript to improve organization and format of th contents for better readability.

Minors:
1.Line 45, page 11, "...predicted using the updated model is It is...
2.Line 30, page 3, "...the system remain"-> remains
3.Line 61, page 4, "... approaches and considering..."-> consider
4.Algorithm 1 has not been explicitly cited throughout the text

Response: Thank you very much for pointing out these problems. We have fixed these problems in the revision. In addition, we have carefully gone through the manuscript to eliminate such minor errors as well as improved the manuscript to make it more readable.

审稿意见模板

We would like to thank all reviewers and editors for their valuable feedbacks. Your comments have helped us greatly in improving the manuscript. We have carefully addressed all your comments and suggestions in the revision. Below, we provide detailed responses to each reviewer. We color all new text in red to highlight the changes that we have made in the manuscript. The references are in the same order, as shown in the manuscript.

Reviewer 1
This paper proposes a …
Strength: 1. 2. 3

Response: We appreciate very much for your effort in reviewing our paper.

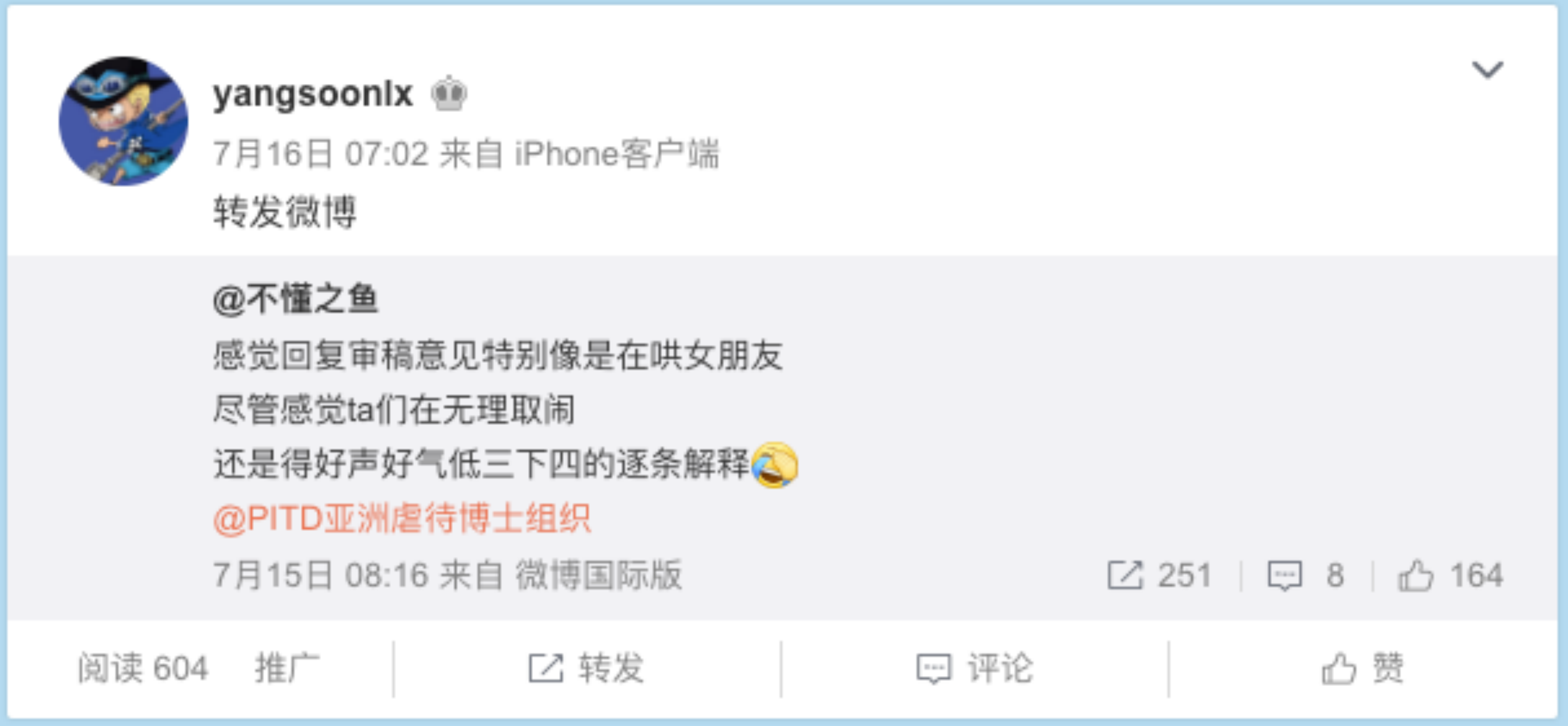
1. There are missing details on how the proposed method has been implemented into Flink.

Response: Thank you very much for these comments. The implementation of our proposed adaptive watermark generation mechanism in Flink can be decomposed into two major parts: 1) …

The relevant text appears in Section 5 on page 7. We reproduce the relevant text below for your convenience.

"To demonstrate the advantage of our approach, we compare with periodic watermark [4,5], and concept drift watermark [6]. On the New York CityBike dataset, we select four different time windows of 0.5min, 1min, 1.5min, and 2min. In the taxi dataset, we select four different time windows of 1min, 2min, 3min, and 4min. The time window is selected based on the application scenario. Through the analysis of the evaluation datasets, we find that the length of delay of the late data is at the second and minute granularity. These late data have the greatest impact on the minute-level window. That is, the window may wait for N time windows of the late data.
"

最后


yangsoonlx
7月16日 07:02 来自 iPhone客户端
转发微博
@不懂之鱼
感觉回复审稿意见特别像是在哄女朋友
尽管感觉ta们在无理取闹
还是得好声好气低三下四的逐条解释
@PITD亚洲虐待博士组织
7月15日 08:16 来自 微博国际版
251
8
164
阅读 604
推广
转发
评论
赞


我今天不舔你两口
你就不知道什么叫舔狗