Specialist English: Assignment 6 (solutions)

Rebecca J. Stones
rebecca.stones82@nbjl.nankai.edu.cn

December 4, 2018

Here's my solutions; your solutions needn't be identical.

Problem 1 The abstract is:

ABSTRACT

With the advent of location-based social media and location-acquisition technologies, trajectory data are becoming more and more ubiquitous in the real world. Trajectory pattern mining has received a lot of attention in recent years. Frequent sub-trajectories, in particular, might contain very usable knowledge. In this paper, we define a new trajectory pattern called frequent sub-trajectories with time constraints (FSTTC) that requires not only the same continuous location sequence but also the similar staying time in each location. We present a two-phase approach to find FSTTCs based on suffix tree. Firstly, we select the spatial information from the trajectories and generate location sequences. Then the suffix tree is adopted to mine out the frequent location sequences. Secondly, we cluster all sub-trajectories with the same frequent location sequence with respect to the staying time using modified DBSCAN algorithm to find the densest clusters. Accordingly, the frequent sub-trajectories with time constraints, represented by the clusters, are identified. Experimental results show that our approach is efficient and can find useful and interesting information from the spatio-temporal trajectories.

— Huang, Luo, Wang, UrbComp, 2013.

In this abstract, we identify the introduction, solution, and demonstration components as follows:

> With the advent of location-based social media and location-acquisition technologies, trajectory data are becoming more and more ubiquitous in the real world. Trajectory pattern mining has received a lot of attention in recent years. Frequent sub-trajectories, in particular, might contain very usable knowledge. In this paper, we define a new trajectory pattern called frequent sub-trajectories with time constraints (FSTTC) that requires not only the same continuous location sequence but also the similar staying time in each location. We present a two-phase approach to find FSTTCs based on suffix tree. Firstly, we select the spatial information from the trajectories and generate location sequences. Then the suffix tree is adopted to mine out the frequent location sequences. Secondly, we cluster all sub-trajectories with the same frequent location sequence with respect to the staying time using modified DBSCAN algorithm to find the densest clusters. Accordingly, the frequent

sub-trajectories with time constraints, represented by the clusters, are identified. Experimental results show that our approach is efficient and can find useful and interesting information from the spatio-temporal trajectories.

There is no implications component in this abstract. (Some students said the implications component is "can find useful and interesting information from the spatio-temporal trajectories", which is a reasonable alternative.)

The introduction component is:

> With the advent of location-based social media and location-acquisition technologies, trajectory data are becoming more and more ubiquitous in the real world. Trajectory pattern mining has received a lot of attention in recent years. Frequent sub-trajectories, in particular, might contain very usable knowledge.

We can extract the important elements and delete the rest, and string the results into a sentence e.g. as follows:

> Social media utilizes location-acquisition technology and hence the use of trajectory data is increasing, so mining frequent sub-trajectories may be useful.

There are many ways of writing this succinctly. It's important to eliminate the noise, e.g., words like "the advent of" and "very" in the original snippet are unnecessary words.

Problem 2 The relevant snippets are the following:

> ⋮
>
> mor datasets collected from Weibo and Twitter. The results demonstrate the effectiveness of the proposed end-to-end att-RNN in detecting rumors with multimodal contents.
>
> — Jin et al., MM (2017).

> ⋮
>
> **of required adders by half. Experimental results show that our proposed scheme improves the energy efficiency by** 45.9%**, the area efficiency by** 93.6% **and achieves 8x of throughput per area compared with the state-of-the-art CMOS-based implementation.**
>
> — Luo et al., ICCAD (2017).

We compare the demonstration components of these two abstracts below:

	how is performance measured?	what is the proposed method compared to?	what dataset is used for testing?	how big is the improvement?
Jin et al.	unstated	unstated	datasets collected from Twitter and Weibo	unstated
Luo et al.	energy efficiency, area efficiency, and throughput per area	a state-of-the-art CMOS-based implementation	unstated	49.5%, 93.6%, and 8×, respectively

Luo et al.'s demonstration component is more informative, since it succinctly describes how performance is measured, what it's measured against (i.e., the baseline), and the magnitude of the improvement, whereas Jin et al.'s demonstration component only describes the dataset.

Looking at the papers, we find that Luo et al. (2017) doesn't actually use a dataset (so it can't be described), whereas Jin et al. (2017) could describe how performance is measured (accuracy, precision, recall, and F_1), what the baselines are (a variety of methods, including VQA* and NeuralTalk), and how big the improvement is (e.g., increases precision by 0.03).

Problem 3 The snippet is:

> **ABSTRACT**
>
> Online social media yields a large-scale corpora which is fairly informative and sometimes includes many up-to-date entities. The challenging task of expanding entity sets on social media text is to extract more uncommon entities only using several seeds already in hand. In this paper, we present an approach which is able to find novel entities by expanding a small initial seed set on Twitter text. Our method first generates candidate sets on the basis of the semantic similarity feature. Then it jointly utilizes 2 text-based features and other 12 ones which carry social media specific information. With the scores on those features, a ranking model is learned by a supervised algorithm to synthetically score each candidate terms and then the final ranked list is taken as the target expanded set. We do experiments with 24 entity classes on the Twitter corpus and in the expanded sets there come many novel entities which have not been completely detected in previous researches. And the experimental results on the datasets of different years can perfectly consist with the objective law that fresh entities change as time goes on.
>
> — Zhao et al., ICTIR (2018).

1. The phrase "a large-scale corpora" is erroneous since "corpora" is plural whereas the article "a" implies singular. (The question didn't ask for how to fix it, but a description of the problem.)

2. The phrase "and other 12 ones" would be grammatically correct if it said "and another 12 ones", however it would be clearer if we explain what "ones" refers to: "and another 12 features" (or "and another twelve features").

3. The "do" in "We do experiments" is how a child speaks: suitable alternatives are "We perform experiments" or "We conduct experiments".

4. The word "research" (as a noun) is uncountable—it is therefore incorrect to use "researches" to mean the plural of "research". (The question also didn't ask for how to fix it.)

5. The word "different" in "of different years" means "various", which is a suitable replacement. In general, I recommend not using "different" to mean "various" (since "different" provokes the question in the reader's mind "different to what?").