TITLE

1. *introduction (includes motivation, potential real-life applications)*
2. *related work (e.g., compare & contrast existing work, point out shortcomings of existing work)*
3. *main body (includes description of your new idea, step-by-step illustrative examples, pseudo codes, highlight of any differences between your work vs. existing works)*
4. *analytical (e.g., theorems) and empirical evaluation (e.g., experiments, tables & figures, explanations on experimental results)*
5. *conclusions and future work*
6. *references (e.g., cite at least 5 books, journals, and/or conference papers--but not websites)*

1. Introduction

- big data blurb

- lots of people use transit, important in urban areas

- always needs to be improved, growing use as population grows

- environmental motivation, part of increased use of transit(?)

- can be applied to any city that collects this kind of transit information

- big cities tend to gather data, they have technology/infrastructure in place for it

- lots of data --> data mining (DM) techniques are suited to this

- we can learn something from past data

- are delay patterns constant, or at least stable enough that old data is a good source of information

- compare diff incremental mining models (exact landmark vs. exact time-fading) to see if accuracy/constantness/relevance of patterns (from older data to newer data) is affected

(No background section? or should this be in intro?)

Background:

- TTC network, what data do they collect, ridership numbers over the years to show increase

- demands, new services, improvement initiatives over the years?

- any higher level transit surveys available? something nation-wide?

2. Related Work

3. Our approach

- data cleaning/pruning/prep:

- location name (keep >= 85 similarity only), direction consistency

- min gap prunning (get rid of < 5, not as meaningful)

- describe time of day binning (?)

- batching approach (3 months ?)

- frequent pattern mining to establish pattern on historical data

(use incremental mining, compare exact landmark to exact time-fading model)

- frequent pattern to establish pattern in newer/recent (historical) data

(same incremental mining model or just whole batch?)

- look for where/when delays tend to occur correlate with severity of delay

- do by time of day, day of week, month, and combinations of

- correlate cause of delay with severity of delay?

4. Evaluation

- compare older patterns to newer patterns (explain methods for doing so and any metrics used)

5. Conclusion & Future Work

6. References