Machine Learning Project 2015-2016 (Feb)

Predictive Web Browsing

1 Problem and Approach

The goal of this project is to implement a machine learning based assistant that helps a user to browse the web more efficiently by predicting the websites the user is trying to reach. This assistant provides an adaptive interface that customizes towards individual users or groups of users by learning from their browsing history and updates while they continue surfing the web.

An example scenario would be to go to kuleuven.be and find the schedule for your courses this semester by only clicking links. Another example scenario is to find news about the Leuven Bears. To find that news on the sporza.be website you would have to perform multiple clicks. An adaptive browser could suggest your courses immediately when you visit kuleuven.be and click through to student information. Or suggest the basketball page immediately when you visit sporza.be again. Maybe even a list of basketball pages on different sports websites you frequent.

The system you will build should be able to predict these sequences of clicks and dynamically offer shortcuts to the endpoint(s) of the sequence of clicks once it detects which sequence the user is executing. The result is that the user saves a number of clicks (and searches over the pages) by being directed immediately to the page the user is interested in. Or if it cannot figure out the endpoint of the sequence it could at least list or highlight links on the current page that you are most likely to click next to make navigation more easy.

There are three types of features that can be taken into account to achieve either type of prediction: (1) temporal features, for example, identify frequent repetitions or morning habits; (2) derivatives of the full url, for example, domain name or variations of the path; (3) potentially, you can also extract features from the webpage content in which the link appears but this is more expensive.

2 Tasks

2.1 Form Groups

By Feb 19th

Mail davide.nitti@cs.kuleuven.be whether you work alone on this project or in a team of two. In the latter case, also mention both team member names. **Do not share your implementation with students outside of your group**. The only exception are fragments that appear in questions and answers on the Toledo forum, which is accessible to everyone.

2.2 Literature Study and Experimental Setup

Familiarize yourself with the concept of machine learning for predictive web browsing [1], clickstreams [3] and user activity streams [2, 4]. Get a feeling for what has been done before in this context by searching for these terms[1]. Divide the work between the two people in the team.

Set up the experimental environment included with this project (`urlStreamHander/README`). Make sure you are able to generate data yourself.

[1] https://scholar.google.com, http://academic.research.microsoft.com

2.3 First Report and Data Collection

By March 26th

In a first phase of the project you explain in a brief report how you will address the task at hand and generate a data set of your own.

Report 1: Mail a report (PDF, $\sim$2 pages) to the aforementioned email address in which you:

- Describe what literature you have read and what you have learned from it.
- Describe the overall pipeline you have in mind.
- Compile a list of questions that you want to find answers for by the end of the project. Examples are:
  - Which data representation (classes and features)?
  - Which machine learning model(s) and parameter settings (hyperparameters)?
  - What is your strategy for incremental learning?
  - How will you evaluate your classifier (methodology and metrics)?
  - What is the computational and memory cost of preprocessing, learning and evaluating?

The quality of this report influences your final score for the project, so try your best to come up with a good plan. After the reports are handed in, you will get feedback on your report, and, where necessary, we may give more concrete guidelines on how to proceed.

Data set: Include a set of csv-files with url streams you have generated yourself. Per person we expect at least 100 'load' entries. Try to generate streams where you try to accomplish a number of repetitive tasks (in contrast to randomly surfing news headlines). For example, checking stats about your favourite team, checking weather forecasts, checking course information, etc, by navigating from the main page. Make sure not to include any privacy-sensitive information as this data will be aggregated and shared among all students one week later.

2.4 Final Report and Prototype

By May 11th

Report 2: Write a report with your findings.

Mail your final report (PDF, $\leq$10 pages) describing details of your approach and the results obtained:

- Clearly state the **questions** you address with your research
- Describe how you try to **answer** them (what experiments were performed, how do you measure success, etc.),
- Write out the **conclusions** you draw from your experiments together with a scientifically supported motivation for these conclusions. Such a motivation should include descriptions of your methodology, used techniques, performed experiments and a report and discussion of the results of these experiments. Be concrete about methods, formulas and numbers.
- Report the total **time you spent** on the project, and how it was divided over the different tasks mentioned.

Prototype: Include the code you used to perform the machine learning and experiments with the final report (also if the final application runs in the cloud). Running and experimenting your application is part of the evaluation. You are free to use any mainstream programming language and any publicly available toolbox. Make sure that:

- The code is intuitive to run and includes a short README file with the necessary steps.
- The default setting is to use the model you have found to be the best.
- Your code is self-contained and includes all dependencies.
- Your code is print-friendly (e.g., max 80 columns).

2.5 Peer assessment

By May 11th, individually

Send by email a peer-assessment of your partner's efforts. This should be done on a scale from 0-4 where 0 means "I did all the work", 2 means "I and my partner did about the same effort", and 4 means "My partner did all the work". Add a short motivation to clarify your score. This information is used only by the professor and his assistants and is not communicated further.

2.6 Discussion

By Week of May 16th

There will be an oral discussion of your project report where you will also get the chance to demo your prototype. The instructors may try a certain set of actions to assess the adaptivity of your system.

Questions

Please direct any questions that you may have about the project to the Toledo forum or the classroom discussion moments.

Good luck!

References

[1] Janez Brank, Natasa Milic-Frayling, Anthony Frayling, and Gavin Smyth. *Predictive Algorithms for Browser Support of Habitual User Activities on the Web*. Tech. rep. MSR-TR-2004-122. Microsoft Research, 2004, p. 11.

[2] Brian D Davison and Haym Hirsh. "Predicting sequences of user actions". In: *Notes of the AAAI/ICML 1998 Workshop on Predicting the Future: AI Approaches to Time-Series Analysis*. 1998, pp. 5–12.

[3] Ron Kohavi, Carla E. Brodley, Brian Frasca, Llew Mason, and Zijian Zheng. "KDD-Cup 2000 organizers' report: Peeling the onion". In: *ACM SIGKDD Explorations Newsletter* 2.2 (2000), pp. 86–93.

[4] Benjamin Korvemaker and Russell Greiner. "Predicting Unix command lines: adjusting to user patterns". In: *Proceedings of AAAI*. 2000, pp. 230–235.