Name: Karina Palyutina
E-mail: kp368@cam.ac.uk
Overseers: Markus Kuhn, Anil Madhavapeddy
Supervisor: Jon Crowcroft
DoS: Sergei Taraskin

“Machine learning inference of search engine heuristics”

The project is one week behind primarily due to a poor original work plan. By nature the project is exploratory, so it wasn’t clear how long each part should take or which methods would be employed until after the research was complete.
The research of machine learning techniques was, therefore, intermittent with the coding process and has taken longer than originally accounted for.

So far I have implemented a crawler and an indexer using existing library tools. PageRank computation is also implemented as described in the original paper. Currently, the search engine returns relevant pages in order of pagerank. This constitutes the basic block of the project: a simple search engine. There are only a few trivial enhancements to this part of the system left.

A natural language processing toolkit library implementation of a Naive Bayes Classifier was used as a prototype for the first learner. The integration involved parsing pages to construct feature sets that are then fed to the learner. The performance of the classifier was evaluated with different parameters and the results were written to persistent storage for further use and analysis. The first results matched the expectations: Bayes performs very well for discrete binary classification. However, once PageRank was incorporated into the search engine to order the results, the classification became almost random. From this I confirmed that Bayes does not work well with continuous features.

An epsilon-Support Vector Machine for regression has been designed and implemented to form the major part of the project. I have used an existing quadratic programming solver, so a large part of the effort went into rewriting the constraints of the regression maximization problem to fit with the library specification (standard QP formulation). Some informal evaluation has been conducted (3 dimensional plots), which showed promising results. More rigorous evaluation of the method will be done using statistical methods allowing us to test further improvements to the search engine.

Throughout the process optimization was an important issue, which was overlooked in the original work plan. Because thousands of pages needed to be parsed quickly and reliably, considerable effort has been invested into profiling and optimizing.

Name:	Karina Palyutina	Supervisor:	Jon Crowcroft
E-mail:	kp368@cam.ac.uk	DoS:	Sergei Taraskin
Overseers:	Markus Kuhn, Anil Madhavapeddy		

Updated work plan

Feb 2 - Feb 20: Complete evaluation of support vector machine learner.
Feb 20 - Mar 5: Add more features to the search engine and evaluate the performance of the SVM learner. Compare to Naive Bayes.
Mar 5 - Mar 25: Implement and evaluate extensions.
Mar 25 - April 10 Write dissertation draft.
Milestone: Dissertation final draft is finished.
April 11 - April 20: Proofreading and submission