Computer Science 6100/4100: Final Project

Due: December 4, 2008

General notes: This assignment is intended to give you some room to explore topics that you think are interesting. I've suggested a couple of ideas which I think are interesting projects to work on, but if you feel like working on something else, I encourage you to do so. However, I ask for two things: (1) make sure that what you are doing makes sense by doing a literature survey and writing about the relevant related research (2) after doing a preliminary literature survey please come and tell me what you're planning to do and what the state of the existing literature is. You have quite a lot of freedom even within the ideas I've given below – these are not intended to be complete specifications of the problems. However, it is absolutely critical to justify the decisions that you take in your final writeup. Working in teams of two on the project is permitted.
A note on the writeup: You must submit a maximum 4-page paper that strictly adheres to the official ACM style file that can be found here (use the "tighter alternate style"):
`http://www.acm.org/sigs/publications/proceedings-templates`
This means you have to make significant editorial decisions about what to include and what not to, but that's part of the challenge of academic writing, at least in Computer Science!

Another note: one of the challenges in both suggested options will be dealing with the fact that the datasets are enormous. You should test your algorithms with a representative subset to figure out what is going on. Either find one someone has made or come up with a good way to construct a representative subset.

1 Option 1: Netflix

Read about the Netflix contest and prize at:
`http://www.netflixprize.com`

Compare at least two approaches to this problem in terms of RMSE on the "probe set." The approaches need not be complicated, but keep in mind that the matrix is very sparse. A first, simple approach, is to predict the following for every (User, Movie) pair: $\alpha\overline{M} + \beta\overline{U}$, where $\overline{M}$ is the average rating for the movie and $\overline{U}$ is the average rating the user has given to all movies she has rated. Learn α and β from the data in some manner. Another approach is to find the k nearest neighbors that also rated the same movie, and then predict the average rating these users gave that movie. You will have to define a meaningful distance metric that does not degenerate because of the sparsity of data, and figure out what to do when no one who is "close" to this user has rated this movie.

2 Option 2: Blogs

Read about the ICWSM 2009 Spinn3r Blog Dataset at:
`http://www.icwsm.org/2009/data/`

Here is one possibility: what is predictive of whether or not a blog post will receive lots of comments (find some reasonable threshold)?

Think about what kinds of features would be useful for this task: in addition to indegree and ranking of the blog, you could try and use the dates and times of the first k comments to try and predict whether the post will receive at least n comments in total. You may also want to use some features of the text in the post, or some more global features of that blog (How often is there a new post on the blog? What is the average number of comments per post on this blog?), or some cross-blog information (does the post link to an item that other posts also link to?). There are lots of possibilites for designing useful features.

Alternatively try to predict some other interesting attribute of blogs or blog posts. This problem is even more open-ended than the previous one. You are welcome to try other tasks on this data.

3 Option 3: Your idea here

If you want to pursue your own idea come and meet with me about it soon. Another well defined possibility is to try and come up with a good prior and a Gittins index strategy that works well on the webpage data from this paper by Vermorel and Mohri (2005):
`http://bandit.sourceforge.net/`