CPSC 445
3/26/07
Al Bashawaty

CPSC 445 – Term Project Proposal

Background

There is a company in California that produces an on-line technology journal. This journal is available through a website where both anonymous and registered users can come and click on stories specifically focused on how IT professionals are solving different challenges in their respective businesses. The content is free. Some people register with the journal to get access to premium services.

The company makes its money based on its ability to get insight into what issues IT professionals are struggling with, and then by selling that aggregate insight to marketing firms that are trying to understand who is most focused on the types of products they are selling and why.

The company proactively sends out emails to new potential readers with headlines and hyperlinks for new content that the journal has put our. The goal of these emails is to drive people to the website.

The journal is only a few years old, but has built up a significant readership. All trend spotting and analysis, to date, has been done manually.

Proposal

For my term project, I propose to take a portion of the company's database (which they are happy to provide me with individual identities removed) to see if I can apply some of the techniques taught in CPSC 445 to data mine for answers to one or more of five questions outlined in the next section. Which questions I focus on will be determined by the exact nature of the dataset that they provide me with.

The Five Questions Data Mining Might Shed Light Upon:

1) Given an email recipient with known attributes, which attributes predict likelihood to visit website;
2) Given a reader attributes, which stories will they be interested in;
3) Given a reader who has read more than one story, what other content are the most likely to also read;
4) Given a reader who we don't have attributes for (an anonymous visitor) and who has read more than one story, which story(ies) will they also read given the first story they read; and
5) Is the Company's taxonomy correct or would some other way of categorizing the similarity between stories be more useful for the purpose of data mining.

The Data:

The Company will provide me with six normalized data tables which will provide me, in aggregate, with the data necessary to identify anonymous and registered users, their known attributes, and the content the found interesting if they went to the web site.

The data tables are of varying sizes and predictably, will have different levels of sparseness. My initial estimate is that, at a minimum, I will have full data on between 3500 and 4000 recipients to analyze.

Useful Attributes for Data Mining

	Reader Attributes	Content Attributes	Format Attributes
Primary Key	Recipient ID IP Address	Content ID Issue ID	
Data Mining Attributes	Title City State Country Zip Phone IT Budget Employees Sales SIC Code Industry Time Sent Time Opened Time of Visit Time Content Click	Abstract Headline Main Content Type Media Type Author Content Hierarchy Click Rate	Template Type Media Type (HTML, Or Video)

Note: Within the scope of this project, I will not have time to look at questions of presentation such as, how do different presentation templates or media types impact readership patterns. I will simply be trying to match reader attributes to content interest.

Understanding with the Company

The Company fully understands that I am doing this project as part of a class and that I will conclude the work with a presentation of the results to the class.

I will not disclose the name of the Company as part of the work, nor will I reveal any proprietary information of the Company as part of this work.

In return for supplying the data sample, I intend to review my work on this project when I complete it with the Company. At that time, the Company will determine whether the project’s results suggest that they should further pursue these types of data mining techniques.