Using SAS® Text Miner for automatic categorization of blog posts on a social networking site dedicated to cyclists

Heramb J. Joshi
SAS & OSU Data Mining certificate program
Introduction

• Cycling
• 42–51 million cyclists per year on an average in US.
• Cyclists discuss about upcoming activities on social media
• Many cyclists post their experience on blogs
• Steep decline in number of cyclists in recent years
Problem statement

• Information is scattered on blogs
• Difficult to access relevant information through blogs
• These blogs need categorization
Data preparation

- Web crawling

- Importing textual data
Approach and methodology

- Two step approach

- Unsupervised classification
- Supervised classification
Unsupervised classification: Text clustering

- Text parsing
- Text filtering
Concept links
## Clustering

![Pie chart and scatter plot showing distance between clusters]

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Descriptive Terms</th>
<th>Frequency (%)</th>
<th>Meaningful Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>+life +feel +bicycle +work +old +find +people +first +year +help</td>
<td>520 (36%)</td>
<td>Ride Experience</td>
</tr>
<tr>
<td>2</td>
<td>+bike +price +sale +sell +interest +buy +contact +condition +rs +frame</td>
<td>170 (12%)</td>
<td>Buy and Sell</td>
</tr>
<tr>
<td>3</td>
<td>+bike +good +buy +budget +gear +suggest mtb +brand +trek +suggestion</td>
<td>191 (13%)</td>
<td>Bicycle Advice</td>
</tr>
<tr>
<td>4</td>
<td>+ride +start +cyclist +join +group +plan +great +day +time +back</td>
<td>565 (39%)</td>
<td>Ride Events</td>
</tr>
</tbody>
</table>
Supervised classification: Text Rule Builder

- Data partitioning (70% – 30%)
- Text parsing
- Text filtering (Mutual information)
- Validation accuracy around 64%
Text rules

- Buy & sell

<table>
<thead>
<tr>
<th>Buy and Sell</th>
<th>Legend</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Interested, Interesting, Interest) AND NOT(helping, help, helped) AND (Selling, sold, sell)</td>
<td>Most Important Words</td>
</tr>
<tr>
<td>(Interested, Interesting, Interest) AND NOT(helping, help, helped) AND (Buying, bought, buy)</td>
<td>Must not contain these words</td>
</tr>
<tr>
<td>Selling</td>
<td></td>
</tr>
<tr>
<td>(Sales, sale) AND NOT(ridding, ridden, ride, rode)</td>
<td>Either of the words must be present</td>
</tr>
<tr>
<td>sale AND accessories</td>
<td></td>
</tr>
</tbody>
</table>
Implementation
Acknowledgements

Mr. Digambar Chaudhari, Founder and CEO at ZERO Emission – Early morning riders (India)
Dr. Goutam Chakraborty for his valuable guidance and motivation for accomplishing this paper.

Heramb J. Joshi (Cycling enthusiast)
Email: - heramb.joshi@okstate.edu Phone: – 512-947-7312