


Text Classification and the Web



Qiang Yang
CMPT 470/882

A Naive Bayesian Classifier for Texts

- Step 1: extract keywords
- Step 2: build a relational table for training
 - Each document → vector of keywords
 - Each keyword → attribute
 - Each attribute =1 iff keyword present in document
- Step 3: build a Classification Model
 - Naive bayesian
 - Association Rules
 - Decision Trees
 - Others (Support Vector Machines, etc)
- Method known as Bag-of-Words

Text Ranking

- Question: given a set of documents and a keyword, how important is the keyword in document search?
- For a given document d and term (keyword) k ,
 - Term Frequency, $tf(k, d)$:
 - The number of times a term appears in a document, normalized on all words

$$tf(k, d) = \frac{count(k, d)}{\sqrt{\sum_{w \in d} count(w)^2}}$$

- Document frequency $df(k, D)$:
 - D : set of documents under consideration
 - $df(k, D)$ =The number of documents containing the term.
 - Inverse document frequency, $idf(k, D)$.

$$idf(k, D) = \log\left(\frac{|D|}{df(k, D)}\right)$$

Example: training data

- Page 1:
 - Text: "Scattered showers and low temperature"
 - Category: Weather
- Page 2:
 - Text: Nortel Networks set the stage for a sharp decline in the tech sector in Nasdaq Wednesday
 - Category: Financial

Keyword Extraction

- Page1: shower, temperature
- Page2: tech, nasdaq

shower	temperature	tech	nasdaq	class
1	1	0	0	weather
0	0	1	1	financial

New Page: Page3="the temperature today is high"
Question: what is the category?

Training the Bayesian Model

- $P(\text{temperature}=1|\text{class}=\text{weather})$
- $P(\text{temperature}=0|\text{class}=\text{weather})$
- $P(\text{temperature}=1|\text{class}=\text{financial})$
- $P(\text{temperature}=0|\text{class}=\text{financial})$
- $P(\text{class}=\text{weather})$
- Likewise for the rest

Applying the Bayesian Model

- Likelihood of "weather"
 - = $P(\text{page3}|\text{weather}=\text{yes}) * P(\text{weather})$
 - = $P(\text{"temperature"}=1, |\text{weather})^*$
 - $P(\text{"shower"}=0|\text{weather})^*$
 - $P(\text{"tech"}=0|\text{weather})^*$
 - $P(\text{"nasdag"}=0|\text{weather}) * P(\text{weather})$
- Likelihood of "financial"
 - = $P(\text{page3}|\text{weather}=\text{yes}) * P(\text{financial})$
 - = $P(\text{"temperature"}=1|\text{financial})^*$
 - $P(\text{"shower"}=0|\text{financial})^*$
 - $P(\text{"tech"}=0|\text{financial})^*$
 - $P(\text{"nasdag"}=0|\text{financial}) * P(\text{financial})$
- Likelihood of Weather > Likelihood of financial
- Thus, Page 3 is a financial page

Application 1: Search Result Categorization

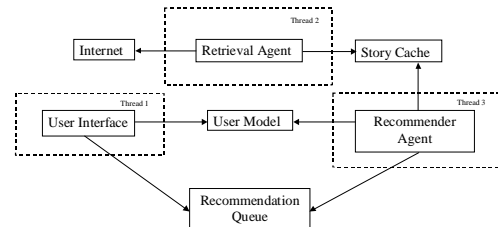
- Susan Dumais Work at Microsoft Research
 - <http://www.research.microsoft.com/~sdumais/>
 - Paper: "bringing order to the web"
 - Training data: web directory (looksmart.com)
 - Classifier: given a page, produce a category
 - Hierarchical categories: from a user interface point of view, offers great results

Application 2: NewsDude @ UCI[*]

[*] D.Billsus, M.Pazzani. *A Hybrid User Model for News Story Classification*. Proc. In UM99, Banff, Canada, June99.
<http://www.ics.uci.edu/~pazzani/Publications/Publications.html> look under (UM99)

- Intelligent agent compiles a daily news program for individual user (info retrieval)
- Architecture: How it works?
 - Short-term vs. Long-term models for user modeling
 - Time-coded feedback to increase prediction accuracy

NewsDude Architecture



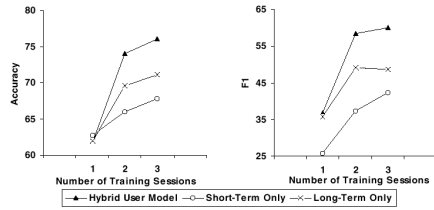
Learning Models

- short-term model (NN)
 - news threads for ongoing recent events
- long-term model (Naïve Bayes Classifier)
 - general news preferences
- hybrid
 - Use short-term model
 - Use long-term model
 - Assign default score

Time-coded feedback

- Use the amount of time a user has listened to a story as implicit feedback
- User's direct binary feedback + Time-coded feedback = Fine-grained scale w/out extra burden on user
 - pl = proportion of a story user has heard
 - If story was rated as uninteresting: score = $0.3 * pl$
 - If story was rated as interesting: score = $0.7 + 0.3 * pl$
 - If user asked for more information: score = 1.0

NewsDude evaluation



NewsDude: Strengths and Limitations

- ✓ Tracks user's changing interests in real-time without sacrificing general interests
- ✓ Simple feedback but accurate prediction
- ✗ Rate enough before personalizing
- ✗ Not flexible: recalculate classifier if adding new keywords
- Similar systems: GroupLens (U.Minnesota)