Text Categorization

Prof. Chris Clifton
10 October 2018

Material adapted from course created by Dr. Luo Si, now leading Alibaba research group

Text Categorization

• Introduction to the task of text categorization
  – Manual vs. automatic text categorization
• Text categorization applications
• Evaluation of text categorization
• K nearest neighbor text categorization method
Text Categorization

- **Tasks**
  - Assign predefined categories to text documents / objects

- **Motivation**
  - Provide an organizational view of the data

- **Large cost of manual text categorization**
  - Millions of dollars spent for manual categorization in companies, governments, public libraries, hospitals
  - Manual categorization is almost impossible for some large scale application (Classification or Web pages)

Text Categorization

- **Automatic text categorization**
  - Learn algorithm to automatically assign predefined categories to text documents / objects
  - automatic or semi-automatic

- **Procedures**
  - **Training**: Given a set of categories and labeled document examples; learn a method to map a document to correct category (categories)
  - **Testing**: Predict the category (categories) of a new document

- Automatic or semi-automatic categorization can significantly reduce manual effort
Text Categorization: Examples

News Categories

Top Stories
World
U.S.
Business
Sci/Tech
Sports
Entertainment
Health
Most Popular

Google News

Search and browse 4,500 news sources updated continuously.

Sci/Tech
Global warming has been a popular topic among scientists
DailyTech: 3 hours ago
The Earth’s average temperature over the past quarter century has been the hottest in four centuries — and part of the world has been warmer during the past 25 years than any period in the past 1,000 years, according to the National Academy of Sciences ...
National panel supports 56 global warming evidence
Boston Globe
No More Dodging Global Temp Threat: Distinct Plus Press
Guardian Unlimited
Seattle Times
Reuters
Forbes
all 441 related »

World's oldest bling: two tiny 100,000-year-old shells
Guardian Unlimited: 5 hours ago
They may not compare with today's diamond-encrusted bling, but in their own way, they are of far greater value. Two tiny shells have been confirmed as the world's oldest known items of jewellery, probably used on a necklace about 100,000 years ago.
Tiny shells may be world's oldest beads
MONEY
Researchers identify What May Be Oldest Known Jewelry
Voice of America
BBC News
New York Times
People's Daily Online
Telegraph.co.uk
all 73 related »

Categories

Computer Science > Human-Computer Interaction (HCI)

CATEGORIES (more this)

- Computer Supported Cooperative Work (CSCW)
- Conferences
- Courses
- Ergonomics
- Information Architecture and Design
- Institutes
- Journals
- Organizations
- Projects
- Web Directories

SITE LISTINGS
By Popularity | Alphabetical | Others Tech

- HCI Bibliography
Features abstracted validated bibliographic entries, along with a variety of reference materials.
www.hcbb.org
Text Categorization: Examples

Example: 1990 US Census

- Included 22 million responses
- Needed to be classified into industry categories (200+) and occupation categories (500+)
- Estimated $15 million if done by hand
- Two alternative automatic text categorization methods evaluated
  - Knowledge-Engineering (Expert System)
  - Machine Learning (k-nearest neighbor method)
Example: 1990 US Census

- Knowledge-Engineering Approach
  - Expert System (Designed by domain expert)
  - Hand-Coded rule
    (e.g., “Professor” and “Lecturer” ➔ “Education”)
  - Development cost: 2 experts, 8 years (192 Person-months)
  - Accuracy = 47%

- Machine Learning Approach
  - k-Nearest Neighbor (KNN) classification
    - “You are like people like you”, details later
  - Fully automatic
  - Development cost: 4 Person-months
  - Accuracy = 60%

Many Applications!

- Web page classification (Yahoo-like category taxonomies)
- News article classification (more formal than most Web pages)
- Automatic email sorting (spam detection; into different folders)
- Word sense disambiguation (Java programming vs. Java in Indonesia)
- Gene function classification (find the functions of a gene from the articles talking about the gene)
- What is your favorite application?...
Techniques Explored in Text Categorization

- Rule-based Expert system (Hayes, 1990)
- Nearest Neighbor methods (Creecy'92; Yang'94)
- Decision symbolic rule induction (Apte'94)
- Naïve Bayes (Language Model) (Lewis'94; McCallum'98)
- Regression method (Furh’92; Yang’92)
- Support Vector Machines (Joachims’98)
- Boosting or Bagging (Schapier‘98)
- Neural networks (Wiener’95)
- ……

Text Categorization: Evaluation

Performance of different algorithms on Reuters-21578 corpus: 90 categories, 7769 Training docs, 3019 test docs, (Yang, JIR 1999)
Text Categorization: Evaluation

Contingency Table Per Category (for all docs)

<table>
<thead>
<tr>
<th></th>
<th>Truth: True</th>
<th>Truth: False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Positive</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Predicted Negative</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>a+c</td>
<td>b+d</td>
</tr>
</tbody>
</table>

- a: number of truly positive docs
- b: number of false-positive docs
- c: number of false negative docs
- d: number of truly-negative docs
- n: total number of test documents

Sensitivity: \( \frac{a}{a+c} \) truly-positive rate, the larger the better
Specificity: \( \frac{d}{b+d} \) truly-negative rate, the larger the better
Depends on decision threshold, trade off between the values
### Text Categorization: Evaluation

**Recall:** $r = \frac{a}{a+c}$  
percentage of positive docs detected

**Precision:** $p = \frac{a}{a+b}$  
how accurate are the predicted positive docs

**Accuracy:** $\frac{a+d}{n}$  
how accurate are all the predicted docs

**F-measure:**

\[
F_\beta = \frac{(\beta^2 + 1)pr}{\beta^2 p + r}
\]

\[
F_1 = \frac{2pr}{p + r}
\]

**Harmonic average:**

\[
\frac{1}{\frac{1}{2} \left( \frac{1}{x_1} + \frac{1}{x_2} \right)}
\]

**Error:** $\frac{b+c}{n}$  
error rate of predicted docs

**Accuracy + Error = 1**

---

**Text Categorization: Evaluation**

- **Micro F1-Measure**
  - Calculate a single contingency table for all categories and calculate F1 measure
  - Treat each prediction with equal weight; better for algorithms that work well on large categories

- **Macro F1-Measure**
  - Calculate a single contingency table for every category; calculate F1 measure separately and average the values
  - Treat each category with equal weight; better for algorithms that work well on many small categories