Overview

- Introduction
  - Web Information Retrieval
  - Machine Learning (ML)
  - ML Methods for Text/Web Data Mining
- Text/Web Data Analysis
  - Text Mining Using Helmholtz Machines
  - Web Mining Using Bayesian Networks
- Summary
  - Current and Future Work
Web Information Retrieval

Preprocessing and Indexing → Text Classification

Text Data

Information Filtering

Information Extraction

Information Filtering System

DB Template Filling & Information Extraction System

user profile

feedback

filtered data

question

answer

DB

Machine Learning

- Supervised Learning
  - Estimate an unknown mapping from known input-output pairs
  - Learn $f_w$ from training set $D=\{(x,y)\}$ s.t. $f_w(x) = y = f(x)$
  - Classification: $y$ is discrete
  - Regression: $y$ is continuous

- Unsupervised Learning
  - Only input values are provided
  - Learn $f_w$ from $D=\{x\}$ s.t. $f_w(x) = x$
  - Density Estimation
  - Compression, Clustering
Machine Learning Methods

- Neural Networks
  - Multilayer Perceptrons (MLPs)
  - Self-Organizing Maps (SOMs)
  - Support Vector Machines (SVMs)
- Probabilistic Models
  - Bayesian Networks (BNs)
  - Helmholtz Machines (HMs)
  - Latent Variable Models (LVMs)
- Other Machine Learning Methods
  - Evolutionary Algorithms (EAs)
  - Reinforcement Learning (RL)
  - Boosting Algorithms
  - Decision Trees (DTs)

ML for Text/Web Data Mining

- Bayesian Networks for Text Classification
- Helmholtz Machines for Text Clustering/Categorization
- Latent Variable Models for Topic Word Extraction
- Boosted Learning for TREC Filtering Task
- Evolutionary Learning for Web Document Retrieval
- Reinforcement Learning for Web Filtering Agents
- Bayesian Networks for Web Customer Data Mining
Preprocessing for Text Learning

From: xxx@sciences.sdsu.edu
Newsgroups: comp.graphics
Subject: Need specs on Apple QT

I need to the specs, or at least a very verbose interpretation of the specs, for QuickTime. Technical articles from magazines and references to books would be nice, too.

I also need the specs in a format usable on a Unix or MS-Dos system. I can't do much with the QuickTime stuff they have on..

Text Mining: Data Sets

- Usenet Newsgroup Data
  - 20 categories
  - 1000 documents for each category
  - 20000 documents in total.

- TDT2 Corpus
  - Target detection and tracking (TDT): NIST
  - Used 6,169 documents in experiments
Text Mining: Helmholtz Machine Architecture

- Input nodes
  - Binary values
  - Represent the existence or absence of words in documents.

- Latent nodes
  - Binary values
  - Extract the underlying causal structure in the document set.
  - Capture correlations of the words in documents.

[Chang and Zhang, 2000]

\[
P(h_i = 1) = \frac{1}{1 + \exp(-b_i - \sum w_{ij}d_j)}
\]

\[
P(d_i = 1) = \frac{1}{1 + \exp(-b_d - \sum w_{id}h_i)}
\]

Text Mining: Learning Helmholtz Machines

- Introduce a recognition network for estimation of a generative network.

\[
\log(D|\theta) = \sum_{t \in T} \log \left( \sum_{\alpha^{(t)}} P(d^{(t)}, \alpha^{(t)}|\theta) \right) = \sum_{t \in T} \log \left( \sum_{\alpha^{(t)}} \frac{Q(\alpha^{(t)}) P(d^{(t)}, \alpha^{(t)}|\theta)}{Q(\alpha^{(t)})} \right) 
\geq \sum_{t \in T} \sum_{\alpha^{(t)}} Q(\alpha^{(t)}) \log \frac{P(d^{(t)}, \alpha^{(t)}|\theta)}{Q(\alpha^{(t)})}
\]

- Wake-Sleep Algorithm
  - Train the recognition and generative models alternately.
  - Update the weight in network iteratively by simple local delta rule.

\[
w_{ij}^{new} = w_{ij}^{old} + \Delta w_{ij}
\]
\[
\Delta w_{ij} = \gamma s_j(s_j - p(s_j = 1))
\]
**Text Mining: Methods**

- **Text Categorization**
  - Train a Helmholtz machine for each category.
  - Total $N$ machines for $N$ categories.
  - Once the $N$ machines have been estimated, classification of a test document proceeds by estimating the likelihood of the document for each machine.

\[
\hat{c} = \arg \max_{c \in \mathcal{C}} [\log P(d \mid c)]
\]

- **Topic Words Extraction**
  - For the entire document sets, train a Helmholtz machine.
  - After training, examine the weights of connections from a latent node to input nodes, that is words.

---

**Text Mining: Categorization Results**

**Usenet Newsgroup Data**

- 20 categories, 1000 documents for each category, 20000 documents in total.

<table>
<thead>
<tr>
<th>category</th>
<th>naive Bayes classifier</th>
<th>Helmholtz machine</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>recall</td>
<td>precision</td>
</tr>
<tr>
<td>1</td>
<td>75.00 %</td>
<td>63.02 %</td>
</tr>
<tr>
<td>2</td>
<td>63.59 %</td>
<td>70.89 %</td>
</tr>
<tr>
<td>4</td>
<td>84.07 %</td>
<td>10.15 %</td>
</tr>
<tr>
<td>3</td>
<td>69.34 %</td>
<td>60.00 %</td>
</tr>
<tr>
<td>1</td>
<td>76.24 %</td>
<td>76.36 %</td>
</tr>
<tr>
<td>5</td>
<td>81.43 %</td>
<td>70.35 %</td>
</tr>
<tr>
<td>6</td>
<td>76.06 %</td>
<td>76.00 %</td>
</tr>
<tr>
<td>7</td>
<td>86.27 %</td>
<td>86.13 %</td>
</tr>
<tr>
<td>8</td>
<td>89.58 %</td>
<td>92.15 %</td>
</tr>
<tr>
<td>9</td>
<td>90.23 %</td>
<td>90.53 %</td>
</tr>
<tr>
<td>10</td>
<td>96.04 %</td>
<td>95.96 %</td>
</tr>
<tr>
<td>11</td>
<td>88.29 %</td>
<td>88.59 %</td>
</tr>
<tr>
<td>12</td>
<td>78.40 %</td>
<td>76.06 %</td>
</tr>
<tr>
<td>13</td>
<td>87.00 %</td>
<td>90.63 %</td>
</tr>
<tr>
<td>14</td>
<td>88.00 %</td>
<td>89.04 %</td>
</tr>
<tr>
<td>15</td>
<td>90.57 %</td>
<td>84.73 %</td>
</tr>
<tr>
<td>16</td>
<td>87.00 %</td>
<td>79.32 %</td>
</tr>
<tr>
<td>17</td>
<td>88.00 %</td>
<td>80.10 %</td>
</tr>
<tr>
<td>18</td>
<td>66.67 %</td>
<td>69.03 %</td>
</tr>
<tr>
<td>19</td>
<td>42.33 %</td>
<td>58.53 %</td>
</tr>
</tbody>
</table>

average: 78.42 % 79.07 % 78.69 % 81.10 % 82.39 % 81.26 %
Text Mining: Topic Words Extraction Results

TD12 Corpus
6,691 documents

1. tobacco, smoking, gingrich, newt, treat, republicans, congressional, republicans, attorney, smokers, lawsuit, senate, cigarette, morris, nicotine

2. warplane, airline, saudi, gulf, Wright, soldiers, Yezhak, tanks, stealth, sabah, stations, kords, merdechai, separatist, governor

3. olympics, nagano, olympic, winter, medal, hockey, athletes, cup, games, slalom, medals, brenze, skating, lillohammer, downhill

4. netanyahu, palestinian, arafat, israeli, yasser, kofi, annan, benjamin, palestinians, mideast, gaza, jerusalem, en, paris, israel

5. India, pakistan, pakistani, delhi, hindu, vajpayee, nuclear, tests, atal, kashmir, indian, janata, bharatiya, Islamabad, bihari

6. suharto, habibie, demonstrators, riots, indonesians, demonstrations, soeharto, resignation, jakarta, rioting, electoral, rallies, wiranto, unrest, megawati

7. IMF, monetary, currencies, currency, rupiah, singapore, bailout, traders, markets, thailand, inflation, investors, fund, banks, baht

8. pope, cuba, cuban, embargo, Castro, lifting, cubans, havana, alan, invasion, reserve, paul, output, vatican, freedom

Web Mining: Customer Analysis

- KDD-2000 Web Mining Competition

  Data: 465 features over 1700 customers
  - Features include friend promotion rate, date visited, weight of items, price of house, discount rate, …
  - Data was collected during Jan. 30 – March 30, 2000
  - Friend promotion was started from Feb. 29 with TV advertisement.

  Aims: Description of heavy/low spenders
Web Mining: Feature Selection

- Features selected by various ways [Yang & Zhang, 2000]

<table>
<thead>
<tr>
<th>Decision Tree + Factor Analysis</th>
<th>Decision Tree</th>
<th>Discriminant Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>V368 (Weight Average)</td>
<td>V243 (OrderLine Quantity Sum)</td>
<td>V240 (Friend)</td>
</tr>
<tr>
<td>V245 (OrderLine Quantity Maximum)</td>
<td>V234 (OrderItemQuantity Sum% HavingDiscountRange(5 : 10))</td>
<td>V229 (Order-Average)</td>
</tr>
<tr>
<td>F1 = 0.94<em>V234 + 0.868</em>V374 + 0.898*V412</td>
<td>V237 (OrderItemQuantitySum% HavingDiscountRange(10.))</td>
<td>V304 (OrderShippingAmtMin.)</td>
</tr>
<tr>
<td>F2 = 0.829<em>V234 + 0.857</em>V240</td>
<td>V240 (OrderLineQuantitySum)</td>
<td>V368 (Weight Average)</td>
</tr>
<tr>
<td>F3 = -0.795<em>V237 + 0.778</em>V304</td>
<td>V245 (OrderLineQuantity Maximum)</td>
<td>V43 (Home Market Value)</td>
</tr>
<tr>
<td></td>
<td>V304 (OrderShippingAmtMin)</td>
<td>V368 (Weight Average)</td>
</tr>
<tr>
<td></td>
<td>V324 (NumLegwearProduct Views)</td>
<td>V43 (Home Market Value)</td>
</tr>
<tr>
<td></td>
<td>V368 (Weight Average)</td>
<td>V377 (NumAccountTemplate Views)</td>
</tr>
<tr>
<td></td>
<td>V243 (OrderLineQuantitySum)</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>V245 (OrderLineQuantity Maximum)</td>
<td>V11 (Which)</td>
</tr>
<tr>
<td></td>
<td>V304 (OrderShippingAmtMin)</td>
<td>DoYouWearMostFrequent</td>
</tr>
<tr>
<td></td>
<td>V324 (NumLegwearProduct Views)</td>
<td>V13 (SendEmail)</td>
</tr>
<tr>
<td></td>
<td>V368 (Weight Average)</td>
<td>V17 (USState)</td>
</tr>
<tr>
<td></td>
<td>V374 (NumMainTemplateViews)</td>
<td>V45 (VehicleLifeStyle)</td>
</tr>
<tr>
<td></td>
<td>V412 (NumReplenishable Stock Views)</td>
<td>V68 (RetailActivity)</td>
</tr>
<tr>
<td></td>
<td>V374 (NumMainTemplateViews)</td>
<td>V19 (Date)</td>
</tr>
</tbody>
</table>

Web Mining: Bayesian Nets

- Bayesian network
  - DAG (Directed Acyclic Graph)
  - Express dependence relations between variables
  - Can use prior knowledge on the data (parameters)

\[
P(A, B, C, D, E) = P(A)P(B|A)P(C|B)P(D|A, B)P(E|B, C, D)
\]

- Examples of conjugate priors:
  - Dirichlet for multinomial data
  - Normal-Wishart for normal data
Web Mining: Results

- A Bayesian net for KDD web data
- V229 (Order-Average) and V240 (Friend) directly influence V312 (Target)
- V19 (Date) was influenced by V240 (Friend) reflecting the TV advertisement.

Summary

- We study machine learning methods, such as
  - Probabilistic neural networks
  - Evolutionary algorithms
  - Reinforcement learning
- Application areas include
  - Text mining
  - Web mining
  - Bioinformatics (not addressed in this talk)
- Recent work focuses on probabilistic graphical models for web/text/bio data mining, including
  - Bayesian networks
  - Helmholtz machines
  - Latent variable models
Bayesian Networks: Architecture

A Bayesian network represents the probabilistic relationships between the variables.

\[ P(L, B, G, M) = P(L)P(B | L)P(G | L, B)P(M | L, B, G) \]
\[ = P(L)P(B)P(G | B)P(M | B, L) \]

- A Bayesian network represents the probabilistic relationships between the variables.

\[ P(X) = \prod_{j=1}^{n} P(X_j | pa_j) \]
\( pa_j \) is the set of parent nodes of \( X_j \).
Bayesian Networks:
Applications in IR – A Simple BN for Text Classification

- $C$: document class
- $t_i$: $i^{th}$ term

- The network structure represents the naïve Bayes assumption.
- All nodes are binary.
- [Hwang & Zhang, 2000]

Bayesian Networks:
Experimental Results

- Dataset
  - The acq dataset from Reuters-21578
  - 8754 terms were selected by TFIDF.
  - Training data: 8762 documents
  - Test data: 3009 documents

- Parametric Learning
  - Dirichlet prior assumptions for the network parameter distributions.
    
    \[ p(\theta_y | S^k) = \text{Dir}(\theta_y | \alpha_{y1}, \ldots, \alpha_{yv}) \]
    
  - Parameter distributions are updated with training data.
    
    \[ p(\theta_y | D, S^k) = \text{Dir}(\theta_y | \alpha_{y1} + N_{y1}, \ldots, \alpha_{yv} + N_{yv}) \]
Bayesian Networks: Experimental Results

- For training data
  - Accuracy: 94.28%

<table>
<thead>
<tr>
<th></th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive examples</td>
<td>96.83</td>
<td>75.98</td>
</tr>
<tr>
<td>Negative examples</td>
<td>93.76</td>
<td>99.32</td>
</tr>
</tbody>
</table>

- For test data
  - Accuracy: 96.51%

<table>
<thead>
<tr>
<th></th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive examples</td>
<td>95.16</td>
<td>89.17</td>
</tr>
<tr>
<td>Negative examples</td>
<td>96.88</td>
<td>98.67</td>
</tr>
</tbody>
</table>

Latent Variable Models: Architecture

- [Shin & Zhang, 2000]
- Maximize log-likelihood

\[
L = \sum_{n=1}^{N} \sum_{m=1}^{M} n(d_n, w_m) \log P(d_n, w_m)
\]

\[
= \sum_{n=1}^{N} \sum_{m=1}^{M} n(d_n, w_m) \log \sum_{z_i=1}^{K} P(z_i) P(w_m | z_i) P(d_n | z_i)
\]

- Update \( P(z_i) \), \( P(w_m | z_i) \), \( P(d_n | z_i) \)
- With EM Algorithm
EM (Expectation-Maximization) Algorithm
- Algorithm to maximize pre-defined log-likelihood

Iteration of E-Step and M-Step

<table>
<thead>
<tr>
<th>E-Step</th>
<th>M-Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(z_i \mid d_s, w_n) = \frac{P(z_i)P(d_s \mid z_i)P(w_n \mid z_i)}{\sum_{i=1}^{K} P(z_i)P(d_s \mid z_i)P(w_n \mid z_i)} )</td>
<td>( P(w_n) = \frac{\sum_{i=1}^{N} n(d_s, w_n)P(z_i \mid d_s, w_n)}{\sum_{i=1}^{M} \sum_{n=1}^{N} n(d_s, w_n)P(z_i \mid d_s, w_n)} )</td>
</tr>
<tr>
<td>( P(d_s \mid z_i) = \frac{\sum_{n=1}^{N} n(d_s, w_n)P(z_i \mid d_s, w_n)}{\sum_{n=1}^{M} \sum_{n=1}^{N} n(d_s, w_n)P(z_i \mid d_s, w_n)} )</td>
<td>( P(z_i) = \frac{1}{R} \sum_{s=1}^{S} \sum_{n=1}^{N} n(d_s, w_n)P(z_i \mid d_s, w_n) )</td>
</tr>
<tr>
<td>( R = \sum_{s=1}^{S} \sum_{n=1}^{N} n(d_s, w_n) )</td>
<td></td>
</tr>
</tbody>
</table>

Latent Variable Models:
Applications in IR – Experimental Results

- Topic Words Extraction and Document Clustering with a subset of TREC-8 data
- TREC-8 adhoc task data
  - Documents: DTDS, FR94, FT, FBIS, LATIMES
  - Topics: 401-450 (401, 434, 439, 450)
    - 401: Foreign Minorities, Germany
    - 434: Estonia, Economy
    - 439: Inventions, Scientific discovery
    - 450: King Hussein, Peace
Latent Variable Models:
Applications in IR – Experimental Results

| Label (assigned to $z_k$ with Maximum $P(d_i|z_k)$) | Topic (#Docs) | $z_1$ | $z_2$ | $z_3$ | $z_4$ | Precision | Recall |
|-------------------------------------------------|--------------|------|------|------|------|-----------|--------|
|                                                 | 401 (300)    | 279  | 1    | 0    | 20   | 0.902     | 0.930  |
|                                                 | 434 (347)    | 238  | 10   | 79   | 203  | 0.996     | 0.686  |
|                                                 | 439 (219)    | 7    | 0    | 9    | 503  | 0.953     | 0.927  |
|                                                 | 450 (293)    | 3    | 0    | 0    | 290  | 0.729     | 0.990  |

**Topics**

- **Cluster 1 ($z_1$)**: jordan, peac, isreal, palestinian, king, isra, arab, meet, talk, husayn, agreem, presid, majesti, negot, minist, visit, region, arafat, secur, peopl, east, washington, econom, sign, relat, jerusalem, rabin, syria, iraq, ...
- **Cluster 2 ($z_2$)**: german, germani, mr, parti, year, foreign, people, countri, govern, asylum, polit, nation, law, minist, europ, state, immig, democrat, wing, social, turkish, west, east, member, attack, ...
- **Cluster 3 ($z_3$)**: percent, estonia, bank, state, privat, russian, year, enterprise, trade, million, trade, estonian, econom, countri, govern, compani, foreign, baltic, polish, loan, invest, fund, product, ...
- **Cluster 4 ($z_4$)**: research, technology, develop, mar, materi, system, nuclear, environment, electr, process, product, power, energi, control, japan, pollution, structur, chemic, plant, ...

**Extracted Topic Words (top 35 words with highest $P(w_j|z_k)$)**

- **Cluster 1 ($z_1$)**: jordan, peac, isreal, palestinian, king, isra, arab, meet, talk, husayn, agreem, presid, majesti, negot, minist, visit, region, arafat, secur, peopl, east, washington, econom, sign, relat, jerusalem, rabin, syria, iraq, ...
- **Cluster 2 ($z_2$)**: german, germani, mr, parti, year, foreign, people, countri, govern, asylum, polit, nation, law, minist, europ, state, immig, democrat, wing, social, turkish, west, east, member, attack, ...
- **Cluster 3 ($z_3$)**: percent, estonia, bank, state, privat, russian, year, enterprise, trade, million, trade, estonian, econom, countri, govern, compani, foreign, baltic, polish, loan, invest, fund, product, ...
- **Cluster 4 ($z_4$)**: research, technology, develop, mar, materi, system, nuclear, environment, electr, process, product, power, energi, control, japan, pollution, structur, chemic, plant, ...

---

**Boosting: Algorithms**

- A general method of converting rough rules into a highly accurate prediction rule
- **Learning procedure**
  - Examine the training set
  - Derive a rough rule (weak learner)
  - Re-weight the examples in the training set, concentrating on the hard cases for previous rules
  - Repeat $T$ times

Importance weights of training documents
**Boosting:**

Applied to Text Filtering

- Naïve Bayes
  - Traditional algorithm for text filtering
    \[ c_{ML} = \arg \max_{c \in \{relevant, irrelevant\}} P(c_i | d_j) \]
    \[ = \arg \max_{c_j} P(c_j) \prod_{k=1}^{n} P(w_{ij} | c_j) \]
    \[ = \arg \max_{c_j} P(c_j) P(w_{i1} = "our" | c_j) P(w_{i2} = "approach" | c_j) \ldots \]
    \[ P(w_{in} = "trouble" | c_j) \]
  - Boosting naïve Bayes
    - Using naïve Bayes classifiers as weak learners
    - [Kim & Zhang, SIGIR-2000]

**TREC** (Text Retrieval Conference)
- Sponsored by NIST
- TREC-7 filtering datasets
  - Training Documents
    - AP articles (1988)
    - 237 MB, 79919 documents
  - Test Documents
    - 471 MB, 162999 documents
- No. of topics: 50

Example of a document

- TREC-8 filtering datasets
  - Training Documents
    - 167 MB, 64139 documents
  - Test Documents
    - 382 MB, 140651 documents
- No. of topics: 50

PORT-NO-PRIME, Haiti (AP) — Opposition leader Lozé
summoned to court next week for a hearing apparently
arrest on charges of inciting the public to revolt, a
told Friday.

...
Boosting:
Applied to Text Filtering – Experimental Results

Compared with the state-of-the-art text filtering systems

TREC-7

<table>
<thead>
<tr>
<th>Averaged Scaled F1</th>
<th>Averaged Scaled F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boosting</td>
<td>ATT</td>
</tr>
<tr>
<td></td>
<td>0.474</td>
</tr>
<tr>
<td></td>
<td>0.467</td>
</tr>
</tbody>
</table>

TREC-8

<table>
<thead>
<tr>
<th>Averaged Scaled LF1</th>
<th>Averaged Scaled LF2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boosting</td>
<td>PLT1</td>
</tr>
<tr>
<td></td>
<td>0.717</td>
</tr>
<tr>
<td></td>
<td>0.722</td>
</tr>
</tbody>
</table>

Evolutionary Learning:
Applications in IR - Web-Document Retrieval

[Kim & Zhang, 2000]
Evolutionary Learning:
Applications in IR – Tag Weighting

- Crossover
  - Chromosome X
  - Chromosome Y
  - Chromosome Z (offspring)
  -\[ z_i = \frac{(x_i + y_i)}{2} \text{ w.p. } P_c \]
  -\[ z_1, z_2, z_3, \ldots, z_n \]

- Mutation
  - Chromosome X
  - Chromosome X’
  - Change value w.p. \( P_m \)

- Truncation selection

Evolutionary Learning:
Applications in IR - Experimental Results

- Datasets
  - TREC-8 Web Track Data
  - 2GB, 247491 web documents (WT2g)
  - No. of training topics: 10, No. of test topics: 10

- Results

![Graph showing experimental results]
Reinforcement Learning:
Basic Concept

1. State $s_t$
2. Action $a_t$
3. Reward $r_{t+1}$
4. State $s_{t+1}$

Reinforcement Learning:
Applications in IR - Information Filtering

[Seo & Zhang, 2000]

1. $State_i$ (user profile)
2. $Action_i$ (modify profile)
3. $Reward_{i+1}$ (relevance feedback)
4. $State_{i+1}$

• retrieve documents
• calculate similarity

User profile
Document filtering
Filtered documents
Reinforcement Learning:
Experimental Results (Explicit Feedback)

Number of filtered HTML documents

Average of explicit relevance feedback

Reinforcement Learning:
Experimental Results (Implicit Feedback)

Number of filtered HTML documents

Average of implicit relevance feedback