Machine Learning Methods for Text / Web Data Mining

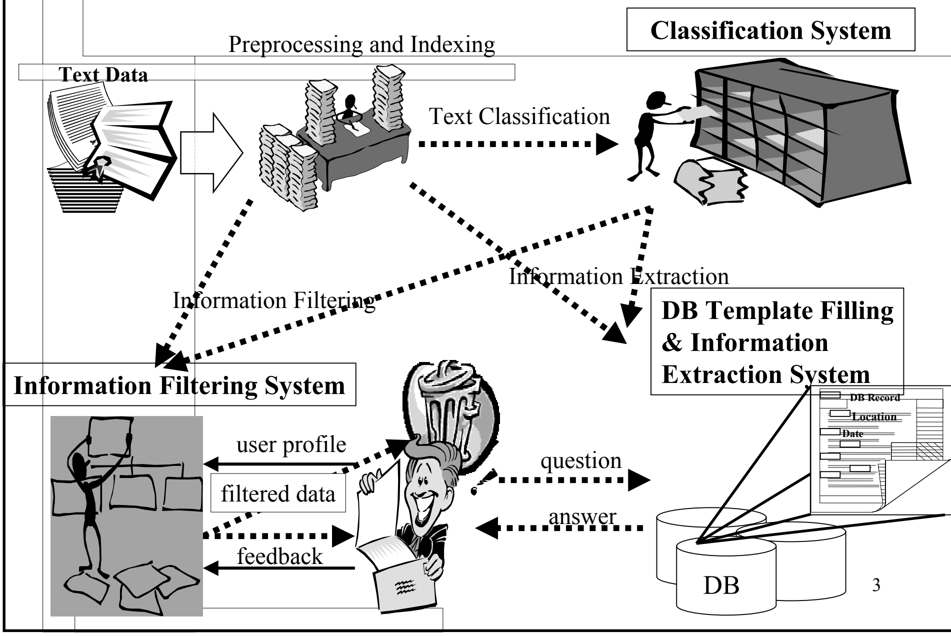
Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University
E-mail: btzhang@cse.snu.ac.kr

This material is available at
http://scai.snu.ac.kr./~btzhang/

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



Machine Learning

- Supervised Learning
  - Estimate an unknown mapping from known input-output pairs
  - Learn $f_{\mathbf{w}}$ from training set $D=\{(\mathbf{x},y)\}$ s.t. $f_{\mathbf{w}}(\mathbf{x}) = y = f(\mathbf{x})$
  - Classification: y is discrete
  - Regression: y is continuous
- Unsupervised Learning
  - Only input values are provided
  - Learn $f_{\mathbf{w}}$ from $D=\{(\mathbf{x})\}$ s.t. $f_{\mathbf{w}}(\mathbf{x}) = \mathbf{x}$
  - Density Estimation
  - Compression, Clustering

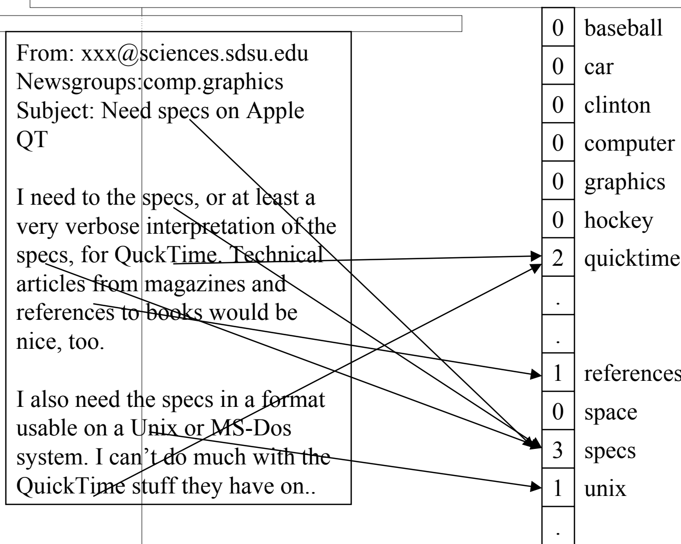
Machine Learning Methods

- 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

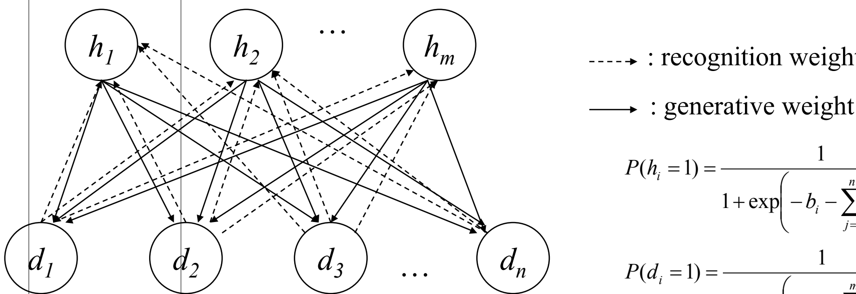
Preprocessing for Text Learning



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



$$P(h_i = 1) = \frac{1}{1 + \exp\left(-b_i - \sum_{j=1}^{n} w_{ij} d_j\right)}$$

$$P(d_i = 1) = \frac{1}{1 + \exp\left(-b_i - \sum_{j=1}^{m} w_{ij} h_j\right)}$$

- [Chang and Zhang, 2000]
- 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.

Text Mining:
Learning Helmholtz Machines

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

$$\log(D \mid \theta) = \sum_{t=1}^{T} \log\left[\sum_{\alpha(t)} P\left(d^{(t)}, \alpha^{(t)} \mid \theta\right)\right] = \sum_{t=1}^{T} \log\left[\sum_{\alpha(t)} Q\left(\alpha^{(t)}\right) \frac{P\left(d^{(t)}, \alpha^{(t)} \mid \theta\right)}{Q\left(\alpha^{(t)}\right)}\right]$$

$$\geq \sum_{t=1}^{T} \sum_{\alpha^{(t)}} Q\left(\alpha^{(t)}\right) \log \frac{P\left(d^{(t)}, \alpha^{(t)} \mid \theta\right)}{Q\left(\alpha^{(t)}\right)}$$

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

$$w_{ij}^{new} = w_{ij}^{old} + \Delta w_{ij}$$

$$\Delta w_{ij} = \gamma\, s_i (s_j - p(s_j = 1))$$

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.

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

- 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.

category	naive Bayes classifier			Helmholtz machine		
	recall	precision	F1	recall	precision	F1
0	75.00 %	63.92 %	69.02 %	68.67 %	73.57 %	71.04 %
1	80.33 %	63.59 %	70.99 %	74.67 %	67.67 %	71.00 %
2	**5.67** %	77.27 %	10.56 %	**76.67** %	77.44 %	77.05 %
3	80.67 %	60.35 %	69.05 %	73.00 %	70.19 %	71.57 %
4	85.00 %	69.11 %	76.24 %	75.33 %	77.93 %	76.61 %
5	81.00 %	70.43 %	75.35 %	81.67 %	79.54 %	80.59 %
6	74.67 %	82.96 %	78.60 %	80.33 %	79.02 %	79.67 %
7	88.00 %	86.27 %	87.13 %	87.00 %	86.71 %	86.85 %
8	94.00 %	90.38 %	92.15 %	91.00 %	94.14 %	92.54 %
9	96.33 %	94.75 %	95.53 %	93.33 %	95.56 %	94.43 %
10	95.00 %	96.94 %	95.96 %	95.00 %	95.64 %	95.32 %
11	88.29 %	88.89 %	88.59 %	88.29 %	94.96 %	91.50 %
12	75.00 %	78.40 %	76.66 %	81.00 %	77.88 %	79.41 %
13	87.00 %	90.63 %	88.78 %	86.00 %	87.76 %	86.87 %
14	88.00 %	90.10 %	89.04 %	90.67 %	87.18 %	88.89 %
15	90.57 %	79.59 %	84.73 %	90.57 %	86.50 %	88.49 %
16	87.00 %	79.32 %	82.90 %	81.33 %	80.79 %	81.06 %
17	88.00 %	90.10 %	89.04 %	88.33 %	90.44 %	89.37 %
18	66.67 %	69.93 %	68.26 %	69.67 %	70.37 %	70.02 %
19	42.33 %	58.53 %	49.13 %	51.33 %	54.41 %	52.83 %
average	78.42 %	79.07 %	76.89 %	81.19 %	82.39 %	81.26 %

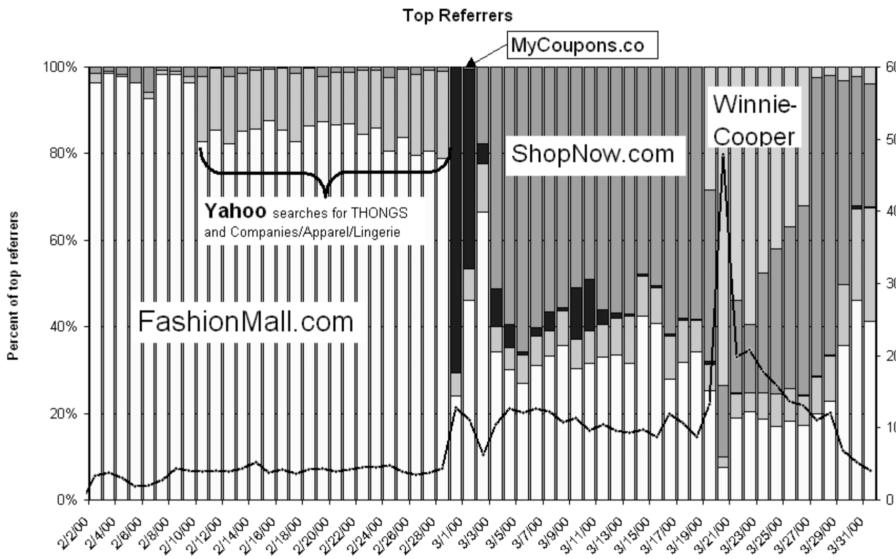
Text Mining: Topic Words Extraction Results

- TDT2 Corpus
- 6,169 documents

1	tabacco, smoking, gingrich, newt, trent, republicans, congressional, republicans, attorney, smokers, lawsuit, senate, cigarette, morris, nicotine
2	warplane, airline, saudi, gulf, wright, soldiers, yitzhak, tanks, stealth, sabah, stations, kurds, mordechai, separatist, governor
3	olympics, nagano, olympic, winter, medal, hockey, atheletes, cup, games, slalom, medals, bronze, skating, lillehammer, downhill
4	netanyahu, palestinian, arafat, israeli, yasser, kofi, annan, benjamin, palestinians, mideast, gaza, jerusalem, eu, 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



?,?,?,?,NULL,?,?,?,?,?,?,NULL,NULL,?,?,NULL,NULL,NULL,?,?,?,NULL,?,?,NULL,NULL,NULL,?,?,?,NULL,NULL,NULL,NULL,NULL,NULL,NU
?,?,Male,4 or more,NULL,?,?,?,?,2,?,NULL,True,?,Friend/Co-worker,Westport,United States,CT,2000-01-27,20₩:48₩:03,1963,Gazell
?,?,Male,0,NULL,?,?,?,?,2,?,NULL,True,?,Friend/Co-worker,Novato,United States,CA,2000-01-29,10₩:20₩:33,1961,Gazelle,0,132,NU
?,?,Male,0,NULL,?,?,?,?,2,?,NULL,True,?,Friend/Co-worker,Cupertino,United States,CA,2000-01-29,13₩:52₩:26,1953,Gazelle,0,168,
?,?,Male,2,NULL,?,?,?,?,2,?,NULL,False,?,Other,San Ramon,United States,CA,2000-01-29,17₩:55₩:38,?,COM,0,184,NULL,NULL,NUL
?,?,Female,2,NULL,?,?,?,?,2,?,NULL,True,?,Other,Scarsdale,United States,NY,2000-01-30,14₩:06₩:43,1956,COM,?,224,False,False,
?,?,Male,0,NULL,?,?,?,?,1,?,NULL,True,?,Other,Novato,United States,CA,2000-01-30,16₩:14₩:14,?,Gazelle,0,236,False,False,False,
?,?,Female,4 or more,NULL,?,?,?,?,2,?,NULL,True,?,Other,Westport,United States,CT,2000-01-30,19₩:00₩:28,1962,COM,0,240,True,
?,?,Male,2,NULL,?,?,?,?,2,?,NULL,True,?,Friend/Co-worker,San Francisco,United States,CA,2000-01-30,20₩:20₩:25,1957,NET,0,25
?,?,Male,0,NULL,?,?,?,?,1,?,NULL,False,?,Friend/Co-worker,San Jose,United States,CA,2000-01-31,10₩:16₩:02,?,COM,0,344,False
?,?,Female,0,NULL,?,?,?,?,1,?,NULL,True,?,Other,NY,United States,NY,2000-01-31,10₩:23₩:12,1975,COM,0,356,NULL,NULL,NULL,?,
?,?,Female,0,NULL,?,?,?,?,2,?,NULL,True,?,Friend/Co-worker,new haven,United States,CT,2000-01-31,10₩:27₩:35,1966,Gazelle,0,
?,?,?,?,NULL,?,?,?,?,?,?,NULL,NULL,?,?,Stamford,United States,CT,2000-01-31,13₩:47₩:50,?,NULL,?,474,False,False,False,?,36,40,N
?,?,Female,2,NULL,?,?,?,?,3 or more,?,NULL,True,?,Friend/Co-worker,Stamford,United States,CT,2000-01-31,14₩:42₩:48,1962,COM
?,?,Female,0,NULL,?,?,?,?,2,?,NULL,False,?,Friend/Co-worker,Menlo Park,United States,CA,2000-01-31,17₩:44₩:11,1968,COM,0,59
?,?,Male,0,NULL,?,?,?,?,1,?,NULL,False,?,Friend/Co-worker,San Francisco,United States,CA,2000-01-31,18₩:44₩:21,?,NET,4,628,F
?,?,Male,0,NULL,?,?,?,?,0,?,NULL,True,?,Other,Oakland,United States,CA,2000-01-31,20₩:28₩:38,?,COM,7,684,False,False,False,?,
?,?,Male,0,NULL,?,?,?,?,2,?,NULL,False,?,Friend/Co-worker,Norwalk,United States,CT,2000-01-31,20₩:44₩:25,?,NET,?,692,False,Fa
?,?,Male,0,NULL,?,?,?,?,0,?,NULL,True,?,Friend/Co-worker,Austin,United States,TX,2000-01-31,20₩:51₩:28,?,COM,?,696,NULL,NUL
?,?,?,?,NULL,?,?,?,?,?,?,NULL,NULL,?,?,Novato,United States,CA,2000-01-31,22₩:09₩:35,?,NULL,?,726,False,False,False,?,34,?,Mar
?,?,Male,0,NULL,?,?,?,?,1,?,NULL,True,?,Friend/Co-worker,San Francisco,United States,CA,2000-01-31,22₩:20₩:28,1968,COM,0,7
?,?,Female,2,NULL,?,?,?,?,1,?,NULL,True,?,Friend/Co-worker,Westport,United States,CT,2000-02-01,07₩:04₩:31,1955,COM,?,790,F
?,?,Male,0,NULL,?,?,?,?,2,?,NULL,True,?,Friend/Co-worker,San Francisco,United States,CA,2000-02-01,09₩:12₩:22,1970,COM,0,84
?,?,Male,0,NULL,?,?,?,?,2,?,NULL,False,?,Friend/Co-worker,San Francisco,United States,CA,2000-02-01,09₩:57₩:56,?,COM,?,848,
?,?,Female,0,NULL,?,?,?,?,0,?,NULL,True,?,Friend/Co-worker,Menlo Park,United States,CA,2000-02-01,10₩:07₩:27,?,Gazelle,0,860,

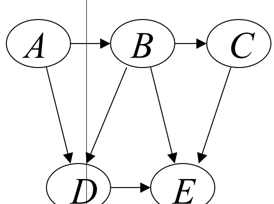
Web Mining: Feature Selection

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

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

Web Mining: Bayesian Nets

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

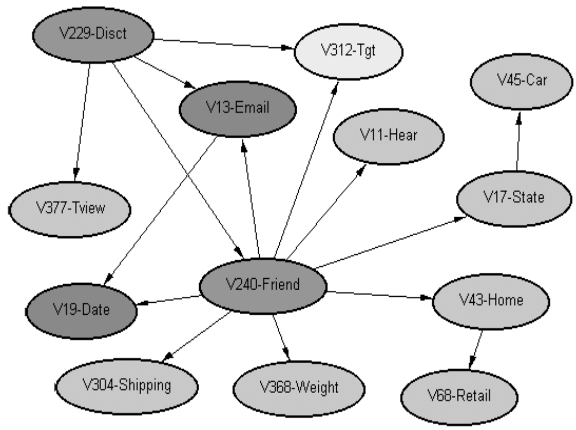


$$P(A,B,C,D,E) = P(A)P(B|A)P(C|B)$$
$$P(D|A,B)P(E|B,C,D)$$

  - Examples of conjugate priors:
    Dirichlet for multinomial data, Normal-Wishart for normal data

Web Mining: Results

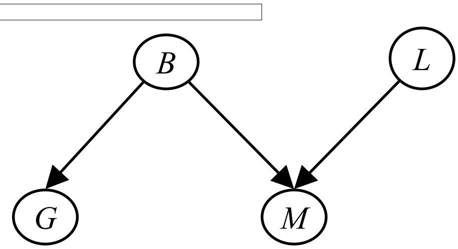
- A Bayesian net for KDD web data
- V229 (Order-Average) and V240 (Friend) directly influence V312 (Target)
- V19 (Date) was influenced by V240 (Friend) reflecting the TV advertisement.



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

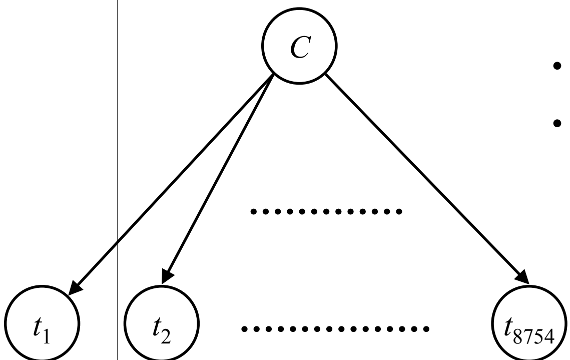


$$P(L,B,G,M) = P(L)P(B \mid L)P(G \mid L,B)P(M \mid L,B,G)$$
$$= P(L)P(B)P(G \mid B)P(M \mid B,L)$$

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

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i \mid \mathbf{pa}_i)$$ $\mathbf{pa}_i$ is the set of parent nodes of X_i.

Bayesian Networks:
Applications in IR – A Simple BN for Text Classification



- C: document class
- t_i: ith 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.

    $$p(\theta_{ij} \mid S^h) = \mathrm{Dir}(\theta_{ij} \mid \alpha_{ij1}, \ldots, \alpha_{ijr_i})$$

  - Parameter distributions are updated with training data.

    $$p(\theta_{ij} \mid D, S^h) = \mathrm{Dir}(\theta_{ij} \mid \alpha_{ij1} + N_{ij1}, \ldots, \alpha_{ijr_i} + N_{ijr_i})$$

Bayesian Networks:
Experimental Results

- For training data
  - Accuracy: 94.28%

	Recall (%)	Precision (%)
Positive examples	96.83	75.98
Negative examples	93.76	99.32

- For test data
  - Accuracy: 96.51%

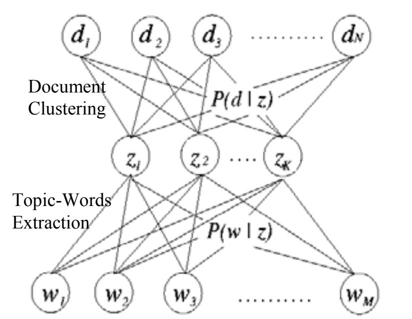
	Recall (%)	Precision (%)
Positive examples	95.16	89.17
Negative examples	96.88	98.67

Latent Variable Models:
Architecture

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

$$L = \sum_{n=1}^{N}\sum_{m=1}^{M} n(d_n, w_m)\log P(d_n, w_m)$$
$$= \sum_{n=1}^{N}\sum_{m=1}^{M} n(d_n, w_m)\log \sum_{k=1}^{K} P(z_k)P(w_m \mid z_k)P(d_n \mid z_k)$$

  - Update $P(z_k)$, $P(w_m \mid z_k)$, $P(d_n \mid z_k)$.
  - With EM Algorithm



Latent Variable Model for Topic Words Extraction and Document Clustering

Latent Variable Models: Learning

- EM (Expectation-Maximization) Algorithm
  - Algorithm to maximize pre-defined log-likelihood
- Iteration of E-Step and M-Step
  - E-Step

$$P(z_k \mid d_n, w_m) = \frac{P(z_k)P(d_n \mid z_k)P(w_m \mid z_k)}{\sum_{k=1}^{K} P(z_k)P(d_n \mid z_k)P(w_m \mid z_k)}$$

  - M-Step

$$P(w_m \mid z_k) = \frac{\sum_{n=1}^{N} n(d_n, w_m)P(z_k \mid d_n, w_m)}{\sum_{m=1}^{M}\sum_{n=1}^{N} n(d_n, w_m)P(z_k \mid d_n, w_m)}$$

$$P(d_n \mid z_k) = \frac{\sum_{m=1}^{M} n(d_n, w_m)P(z_k \mid d_n, w_m)}{\sum_{m=1}^{M}\sum_{n=1}^{N} n(d_n, w_m)P(z_k \mid d_n, w_m)}$$

$$P(z_k) = \frac{1}{R}\sum_{m=1}^{M}\sum_{n=1}^{N} n(d_n, w_m)P(z_k \mid d_n, w_m),$$

$$R \equiv \sum_{m=1}^{M}\sum_{n=1}^{N} n(d_n, w_m)$$

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, <u>FT</u>, <u>FBIS</u>, <u>LATIMES</u>
  - 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

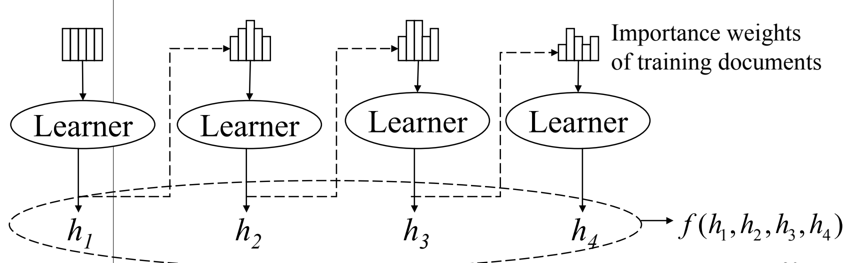
Label (assigned to z_k with *Maximum $P(d_i|z_k)$*)						
Topic (#Docs)	z_2	z_4	z_3	z_1	Precision	Recall
401 (300)	**279**	1	0	20	0.902	0.930
434 (347)	20	**238**	10	79	0.996	0.686
439 (219)	7	0	**203**	9	0.953	0.927
450 (293)	3	0	0	**290**	0.729	0.990

Topics	**Extracted Topic Words (top 35 words with highest *$P(w_i|z_k)$*)**
Cluster 2 (z_2)	german, germani, mr, parti, year, foreign, people, countri, govern, asylum, polit, nation, law, minist, europ, state, immigr, democrat, wing, social, turkish, west, east, member, attack, …
Cluster 4 (z_4)	percent, estonia, bank, state, privat, russian, year, enterprise, trade, million, trade, estonian, econom, countri, govern, compani, foreign, baltic, polish, loan, invest, fund, product, …
Cluster 3 (z_3)	research, technology, develop, mar, materi, system, nuclear, environment, electr, process, product, power, energi, countrol, japan, pollution, structur, chemic, plant, …
Cluster 1 (z_1)	jordan, peac, isreal, palestinian, king, isra, arab, meet, talk, husayn, agreem, presid, majesti, negoti, minist, visit, region, arafat, secur, peopl, east, washington, econom, sign, relat, jerusalem, rabin, syria, iraq, …

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



Boosting:
Applied to Text Filtering

- Naïve Bayes
  - Traditional algorithm for text filtering

$$c_{NM} = \operatorname*{arg\,max}_{c_j \in \{relevant,\ irrelevant\}} P(c_j)P(d_i \mid c_j)$$

$$= \operatorname*{arg\,max}_{c_j} P(c_j)\prod_{k=1}^{n} P(w_{ik} \mid c_j)$$

$$= \operatorname*{arg\,max}_{c_j} P(c_j)P(w_{i1} = \text{"}our\text{"} \mid c_j)P(w_{i2} = \text{"}approach\text{"} \mid c_j)\ \ldots$$

$$P(w_{in} = \text{"}trouble\text{"} \mid c_j)$$

Assume independence among terms

- Boosting naïve Bayes
  - Using naïve Bayes classifiers as weak learners
  - [Kim & Zhang, SIGIR-2000]

Boosting:
Applied to Text Filtering – Experimental Results

- TREC (Text Retrieval Conference)
  - Sponsored by NIST
- TREC-7 filtering datasets
  - Training Documents
    - AP articles (1988)
    - 237 MB, 79919 documents
  - Test Documents
    - AP articles (1989~1990)
    - 471 MB, 162999 documents
  - No. of topics: 50
- TREC-8 filtering datasets
  - Training Documents
    - Financial Times (1991~1992)
    - 167 MB, 64139 documents
  - Test Documents
    - Financial Time (1993~1994)
    - 382 MB, 140651 documents
  - No. of topics: 50

Example of a document

```
<DOC>
<DOCNO> AP880213-0001 </DOCNO>
<FILEID>AP-NR-02-13-88 0549EST</FILEID>
<FIRST>r i AM-Haiti-Opposition     02-13 0245</FIRST>
<SECOND>AM-Haiti-Opposition,0253</SECOND>
<HEAD>Opposition Leader Faces Court Appearance</HEAD>
<TEXT>
   PORT-AU-PRINCE, Haiti (AP) _ Opposition leader Lou
summoned to court next week for a hearing apparently
arrest on charges of inciting the public to revolt, a
said Friday.
...
</TEXT>
</DOC>
```

Boosting:
Applied to Text Filtering – Experimental Results

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

TREC-7

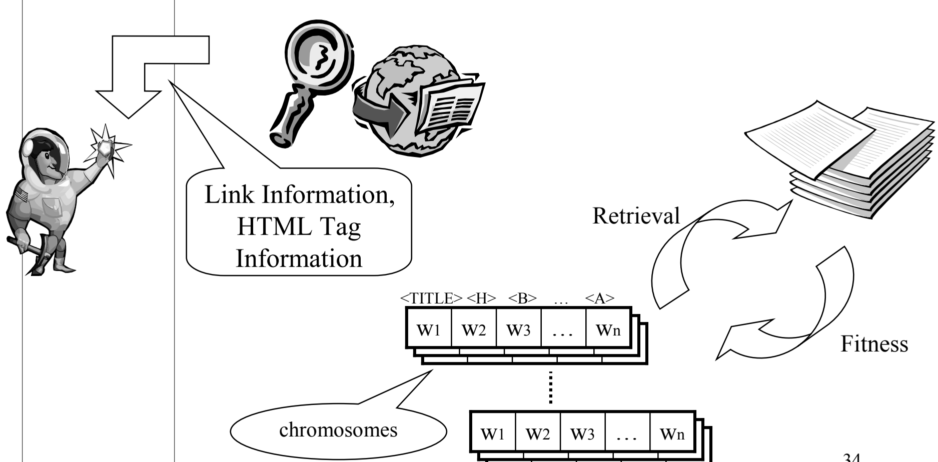
Averaged Scaled F1				Averaged Scaled F3			
Boosting	ATT	NTT	PIRC	Boosting	ATT	NTT	PIRC
0.474	0.461	0.452	0.500	0.467	0.460	0.505	0.509

TREC-8

Averaged Scaled LF1				Averaged Scaled LF2			
Boosting	PLT1	PLT2	PIRC	Boosting	CL	PIRC	Mer
0.717	0.712	0.713	0.714	0.722	0.721	0.734	0.720

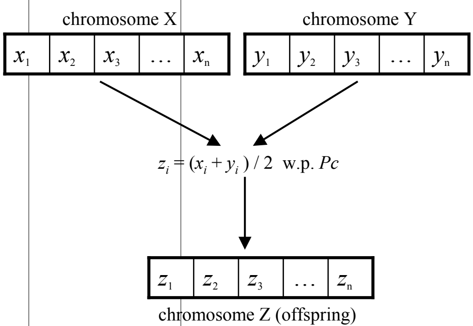
Evolutionary Learning:
Applications in IR - Web-Document Retrieval

- [Kim & Zhang, 2000]

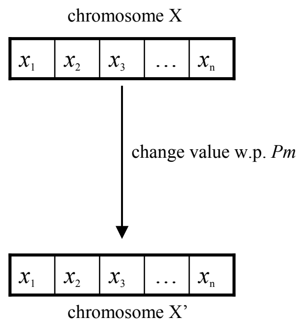


Evolutionary Learning:
Applications in IR – Tag Weighting

- Crossover



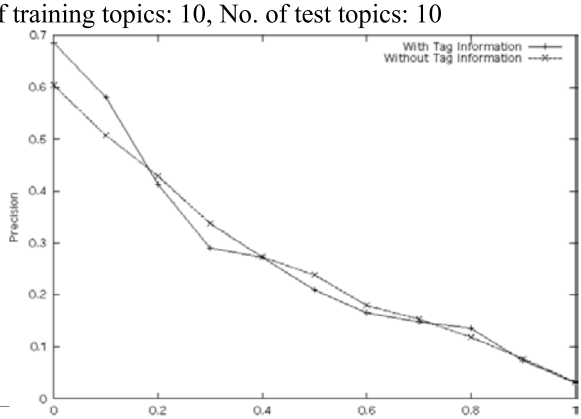
- Mutation



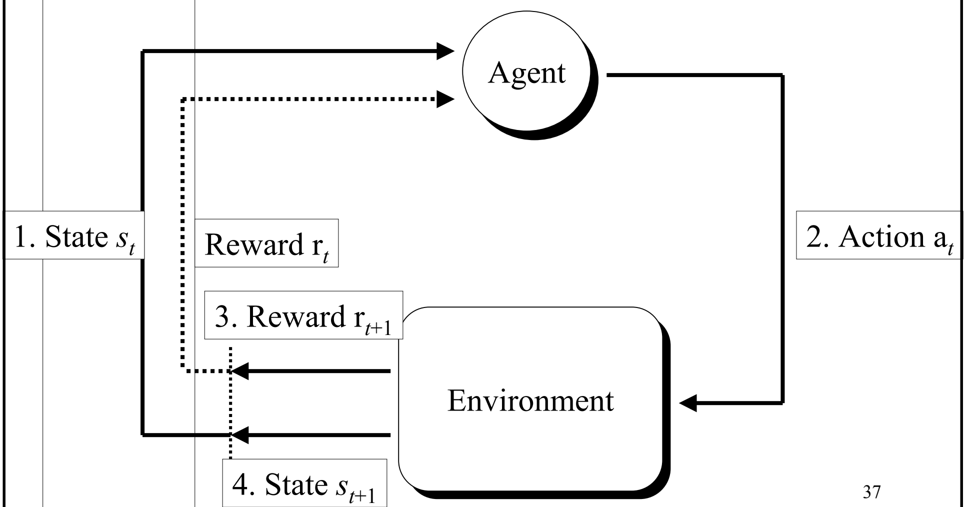
- Truncation selection

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

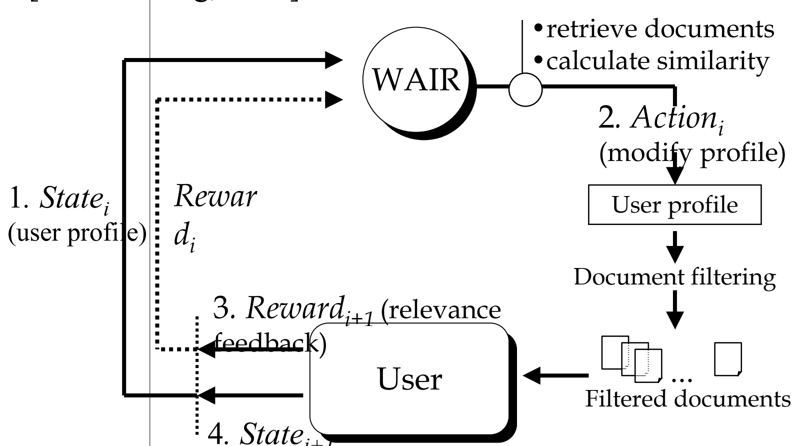


Reinforcement Learning:
Basic Concept

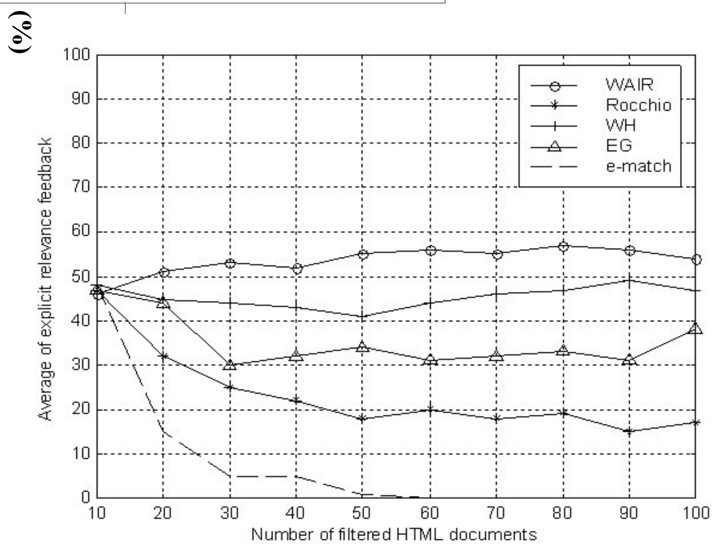


Reinforcement Learning:
Applications in IR - Information Filtering

[Seo & Zhang, 2000]



Reinforcement Learning:
Experimental Results (Explicit Feedback)



Reinforcement Learning:
Experimental Results (Implicit Feedback)