Utilizing Machine Learning in Information Retrieval:

- Text Classification

(COSC 488)

Nazli Goharian
nazli@cs.georgetown.edu

Literatures used to prepare the slides: See last page!

What is Text Classification?

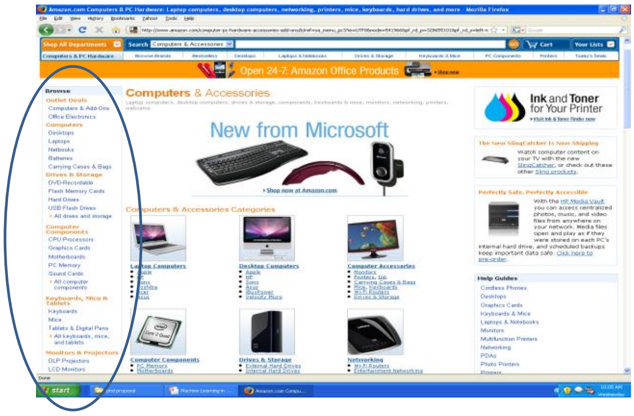
Text classification also known as *text categorization*, *topic classification*, or *topic spotting* is the process of assigning predefined categor(ies)/topic(s)/class(e)s/label(s) to a document that reflect its overall contents.

Application of Text Classification

- News Classification
  - *"Politics", "Sports", "Business"*

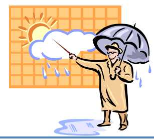
Application of Text Classification

- Shopping Products Classification
  - *"Electronics", "Home Appliances", "Books"*



Application of Text Classification

- News Routing/Filtering

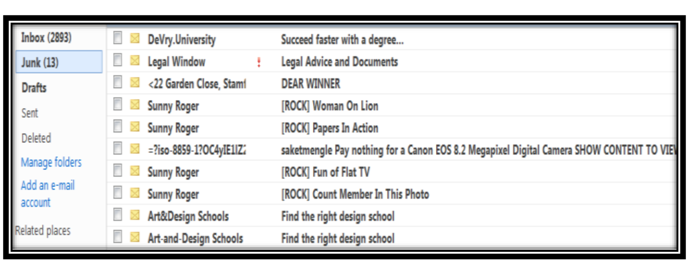


Tropical storms are building up in the south Pacific due to high pressure belts. The rains may continue for few more days.

Users interested in weather news *(standing queries)*

Application of Text Classification

- Spam Filtering
  - *"Spam", "Not Spam"*



Improving Search Results via Text Classification

- Query is searched in the user selected categories in web directories
- Categorized result set is presented to user
- Learning to rank -- (more recent efforts)
  Using various document features such as document length, age, etc. and their relevance to a query, build a model to rank/re-rank the documents
- Query category is searched against categorized pages (vertical search, advertisement search,…)

Web Directories

Constructing Web directories to be able to browse information via predefined set of categories:

- Yahoo
- dmoz Open Directory Project (ODP)
- Existing directories are based on human efforts
  - 80,000 editors involved to maintain ODP; www.dmoz.org

Using Web directories (Yahoo,ODP, Wikipedia,…) as training data, the classifier classifies new web pages into categories

Application of Text Classification

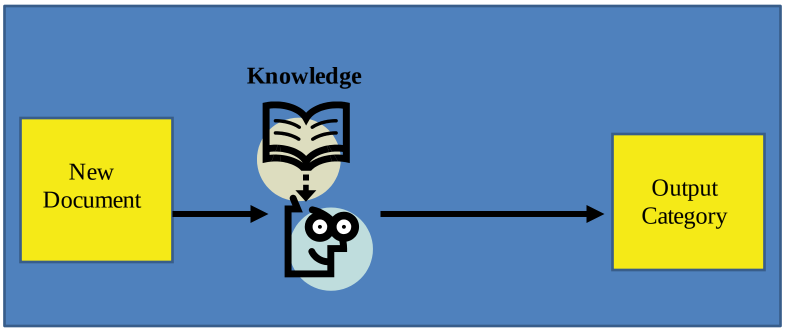
- Blog identification (Identifying blogs vs. non-blogs; using blog title, content, tags)
- Mood/Sentiment classification
  - Individual posts
  - Aggregate moods across posts
- Genre classification
  - Individual posts (ex: news, commentary, journals, personal, political, sports…)
- Words Sense Disambiguation (Identifying *meaning for words in context)*

Classification Methods

- Manual Classification
- Hand-crafted rules (Knowledge Engineering/semi-automatic) (80's)
- Supervised Learning
  – *Naïve Bayes, kNN, Rochio, SVM…& more*
- Semi/partial-Supervised

- *Note: Clustering is an unsupervised learning approach to grouping text into categories and will be discussed separately!*

Manual Classification



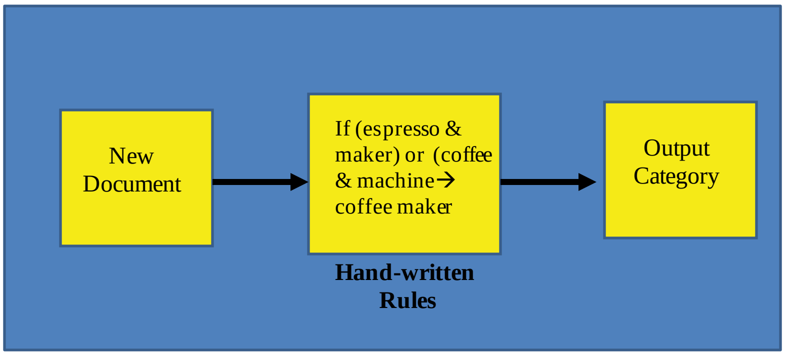
Manual Classification

- Domain experts label data
- Very accurate if done by experts

Examples:

- US Census Bureau's decennial census (1990: 22 million responses)
  – 232 industry categories and 504 occupation categories
  – $15 million estimated cost
- Librarians
- ODP (Open Directory Project) www.dmoz.org
- …….

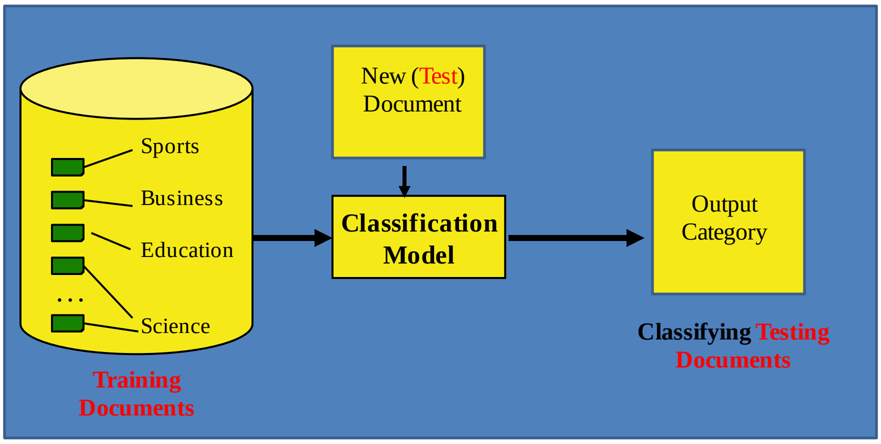
Knowledge Engineering/Semi-automatic



Knowledge Engineering

- A Knowledge Engineering (KE) approach
- Hand written rules to define each category (rule-based expert systems)
- Hand-written rules are then automatically applied to categorize new documents
- Accuracy is often very high if a rule has been carefully refined over time by a domain expert
- Building/maintaining these rules is expensive

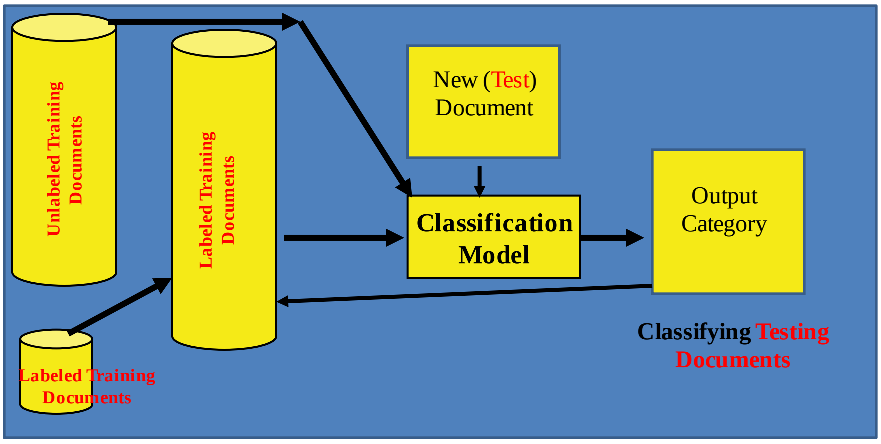
Supervised Learning (Classification)



Supervised Learning (Classification)

- Learning a model (classifier), using annotated training samples (documents) to classify any new incoming document into pre-defined set of topics
- Each Training document has one/more label(s)
- Various learning algorithms exists, examples:
  - Example: *Naïve Bayes, decision tree, support vector machine, neural network, regression, K-nearest neighbor, ...*
- Model/Classifier is used to classify incoming (test) documents

Semi/Partial-Supervised Learning (Bootstrap Classification)



Semi/Partial-Supervised Learning (Bootstrap Classification)

- Bootstrapping
  - Works well with small data sets
  - Samples the given training tuples uniformly *with replacement*
    - i.e., each time a tuple is selected, it is equally likely to be selected again and re-added to the training set
  - Expand "seed patterns/rules" with techniques of unsupervised learning and/or external knowledge resources
- Several bootstrap methods, and a common one is **.632 bootstrap**
  - Suppose we are given a data set of *d* tuples. The data set is sampled *d* times, with replacement, resulting in a training set of *d* samples. The data tuples that did not make it into the training set end up forming the test set. Repeat the sampling procedure *k* times.

Types of Classification: Binary vs. Multi-Class

- Binary: Only one of the two predefined categories are assigned to each document by a text classifier
- Multi-Class: Classification may involve more than two predefined categories
  - Single Label: Each document is assigned only one category (out of the *n* categories) by the text classifier
  - Multi-label: Each document is assigned one or more than one category by the text classifier

Example: Single-labeled Document

The Dow Jones industrial average lost 26 points, or 0.3%. The S&P 500 index fell 6 points, or 0.6%. The Nasdaq composite was little changed. Stocks slipped through most of the session as investors mulled the implications of a weaker-than-expected reading on the services sector of the economy, and mixed reports on the jobs market, ahead of Friday's big monthly payrolls report.

Source: CNN (http://money.cnn.com/2010/02/03/markets/markets_newyork/index.htm?postversion=2010020318)

- Politics
- Business
- Sports
- Entertainment

Example: Multi-labeled Document

President Obama, in his proposed 2011 budget, is calling on Congress to make a number of tax changes for individuals. Some ideas are new. Many others were made last year, but not enacted by Congress. So the estimates of the revenue that may be raised by his proposals may be overly optimistic.

Source: CNN (http://money.cnn.com/2010/02/01/pf/taxes/obama_budget_tax_changes/index.htm)

- Politics
- Business
- Sports
- Entertainment

Hard Categorization vs. Ranking Categorization

Hard Categorization

Complete decision of True or False for each pair $< d_j, c_i >$

Document	Category Assigned
d_1	c_1, c_2
d_2	c_2
d_3	c_3, c_4
d_4	c_4

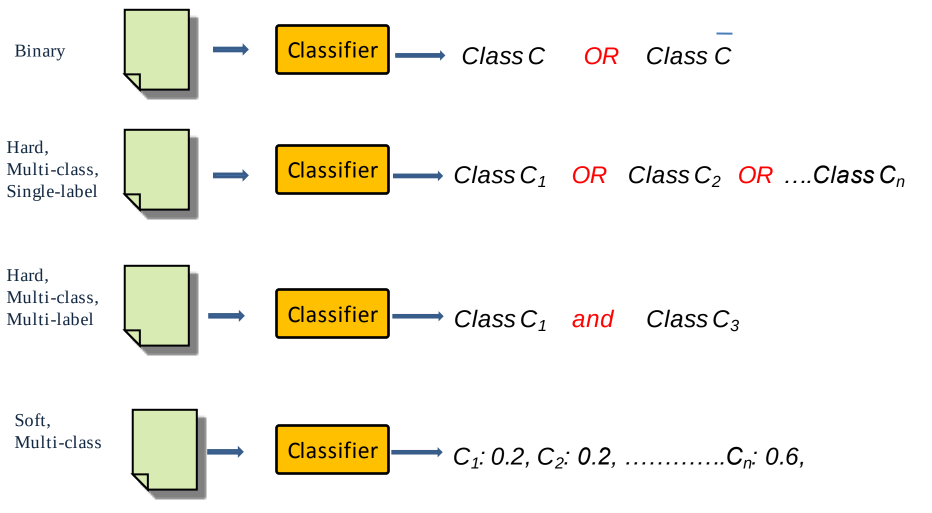
Ranking (Soft) Categorization

Given $d_j \in D$, rank the categories according to their estimated appropriateness to d_j

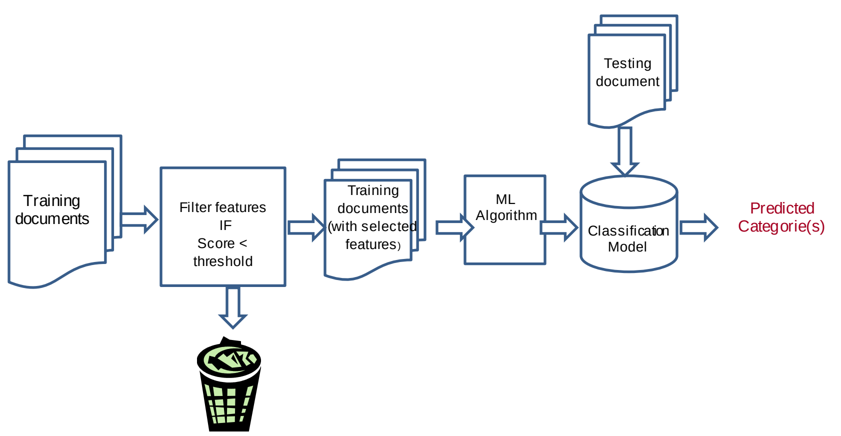
Document	Category	Estimated appropriateness
d_1	c_2	0.6
	c_1	0.3
	c_3	0.05
	c_4	0.05

Types of Classification

from: X. Qi and B. Davison, ACM Computing Surveys, 2009



Text Classification Process



Feature Selection

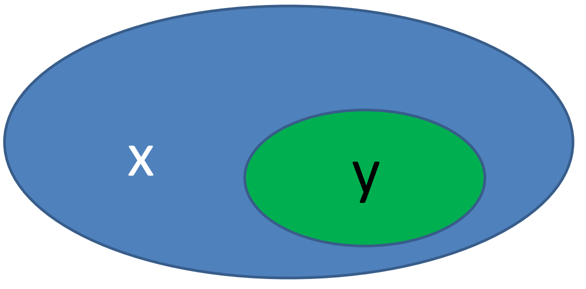
- Feature Selection in text classification refers to selecting a subset of the collection terms and utilize them in the process of text classification.
- Good features are better indicators of a class label
- Feature reduction tends to:
  - Reduce *overfitting* -- *as it makes it less specific*
  - Improve performance due to reducing dimensionality
- Feature Extraction provides more detailed features and feature relationships ***(not covered in this course)***

Model Overfitting

- Caused by:
  - Presence of noise
  - Lack of representative samples
  - Complexity of model (for example in decision tree)
- *Leads to:*
  - High *generalization error* (high number of misclassifications on unseen data)

Feature Selection

- Given a feature set $X=\{x_i \mid i=1\ldots N\}$, find a subset $Y=\{x_{i1}, x_{i2}, \ldots, x_{iM}\}$, with $M<N$, that increases the probability of correct classification



Text Features

- Feature space in text may include:
  - Lexical features (words, phrases)
  - Part-of-Speech (POS)
  - N-grams
  - Synonyms
  - ….
- General feature types may be:
  - Numeric
  - Nominal
  - Ordinal
  - Ratio

Web Page Features

- Additional features are utilized in Web page classification task:
  - On-Page Features
  - Neighboring Page Features (External Links)

On-Page Features

HTML tags:

- title
- headings
- metadata
- main text

HTML tags usually removed in pre-processing; the content of tags preserved

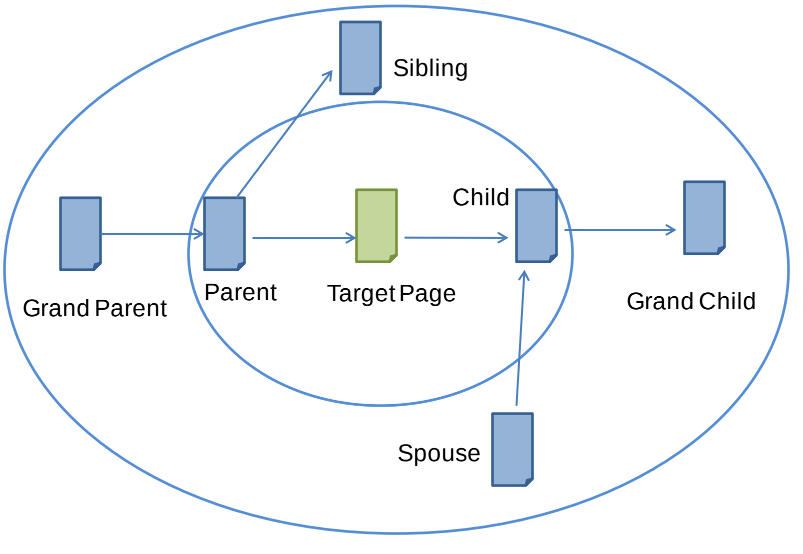
URL – classify without using page content

Neighboring-Page Features

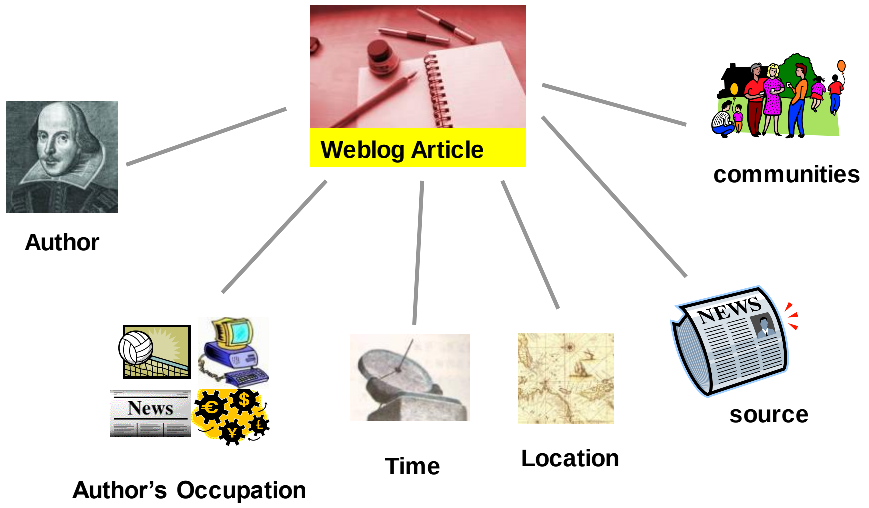
- Neighbors (linked pages) have similar topics and categories
- Number of steps from a page --shown as 2 (parent, child, sibling, grand parent, grand child); more steps more expensive & less effective
- Although all useful, but sibling is shown to be more effective
- Using only portion of neighboring content: title, anchor text, text closer to hyperlink to train a classifier
- Voting -- majority class of neighbors used

Neighboring-Page Features

from: X. Qi and B. Davison , ACM Computing Surveys, 2009



Context Features of a Document



Slide from: Cheng Xiang Zhai, keynote, SIGIR, 2011

Feature Selection Algorithms

Example of some of the feature selection methods:

- df
- tf-idf
- Tf-icf
- Mutual Information
- Information Gain
- χ^2 Statistic (CHI)
- Odds Ratio

(Note: There are some more FS algorithms!)

Feature Selection

- DF (Document Frequency): *Frequency of a term in the collection*
  - Retain terms that are not *stop terms* (high *df*) and do not have very low *df* (noise, not of interest)
- TF-IDF

  tf: frequency of a term in a document -- *commonly normalized*

  idf: inverse document frequency $tfidf(t_k, d_i) = TF(t_k, d_i) * \log\left(\frac{|D|}{df(t_k)}\right)$
  - Retain terms with high *tf-idf* in a document
- TF-ICF
  - Analogous to *tf-idf* but considering the frequency of term in the category. $tficf(t_k, c_i) = TF(t_k, c_i) * \log\left(\frac{|C|}{cf(t_k)}\right)$

Feature Selection (FS)

Consider the Term-Class incidence table:

Case	Docs in class: c_p	Docs not in class: $\overline{c_p}$	Total
Docs that contain term k_i	$n_{i,p}$	$n_i - n_{i,p}$	n_i
Docs that do not contain term $\overline{k_i}$	$n_p - n_{i,p}$	$N_t - n_i - (n_p - n_{i,p})$	$N_t - n_i$
All docs	n_p	$N_t - n_p$	N_t

The notations used in this table are used in the FS algorithms of the next few pages!

From: *Modern Information retrieval, R. Baeza-Yates & B. Ribeiro-Neto, 2011*

FS: Mutual Information (MI)

Measuring the amount of information the presence of a term contributes to the classification

MI between term k_i and set of classes C is expressed as expected value of:

$$I(k_i, c_p) = \log \frac{P(k_i, c_p)}{p(k_i) P(c_p)} = \log \left(\frac{\frac{n_{i,p}}{N_t}}{\frac{n_i}{N_t} \cdot \frac{n_p}{N_t}} \right)$$

Two alternates: 1) across all classes; 2) maximum term information:

$$MI(k_i, C) = \sum_{p=1}^{L} p(c_p)\, I(k_i, c_p) = \sum_{p=1}^{L} \frac{n_p}{N_t} \log \left(\frac{\frac{n_{i,p}}{N_t}}{\frac{n_i}{N_t} \cdot \frac{n_p}{N_t}} \right)$$

$$I_{\max}(k_i, C) = \max_{p=1}^{L} I(k_i, c_p) = \max_{p=1}^{L} \log \left(\frac{\frac{n_{i,p}}{N_t}}{\frac{n_i}{N_t} \cdot \frac{n_p}{N_t}} \right)$$

FS: Information Gain (IG)

Measuring the amount of information both the presence and the absence of a term contribute to the classification.

Terms with *IG >= threshold* are kept.

$$IG(k_i, C) = -\sum_{p=1}^{L} P(c_p) \log P(c_p) - (-\sum_{p=1}^{L} P(c_p, k_i) \log P(c_p \mid k_i)) - (-\sum_{p=1}^{L} P(c_p, \bar{k}_i) \log P(c_p \mid \bar{k}_i))$$

$$IC(k_i, C) = -\sum_{p=1}^{L} \left(\left(\frac{n_p}{N_t} \log \frac{n_p}{N_t}\right) - \left(\frac{n_{i,p}}{N_t} \log \frac{n_{i,p}}{n_i}\right) - \left(\frac{n_p - n_{i,p}}{N_t} \log \frac{n_p - n_{i,p}}{N_t - n_i}\right) \right)$$

FS: Chi Square (χ^2)

- Chi Square measures the *dependency* between the term and the class (*value of zero indicates independency*)

$$\chi^2(k_i, c_p) = \frac{N_t\left(P(k_i, c_p)P(\bar{k}_i, \bar{c}_p) - P(k_i, \bar{c}_p)P(\bar{k}_i, c_p)\right)^2}{P(c_p)P(\bar{c}_p)P(k_i)P(\bar{k}_i)}$$

- Calculate χ^2 of a term over all categories and retain the term if the value meets a threshold. Two alternatives:

1) Averaging over all categories: $\chi^2_{avg}(k_i) = \sum_{p=1}^{L} P(c_p)\chi^2(k_i, c_p)$

2) Considering the largest value: $\chi^2_{\max}(k_i) = \max_{p=1}^{L} \chi^2(k_i, c_p)$

FS: Chi Square (χ^2) (Cont'd)

- Chi Square measures the *dependency* between the term and the class (*value of zero indicates independency*)

$$\chi^2(k_i, c_p) = \frac{N_t\left(P(k_i, c_p)P(\bar{k}_i, \bar{c}_p) - P(k_i, \bar{c}_p)P(\bar{k}_i, c_p)\right)^2}{P(c_p)P(\bar{c}_p)P(k_i)P(\bar{k}_i)}$$

$$\chi^2(k_i, c_p) = \frac{N_t\left(n_{i,p}\left(N_t - n_i - n_p + n_{i,p}\right) - \left(n_i - n_{i,p}\right)\left(n_p - n_{i,p}\right)\right)^2}{n_p\left(N_t - n_p\right)n_i\left(N_t - n_i\right)}$$

$$= \frac{N_t\left(n_{i,p}N_t - n_p n_i\right)^2}{n_p n_i\left(N_t - n_p\right)\left(N_t - n_i\right)}$$

FS: Odds Ratio

- Odds Ratio reflects the odds of the word occurring in the positive class normalized by that of the negative class.
- Odds Ratio for a term t_k in category c_i

$$OR(t_k, c_i) = \frac{P(t_k \mid c_i).[1 - P(t_k \mid \overline{c_i})]}{P(\overline{t_k} \mid \overline{c_i}).[1 - P(\overline{t_k} \mid c_i)]}$$

Supervised Learning Algorithms

- Naïve Bayes
- K-Nearest Neighbor (KNN)

Only these two are covered in this course!

- Support Vector Machines (SVM)
- Decision-tree
- Decision-Rule classifiers
- Neural Networks
- Rocchio
- HMM
- CRF

Representation of Text

> This week, the United Nations created the position of czar in the global fight against a possible avian influenza pandemic. Meanwhile, officials here in the United States acknowledged the country is unprepared if this never-before-seen strain of flu, known to scientists as H5N1 virus, were to hit this winter.

- Commonly used pre-processing: stop word removal, stemming,…

d1:<*week, united, nations, create, position, czar, global, fight, against, possible,…..*>

Term	Frequency
Week	1
united	2
nation	1
………	

Phrases:
United nations
Avian influenza
………..

Bayes Theorem

$$\underbrace{P = (H \mid X)}_{\text{Posterior Probability of class } C_i} = \frac{\overbrace{P(X \mid H)}^{\text{Posterior Probability of X}}\,\overbrace{P(H)}^{\text{Prior Probability of class } C_i}}{P(X)}$$

As P(X) is constant, it is ignored in the calculations.

Naïve Bayes Text Classifier

- Text as "bag-of-words"
- Independent assumption -- *occurrence of terms and their positions*
- Building Model:
  - For each category c_i build a probabilistic model

    T: text in class c_i
    n: size of the vocabulary

    $$P(T: t_1, t_2, \ldots. t_n \mid c_i)$$

  - Calculate the prior probability $P(C_i)$

Naïve Bayes Text Classifier

- Classify Text:
  - Calculate probability of each category for a given text

    $$P(c_i \mid d_j) = p(c_i) P(d_j \mid c_i)$$

  - The *category* c_i with the highest score among all categories C is the one that is most probable to generate the text d_j

    $$c_{\max a\ posteriori} = \arg\max_{c_i \in C} p(c_i) P(d_j \mid c_i)$$

Naïve Bayes Text Classifier

Posterior Probability of d

Prior Probability of class C_i

$$P(c_i \mid d_j) = p(c_i)\, \underbrace{P(d_i \mid c_i)}$$

Posterior Probability of class C_i

$$\prod_{k=1}^{|T|} P(t_{kj} \mid c_i) = \sum_{i=1}^{|T|} \log P(t_{kj} \mid c_i)$$

Naïve Bayes Text Classifier

- Need to estimate the probability: $P(t_{kj} \mid c_i)$
  - Multinomial model:

$$\frac{number\ of\ times\ term\ t_{kj}\ appears\ in\ category\ c_i + 0.5}{total\ terms\ in\ c_i + 1}$$

  - binomial or Bernoulli model:

$$\frac{number\ of\ documents\ in\ category\ c_i\ that\ term\ t_{kj}\ appears}{total\ documents\ in\ c_i}$$

Naïve Bayes Text Classifier

Multinomial model:

$$P(c_i \mid d_j) = \underbrace{p(c_i)}_{\log\left(\frac{docs\ in\ c_i}{total\ docs}\right)} \underbrace{P(\vec{d}_j \mid c_i)}$$

$$\log\left(\frac{docs\ in\ c_i}{total\ docs}\right)$$

$$\prod_{k=1}^{|T|} P(t_{kj} \mid c_i) = \sum_{i=1}^{|T|} \log \underbrace{P(t_{kj} \mid c_i)}_{\frac{number\ of\ times\ term\ t_{kj}\ appears\ in\ category\ c_i + 0.5}{total\ terms\ in\ c_i + 1}}$$

To avoid a zero if a new term appears → Smoothing

- Various approaches: *Dirchelet prior, Laplace,..*

Example

Doc-1
Category: Computers
The sales of laptops in 2009 was high as many OS were released

Doc-2
Category: Computers
Many OS provide varying level of securities for laptops as they tend to switch networks. This makes the laptops more secure from computer viruses

Doc-3
Category: Epidemic
A new virus called H1N1 causes Swine Flu.

Doc-4
Category: Epidemic
Bird flu is caused by a virus called H5N1. The disease is of concern to humans, who have no immunity against it.

Example

Assume that red terms are the selected features:

Doc-1
Category: Computers
The sales of laptops in 2009 was high as many OS were released

Doc-2
Category: Computers
Many OS provide varying level of securities for laptops as they tend to switch networks. This makes the laptops more secure from computer viruses

Doc-3
Category: Epidemic
A new virus called H1N1 causes Swine Flu.

Doc-4
Category: Epidemic
Bird flu is caused by a virus called H5N1. The disease is of concern to humans, who have no immunity against it.

Example: Naïve Bayes Text Classifier

Task: Classify D5: "*A deadly virus called H1N1 was detected in various parts of the world*"

- **P(Computers|D5)** = P(Computers) P(Virus|Computers) P(H1N1|Computers)
- **P(Epidemic|D5)** = P(Epidemic) P(Virus|Epidemic) P(H1N1|Epidemic)

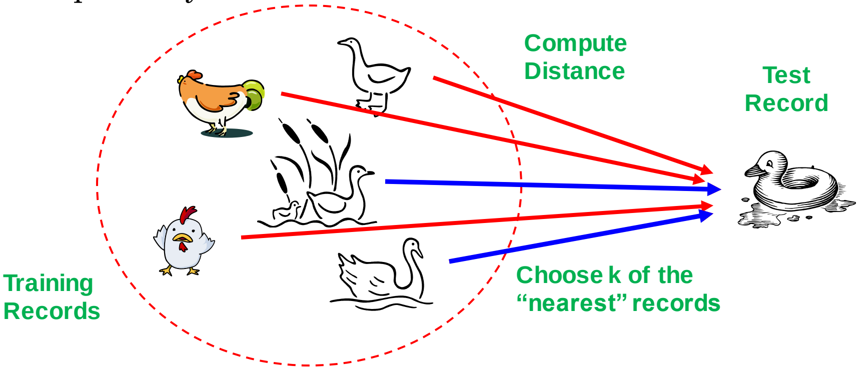
P(Epidemic|D5) > P(Computers|D5)

Thus, class of D5 is Epidemics

Nearest Neighbor Classifiers

Slide from: *Tan, Steinback, Kumar, 2004*

- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



K-Nearest Neighbor Classifier

- No model is built (lazy learner) a priori
  (Classification done based on raw training data)
- The class of a document will be the class of the majority class of the *k* nearest neighbor (*majority voting*)
- The relatedness/nearness of two documents can be quantified in terms of *similarity* (eg. *Cosine measure*) or *distance* (*eg. Euclidean distance*)
  - Different weight for different features
  - Feature values can be normalized to prevent different handling (may prefer different handling!)
- Sensitivity to value of *K*
  - Picked empirically, domain knowledge

Distance/Similarity Measures

Euclidean Distance:

$$dist(d_i, d_j) = \sqrt{(|d_{i1} - d_{j1}|^2 + |d_{i2} - d_{j2}|^2 + \ldots + |d_{ip} - d_{jp}|^2)}$$

Cosine Similarity:

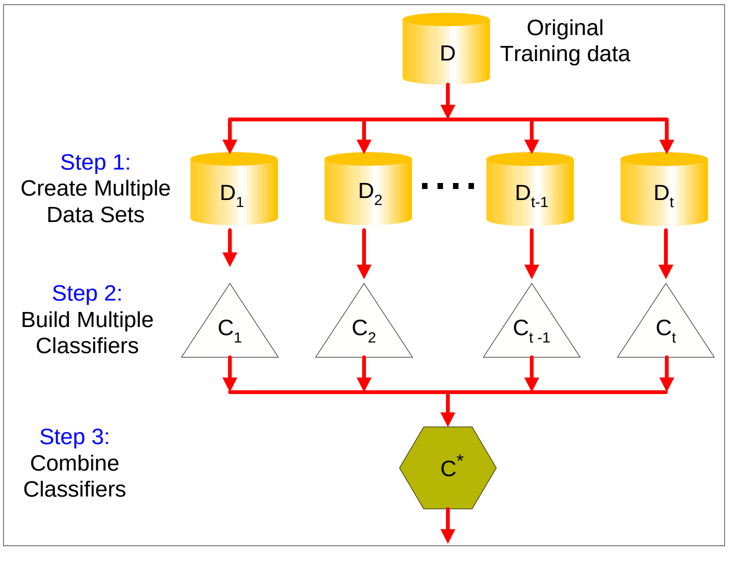
$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{t} d_{ik} \; x \; d_{jk}}{\sqrt{\sum_{k=1}^{t} (d_{ik})^2 \; \sum_{k=1}^{t} (d_{jk})^2}}$$

Term weight:

$$w_{ij} = \frac{(\log tf_{ij} + 1.0) * idf_j}{\sum_{j=1}^{t} [(\log tf_{ij} + 1.0) * idf_j]^2}$$

Ensemble Classifier: General Idea

from: Data Mining book



Bagging: Bootstrap Aggregation

from: Data Mining book

- Analogy: Diagnosis based on multiple doctors' majority vote
- Training
  - Given a set D of d tuples, at each iteration i, a training set D_i of d tuples is sampled with replacement from D (i.e., bootstrap)
  - A classifier model M_i is learned for each training set D_i
- Classification: classify an unknown sample **X**
  - Each classifier M_i returns its class prediction
  - The bagged classifier M* counts the votes and assigns the class with the most votes to **X**
- Accuracy
  - Often better than a single classifier derived from D

Bagging (Bootstrap Aggregating)

from: Data Mining book

- Sampling with replacement

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- Build classifier on each bootstrap sample (same size as the original training data)
- Classify data by taking majority vote among the predictions made by each base classifier

Boosting

from: Data Mining book

- An iterative procedure to adaptively change distribution of training data by focusing more on previously misclassified records
  - Initially, all N records are assigned equal weights, 1/N
  - Unlike bagging, weights may change at the end of boosting round
- Instead of using majority voting, the prediction by each classifier is weighted base on classifier error rate.

Boosting

from: Data Mining book

- Records that are wrongly classified will have their weights increased
- Records that are classified correctly will have their weights decreased

Original Data	1	2	3	4	5	6	7	8	9	10
Boosting (Round 1)	7	3	2	8	7	9	4	10	6	3
Boosting (Round 2)	5	4	9	4	2	5	1	7	4	2
Boosting (Round 3)	4	4	8	10	4	5	4	6	3	4

- Example 4 is hard to classify
- Its weight is increased, therefore it is more likely to be chosen again in subsequent rounds

Evaluation Metrics

	PREDICTED CLASS	
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	TP	FN
	Class=No	FP	TN

$$\text{Precision (p)} = \frac{tp}{tp+fp}$$

$$\text{Recall (r)} = \frac{tp}{tp+fn}$$

$$\text{F1-measure(F1)} = \frac{2rp}{r+p}$$

Macro-Averaging

- Macro-average:
  - Equal weight to each category

$$\text{Macro-Precision} = \frac{\text{Precision(A)} + \text{Precision(B)} + \text{Precision(C)}}{3}$$

$$\text{Macro-Recall} = \frac{\text{Recall(A)} + \text{Recall(B)} + \text{Recall(C)}}{3}$$

$$\text{Macro-F1Measure} = \frac{\text{F1Measure(A)} + \text{F1Measure(B)} + \text{F1Measure(C)}}{3}$$

Micro-Averaging

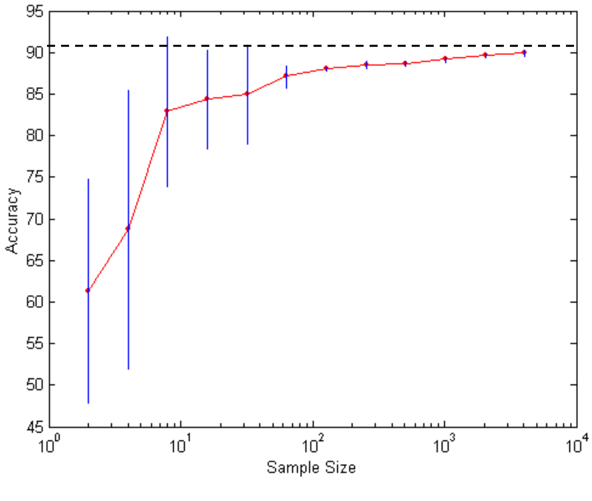
- Micro-average:
  - Equal weight to each sample (record, document)

$$\text{Micro-Precision} = \frac{TP_A + TP_B + TP_C}{TP_A + TP_B + TP_C + FP_A + FP_B + FP_C}$$

$$\text{Micro-Recall} = \frac{TP_A + TP_B + TP_C}{TP_A + TP_B + TP_C + FN_A + FN_B + FN_C}$$

$$\text{Micro-F1 Measure} = \frac{2 * \text{Micro-Precision} * \text{Micro-Recall}}{\text{Micro} - \text{Precision} + \mathit{Micro} - \text{Recall}}$$

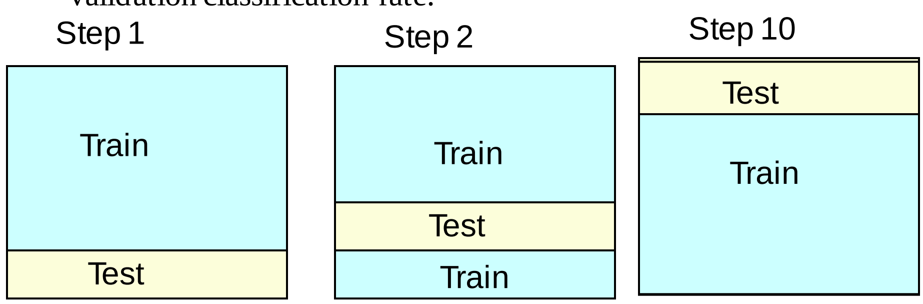
Learning Curve



- Learning curve shows how accuracy changes with varying sample size

10-fold cross validation

- Training data: 90%
- Test data: 10%
- Stratified validation: same label distribution in training & test
- Each run will result in a particular classification rate.
- Average the ten classification rates for a final 10-fold cross validation classification rate.



Evaluation Dataset

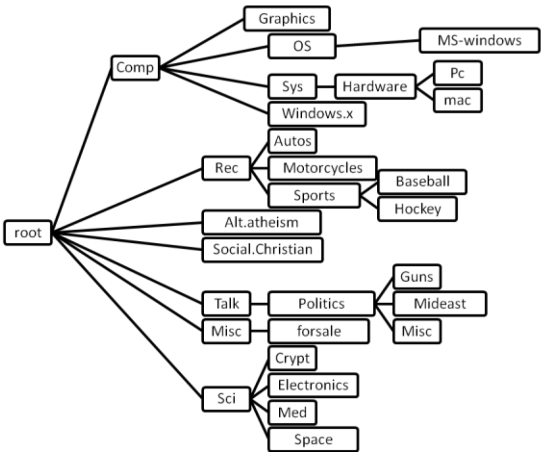
- Manual labeling needs excessive effort
- Available Web directory: *Yahoo directory* & *dmoz ODP (Open Directory Project)*
- Several other sources available – nowadays *Wikipedia*
- Problem – not one given benchmark!
- Not one given domain!

Some of the Text Classification Benchmark Datasets

Datasets	No. of documents	No. of Categories	Size of dataset	Domain
Reuters 21578	21,578	108 Categories (we used top 10)	28 MB	News Articles
20 News Group	20,000	20 categories	61 MB	News Articles
WebKB	8,282	7 categories	43 MB	Web Pages (University websites)
OHSUMED	54,710 (Total) 39,320 (Subset)	4,308 (we used top 50)	382 MB	Bio-medical Documents
GENOMICS (TREC 05)	4.5 million (Total) 591,689 (Subset)	20,184 (we used top 50)	15.5 GB	Bio-medical Documents

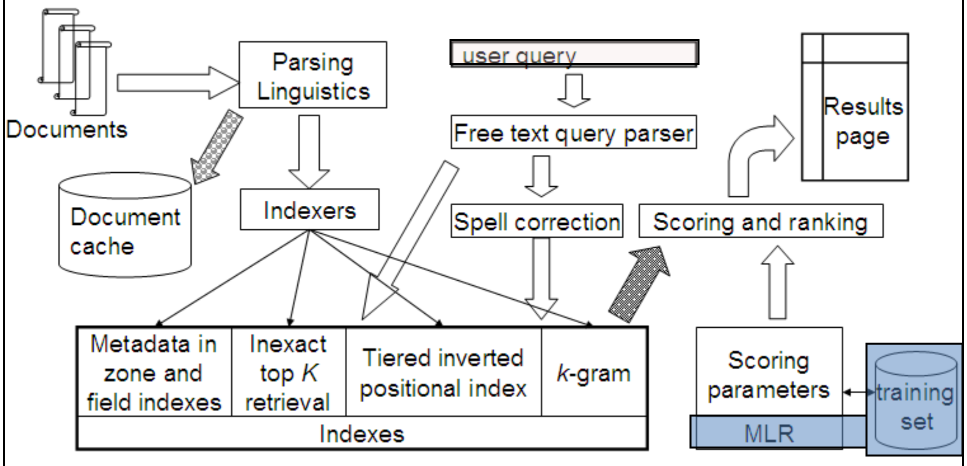
More benchmark datasets exist!

Sample Dataset: *20 Newsgroups Hierarchy*



Putting it all together

©D. Manning, P. Raghavan, H. Schutze, *Introduction to Information retrieval*, p 135, Cambridge University Press., 2008.



Learning to Rank

- Retrieval models need tuning parameters
  - Not a trivial task
  - may lead to overfitting
- Not one retrieval model outcome may suffice for ranking, a combination maybe helpful
  - Thus, using ML to automatically
    - Tune parameters
    - Combine ranking features

"Learning-to-rank" methods are those ranking methods that use ML for ranking!

91

Learning to Rank: Sample Learning Features (Trec)

1 Term frequency (TF) of body
2 TF of anchor
3 TF of title
4 TF of URL
5 TF of whole document
6 Inverse document frequency (IDF) of body
7 IDF of anchor
8 IDF of title
9 IDF of URL
10 IDF of whole document
11 TF*IDF of body
12 TF*IDF of anchor
13 TF*IDF of title
14 TF*IDF of URL
15 TF*IDF of whole document
16 Document length (DL) of body
17 DL of anchor
18 DL of title
19 DL of URL
20 DL of whole document
21 BM25 of body
22 BM25 of anchor
23 BM25 of title
24 BM25 of URL
25 BM25 of whole document
26 LMIR.ABS of body
27 LMIR.ABS of anchor
28 LMIR.ABS of title
29 LMIR.ABS of URL
30 LMIR.ABS of whole document
31 LMIR.DIR of body
32 LMIR.DIR of anchor
33 LMIR.DIR of title
34 LMIR.DIR of URL
35 LMIR.DIR of whole document
36 LMIR.JM of body
37 LMIR.JM of anchor
38 LMIR.JM of title
39 LMIR.JM of URL
40 LMIR.JM of whole document
41 Sitemap based term propagation
42 Sitemap based score propagation
43 Hyperlink base score propagation: weighted in-link
44 Hyperlink base score propagation: weighted out-link
45 Hyperlink base score propagation: uniform out-link
46 Hyperlink base feature propagation: weighted in-link
47 Hyperlink base feature propagation: weighted out-link
48 Hyperlink base feature propagation: uniform out-link
49 HITS authority
50 HITS hub
51 PageRank
52 HostRank
53 Topical PageRank
54 Topical HITS authority
55 Topical HITS hub
56 Inlink number
57 Outlink number
58 Number of slash in URL
59 Length of URL
60 Number of child page
61 BM25 of extracted title
62 LMIR.ABS of extracted title
63 LMIR.DIR of extracted title
64 LMIR.JM of extracted title

T. Liu, "Learning to Rank for Information Retrieval", Foundations & Trends in Information Retrieval, 2009

Sample of related Research Projects

- Passage detection: *Identifying Leakage of information within text*
  - S. Mengle, N. Goharian, "Detecting Hidden Passages from Documents", *SIAM Conference on Data Mining (SIAM - SDM) Workshop*, 2008.
  - N. Goharian, S. Mengle, "On Document Splitting in Passage Detection", *SIGIR*, 2008. (short)
  - S. Mengle and N. Goharian, "Passage Detection Using Text Classification", *Journal of American Society for Information Science and Technology (JASIST)*, 60 (4), March 2009.
- Feature selection: *Ambiguity Feature Selection Algorithm*
  - S. Mengle, N. Goharian, "Using Ambiguity Measure Feature Selection Algorithm for Support Vector Machine Classifier", *ACM 23rd Symposium on Applied Computing (SAC)*, March 2008.
  - S. Mengle and N. Goharian, "Ambiguity Measure Feature Selection Algorithm", *Journal of American Society for Information Science and Technology (JASIST)*, 60 (5), April 2009.
- Using misclassification information to identify topic/label/category relationships
  - S. Mengle and N. Goharian, "Detecting Relationships among Categories using Text Classification", *Journal of American Society for Information Science and Technology (JASIST)*, 61 (5), May 2010
  - N. Goharian, S. Mengle "Networked Hierarchies for Web Directories", *20th International World Wide Web conference (WWW)*, March 2011. (short)
- Analyzing query session / user intent
  - N. Goharian, S. Mengle, "Context Aware Query Classification Using Dynamic Query Window and Relationship Net", *In proceedings of ACM 33rd Conference on Research and Development in Information Retrieval (SIGIR)*, July 2010. (short)
- SMS spam detection
  - Z. Tan, N. Goharian, M. Sherr, "$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam", *In proceedings of ACM 35th Conference on Research and Development in Information Retrieval (SIGIR)*, August 2012. (short)

Passage Detection:

A Football story

Former Italy coach Azeglio Vicini has said the Azzurri have as good a chance as ever of winning the World Cup for a fifth time. There is plenty of expectation from Marcello Lippi's men and the big question is whether they are good enough to retain the trophy they won in 2006. And it's a simple answer for Vicini. "Italy, for the titles they have won, are a very competitive national team, and they always have been," he told Calciomercato.com." They are among the favourites to win it. I think Brazil are the outright favourites, but it doesn't mean that they will win it."Vicini believes Lippi has the best group of players at his disposal, despite the exclusions of Antonio Cassano and Fabrizio Miccoli, Lehman Brothers investment bank announces it's filing for bankruptcy two of Serie A's best players this term. "I think Lippi has the best of Italian football in his ranks, even though there is no Cassano.

97

What about this passage? Not a Football story

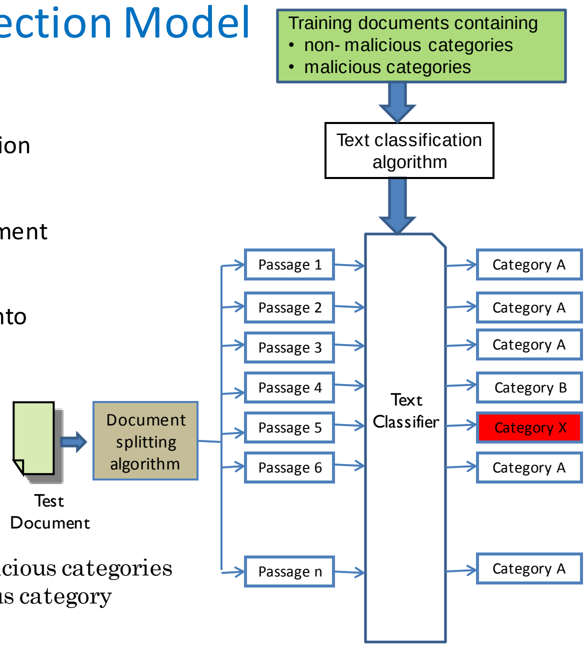
Detecting Leakage of Information

Former Italy coach Azeglio Vicini has said the Azzurri have as good a chance as ever of winning the World Cup for a fifth time. There is plenty of expectation from Marcello Lippi's men and the big question is whether they are good enough to retain the trophy they won in 2006. And it's a simple answer for Vicini. "Italy, for the titles they have won, are a very competitive national team, and they always have been," he told Calciomercato.com."They are among the favourites to win it. I think Brazil are the outright favourites, but it doesn't mean that they will win it."Vicini believes Lippi has the best group of players at his disposal, despite the exclusions of Antonio Cassano and Fabrizio Miccoli, **Lehman Brothers investment bank announces it's filing for bankruptcy** two of Serie A's best players this term. "I think Lippi has the best of Italian football in his ranks, even though there is no Cassano.

98

Passage Detection Model

1. Building text classification model
2. Splitting testing document into passages
3. Classifying passages into [related] categories



Category A & B: Non-malicious categories
Category X: Malicious category

S. Mengle & N. Goharian, "Passage Detection Using Text Classification", *Journal of American Society for Information Science and Technology (JASIST)*, 60 (4), March 2009.

Discourse Passage (DP)

- Discourse passages are based on logical components such as discourse boundaries like a sentence

The sky is blue. How beautiful! It was cloudy yesterday.

Non-Overlapping Window Passage (NWP)

- Window based passage approach defines a passage as *n* number of words

The sky is blue. However, it is raining continuously since morning.

101

Overlapping Window Passage (OWP)

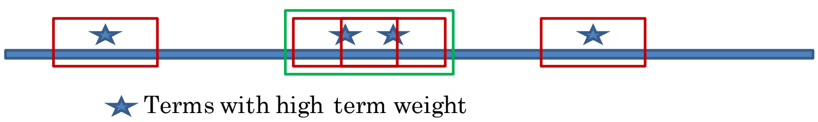
- Document is divided into passages of evenly sized blocks by overlapping *n/2 from the prior passage and n/2 from the next passage.*

The sky is blue. However, it is raining continuously since morning.

102

Keyword Based Dynamic Passage (KDP)

- Calculate term weight for all the terms in the training documents (labeled documents)
- Select terms (keywords) with high term weight in a testing document
- Select passages of *n* words with *n/2* words before and *n/2* words after the keyword
- Classify the identified passage into a category



S. Mengle & N. Goharian, "Passage Detection Using Text Classification", *Journal of American Society for Information Science and Technology (JASIST)*, 60 (4), March 2009.

103

Term Weighting Algorithm: Ambiguity Measure

1. Counting the number of occurrences of terms in every category
2. Calculating AM for each term

$$AM(t_k, C_i) = \left(\frac{tf(t_k, c_i)}{tf(t_k)} \right)$$

$$AM(t_k) = \max(AM(t_k, C_i))$$

Training documents

Term / Category	H5N1 Count	Virus Count	Officials Count
Pornography	10	1000	280
Epidemic	990	1500	320
Drug trafficking	0	0	600
Terrorism	0	0	400
Total	1000	2500	1600

Term / Category	H5N1 AM	Virus AM	Officials AM
Pornography	0.01	0.40	0.175
Epidemic	0.99	0.60	0.2
Drug trafficking	0.00	0.00	0.375
Terrorism	0.00	0.00	0.25

- S. Mengle and N. Goharian, "Ambiguity Measure Feature Selection Algorithm", *Journal of American Society for Information Science and Technology (JASIST)*, 60 (5), 2009.
- S. Mengle, N. Goharian, "Using Ambiguity Measure Feature Selection Algorithm for Support Vector Machine Classifier", *ACM 23rd Symposium on Applied Computing (SAC)*, March 2008.

What are Related Categories?

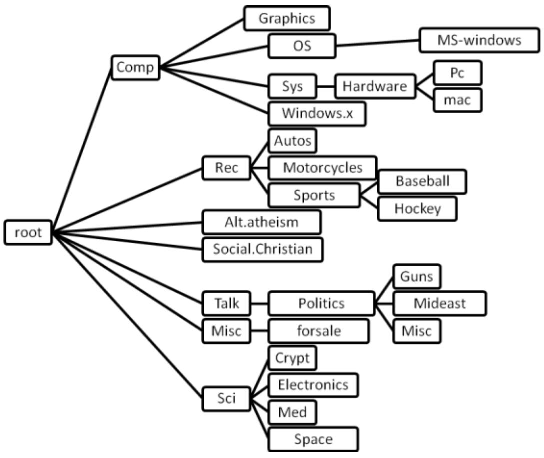
- Related categories are categories that overlap with each other in terms of subject/theme
- We present the relationships among categories using a graph structure called relationship-net *G(V,E)* , where

  V : set of all categories

  E: set of edges representing relationship among categories.

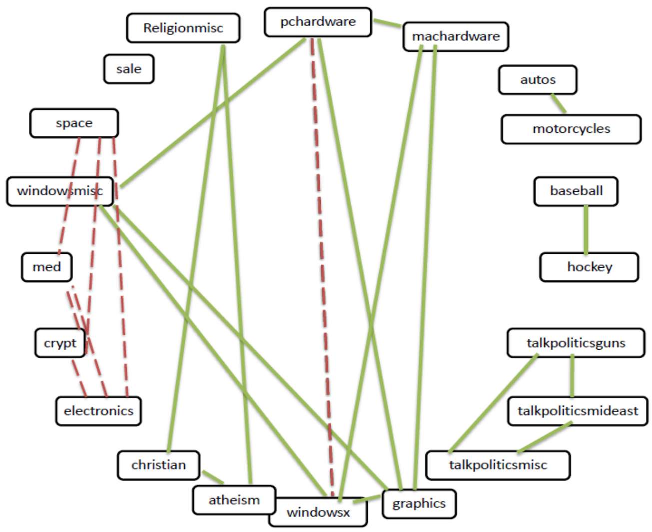
105

Example: 20 Newsgroups Hierarchy



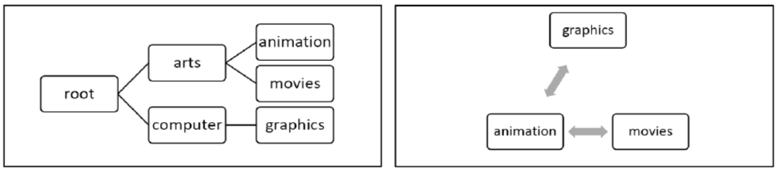
106

20 Newsgroups Relationship-net



107

Category Hierarchies vs. Relationship Net



Category Hierarchy	Relationship-Net
Represents Generalization Relationships	Flat hierarchy that does not represent generalization
Non sibling relationships are not be represented	Non-sibling relationships can be represented
Useful when hierarchy structure between category hierarchy is important	Useful when knowledge of relationships among categories is important

108

Finding Relationships using Misclassification Information

1. Classifying documents and generate a confusion matrix

	Predicted				
Actual	A	R	HP	HM	M
A (Atheism)	843	43	2	10	4
R (Religion)	53	925	4	5	3
HP (Hardware.pc)	4	0	793	31	9
HM (Hardware.mac)	0	0	13	843	4
M (Misc.forsale)	0	6	12	9	872

2. Calculating *relationship weights*

Relationship weight

$$M_N(j,k) = \frac{M(j,k)}{\sum_{i=1}^{n} M(i,k)}$$

	Predicted				
Actual	A	R	HP	HM	M
A (Atheism)	0.000	0.878	0.074	0.182	0.200
R (Religion)	0.930	0.000	0.129	0.091	0.094
HP (Hardware.pc)	0.070	0.000	0.000	0.564	0.281
HM (Hardware.mac)	0.000	0.000	0.419	0.000	0.125
M (Misc.forsale)	0.000	0.122	0.258	0.164	0.000

109

S. Mengle and N. Goharian, "Detecting Relationships among Categories using Text Classification", *Journal of American Society for Information Science and Technology (JASIST)*, 61 (5), May 2010

Finding Relationships using Misclassification Information

3. Assigning relationship weights to relationships

Relationship weight of relationships between categories and their corresponding C_{FN_max}

Category	C_{FN_max}	*Relationship weight*
Atheism	Religion	0.878
Religion	Atheism	0.930
Hardware.pc	Hardware.mac	0.564
Hardware.mac	Hardware.pc	0.419
Misc.forsale	Hardware.pc	0.258

Relationship weight of relationships between categories and their corresponding C_{FP_max}

Category	C_{FP_max}	*Relationship weight*
Atheism	Religion	0.930
Religion	Atheism	0.878
Hardware.pc	Hardware.mac	0.419
Hardware.mac	Hardware.pc	0.564
Misc.forsale	Hardware.pc	0.281

4. Predicting relationship between categories

Relationship Weight Threshold >0.3

Atheism ←→ Religion
Hardware.pc ←→ Hardware.mac

110

S. Mengle and N. Goharian, "Detecting Relationships among Categories using Text Classification", *Journal of American Society for Information Science and Technology (JASIST)*, 61 (5), May 2010

Passage Detection Dataset Summary

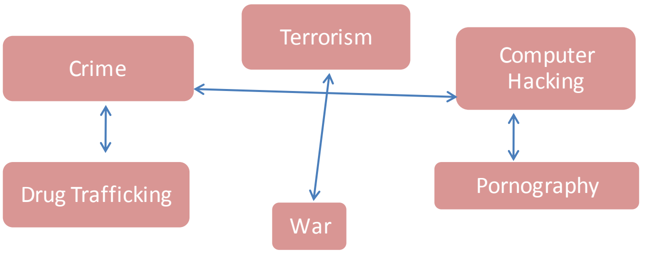
Purpose	Modified 20 Newsgroups dataset (20 Categories)				Modified Reuters 21578 dataset (10 Categories)			
	Dataset	**Number of documents**	**Is the document infected?**	**Length of passage**	**Dataset**	**Number of documents**	**Is the document infected?**	**Length of passage**
Training	20 NG	18,000	-	-	Reuters 21578	9900	-	-
	Security Dataset	3067	-	-	Security Dataset	3067	-	-
Testing	20 NG	1000	No	-	Reuters 21578	550	No	-
	20 NG	200	Yes	10 words	Reuters 21578	110	Yes	10 words
	20 NG	200	Yes	20 words	Reuters 21578	110	Yes	20 words
	20 NG	200	Yes	30 words	Reuters 21578	110	Yes	30 words
	20 NG	200	Yes	40 words	Reuters 21578	110	Yes	40 words
	20 NG	200	Yes	50 words	Reuters 21578	110	Yes	50 words

- Four variations of testing dataset: [1 – 4] infected passages

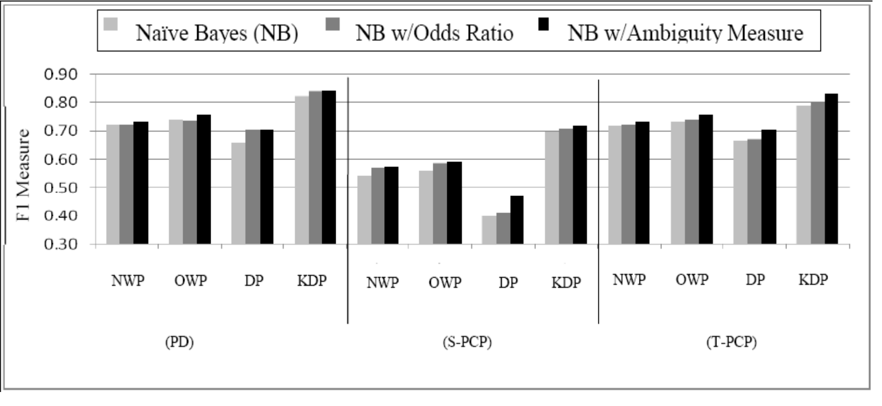
Passage Detection Security Dataset (articles from cnn.com)

Category (6)	Number of documents (3067)	Description
Computer Crimes	329	Computer crimes such as hacking and viruses.
Terrorism	920	Terrorist attacks and counter measures to prevent terrorism
Drugs Crimes	601	Drug trafficking and crimes related to drugs
Pornography	344	Issues related to pornography
War Reports	342	Reports on wars
Nuclear Weapons	531	Reports on nuclear programs in various countries

Relationships among Malicious Categories in our Dataset



Effects of Various Document Splitting Approaches



- Three evaluation tasks
- 20NG dataset (similar results on Reuters dataset)
- PD: passage detection; S/T-PCP: stringent/ Tolerent passage category detection;

S. Mengle & N. Goharian, "Passage Detection Using Text Classification", *Journal of American Society for Information Science and Technology (JASIST)*, 60 (4), March 2009.

A Networked Hierarchy

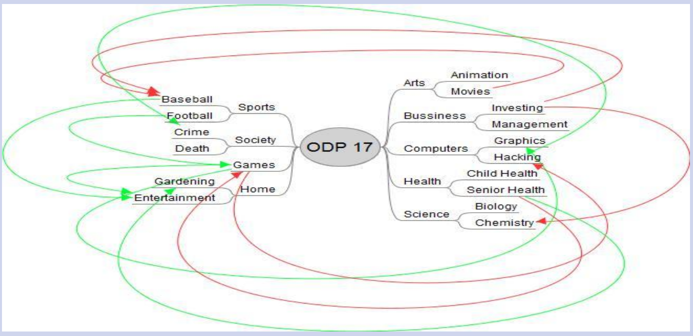
- A *networked hierarchy* is a hierarchy that not only maintains the characteristics of a hierarchy, i.e., *parent, child, sibling*, but also provides links between those *non-sibling categories* (nodes) that are, indeed to a degree, relevant.

- **Goal:** Automatically identifying and constructing relationships between categories of documents to represent all the following relationships:
  - Parent-child
  - Sibling
  - Non-sibling

115

N. Goharian & S. Mengle "Networked Hierarchies for Web Directories", *20th International World Wide Web conference (WWW)*, March 2011.

A Networked Hierarchy



- **Parent-child**
- **Sibling**
- **Non-sibling**

116

N. Goharian & S. Mengle "Networked Hierarchies for Web Directories", *20th International World Wide Web conference (WWW)*, March 2011.

A Networked Hierarchy

Association Rule Mining

Calculate support between each two categories in the hierarchy

$$Support\ (c_{actual}, c_{predicted}) = \frac{\sigma\ (c_{actual} \cup c_{predicted})}{N}$$

$$Confidence\ (c_{actual} \Rightarrow c_{predicted}) = \frac{\sigma\ (c_{actual} \cup c_{predicted})}{\sigma\ (c_{actual})}$$

If *Support(C_{actual}, $C_{predicted}$) >= S-Threshold* *and Confidence(C_{actual}, $C_{predicted}$) >= C-Threshold,* predict relationship between C_{actual} & $C_{predicted}$

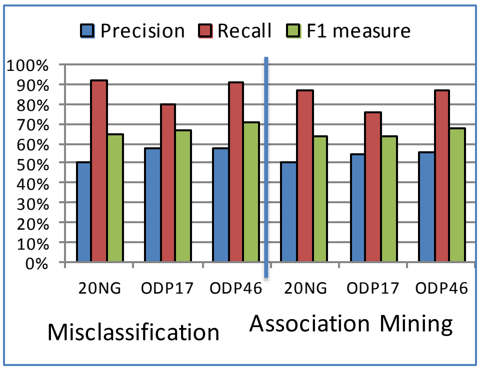
	20 NG	ODP17	ODP46
S-Threshold	0.08	0.17	0.03
C-Threshold	0.04	0.14	0.02

117

N. Goharian & S. Mengle "Networked Hierarchies for Web Directories", *20th International World Wide Web conference (WWW)*, March 2011.

A Networked Hierarchy

- 20 Newsgroups (20 NG)
  - 19,996 docs
- ODP (17 category)
  - 8,500 docs
- ODP (46 category)
  - 23,000 docs
- Manual evaluation by six assessors with a Pearson's correlation of 82%



N. Goharian & S. Mengle "Networked Hierarchies for Web Directories", *20th International World Wide Web conference (WWW)*, March 2011.

Understanding User Intent via Context-Aware Query Session Analysis

Step 1: Determining if context information is needed

Ambiguous query Vs. **Unambiguous query**

❖Only use context information for ambiguous/low weight queries

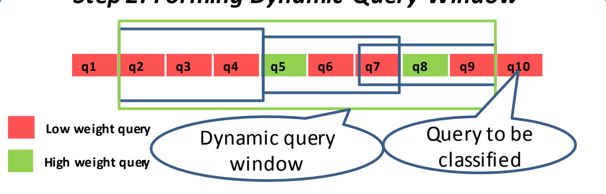
❖If query weight is lower than threshold (empirically determined as 0.7 for this work) then the query is marked as ambiguous

$$Weight(Query) = \sum_{i=1}^{\#Terms} \max\left(\forall_{j=1}^{\#Categories} \left(\frac{tf(t_i, c_j)}{tf(t_i)}\right)\right)$$

N. Goharian, S. Mengle, "Context Aware Query Classification Using Dynamic Query Window and Relationship Net", In proceedings of ACM 33rd SIGIR, July 2010.

Understanding User Intent via Context-Aware Query Analysis

Step 2: Forming Dynamic Query Window



❖Initially, create a static query window (Window size: 3)

❖If the static window contains a high-weight query, recursively expand the window for each unambiguous query

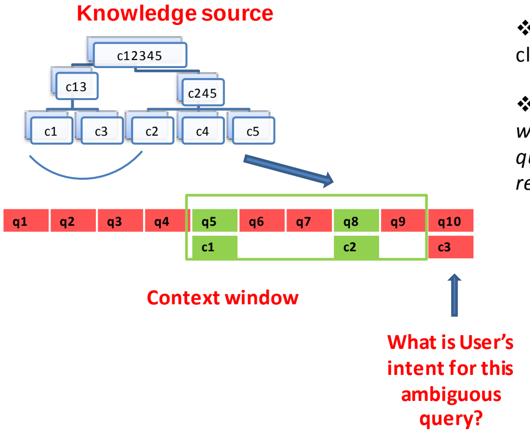
❖Terminate when all queries in window are ambiguous

❖Create the dynamic window by combining all the recursively generated query windows

N. Goharian, S. Mengle, "Context Aware Query Classification Using Dynamic Query Window and Relationship Net", *In proceedings of ACM 33rd SIGIR*, July 2010.

Understanding User Intent via Context-Aware Query Analysis

Step 3: Query weight adjustment using category relationship



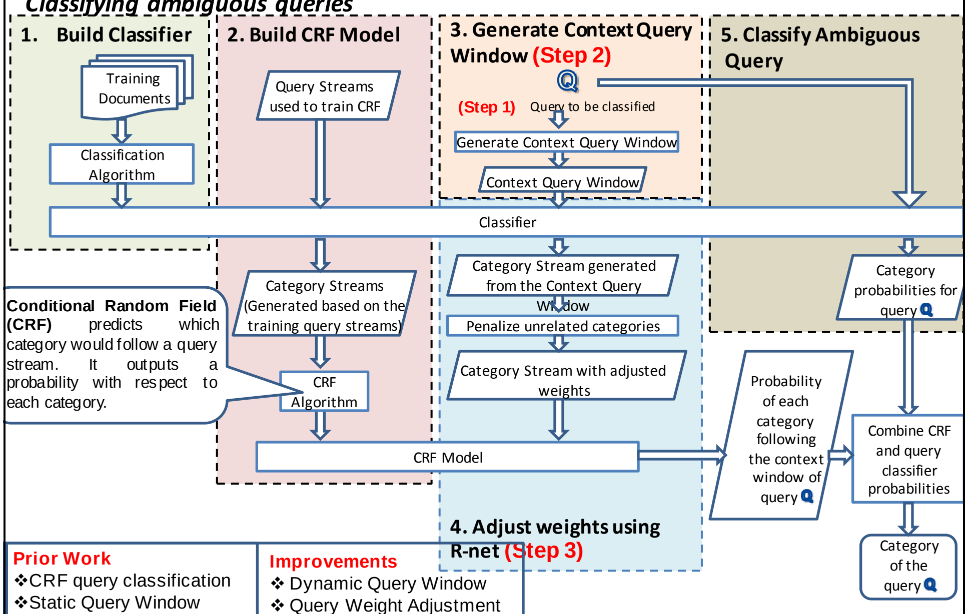
❖ Each unambiguous (high weight) query is classified to a category

❖ *Weight adjustment for each query in query window, by considering the category of that query and the target query as to their relationship in the knowledge source.*

N. Goharian, S. Mengle, "Context Aware Query Classification Using Dynamic Query Window and Relationship Net", *In proceedings of ACM 33rd SIGIR*, July 2010.

N. Goharian, S. Mengle, "Context Aware Query Classification Using Dynamic Query Window and Relationship Net", *In proceedings of ACM 33rd SIGIR*, 2010.

Classifying ambiguous queries



Prior Work
❖CRF query classification
❖Static Query Window
❖ Query Weight Adjustment using taxonomy

Improvements
❖ Dynamic Query Window
❖ Query Weight Adjustment using Relationship-net

Context-Aware Query Classification

- ❖ 67 categories from *KDD Cup 05* (predefined set of categories)
- ❖ 500 documents from *ODP dataset* for each category (as training documents for training a classifier)
- ❖ 500 query streams from *Excite query log*
  Query stream length : min:5; max:18; median:9
  Query length: Avg.: 2.7, Median: 3
- ❖ Taxonomy: KDD Cup 05 (7 level 1; 67 level 2; 306 sibling)
- ❖ R-net: 227 sibling & 58 non-sibling
- ❖ Used 10-fold cross validation to predict the category of the last query in each stream

SQW: Static Query Window
DQW: Dynamic Query Window
Rnet: Relationship Net

	% Improvements		
	Precision	Recall	F1
DQW over SQW	1.88%	3.22%	2.53%
DQW+Taxonomy over SQW+Taxonomy	3.51%	3.03%	3.28%
DQW+Rnet over SQW+Rnet	3.86%	5.72%	4.74%
SQW+Rnet over SQW+Taxonomy	6.71%	7.13%	6.91%
DQW+Rnet over DWQ+Taxonomy	7.06%	9.92%	8.42%
DQW+Rnet over SQW+Taxonomy	10.82%	13.26%	11.98%

N. Goharian, S. Mengle, "Context Aware Query Classification Using Dynamic Query Window and Relationship Net", *In proceedings of ACM 33rd SIGIR*, July 2010.

123

$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam

- Studying potential of *Content-based* approaches to *short text* spam detection.
- Research Questions:
  - Which features are more useful?
  - What are the effects of combining multiple features?
  - Do rule based features (tailored to spam) perform relatively well?

Z. Tan, N. Goharian, M. Sherr, "$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam", *In proceedings of ACM 35th Conference on Research and Development in Information Retrieval (SIGIR)*, August 2012. (short)

$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam (Cont'd)

Feature sets used by earlier research (our baseline):

- ***Cormack et al.*** – union of of orthogonal sparse word bigrams , character 2-grams and 3-grams, and words. We consider two versions for completeness:
  - *Cormack AlphaNum* – alphanumeric symbols only
  - *Cormack Fulltext* – all symbols included
- ***Almeida et al.*** – Two simple tokenization techniques:
  - *tok1* – tokens starting with any printable character followed by alphanumeric characters. Dots/commas/colons treated as separators
  - *tok2* – tokens are a series of any characters except for blanks, tabs, returns, dots, commas, colons, dashes which are treated as separators.

$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam (Cont'd)

Evaluated:

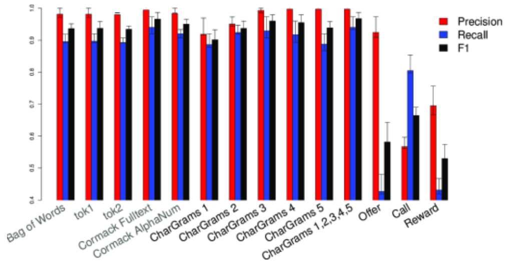
- Rule Based Features (RegEx form)
  - **rate**: (/|per)(|)(year|month|hour|week|call)
  - **reward**: free|award|prize|win|reward
  - **website**: .co|.org|.net
  - **call**: call|text|txt|msg|contact
  - **offer**: (**call** ∪ **website**) ∩ (**reward** ∪ **rate**)
- N-grams
  - Character [1,5]-grams ← *CharGrams#*
  - Word grams
  - Alphanumeric-only versions of previous n-grams
- Statistical features
  - Length, in characters and words
  - Proportion of upper-case letters
  - Proportion of punctuation

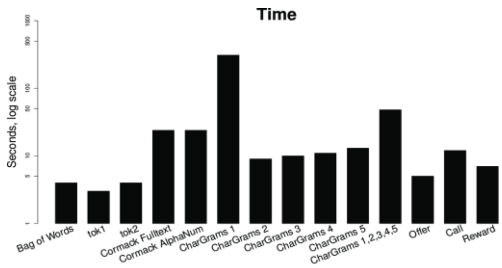
Z. Tan, N. Goharian, M. Sherr, "$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam", *In proceedings of ACM 35th Conference on Research and Development in Information Retrieval (SIGIR)*, August 2012. (short)

Z. Tan, N. Goharian, M. Sherr, "$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam", *In proceedings of ACM 35th Conference on Research and Development in Information Retrieval (SIGIR),* August 2012. (short)

$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam (Cont'd)

- SVMLight
- Stratified 10-fold cross validation
- Labelled data set *[Cormack et al]:*
  Total: 5574
  Ham: 4827
  Spam: 747
- Potential short-comings of dataset:
  ham & spam from different geographic area





$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam (Cont'd)

- Simple is better than composite
  CharGrams3: F1: 95.97 (9 sec) vs. ***Cormack Fulltext**:* F1: 96.62 (24 sec)

- RegEx & statistical features - generally poor performance
  Offer → high *precision, very low recall & F1*
  Call → ok *recall, low precision*

- Mutual Information Study
  Numbers → generally good indicators of spam
  Slang words → are specific to areas appearing in ham
  Words (**claim, won, yes, price**) specific to spam
  – *may lead to better **offer rules**.*

Z. Tan, N. Goharian, M. Sherr, "$100,000 Prize Jackpot. Call Now! Identifying the Pertinent Features of SMS Spam", *In proceedings of ACM 35th Conference on Research and Development in Information Retrieval (SIGIR),* August 2012. (short)

References used to prepare this set of slides

TEXTBOOK:

- Introduction to Information Retrieval, Manning, Raghavan and Schütze, 2008
- Introduction to Data Mining, Tan, Steinbach, Kumar, Addison Wesley, 2006
- Data Mining Concepts and Techniques, Han, Kamber, Pei, Morgan Kaufmann, 2011

RESEARCH LITERATURE:

- Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text problem. In Proceedings of the 14th international conference on machine (ICML), 1997.
- F. Sebastiani , *Machine Learning in Automated Text Categorization, 2002*
- X. Qi and B. Davison, *ACM Computing Surveys article, 2009*
- *Forman, G. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research,2003.*
- S. Mengle, N. Goharian, "Detecting Hidden Passages from Documents", *In proceedings of SIAM Conference on Data Mining (SIAM - SDM) Workshop*, April 2008.
- N. Goharian, S. Mengle, "On Document Splitting in Passage Detection", *In proceedings of ACM 31st Conference on Research and Development in Information Retrieval (SIGIR),* July 2008. (short)
- S. Mengle and N. Goharian, "Passage Detection Using Text Classification", *Journal of American Society for Information Science and Technology (JASIST),* 60 (4), March 2009.
- S. Mengle, N. Goharian, "Using Ambiguity Measure Feature Selection Algorithm for Support Vector Machine Classifier", *ACM 23rd Symposium on Applied Computing (SAC),* March 2008.
- S. Mengle and N. Goharian, "Ambiguity Measure Feature Selection Algorithm", *Journal of American Society for Information Science and Technology (JASIST),* 60 (5), April 2009.
- S. Mengle and N. Goharian, "Detecting Relationships among Categories using Text Classification", *Journal of American Society for Information Science and Technology (JASIST),* 61 (5), May 2010.
- N. Goharian, S. Mengle "Networked Hierarchies for Web Directories", *20th International World Wide Web conference (WWW),* March 2011. (short)
- N. Goharian, S. Mengle, "Context Aware Query Classification Using Dynamic Query Window and Relationship Net", *In proceedings of ACM 33rd Conference on Research and Development in Information Retrieval (SIGIR),* July 2010. (short)

INVITED TALKS (These are my invited talks, from which I have included some slides):

- CNR, Pisa, Italy, June 2010
- Fu-Jen University, Taiwan, December 2012
- Tsinghua Taiwan, December 2012