Information Retrieval

Andreas Rauber
http://www.ifs.tuwien.ac.at/~andi

Outline

- retrieval
- relevance feedback
- automatic text classification (ATC)
- evaluation
- clustering

Retrieval

- documents and queries
- retrieval: find documents satisfying query
- different retrieval models
  - boolean (exact match)
  - vector space model
  - probabilistic models

Retrieval: exact vs. best match

- exact match:
  - precise query
  - retrieved documents satisfy query criteria
  - result: unordered set of documents
  - efficient
  - predictable, clear evaluation criteria
  - works with clearly identifiable goals
  - good performance in specific domains
  - queries difficult to formulate
  - does not work too well in many general-purpose applications

Retrieval: exact vs. best match

- best-match:
  - query describes optimal document
  - retrieved documents satisfy criterium as far as possible
  - result: ranked set of documents
  - works well with unclear criteria
  - may return unrelevant documents
  - harder to evaluate wha a certain document was returned
  - seems to outperform exact match in many application scenarios

Retrieval: boolean model

- most common exact match model
- still widely used
- supports a range of boolean operators:
  - and, or, not
  - proximity
  - position
  - regular expressions

Retrieval: WESTLAW

- (slides taken from lecture "Retrieval Models" Marten de Rijke et al., Univ. Amsterdam)
- large commercial system
- legal material, news, stock exchange data
- operational since 1974
- approx. 700.000 users, 5-7 TB data
- supports exact-match
- best-match added in 1992

Retrieval: WESTLAW

- supports range of operators:
  - phrases: "West Publishing"
  - word proximity: West /5 Publishing
  - same sentence: Massachusetts /s technology
  - same paragraph: "information retrieval" /p "exact match"
  - restrictions: (DATE(AFTER 1992 & BEFORE 1995)
- term expansion:
  - wild card: THOM*ON
  - truncation: THOM!
- queries according to document structure (fields)

Retrieval: WESTLAW

- query examples:
  - What is the statute of limitations in cases involving the federal tort claims act?
    - LIMIT! /3 STATUTE ACTION /S FEDERAL /2 TORT /3 CLAIM
  - What factors are important in determining what constitutes a vessel for purposes of determining liability of a vessel owner for injuries to a seaman under the "Jones Act" (46 USC 688)?
    - (741 +3 824) FACTOR ELEMENT STATUS FACT /P VESSEL SHIP BOAT /P (46 +3 688) "JONES ACT" /P INJUR! /S SEAMAN CREWMAN WORKER

Retrieval: WESTLAW

- query examples:
  - Are there any cases which discuss negliogent maintenance or failure to maintain aids to navigation such as lights, buoys, or channel markers?
    - NOT NEGECT! FAIL! NEGLIG!/5 MAINT! REPAIR! /P NAVIGAT! /5 AID EQUIP! LIGHT BUOY "CHANNEL MARKER"
  - What cases have discussed the concept of excusable delay in the application of statutes of limitations or the doctrine of laches involving actions in admiralty or under the "Jones Act" or the "Death on the High Seas Act"?
    - EXCUS! /3 DELAY /P (LIMIT! /3 STATUTE ACTION) LACHES / P "JONES ACT" "DEATH ON THE HIGH SEAS ACT" (46 +3 761)

Retrieval: boolean model

- develop query incrementally
- add new terms until results satisfy information need
- relatively long queries
- rather complex queries
- require detailed knowlede on query and document domain
- mostly for trained experts
- easier to control

Retrieval: vector space model

- best-match
- indexing: high-dimensional feature space
- query: describes "optimal" document, information need
- documents are vectors
- queries are vectors
- return documents that are closest to query in high-dimensional feature space
- measure similarity in two ways:
  - distance between points
  - angle between vectors

Retrieval: vector space model

- similarity: distance between end points
- range of metrics
- L1 (city block):
- L2 (euclidean distance): $m=\sqrt{\sum_{i=1}^{n}(d_i-q_i)^2}$
- minkovsky metric (general): $m=\left(\sum_{i=1}^{n}(d_i-q_i)^r\right)^{1/r}$

Retrieval: vector space model

- vector normalization
- normalize to unit length
- vector length $\|d\|=\sqrt{\sum_{i=1}^{n} d_i^2}$
- normalize: $\frac{d_i}{\|d\|}$

Retrieval: vector space model

- given unit vectors d, q: $\cos = d \cdot q$
  - cos = 0: no similarity
  - cos = 1: identical docs
- cosine similarity

$$sim(d,q)=\frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2}\cdot\sqrt{\sum_{i=1}^{n} d_i^2}}=\left(\|q_i\|^{-1}\cdot q\right)\cdot\left(\|d_i\|^{-1}\cdot d\right)$$

Retrieval: vector space model

- other similarity measures
  - Jacquard measure
  - Dice's coefficient

Retrieval: probabilistic models

- goal: identifying documents that are relevant
- 2-class classification problem
- model probability that document belongs to relevant class
- different ways of modeling probability
- basic assumption: relevance of document is independent of other docs in collection
- use Bayes rule: P(R|d) = P(d|R)*P(R)/P(d)

Outline

- retrieval
- relevance feedback
- automatic text classification (ATC)
- evaluation
- clustering

Relevance Feedback

- retrieval: one-stop shop
  if results not satisfactory: re-try
  re-fine: add/remove terms
- interactive retrieval
- relevance feedback: have computer add/remove terms automatically
- from result set: show which ones are relevant/not relevant
- algorithm infers how to weight terms
- active learning

Relevance Feedback

- problem: not possible to show many
  - select few documents for user to rate
- problem: which examples to choose?
  - most positive:
    probably little additional information
    not useful to capture range of positives
  - most borderline:
    most difficult to decide
    may not lead to most relevant returned

Relevance Feedback: rocchio

- re-write query be adding terms from relevant documents and subtracting terms from non-relevant documents

$$w_{ki} = \beta \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|POS_i|} - \gamma \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|NEG_i|}$$

- typical values: =4, =1

Relevance Feedback: blind

- normally, user defines relevant docs
- blind relevance feedback: top-n documents are considered relevant
- Rocchio: =1, =0,
  only use pos. examples

Outline

- retrieval
- relevance feedback
- automatic text classification (ATC)
- evaluation
- clustering

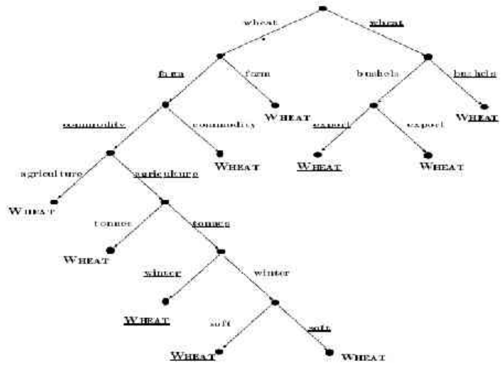
ATC

- assigning documents to pre-defined classes
- relevant vs. not-relevant
- target function $:D \times C \rightarrow \{t, f\}$
- approximated by classifier:
- categories are just symbolic labels
- no exogenous knowledge
- binary, single label, multi-label
- document pivoted vs. category pivoted

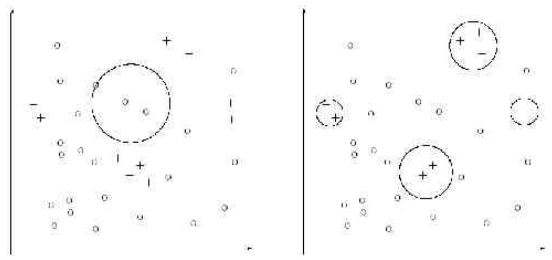
ATC

- range of machine learning algorithms
  - knn-classifiers
  - decision trees
  - rocchio
  - naive bayes
  - support vector machines
  - ...
- multi-label: set of individual binary classifiers
- classifier committees

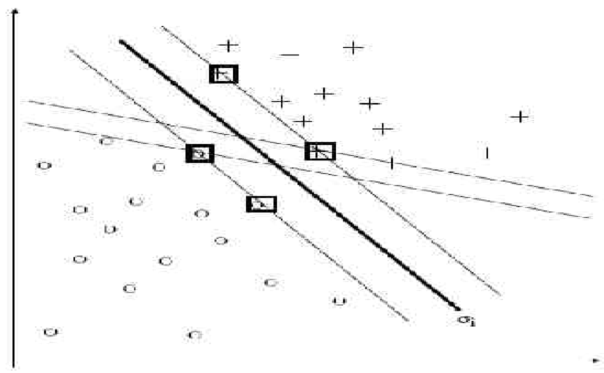
ATC - decision trees



ATC: Rocchio vs. knn



ATC - SVM



ATC - dimensionality reduction

- dimensionality reduction
- global vs. local
- selection vs. extraction
  - information gain
  - chi-squared
  - mutual information
  - ...

ATC: dimensionality reduction

Function	Denoted by	Mathematical form
Document frequency	$\#(t_k, c_i)$	$P(t_k \mid c_i)$
DIA association factor	$z(t_k, c_i)$	$P(c_i \mid t_k)$
Information gain	$IG(t_k, c_i)$	$P(t_k, c_i) \log \frac{P(t_k, c_i)}{P(c_i) \cdot P(t_k)} + P(\overline{t}_k, c_i) \log \frac{P(\overline{t}_k, c_i)}{P(c_i) \cdot P(\overline{t}_k)}$
Mutual information	$MI(t_k, c_i)$	$\log \frac{P(t_k, c_i)}{P(t_k) \cdot P(c_i)}$
Chi-square	$\chi^2(t_k, c_i)$	$\frac{\lvert Tr \rvert \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)]^2}{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}$
NGL coefficient	$NGL(t_k, c_i)$	$\frac{\sqrt{\lvert Tr \rvert} \cdot [P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)]}{\sqrt{P(t_k) \cdot P(\overline{t}_k) \cdot P(c_i) \cdot P(\overline{c}_i)}}$
Relevancy score	$RS(t_k, c_i)$	$\log \frac{P(t_k \mid c_i) + d}{P(\overline{t}_k \mid \overline{c}_i) + d}$
Odds Ratio	$OR(t_k, c_i)$	$\frac{P(t_k \mid c_i) \cdot (1 - P(t_k \mid \overline{c}_i))}{(1 - P(t_k \mid c_i)) \cdot P(t_k \mid \overline{c}_i)}$
GSS coefficient	$GSS(t_k, c_i)$	$P(t_k, c_i) \cdot P(\overline{t}_k, \overline{c}_i) - P(t_k, \overline{c}_i) \cdot P(\overline{t}_k, c_i)$

ATC: dimensionality reduction

- term extraction:
  - term clustering
    replace with cluster centroids
    "synonyms"
  - latent semantic indexing
    singular value decomposition
    creates new feature space
    orthogonal axes

ATC results

System	Type	Results reported by	#1	#2	#3	#4	#5
		# of documents	21,450	14,347	13,272	12,902	12,902
		# of training documents	14,704	10,667	9,610	9,603	9,603
		# of test documents	6,746	3,680	3,662	3,299	3,299
		# of categories	135	93	92	90	10
Word	(non-learning)	[Yang 1999]	.150	.310	.290		
	probabilistic	[Dumais et al. 1998]				.752	.815
	probabilistic	[Joachims 1998]					.720
	probabilistic	[Lam et al. 1997]	.443 (MF$_1$)				
PropBayes	probabilistic	[Lewis 1992a]	.650				
Bim	probabilistic	[Li and Yamanishi 1999]				.747	
	probabilistic	[Li and Yamanishi 1999]				.773	
Nb	probabilistic	[Yang and Liu 1999]				.795	
	decision trees	[Dumais et al. 1998]					.884
C4.5	decision trees	[Joachims 1998]					.794
Ind	decision trees	[Lewis and Ringuette 1994]	.670				
Swap-1	decision rules	[Apté et al. 1994]		.805			
Ripper	decision rules	[Cohen and Singer 1999]	.683	.811		.820	
SleepingExperts	decision rules	[Cohen and Singer 1999]	.753	.759		.827	
Dl-Esc	decision rules	[Li and Yamanishi 1999]				.820	
Charade	decision rules	[Moulinier and Ganascia 1996]		.738			
Charade	decision rules	[Moulinier et al. 1996]		.783 (F$_1$)			
Llsf	regression	[Yang 1999]		.855	.810		
Llsf	regression	[Yang and Liu 1999]				.849	
BalancedWinnow	on-line linear	[Dagan et al. 1997]	.747 (M)	.833 (M)			
Widrow-Hoff	on-line linear	[Lam and Ho 1998]				.822	
Rocchio	batch linear	[Cohen and Singer 1999]	.660	.748		.776	
FindSim	batch linear	[Dumais et al. 1998]				.617	.646
Rocchio	batch linear	[Joachims 1998]					.799
Rocchio	batch linear	[Lam and Ho 1998]				.781	
Rocchio	batch linear	[Li and Yamanishi 1999]				.625	
Classi	neural network	[Ng et al. 1997]		.802			
Nnet	neural network	[Yang and Liu 1999]				.838	
	neural network	[Wiener et al. 1995]			.820		
Gis-W	example-based	[Lam and Ho 1998]				.860	
k-NN	example-based	[Joachims 1998]					.823
k-NN	example-based	[Lam and Ho 1998]				.820	
k-NN	example-based	[Yang 1999]	.690	.852	.820		
k-NN	example-based	[Yang and Liu 1999]				.856	
	SVM	[Dumais et al. 1998]				.870	.920
SvmLight	SVM	[Joachims 1998]					.864
SvmLight	SVM	[Li and Yamanishi 1999]				.841	
SvmLight	SVM	[Yang and Liu 1999]				.859	
AdaBoost.MH	committee	[Schapire and Singer 2000]		.860			
	committee	[Weiss et al. 1999]				.878	
	Bayesian net	[Dumais et al. 1998]				.800	.850
	Bayesian net	[Lam et al. 1997]	.542 (MF$_1$)				

© F. Sebastiani: Machine Learning in Automated Text Categorization,
ACM Computing Surveys, 34(1):1-47, 2001

ATC collections

- Reuters
  (22173, 21578, ModLewis, ModApte)
- OHSUMED
- TREC
- 20-newsgroups

Outline

- retrieval
- relevance feedback
- automatic text classification (ATC)
- evaluation
- clustering

Evaluation

- training set - validation set - test set
- same split
- specific task or cross-validation
- same (or: specified) pre-processing
- same (or specified) indexing
- same evaluation measures

Evaluation

- contingency table:

Category set $C = \{c_1, \ldots, c_{\lvert C \rvert}\}$		expert judgments: YES	expert judgments: NO
classifier	YES	$TP = \sum_{i=1}^{\lvert C \rvert} TP_i$	$FP = \sum_{i=1}^{\lvert C \rvert} FP_i$
judgments	NO	$FN = \sum_{i=1}^{\lvert C \rvert} FN_i$	$TN = \sum_{i=1}^{\lvert C \rvert} TN_i$

Evaluation

- precision: $\hat{\pi}^{\mu} = \frac{TP}{TP+FP} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i+FP_i)}$
- recall: $\hat{\rho}^{\mu} = \frac{TP}{TP+FN} = \frac{\sum_{i=1}^{|C|} TP_i}{\sum_{i=1}^{|C|} (TP_i+FN_i)}$
  (micro-averaged)
- macro averaged: $\hat{\pi}^{M} = \frac{\sum_{i=1}^{|C|} \hat{\pi}_i}{|C|}$
  (first per category, $\hat{\rho}^{M} = \frac{\sum_{i=1}^{|C|} \hat{\rho}_i}{|C|}$
  then average)

Evaluation

- accuracy: $A = \frac{TP+TN}{TP+TN+FP+FN}$
- error: $E = \frac{FP+FN}{TP+TN+FP+FN} = 1 - A$
- utility:

Category set $C = \{c_1, \ldots, c_{\|C\|}\}$		expert judgments YES	expert judgments NO
classifier judgments	YES	u_{TP}	u_{FP}
	NO	u_{FN}	u_{TN}

Evaluation

- combined effectiveness measures
  - break-even: point at which =
  - F : $F_{\beta} = \frac{(\beta^2+1)\pi\rho}{\beta^2\pi - \rho}$

Outline

- retrieval
- relevance feedback
- automatic text classification (ATC)
- evaluation
- clustering

Clustering

- no class information available
- unsupervised learning
- identify groups of documents with similar characteristics
- used in a range of applications
  - pre-processing
  - novelity detection
  - document summarization
  - text analysis

Clustering

- range of clustering techniques
- clustering vs.
  topology-preserving mapping
- visualization
- self-organizing maps
- more difficult to evaluate
- cluster validity measures and interpretations