A Hidden Markov Model Information Retrieval System

Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999.

David R.H. Miller, Tim Leek, and Richard M. Schwartz

Time-Based Language Models

Proceedings of the Twelfth International Conference on Information and Knowledge Management. ACM, 2003.

Xiaoyan Li and W. Bruce Croft

Presented by Matt Chaney

CS 834 - Presentation 3

History and Motivation

- Information Retrieval with Probabilistic Models are not new
  - Maron and Kuhns, 1960[1]

- Recent developments involve heuristics or smoothing, deviating from core probabilistic concepts

- Hidden Markov Model (HMM) more closely tied to formal foundation making extension and study easier

[1] Maron, Melvin Earl, and John L. Kuhns. "On Relevance, Probabilistic Indexing and Information Retrieval." Journal of the ACM (JACM) 7.3 (1960): 216-244.

Hidden Markov Model (HMM)

- Set of states

- State transition probabilities

- Set of output symbols

- Output probability distribution for each state
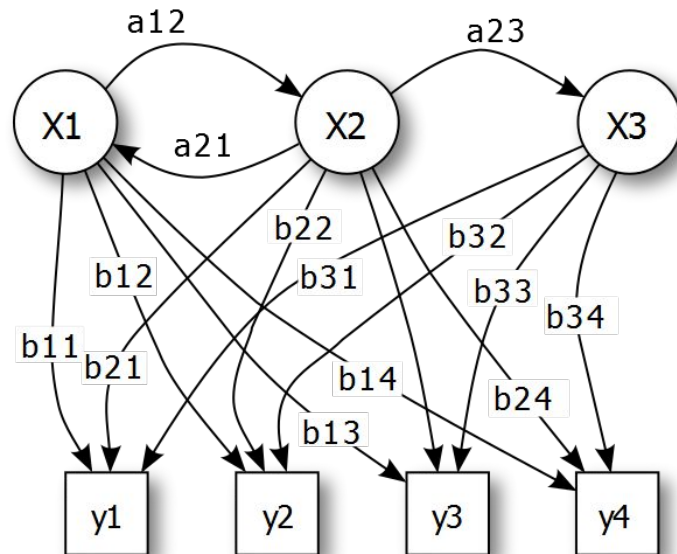
- Internal States Invisible

Fig 1. Hidden Markov Model with Output.
Tdunning. Wikimedia Commons. Public Domain.

HMM Application to Ad Hoc Search

- Observed data ➜ Query (Q)

- Unknown ➜ Desired relevant document (D)

- State Transitions ➜ Query Word Generation

HMM Application to Ad Hoc Search

- Observed data ➜ Query (Q)

- Unknown ➜ Desired relevant document (D)

- State Transitions ➜ Query Word Generation

$$P(D \text{ is } R|Q) = \frac{P(Q|D \text{ is } R) \cdot P(D \text{ is } R)}{P(Q)}$$

HMM Application to Ad Hoc Search

- Observed data ➜ Query (*Q*)

- Unknown ➜ Desired relevant document (*D*)

- State Transitions ➜ Query Word Generation

$$P(D \text{ is } R|Q) = \frac{\boxed{P(Q|D \text{ is } R)} \cdot \cancel{P(D \text{ is } R)}}{\cancel{P(Q)}}$$

- *P(D is R)* and *P(Q)* can be ignored (for now)

Query Term Generation

- Two states
  - $P(q \mid GE)$
  - $P(q \mid D)$

- Two state transition probabilities, a_0 and a_1

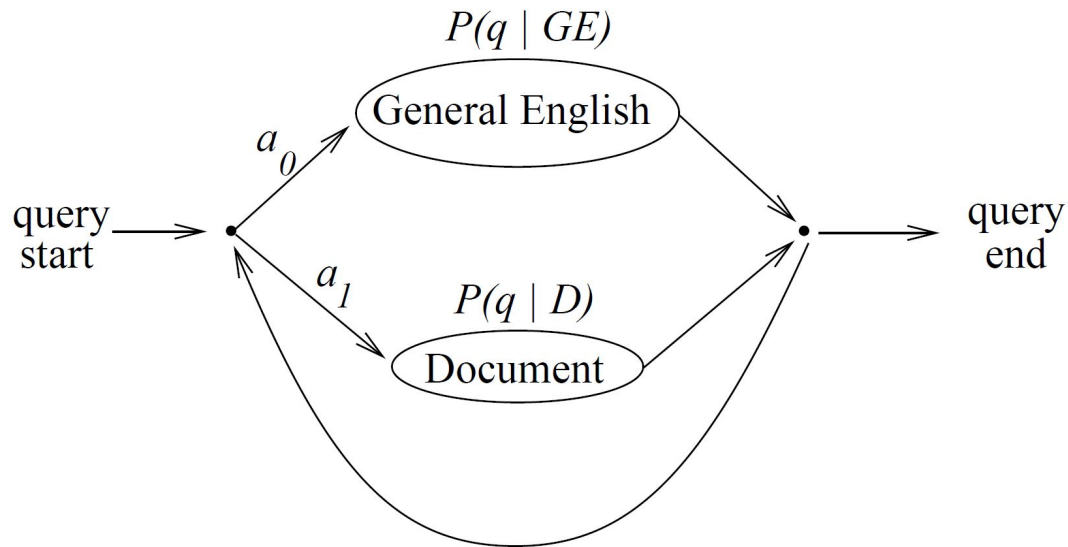- Each document has distinct output symbol probabilities

Fig 1. A simple two-state HMM.
Miller et al. 1999.

Output Probabilities

- From the Document state

$$P(q|D_k) = \frac{\text{number of times } q \text{ appears in } D_k}{\text{length of } D_k}$$

- From the General English state

$$P(q|GE) = \frac{\sum_k \text{number of times } q \text{ appears in } D_k}{\sum_k \text{length of } D_k}$$

Ranking

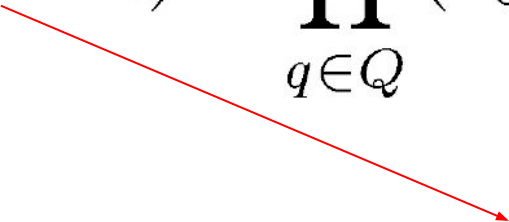- With previous two calculations, derive the following

$$P(Q|D_k \text{ is } R) = \prod_{q \in Q} \left(a_0 P(q|GE) + a_1 P(q|D_k)\right)$$

Ranking

- With previous two calculations, derive the following

$$P(Q|D_k \text{ is } R) = \prod_{q \in Q} (a_0 P(q|GE) + a_1 P(q|D_k))$$

- Recall

$$P(D \text{ is } R|Q) = \frac{P(Q|D \text{ is } R) \cdot P(D \text{ is } R)}{P(Q)}$$

- *P(Q|D is R)* is the ranking score, highest probability "wins"

Experiment Details

- Two corpora
  - TREC-6: 500K Documents from news and governmental agencies
  - TREC-7: Subset of TREC-6

- Each corpus contains 50 topics (queries) with relevant documents for each
  - Topics contain *Title*, *Description* and *Narrative* sections

- Compared HMM precision with familiar *tf.idf* measure with non-interpolated average precision

$$\frac{1}{|Rel|} \sum_{D \in Rel} \frac{|\{D' \in Rel, r(D') \leq r(D)\}|}{r(D)}$$

The old standby

$$tf.idf(Q, D) = \sum_{q_i \in Q} wtf(q_i, D) \cdot idf(q_i)$$

$$wtf(q, D) = \frac{tf(q, D)}{tf(q, D) + 0.5 + 1.5\frac{l(D)}{al}}$$

$$idf(q) = \frac{\log \frac{N}{n_q}}{N + 1}$$

N = number of documents in the corpus

n_q = number of documents in the corpus containing q

$tf(q, D)$ = number of times q appears in D

$l(D)$ = length of D in words

al = avg length in words of a D in the corpus

Fig 3. Comparison *tf.idf* formula.
Miller et al. 1999.

Results

	TREC-6			TREC-7		
	HMM	*tf.idf*	Diff	HMM	*tf.idf*	Diff
Title	21.6	15.9	+5.8	16.1	11.6	+4.5
Desc	18.1	11.9	+6.2	18.3	14.2	+4.1
Narr	21.5	15.8	+5.7	17.7	14.7	+3.0
Full	27.1	18.9	+8.2	23.9	19.0	+4.9

Table 1. HMM scoring vs. *tf.idf* on TREC-6 and TREC-7.
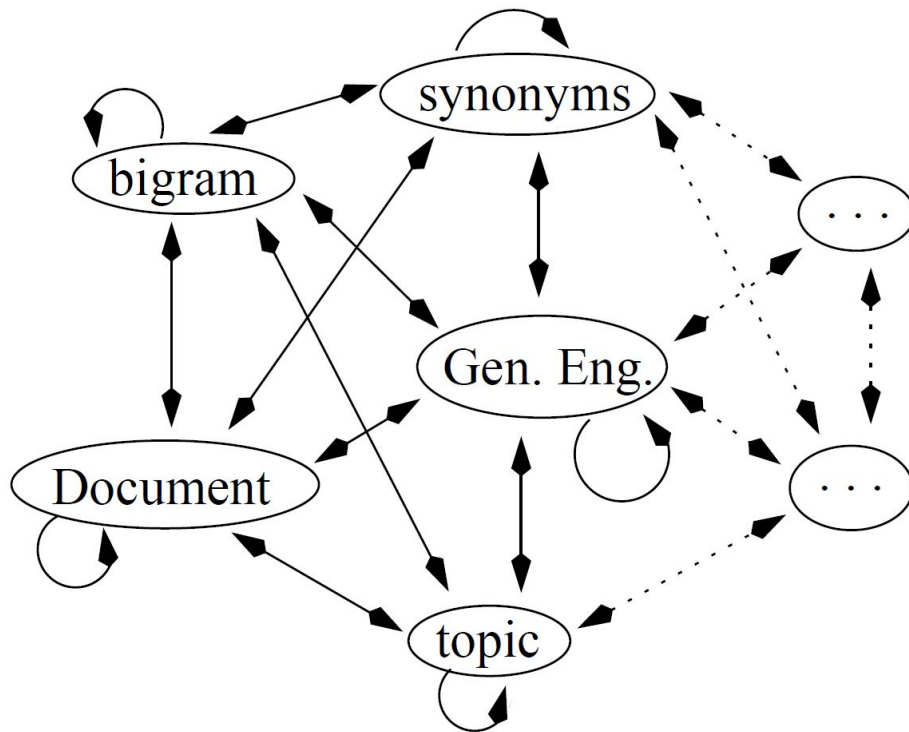*Miller et al. 1999.*

Expanded HMM Example

Fig 2. An expanded multi-state HMM.
Miller et al. 1999.

Refinements

- Blind Feedback

- Bigram Modeling

- Feature Dependent Priors

- Query Section Weighting

Blind Feedback

- Run query and then recalculate relevance based on results
  - Ex: The Rocchio Algorithm [2]

- Augment initial query with words in two or more of the top *N* documents
  - Distinction between stop words and valid terms

- Transition probabilities estimated using training data

[2] Rocchio, Joseph John. "Relevance Feedback in Information Retrieval." (1971): 313-323.

Estimating Transition Probabilities

- Proportionally distribute observed count to either Document or General English state transitions

- If the word exists in the document then a_1 is far greater than a_0, since it is in proportion to the size of the set; consider:
  - Documents contain only hundreds of words
  - General English contains hundreds of thousands of words

- Estimating state transition a_1 becomes very close to

$$P(q' \in D' | D' \text{ is rel. to } Q') =$$

$$\frac{1}{|\mathcal{Q}|} \sum_{Q_i \in \mathcal{Q}} \sum_{w \in Q_i} \frac{|D \ s.t. \ w \in D, D \text{ is rel. to } Q_i|}{|Q_i| \cdot |D \text{ is rel. to } Q_i|}$$

Developing Blind Feedback

- *M*-intersections

$$I_{m,Q} = \left\{ \begin{array}{c} w \text{ appearing in exactly } m \text{ of the} \\ \text{top } N \text{ documents for query } Q \end{array} \right\}$$

- Define

$$\gamma_{m,Q',x} = P\left(q' \in D' \;\middle|\; \begin{array}{l} D' \text{ is rel. to } Q', \\ q' \in I_{m,Q'},\, df(q') = x \end{array} \right)$$

- Finally arriving at

$$a_1 = \gamma_{m,Q,df(q)} - df(q)$$

Blind Feedback Results

	TREC-6	TREC-7
basic HMM	27.1	23.9
w/blind feedback	30.6	27.4
improvement	+3.5	+3.5

Table 2.Performance gain from blind feedback.
Miller et al. 1999.

Bigrams

- Add context to probability of word being generated

$$P(q_n|D_k, q_{n-1}) =$$

$$\frac{\text{number of times } q_{n-1}q_n \text{ appears in } D_k}{\text{number of times } q_{n-1} \text{ appears in } D_k}$$
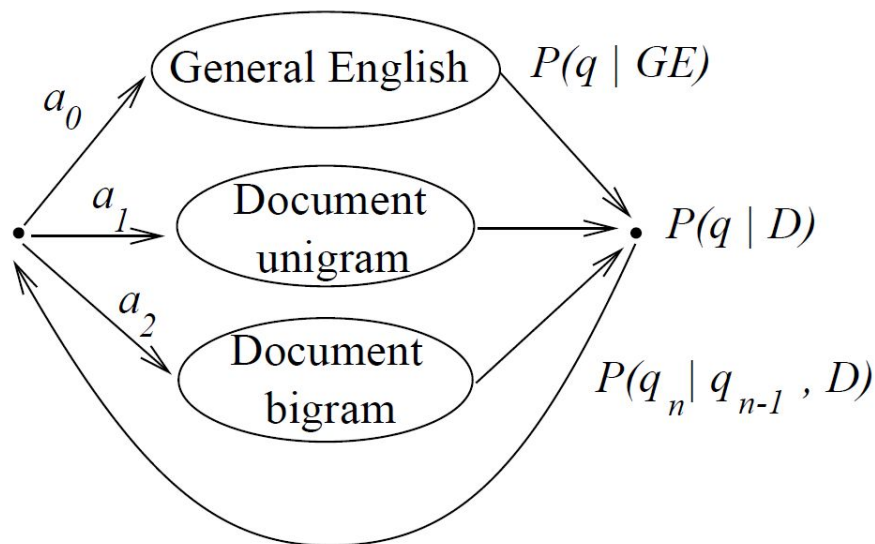
Bigram Results

	TREC-6	TREC-7
basic HMM	27.1	23.9
w/bigrams	28.1	24.4
improvement	+1.0	+0.5

Table 3. Performance gains from adding a bigram state.
Miller et al. 1999.

Query Section Weighting

- Original ranking function

$$P(Q|D_k \text{ is } R) = \prod_{q \in Q} (a_0 P(q|GE) + a_1 P(q|D_k))$$

- $v_{s(q)}$ ➡ weight (# repetitions) for section in which query term *q* appears

$$P(Q|D_k \text{ is } R) = \prod_{q \in Q} (a_0 P(q|GE) + a_1 P(q|D_k))^{v_{s(q)}}$$

Query Section Weighting Results

	TREC6	TREC7
basic HMM	27.1	23.9
w/query weights	30.0	25.1
improvement	+2.9	+1.2

Table 4. Performance gains query section weighting.
Miller et al. 1999.

Document Priors

- Recall prior probability $P(D\ is\ R)$ assumed constant for all documents, but this may not be true
  - Scientific journals likely more relevant than tabloids
  - Longer documents may be more relevant than shorter ones

- Empirically found following features correlate with prior relevance
  - Source
  - Document length
  - Average word length

- Added this to ranking

Document Priors Results

	TREC6	TREC7
basic HMM	27.1	23.9
w/non-constant prior	27.6	24.0
improvement	+0.5	+0.1

Table 5. Performance with non-constant prior.
Miller et al. 1999.

Results Combined

	TREC-6	TREC-7
basic HMM	27.1	23.9
w/blind feedback	+3.5	+3.5
w/query weights	+2.9	+1.2
w/non-constant prior	+0.5	+0.1
w/bigrams	+1.0	+0.5
HMM w/all refinements	33.2	28.0

Table 6. Performance gains with refinements to the HMM system.
Miller et al. 1999.

A Hidden Markov Model Information Retrieval System

Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 1999.

David R.H. Miller, Tim Leek, and Richard M. Schwartz

Time-Based Language Models

Proceedings of the Twelfth International Conference on Information and Knowledge Management. ACM, 2003.
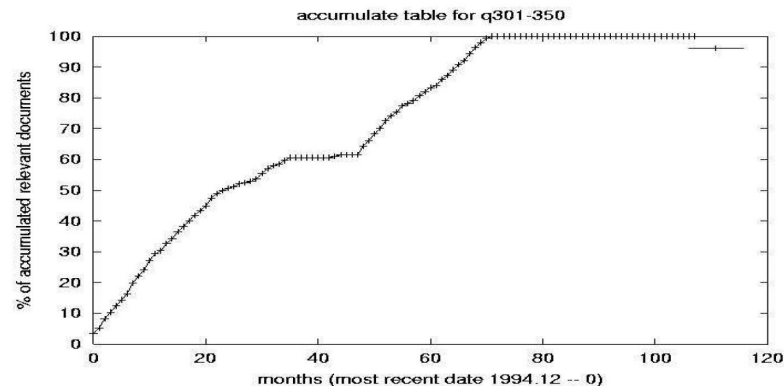
Xiaoyan Li and W. Bruce Croft

Presented by Matt Chaney

CS 834 - Presentation 3

Motivation

- Relevancy may be skewed based on the time the document was created

- Some information needs may be matched up closer to more recent documents

- Others may be more closely aligned with documents from a specific period in time

Time-Based Query Trends

accumulate table for q101-150

accumulate table for q151-200

accumulate talble for q201-249

accumulate table for q301-350
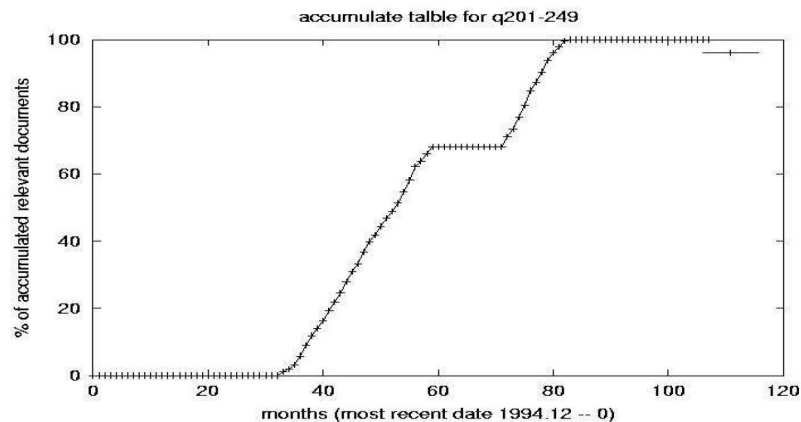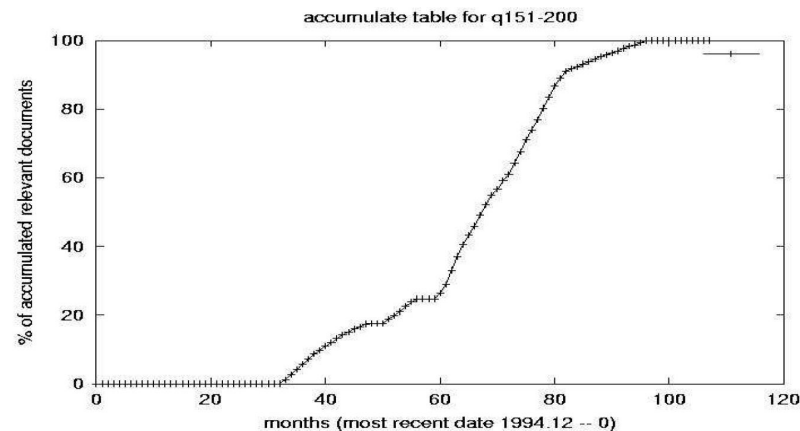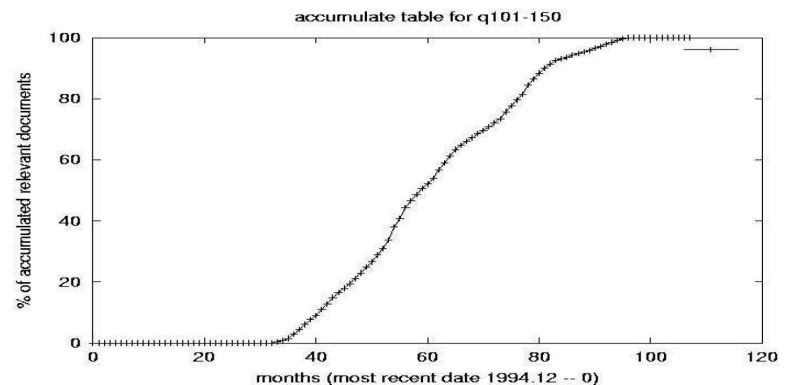
Retrieval Language Models

- Query Likelihood Model
  - Miller et al. HMM model

$$p(d \mid q) \propto p(q \mid d)\, p(d)$$

- Relevance Models
  - KL Divergence

$$p(w, q_1, \ldots, q_m) = \sum_{M \in \mathrm{M}} p(M)\, p(w \mid M) \prod_{i=1}^{m} p(q_i \mid M)$$

Retrieval Language Models

- Query Likelihood Model
  - Miller et al. HMM model

- **Time-Based Relevance Models**

- Replace $p(d)$ with $p(d|T_d)$

$$p(d/q) \propto p(q/d)\boxed{p(d)} \longrightarrow p(d/q) \propto p(q/d)\boxed{p(d/T_d)}$$

- Replace $p(M)$ with $p(M|T_D)$

- Relevance Models
  - KL Divergence

$$p(w,q_1,...,q_m) = \sum_{M \in \mathrm{M}} \boxed{p(M)} p(w/M) \prod_{i=1}^{m} p(q_i/M) \longrightarrow p(w,q_1,...,q_m) = \sum_{M \in \mathrm{M}} \boxed{p(M/T_D)} p(w/M) \prod_{i=1}^{m} p(q_i/M)$$

Time-Related Prior Probability

- Estimate both $p(d|T_d)$ and $p(M|T_D)$ with $P(D|T_D)$

- Boosting recent documents
  - Exponential distribution
  - Parameter λ

$$p(D/T_D) = P(T_D) = \lambda e^{-\frac{(T_C - T_D)^2}{2}}$$

- Boosting documents related to specific time period
  - Normal distribution
  - Parameter σ

$$p(D/T_D) = P(T_D) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(T_C - T_D - \mu)^2}{2\sigma^2}}$$

Experimental Setup

- Training Data ➜ Empirically Set parameters λ and σ
  - Recency-Related Queries
    - 25 queries from TREC queries 301-350 over TREC volumes 4 and 5
  - Time Period Based Queries

    - 5 queries from TREC queries 151-200 over TREC volumes 1, 2 and 4

- Test Data
  - Recency-Related Queries
    - 25 queries from TREC queries 351-400 over TREC volumes 4 and 5
  - Time Period Based Queries
    - 5 queries from TREC queries 151-200 over TREC volumes 1, 2, 4 and 5
    - 5 queries from TREC queries 251-300 over TREC volumes 2 and 4

Results

	LM	TB1-.01	% Chg		RM	TB2-.01	% Chg
Rel	2804	2804			2804	2804	
Rret	1043	1088	+4.3		1529	1552	+1.5
0.00	0.588	0.645	+9.8	0.00	0.582	0.595	+2.1
0.10	0.286	0.312	+9.2	0.10	0.443	0.460	+4.0
0.20	0.244	0.265	+8.7	0.20	0.397	0.412	+4.5
0.30	0.206	0.226	+9.6	0.30	0.321	0.340	+6.0
0.40	0.140	0.159	+12.9	0.40	0.250	0.270	+8.1
0.50	0.107	0.115	+7.5	0.50	0.193	0.213	+10.5
0.60	0.058	0.057	+1.3	0.60	0.143	0.154	+7.5
0.70	0.037	0.042	+12.1	0.70	0.110	0.116	+4.7
0.80	0.025	0.026	+4.5	0.80	0.084	0.083	-0.4
0.90	0.018	0.020	+12.6	0.90	0.045	0.051	+12.7
1.00	0.010	0.010	-3.9	1.00	0.009	0.011	+23.9
Avg	0.134	0.142	+6.2		0.220	0.235	+6.9

	LM	TB1-20	% Chg		RM	TB2-15	% Chg
Rel	1567	1567			1567	1567	
Rret	626	669	+6.9		783	779	-0.5
0.00	0.525	0.560	+6.7	0.00	0.633	0.8	+26.4
0.10	0.468	0.510	+9.0	0.10	0.579	0.752	+29.9
0.20	0.449	0.495	+10.2	0.20	0.575	0.686	+19.3
0.30	0.334	0.472	+41.3	0.30	0.562	0.600	+6.8
0.40	0.288	0.433	+50.3	0.40	0.526	0.569	+8.2
0.50	0.126	0.338	+168.3	0.50	0.474	0.502	+5.9
0.60	0.036	0.216	+500	0.60	0.412	0.430	+4.4
0.70	0	0.035		0.70	0.320	0.330	+3.1
0.80	0	0		0.80	0.221	0.137	-38.0
0.90	0	0		0.90	0.114	0.097	-14.9
1.00	0	0		1.00	0	0	
Avg	0.241	0.270	+12.0		0.395	0.448	+13.4

Results

	LM	TB1-10	% Chg		RM	TB2-7.5	% Chg
Rel	1003	1003			1003	1003	
Rret	122	124	+1.6		294	409	+39.1
0.00	0.540	0.587	+8.7	0.00	0.585	0.572	-2.3
0.10	0.045	0.084	+86.7	0.10	0.231	0.262	+13.4
0.20	0.018	0.023	+27.8	0.20	0.182	0.196	+7.7
0.30	0.010	0.019	+90.0	0.30	0.121	0.157	+29.8
0.40	0.010	0.014	+40.0	0.40	0.054	0.122	+125.9
0.50	0	0.014		0.50	0.028	0.114	+307.1
0.60	0	0		0.60	0.024	0.027	+12.5
0.70	0	0		0.70	0.013	0.015	+15.4
0.80	0	0		0.80	0	0	
0.90	0	0		0.90	0	0	
1.00	0	0		1.00	0	0	
Avg	0.020	0.032	+60.0		0.086	0.113	+31.4