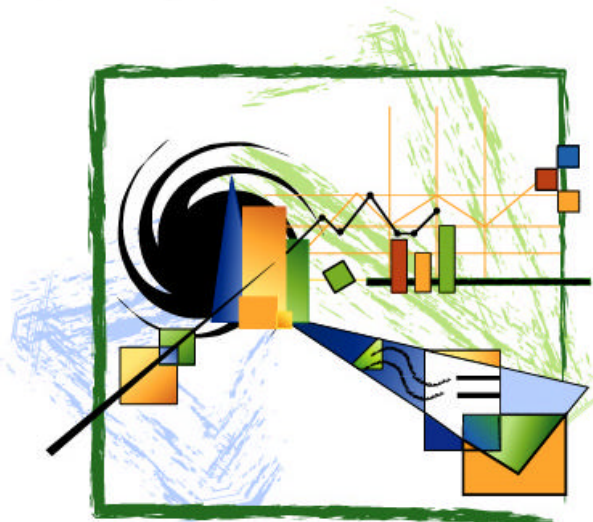


TREC 9 and Commercial Search Engines

SearchEngine 
M E E T I N G

Christopher Cardinal
Vice President,
R&D, Portal/KM



Connecting to Content in Context



Agenda today...

- What is TREC?
- Technical Approach to TREC
- Technical Findings in TREC
- Business Costs
- Benefits
- Why so few commercial vendors?

Text Retrieval Conference (TREC)

- Sponsored by NIST and DARPA
- A recall/precision performance proving ground for information retrieval engines
- Academic, governmental and (more rarely) commercial participants
- Divided into various “tracks” or problem domains
- Each track provides a data set, queries and a relevant result set



Technical Approach to TREC

- Two main tasks, Main Web Task and Large Web Task
- Used new SearchServer 5.0 search kernel with:
 - TREC specific data filter
 - Simple full-text only data schema + title column
 - Minimal stopwords list



Technical Approach to TREC

- Main Web Task:
 - 50 queries against 10 gig database
 - Automatic Title-only and full topic runs
 - All queries were interpreted as Intuitive Searches (a form of document similarity vector-space algorithm)
 - Method incorporates concepts from the Okapi approach to term dampening
 - Various additional search options were tested for effect on precision/recall curves

TREC Technical Findings

- Main Web Task. Title-only results:

SearchServer run	AvgP	P@5	P@1	P@2	Rec0	Rec3	Avg H	H@5	H0
2a: V2:3 ¹	0.1949	32.0 %	26.2 %	22.0 %	0.522 3	0.269 6	0.194 8	15.7 %	0.326 4
2b: V2:3 + exp	0.197 0	32.4%	25.4 %	21.5 %	0.480 2	0.2808	0.170 3	13.5 %	0.273 2
2c: V2:4 1	0.193 1	29.6 %	25.0 %	21.5 %	0.517 0	0.267 4	0.207 8	15.7 %	0.344 1
2d: V2:4 + exp	0.190 9	29.6 %	23.8 %	21.1 %	0.475 6	0.277 4	0.177 8	14.8 %	0.261 8
Median (18 grps)	0.146 4	21.6 %	21.2 %	17.4 %	0.401 5	0.199 3	n/a	n/a	n/a

TREC Technical Findings

- Tables notes:
 - Vector space algorithm V2:3 was slightly better at finding relevant documents
 - Vector space algorithm V2:4 (squared IDF) was better at finding *highly* relevant documents

TREC Technical Findings

- Main Web Task. Title-only results:
 - Using approximate text searching and linguistic expansion

SearchServer Run	AvgP	P@5	P@10	P@20	Rec0	Rec3	Avg H	H@5	H0
3a: ling only	0.191 9	30.4 %	25.6 %	21.5 %	0.526 5	0.268 6	0.234 3	15.7 %	0.354 8
3b: apx, ling	0.201 9	32.0 %	27.2 %	22.9 %	0.551 6	0.276 9	0.250 9	16.5 %	0.364 7
3c: apx only	0.191 4	32.4 %	28.0 %	22.8 %	0.558 6	0.269 3	0.227 3	17.0 %	0.389 8
3d: neither	0.180 5	30.4 %	26.4 %	21.6 %	0.525 8	0.256 2	0.208 9	15.7 %	0.353 5

TREC Technical Findings

- Tables notes:
 - Spell correction (approximate string matching) had a small effect (1-2 points) on effectiveness
 - “tartin” -> “tartan”
 - “1920’s” -> “1920”
 - Linguistic expansion increased average precision but hurt precision @10

TREC Technical Findings

- Main Web Task. Title-only results:
 - Using various document length values

DLen Importance	AvgP	P@5	P@1 0	P@2 0	Rec0	Rec3 0	Avg H	H@5	H0
4a: 0	0.159 5	26.0 %	21.2 %	17.7 %	0.439 3	0.222 2	0.152 1	10.9 %	0.255 4
4b: 250	0.190 8	29.6 %	24.2 %	21.2 %	0.496 0	0.267 7	0.192 6	14.8 %	0.308 4
4c: 500	0.205 0	31.2 %	26.0 %	21.8 %	0.541 0	0.282 8	0.228 2	16.1 %	0.342 7
4d: 750	0.199 2	30.0 %	24.6 %	21.7 %	0.553 9	0.278 8	0.252 8	16.1 %	0.387 8
4e: 1000	0.174 4	27.6 %	20.8 %	18.7 %	0.489 2	0.234 1	0.235 0	16.1 %	0.331 8



TREC Technical Findings

- Tables notes:
 - Approximate string matching and linguistic expansion both turned on
 - Ignoring document length hurt precision, especially amongst highly relevant documents
 - Higher values of document length seem to help identify the highly relevant documents

TREC Technical Findings

- Main Web Task.Full Topic results:
 - Using row expansion, different relevance methods, including Narrative, multiple tables

SearchServer run	AvgP	P@5	P@10	P@20	Rec0	Rec30	AvgH	H@5	H0
5a: T+D, V2:3	0.237 4	39.2 %	30.8 %	25.3 %	0.620 2	0.333 6	0.239 7	17.0 %	0.393 0
5b: 5a + exp	0.221 7	37.2 %	29.4 %	24.0 %	0.521 7	0.308 2	0.178 3	14.8 %	0.272 2
5c: SS4, V2:3:15	0.211 5	37.6 %	30.8 %	24.8 %	0.599 0	0.305 1	0.205 3	15.2 %	0.362 8
5d: 5a + Narr	0.218 4	39.2 %	34.0 %	26.3 %	0.605 9	0.320 1	0.200 5	15.2 %	0.331 3
5e: T+D, V2:4	0.234 7	36.0 %	30.6 %	25.2 %	0.561 7	0.319 0	0.237 2	17.8 %	0.363 5
5f: 5e + exp	0.222 8	34.0 %	28.6 %	24.9 %	0.489 5	0.311 4	0.186 2	14.3 %	0.267 5
5g: SS4, V2:4:15	0.238 0	36.0 %	30.8 %	25.0 %	0.573 5	0.319 7	0.230 1	17.4 %	0.365 9
5h: 5e + Narr	0.233 5	42.0 %	35.2 %	27.4 %	0.639 1	0.334 1	0.215 8	16.5 %	0.379 0



TREC Technical Findings

- Tables notes:
 - Vector space, critical terms squared algorithm seems slightly better
 - Including Narrative data hurt average precision

TREC Technical Findings

- Main Web Task.Full Topic results:
 - Using document length normalization

DLen Importance	AvgP	P@5	P@10	P@20	Rec0	Rec30	Avg H	H@5	H0
6a: 0	0.113 6	20.8 %	18.4 %	15.9 %	0.368 4	0.167 2	0.107 7	7.8%	0.184 6
6b: 250	0.223 3	40.0 %	34.8 %	28.0 %	0.632 0	0.321 9	0.200 7	16.5 %	0.362 8
6c: 500	0.243 5	44.8 %	35.2 %	29.5 %	0.683 3	0.333 0	0.244 7	19.1 %	0.426 4
6d: 750	0.256 9	43.2 %	36.8 %	30.1 %	0.695 8	0.338 4	0.284 3	20.9 %	0.484 5
6e: 1000	0.245 4	42.0 %	36.6 %	28.0 %	0.689 4	0.338 8	0.290 8	20.9 %	0.482 5



TREC Technical Findings

- Tables notes:
 - Again, document length normalization had a very significant impact on precision (about 100% better in rows with it turned on)
 - Impact even larger with highly relevant documents (and higher DOC_LEN values)



TREC Technical Findings

- Large Web Task
 - 100 gig data divided into 12 tables
 - Only 84 of the track's 10000 queries were judged
 - Only the top 10 documents for each query were judged

TREC Technical Findings

- Large Web Task

SearchServer Run	Reciprocal Rank of First Satisfactory	<u>Precision @1</u>	Precision @5	<u>Precision@10</u>
hum9w1	0.4381	30.95%	32.62%	32.50%
hum9w2	0.4262	30.95%	31.43%	31.67%
hum9w3	0.4174	28.57%	30.24%	29.40%



TREC Technical Findings

- Tables notes:
 - Data collection was divided into 12 tables for searching purposes
 - Approximate text searching made a small impression (hum9w1)
 - disabling document length normalization and linguistics lowered precision by 2-3 points (hum9w3)

TREC Technical Findings

- Query Track

Run	AvgP	Experiment (i.e. what was different from baseline)
humB*	0.1732	baseline
humK*	0.1713	keyword fields were not indexed (/k option of cTREC text reader was not used, see section 3.2)
humD*	0.1771 ¹	document length importance was set low (RELEVANCE_DLEN_IMP was set to 200 (baseline was 750))
humV*	0.1648	inverse document frequency not squared (RELEVANCE_METHOD was 'V2:3:15' (baseline was 'V2:4:15'))
humA*	0.1741	approximate text searching added fixes for spelling errors (algorithm of section 4.2 except the table used to index TREC Disk 1 with keywords was used)
hum4*	0.1713	SearchServer 4.0 was used (baseline used experimental SS 5.0 which contained a new linguistic expansion package which was known to still have a few glitches)
humI*	0.1736	terms in more than 15% of rows not discarded (RELEVANCE_METHOD was 'V2:4:100' (baseline was 'V2:4:15'))



TREC Technical Findings

- Tables notes:
 - Searching keyword fields appears to have little impact
 - Again, vector space, squared IDF value seems to be slightly better
 - Approximate string matching produces a marginal improvement (few misspellings in queries themselves)



Business Cost

- The cost of TREC is essentially manpower
 - 2 people working ~ fulltime for 3 months assuming search engine can be submitted as is
 - Define/prepare test environment
 - Develop test scripts
 - Define analysis tools and mechanics
 - Execute scripts and gather results
 - Analyze and prepare TREC submission paper(s)
 - Hummingbird Personnel Cost: ~60K USD

Business Cost

- Hardware costs are very small
 - Workstation and disk space
 - Software for development and execution of scripts
 - Hummingbird total: ~15K USD
- TREC attendance is another ~5K USD.
- Therefore:

Personnel:	60K
Hardware/Software	15K
Attendance	5K
Total	80K + Opportunity Cost



Benefits

- Validation of concepts and algorithms
- Insight into own product's strength and weaknesses
 - Mechanism to continue improving
 - Configuration parameters
- Insight into state-of-the-art technologies in various difficult problem domains
- Excellent access to data/query/result set information for performance analysis
- A good way of attracting and retaining scientifically-oriented staff

Why so few commercial vendors?

- Opportunity cost, opportunity cost
- Weaknesses in the engine
- Inappropriateness of TREC tracks and mechanics for engine in question
- No Marketing Bounce...
- Expense (especially if a lot of customization is required for TREC suites)
- Customers don't know to ask for it...
- Lack of interest

References

- TREC-9 paper on SearchServer 5.0 at http://trec.nist.gov/pubs/trec9/t9_proceedings.html
- In the TREC-9 Main Web Task Title-only category results <http://trec.nist.gov/presentations/TREC9/overview/sld019.htm>
- Full Topic version category results <http://trec.nist.gov/presentations/TREC9/overview/sld020.htm>
- More general slide sets: <http://trec.nist.gov/presentations/TREC9/overview/sld034.htm>

Questions?

