



# **Adaptive Filtering with Entities and Expansion**

David Eichmann  
Padmini Srinivasan

The University of Iowa  
School of Library and Information Science



# Our General Filtering Approach

---

- ❑ Given a stream of documents, create a set of clusters based upon document-cluster similarity
  - Similarity is based upon a straight-forward cosine-vector similarity measure
    - Document vectors are pruned to the 100 most weighty terms
    - Cluster vectors are pruned to the 200 most weighty terms
    - Terms are stemmed with Porter's algorithm
    - TF-IDF is used to weight individual terms
- ❑ New documents are merged into the cluster with the highest similarity, if above  $\alpha$ , the membership threshold



# Two Level Clustering

---

- ❑ Primary clusters correspond to topic definitions
  - Fixed at initialization of run
  - Full (static) text vector (complete topic vocabulary)
  - Positive / Negative example vocabulary Vectors
- ❑ Two options for primary adaptation
  - “Pure” Rocchio scheme with modifications relating to incremental nature of adaptive filtering
  - Differential Rocchio scheme where example vocabularies are distinct
- ❑ Primaries act as ‘gatekeepers’
  - The primary membership threshold allows for gross tuning of recall



# Two Level Clustering

---

- ❑ Secondary clusters form behind a specific primary
  - No fixed number of secondaries for a given primary
  - $\beta$ , the secondary membership threshold allows for gross tuning of cluster coherence, and hence precision
  - Additional declaration threshold allows for control of document declaration similarity level distinct from within-primary clustering
- ❑ Most learning occurs at this level
  - When a secondary exceeds  $\gamma$ , a similarity threshold with its primary, it declares its current document, updates the example vocabulary and is then colored appropriately
  - White clusters continue to declare new arrivals
  - Black clusters continue to grow but never declare again



# Monotonic Adaptation

---

- ❑  $\gamma$ , the secondary declaration threshold has proven to be very useful for adaptation to the specifics of a given topic
- ❑ We currently monitor topic scores and if the trend for a topic is negative, we raise  $\gamma$  by a moderate amount: 0.01 - 0.05, depending upon conditions
- ❑ This provides a gradual shutdown of seriously misbehaving topics while preserving the performance of topics only slightly mistuned
  - This was very successful in a retrospective study of the TREC-7 adaptive filtering task (positive F1 and T9U scores!)



# Lexical Architecture

---

- ❑ Primary lexical scanner is custom written for TREC-style document formats
  - Dictionary-driven phrase recognition a clickable option
    - WordNet
    - Moby database
    - local instance generated from bibliographic citation keywords
- ❑ Alternative lexers implement
  - Wu's Mandarin segmenter
  - Peterson's Mandarin segmenter
  - Brill's rule-driven POS tagger



# Lexical Architecture, con't.

---

- ❑ Lexical analysis is now supported as a cascade of filters
  - Initial token acquisition (including mapping encodings to Unicode)
  - Word segmentation (when necessary)
  - Language transformation (optional)
  - Part-of-speech tagging (optional)
  - Entity extraction (optional)



# Named Entity Recognition

---

- ❑ We have five categories currently being recognized
  - Persons
  - Organizations
  - Locations
  - Event (preliminary)
  - MeSH
- ❑ All categories are driven through examination of noun phrases recognized by the POS tagger (with special handling of certain glue words: ‘and,’ ‘of,’ ‘the,’ etc.)
- ❑ Named entity vectors are maintained separately from the term vector, weighted by their length and the frequency of the constituent terms



# Person Recognition Resources

---

- ❑ Various Web lists of cultural names
  - Anglo, Chinese, Arab, Hebrew, Hindi, Indian, Japanese, Latino, Muslim, Russian
  - World leaders
- ❑ This is enriched with a set of pattern expressions for other instances
  - “President” <proper name>
  - <proper name> “III”



# Organization & Event Recognition Resources

---

- ❑ International political organizations (from CIA Fact Book)
- ❑ Fortune 500 company list
- ❑ Global 500 company list
- ❑ This is enriched with a set of pattern expressions for other instances
  - <proper name> “Incorporated”
  - <proper name> “&” “Sons”



# Location Recognition Resources

---

- ❑ We mine the text of the CIA Fact Book for variants of country names, administrative divisions, capitals, harbors, etc.
- ❑ Various Web lists of
  - World cities
  - U.S. Cities
  - Rivers
  - Lakes
- ❑ This is enriched with a set of pattern expressions for other instances
  - <proper name> “Street”
  - “Mount” <proper name>



# MeSH Recognition Resources

---

- ❑ We first load and reconstruct the MeSH term tree
- ❑ We then load the concept descriptors, binding them into the tree and adding the synonyms
- ❑ Finally the supplements are added to support drugs and compounds



# Newsire Entity Recognition Sample #1

---

- ❑ APW19981001.0262 [Israel(0.271), Jonathan Pollard(0.153), Benjamin Netanyahu(0.102), Bill Clinton(0.102), United States(0.055), ...]
- ❑ Persons
  - Bill Clinton (3)
  - Jonathan Pollard (8)
  - Moshe Fogel (2)
  - Benjamin Netanyahu (2)
  - Esther (1)
  - Israeli Embassy (1)
- ❑ Organizations
  - Cabinet (1)
- ❑ Places
  - Israel (16)
  - United States (5)
  - Washington (2)



# Newsire Entity Recognition Sample #2

---

- ❑ APW19981001.0303 [Vladimir Meciar(0.119), Slovak Democratic Coalition(0.065), Slovakia(0.043), United States and Germany(0.043), ...]
- ❑ Persons
  - Vladimir Meciar (8)
  - Jozef Moravcik (2)
  - God (1)
  - Kalman Petocz (2)
- ❑ Organizations
  - Slovak Democratic Coalition (2)
  - Organization (1)
  - United States and Germany (1)
  - NATO (1)
  - European Union (1)
  - Hungarian Coalition Party (1)
- ❑ Places
  - Slovakia (4)
  - Europe (1)



# Enhancements for Reuters Topics

---

- ❑ Given the extremely brief and generic nature of Reuters topics, we decided to extend topic initialization with three distinct schemes:
  - Synonym expansion via WordNet
  - Reuters topic mapping into IPCT topics
  - Descendent agglomeration in term hierarchies
- ❑ Each scheme is independently switched



# Example Mappings - Topic 2

---

- ❑ Reuters C21: Legal / Judicial
  - IPCT 04006007: Legal services
  - IPCT 02002000: Judiciary (system of justice)
  
- ❑ WordNet expansions (with hyponyms)
  - ADP system (1)
  - articulatory system (1)
  - ...
  - chief justice (1)
  - classification system (1)
  - ...
  - ethical code (1)
  - judicial system (2)
  - justice of the peace (1)
  - language system (1)
  - legal / judicial (1)
  - legal system (1)
  - living arrangement (1)
  - ...



# Example Mappings - Topic 27

---

- ❑ Reuters C42: Labour
  - IPCT 09000000: Labour :: Social aspects, organisations, rules and conditions affecting the employment of human effort for the generation of wealth or provision of services and the economic support of the unemployed.
    - 09001000: Apprentices
    - 09002000: Collective contracts
    - 09003000: Employment
    - 09004000: Labour dispute
    - 09005000: Labour legislation
    - 09006000: Retirement
    - 09007000: Retraining
    - 09008000: Strike
    - 09009000: Unemployment
    - 09010000: Unions
    - 09011000: Wages & Pensions
    - 09012000: Work Relations
    - 09013000: Health & Safety at Work
    - 09014000: Advanced Training
    - 09015000: Employers
    - 09016000: Employees



# Example Mappings - Topic 27

---

- ❑ WordNet expansions
  - Pullman porter (1)
  - acquisition agreement (1)
  - adhesion contract (1)
  - aleatory contract (1)
  - antitrust law (1)
  - antitrust legislation (1)
  - articles of agreement (1)
  - assortative mating (1)
  - base hit (1)
  - bilateral contract (1)
  - body of work (2)
  - brokerage house (2)
  - coaching job (3)
  - collective agreement (1)
  - collective farm (1)
  - common-law marriage (1)
  - company man (1)
  - ...



# Training - Topic 2 (Legal/Judicial)

---

## ☐ Persons

- Lloyd (0.667)
- Robert Payne (0.185)
- Harvey Pitt (0.074)
- David Rowland (0.037)
- Ronald Sandler (0.037)

## ☐ Organizations

- Court of Appeals (0.143)
- PXRE (0.143)
- Securities and Exchange Commission (0.143)
- U.S. 4th Circuit Court (0.143)
- U.S. District Court (0.143)
- U.S. Names (0.143)
- Virginia District Court (0.143)



# Training - Topic 2 (Legal/Judicial)

---

- ❑ Noun Phrases
  - order (0.051)
  - appeal (0.038)
  - Equitas (0.032)
  - consent (0.026)
  - date (0.026)
  - decision (0.026)
  - harm (0.026)
  - lawsuit (0.026)
  - motion (0.026)
  - plaintiff (0.026)



# Training - Topic 27 (Labour)

---

- ❑ Persons
  - Avis (0.545)
  - Serbia (0.182)
  - Motors Corp (0.091)
  - Vukasin Filipovic (0.091)
  - Zoran Nedeljkovic (0.091)
- ❑ Organizations
  - HFS (1.000)



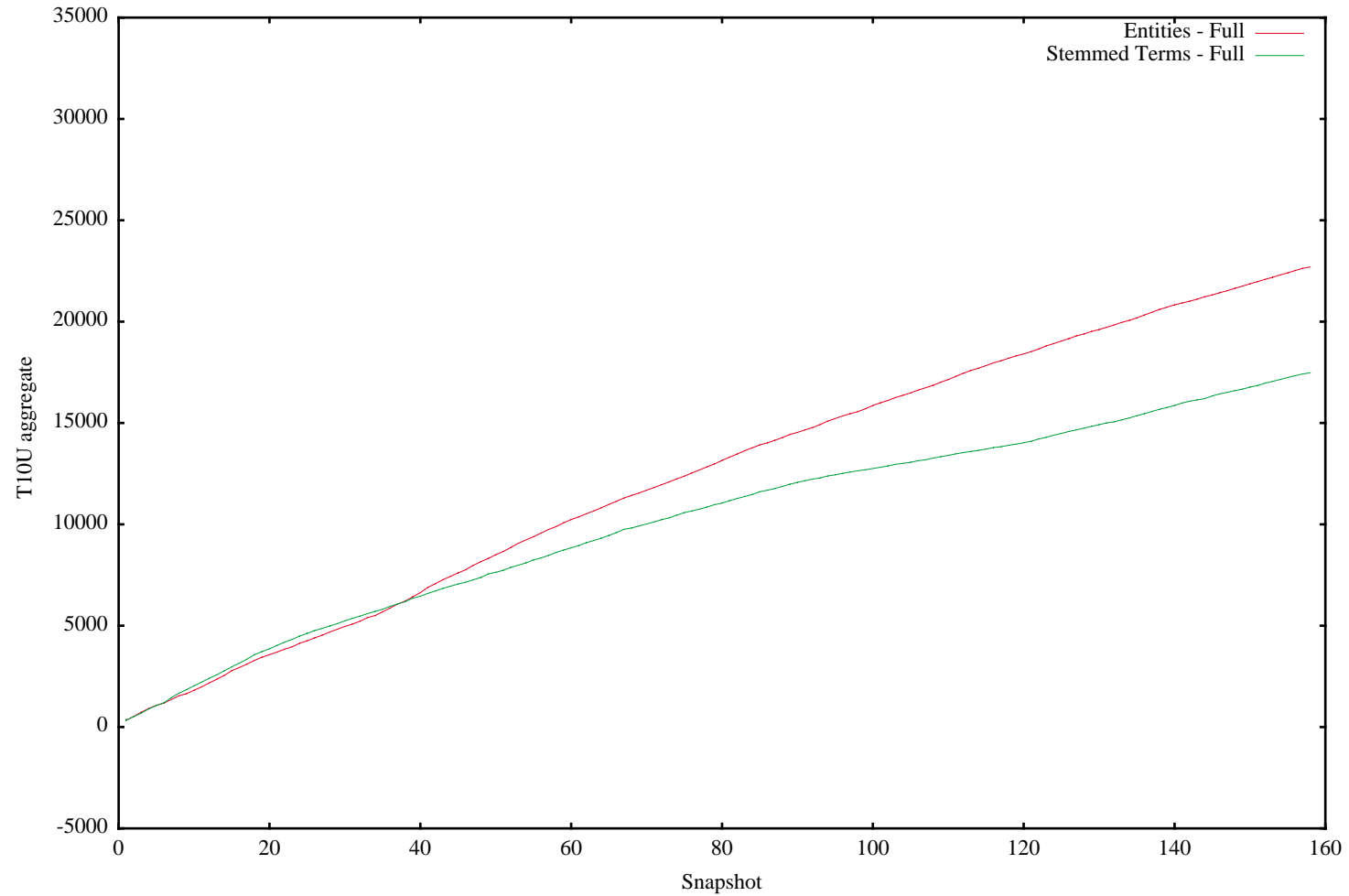
# Training - Topic 27 (Labour)

---

- ❑ Noun Phrases
  - decision (0.073)
  - labour (0.073)
  - town (0.073)
  - contract (0.054)
  - time (0.054)
  - Teamster (0.054)
  - Zastava (0.054)
  - wages (0.054)
  - union (0.036)
  - 3,500 production workers (0.027)



# Temporal Performance - Official Runs





# Official Runs

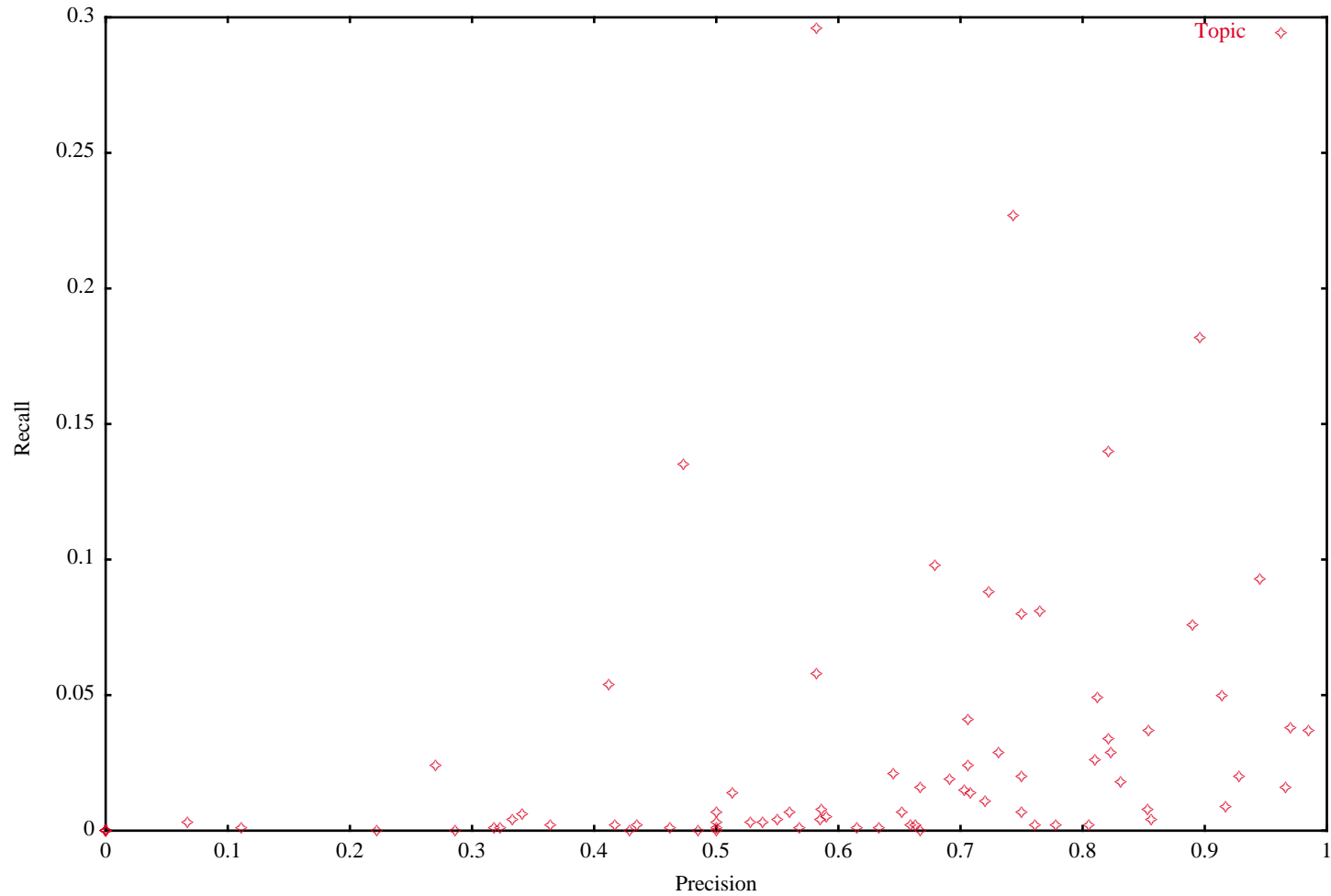
---

Run	Similarity	Expansion	T10SU Score <sup>a</sup>	Fbeta Score	Ave. Precision	Ave. Recall
UIowa01 AF01	Entities	Full	0.051	0.090	0.565	0.028
UIowa01 AF02	Stemmed Terms	Full	0.047	0.069	0.535	0.023

a. Optimized for this score

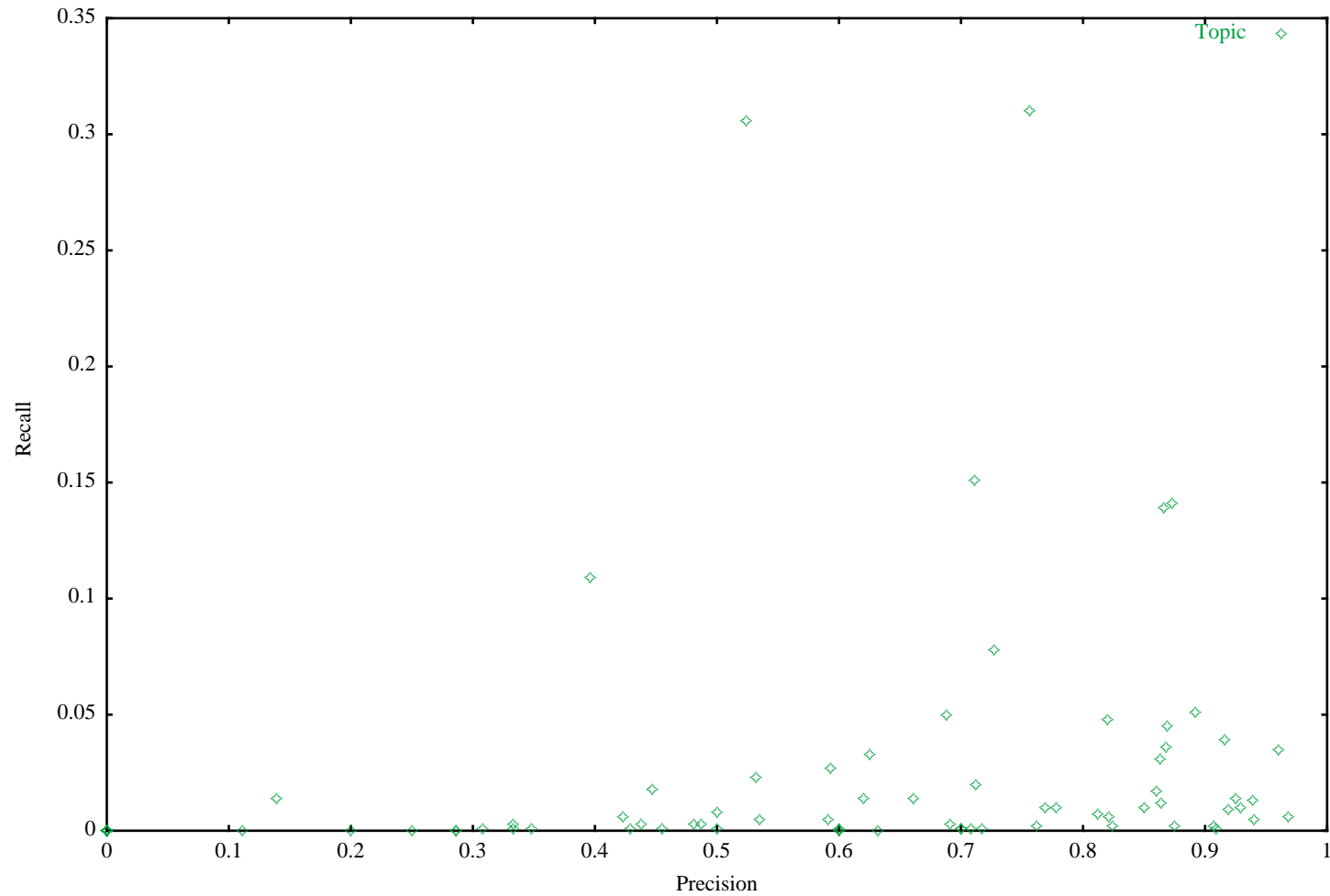


# Entity Precision & Recall





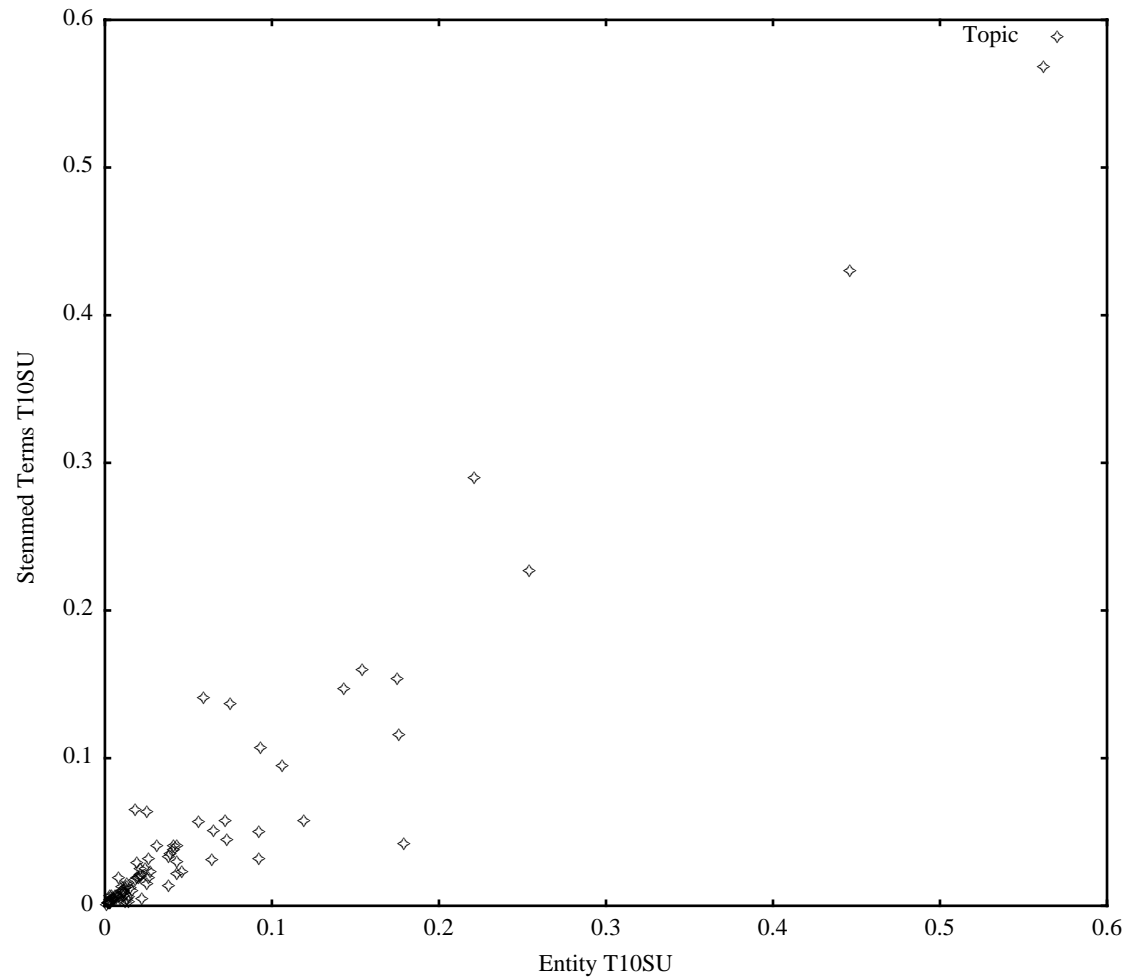
# Stemmed Terms Precision & Recall





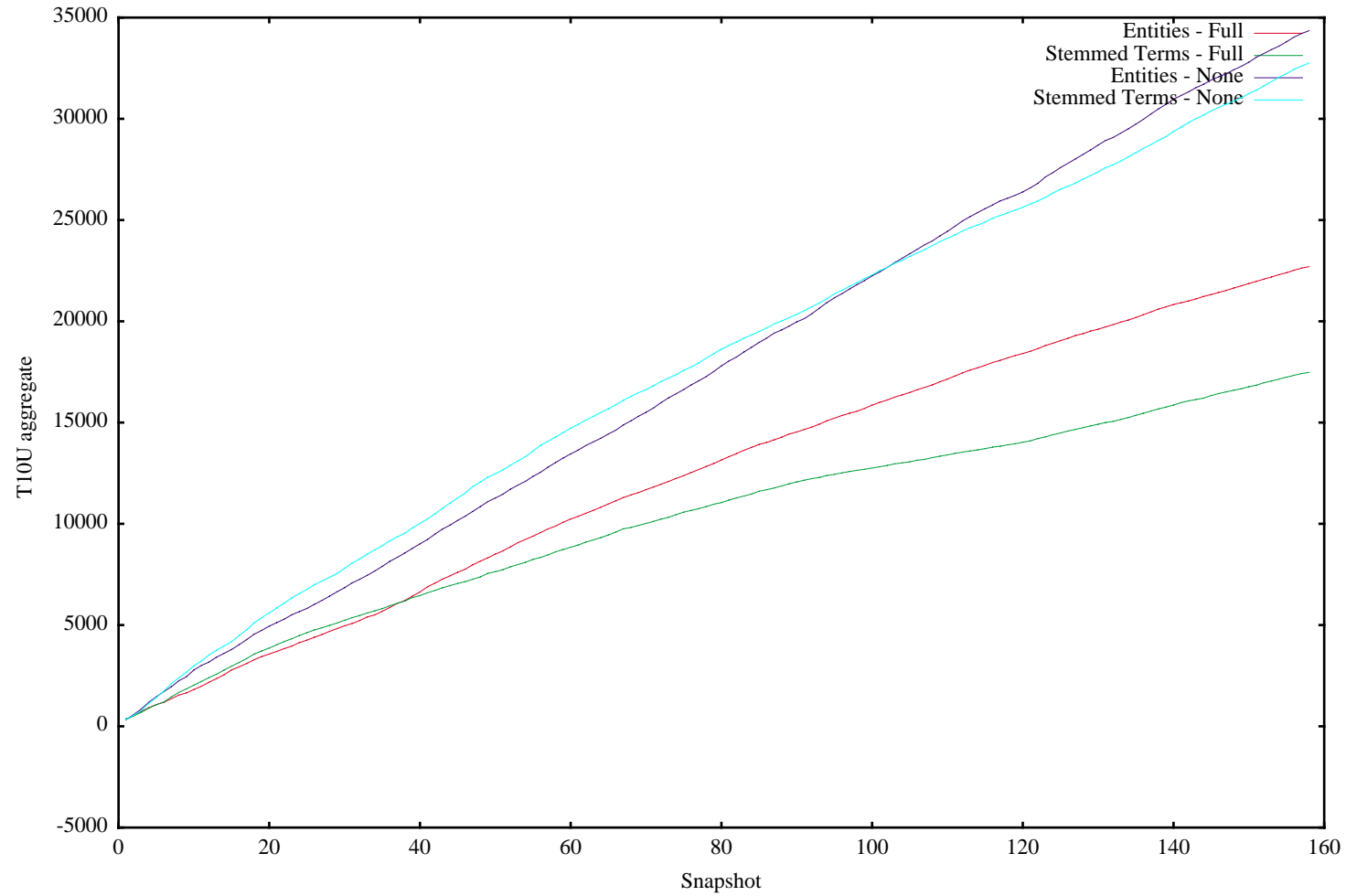
# Entity vs. Stemmed Terms, T10SU

---



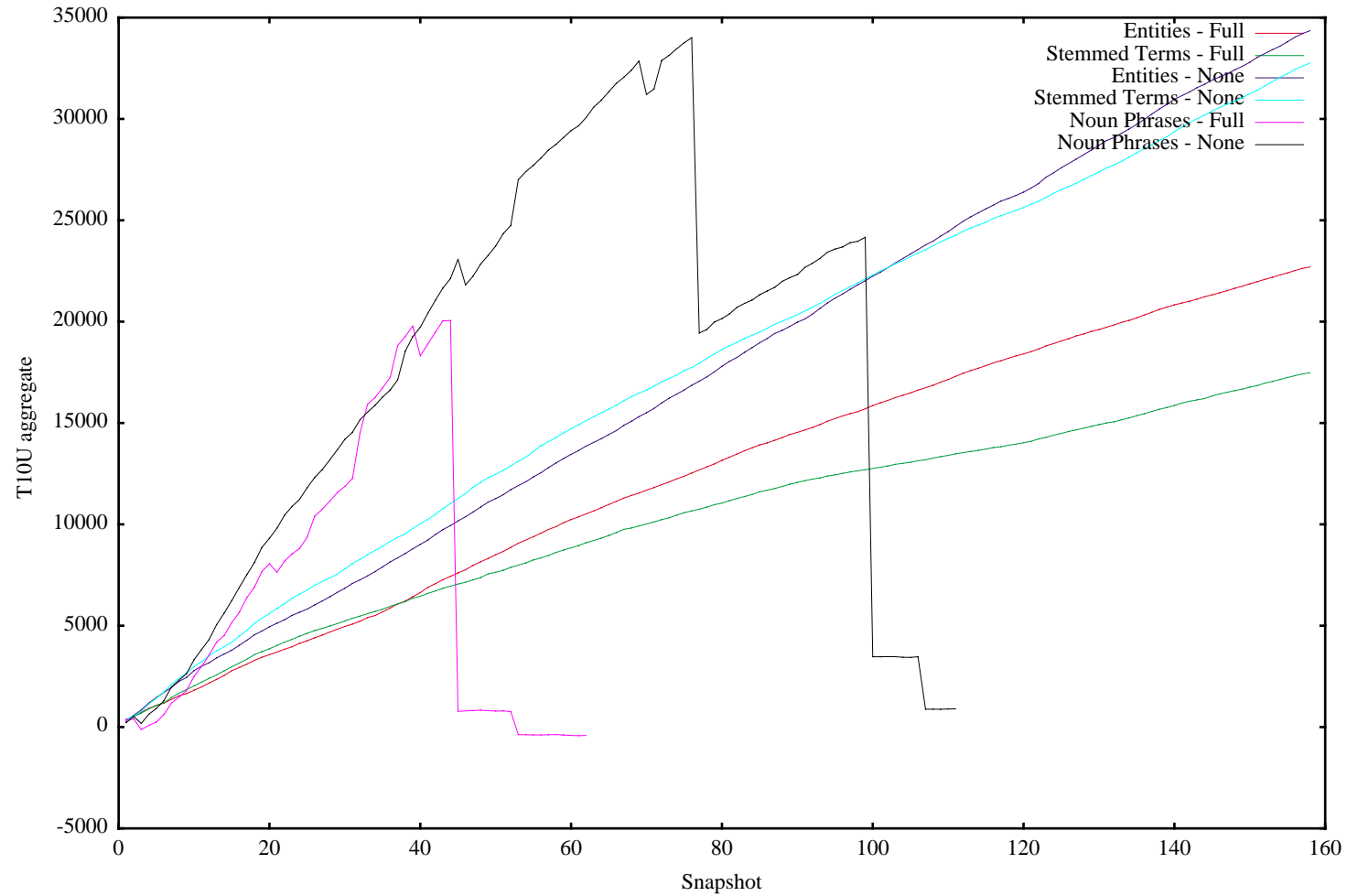


# Temporal Performance - Additional Runs





# Temporal Performance - Additional Runs





## (Some) Answers in Thresholds

---

Similarity Scheme	$\alpha$	$\beta$	$\gamma$
Stemmed Terms	0.10	0.50	0.20
Entities	0.15	0.40	0.25
Noun Phrases	0.15	0.40	0.15



# Conclusions

---

- ❑ Our entity scheme has matured to the point where, even when mistuned, it can out-perform stemmed terms
- ❑ Threshold adaptation works - sort of...
  - Primarily a blindspot relating to only allowing for monotonic increases in threshold
- ❑ For this corpus / topic set, noun phrases rock
  - Note that the topics are basically noun phrases themselves
  - This is not true for our results on our TREC-7 retrospective runs
- ❑ As far as our query expansion scheme goes...
  - CPU cycles will continue to be burned exploring this for the proceedings