SCALING UP TEXT CLASSIFICATION FOR LARGE FILE SYSTEMS

George Forman Shyamsundar Rajaram
KDD '08

Presented by: Yi Zhang
for CPS 296.3 @DUKE

Problem

- Search for files from large file systems using a trained document classifier
  - Class label is binary
  - A labeled training set is given
  - # total files is LARGE: I/O !
  - Positive class is rare (<1%)
- Objective: fast and accurate
  - Cares about both precision (p) and recall (r)
  - The F-measure: 2pr/(p+r)

Solution Overview

- Apply a cheap filter first
  - Utilize a full-text index over all documents
  - Obtain a much smaller subset of files likely to be positive
- A traditional classifier follows
  - Workload reduces to the result set from first step
  - Less I/O, faster runtime
  - However, should be careful about accuracy, recall particularly

Phase 1

- Task: Query a full-text index to produce a set of likely positive docs
- What query terms?
  - Words vs. words+phrases
- How many terms (Q)?
  - Use Q best terms; goodness measured by BNS or IG
- What form of query?
  - Boolean: disjunction of terms
  - Weighted: each term associated with a weight
    - Weight chosen by using a linear SVM
- What objective?
  - Just F-measure vs. higher recall (let Phase 2 restore precision)

Phase 2

- Task: Fetch docs selected in Phase 1, extract their features, and do classification
- What features?
  - Words vs. words+phrases
- How many features (C)?
  - Defaults: C=16384, selected via BNS from all words and two-word phrases
- What training set?
  - Option 1: Training docs that Phase 1 finds positive. ⇨ Too few negatives
  - Option 2: The full training set

Experiments Setup

- Dataset: Reuters RCV1
  - 806,791 news articles in XML format
  - Pre-labeled
  - Tags revealing true class label removed from file
- Full-text index generated by Lucene
- 140 classes picked, each having >1000 docs but <=5% of overall docs
- Each training set has 500 positive + 5000 negative docs
  - Positive rate=9%, higher than actual prevalence
  - Adjusted by weights for Phase 1

Two-phase vs. Test-all

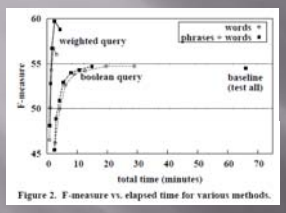
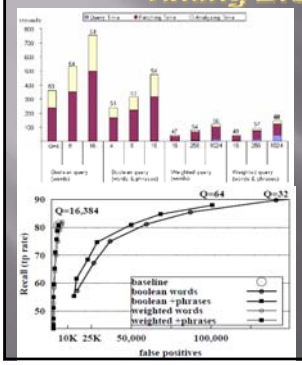


Figure 2. F-measure vs. elapsed time for various methods.

- Two-phase improves both speed and accuracy
- More terms & using weights give Phase 1 more control: less negatives, less time

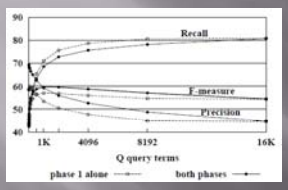
Baseline: only Phase 2, applied to every file
Q for Boolean: 1,2,4,8,16
Q for Weighted:16,64,256,1024,4096

Timing Breakdown



- Boolean:
  - Phase 1 has high recall but low precision, resulting in too much fetching & analyzing
- Weighted:
  - Low cost of adding terms
  - Can have lots of terms—query time increases
  - Much fewer false positives

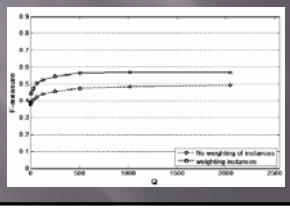
Single-Phase vs. Two-Phase



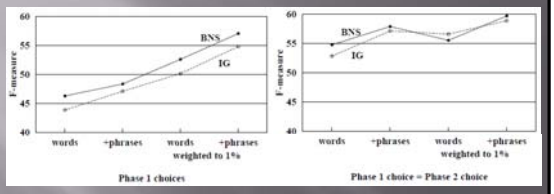
- Single-phase has better recall
- Two-phase has better precision
- Two-phase has better overall F-measure
  - Two-phase: increase in precision > decrease in recall
- As Q → C
  - Difference diminishes
  - No difference when Q=C

Weighting Training Set

- Recall in Phase 1:
  - % of positive samples in training set is higher than in test set
  - Positive samples weighted to 1%



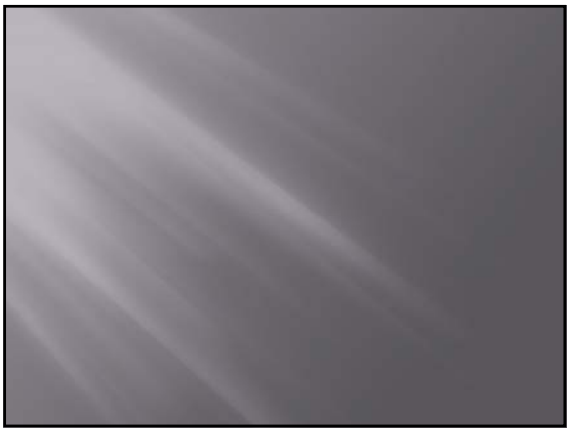
Design Choices



- Phase 1 only, Q=1K
- BNS better than IG
- Weighting helps
- Adding phrases helps
- Both phases, Q=1K, C=16K
- All better than Phase 1 only

Discussion

- Assumes index is available
  - Building takes time
- Requires relatively large training set
  - Active learning & user feedback might help
- Is two-phase strictly necessary
  - Could have one classifier that utilizes the index and does pruning on the fly
- SVM tends to use index well; what about other models?
- Principled way of choosing parameters?



SpotSigs: Discussion

- If you can prune by simply checking cardinality:
  - Either you have a really simple problem
  - Or you should choose a better similarity measure! (Consider a doc as a subset of another: Jaccard is bad)
- Could be worse than LSH
  - When similarity threshold is low
  - When doc lengths have very skewed distribution
- Really tuning-free?
  - Spot signature params: spot distance, chain length
- Can natural language phrases do better than Spot signatures?