Understanding Temporal Aspects in Document Classification

Fernando Mourão Leonardo Rocha Renata Araújo
Thierson Couto Marcos Gonçalves **Wagner Meira Jr.**

Universidade Federal de Minas Gerais
Brazil

February 12, 2008
WSDM'08 - California, USA

Introduction

Motivation

- Automatic Document Classification (ADC): associates documents with semantically meaningful categories.
  - ADC improves tasks such as automated topic tagging, identification of writing styles, creation of digital libraries, and Web searching.
  - ADC is also applicable to spam filtering, detection of adult content, and plagiarism.
- But the characteristics of the documents may change over time:
  - New information is created, new terms are introduced, new fields emerge and large fields are divided into more specialized subfields.
    - ★ Example: Pluto that was considered to be a planet until the middle of 2006 besides being the god of hell in Roman mythology.
- Most of the current techniques for ADC do not take into account the temporal evolution of the collection of documents.

Temporal Effects

Overview

- Temporal-related issues:
  - How does the temporal evolution affect the performance of classifiers?
  - What are the temporal-related characteristics that affect the classification's quality?
- We consider three factors that may affect the quality of classifiers:
  - Class Distribution;
  - Terms Distribution;
  - Class Similarity;

Temporal Effects

Class Distribution

- How does the class distribution vary across time?
  - Classes may appear and disappear, which may happen as a consequence of splits and joins, respectively, of existing classes.
- Example
  - Information Retrieval and Artificial Intelligence used to belong to the same ACM-DL class: Applications. In the new ACM-DL classification schema, they were assigned to two classes: Information Systems and Computing Methodologies.

Temporal Effects

Term Distribution

- How does the discriminative power of a term vary across time?
  - Terms appear, disappear, migrate among classes, or simply loose or gain discriminative power.
- Example
  - Pluto, which lost discriminative power in class astronomy and gained in class mythology.

Temporal Effects

Class Similarity

- How does the similarity among classes vary across time?
  - The more similar are the classes, the harder is the process of building a classifier.
- Example
  - In the past, classes crime and biology were quite disjoint but they became more related as DNA analysis started to be used as legal evidence.

Temporal Effects

Trade-off

- Temporal effects $\times$ sampling effect
  - In order to handle temporal effects we should sample our training set so that we minimize those effects.
  - However, the sampling effect may also affect the classifier precision.
- Challenge
  - Determine a training sample where classification concepts are quite stable and the sample is large enough for building a classifier.

Related Work

Overview

- Handling temporal aspects has been addressed just recently.
- We may divide the efforts into four broad groups
  - Adaptive document classification
    (Liu&Lu 2002, Lawrence&Giles 1998, Caldwell et.al. 2000)
  - Adaptive information filtering
    (Belkin&Croft 1992, Dumaisẽt.al. 1998, Klinkenberg&Renz 1998)
  - Concept drift
    (Forman 2006, Scholz&Klinkenberg 2006)
  - *Temporal adaptation*
    (Jones&Diaz 2007, Lanquillon&Renz 1999, Lewis 1995)

Related Work

Analysis

As far as we know, none of the previous efforts

- considered the three temporal factors simultaneously.
- were able to classify a document in its original temporal context.
- perfomed an in-depth investigation about the impact of temporal effects on classifiers' performance.

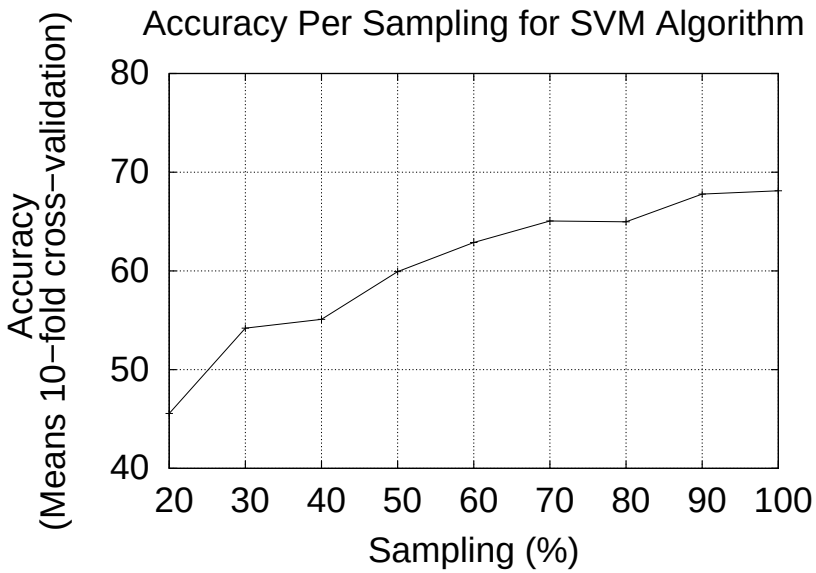
Characterizing the Temporal Effects

Experimental Setup

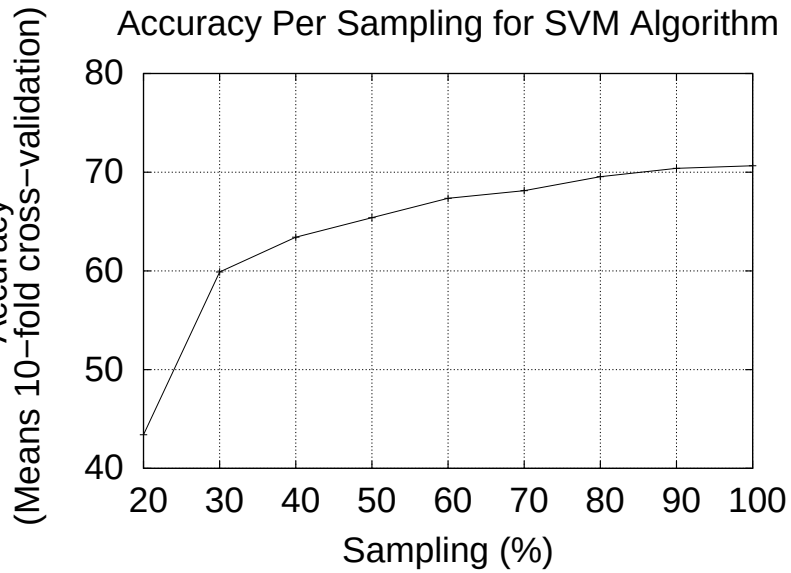
- Document collections:
  - ACM-DL: 30k documents from 1980 to 2002, 11 categories.
  - Medline: 4,112,069 documents from 1970 to 2006, 7 categories.
- Technique
  - Support Vector Machine
    (C-SVM using the Radial Basis Function Kernel)
- Evaluation Strategy:
  1. Assess both sampling and overall temporal effects.
  2. Characterize the three temporal effects.

Characterizing the Temporal Effects

Assessing the Sampling Effect



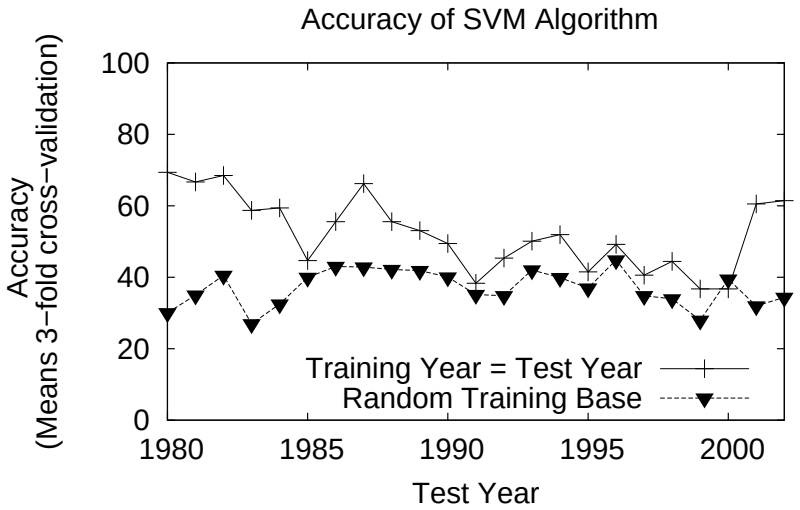
(a) ACM Collection



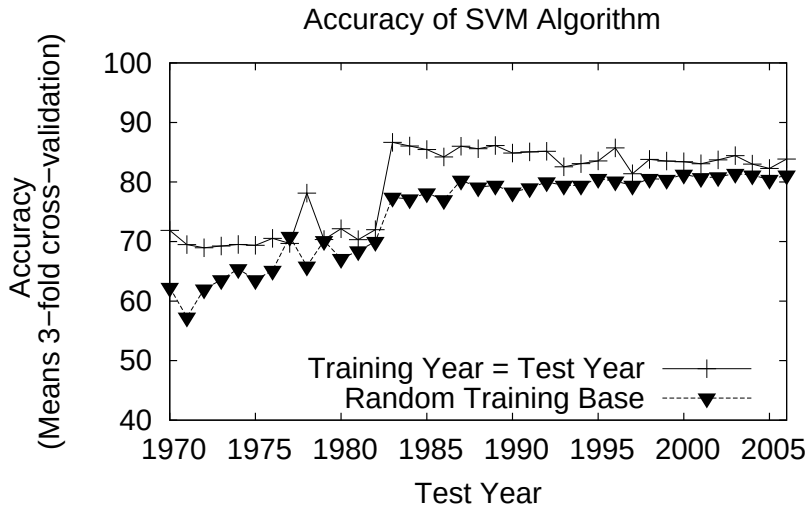
(b) Medline Collection

Characterizing the Temporal Effects

Assessing Overall Temporal Effects



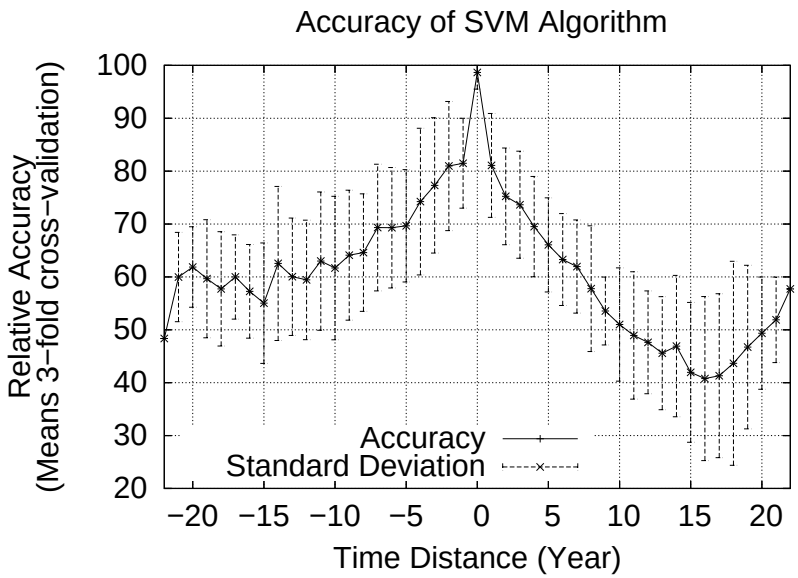
(c) ACM Collection



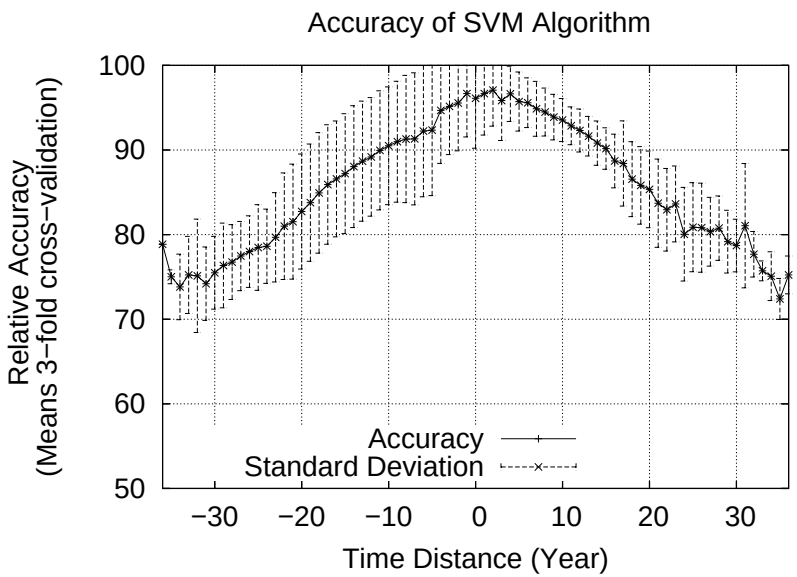
(d) MedLine Collection

Characterizing the Temporal Effects

Accuracy as a Function of *Time Distance*



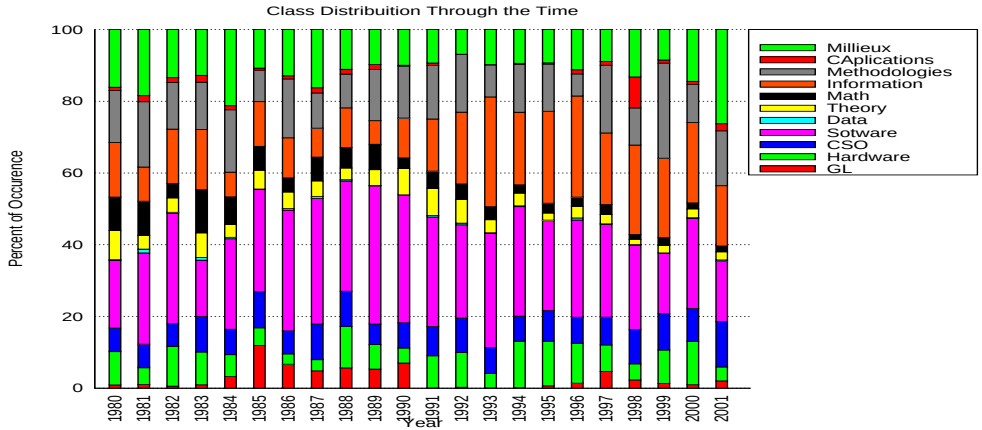
(e) ACM Collection



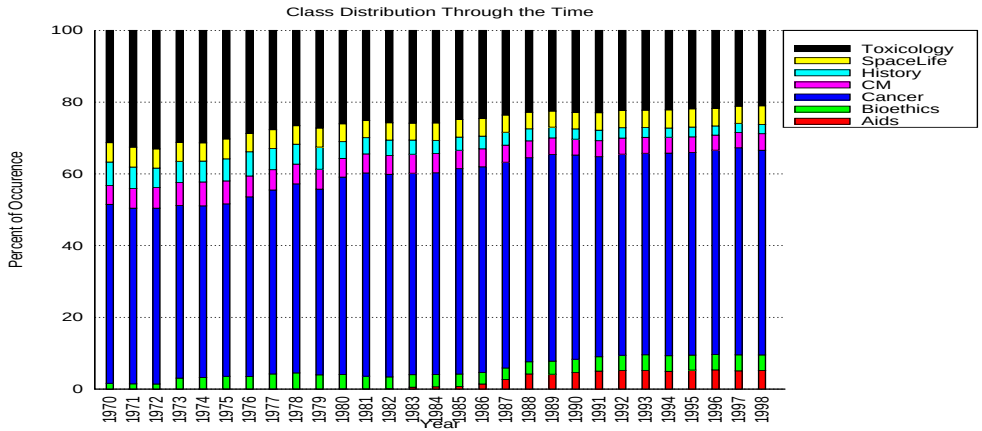
(f) MedLine Collection

Characterizing the Temporal Effects

Class Distribution



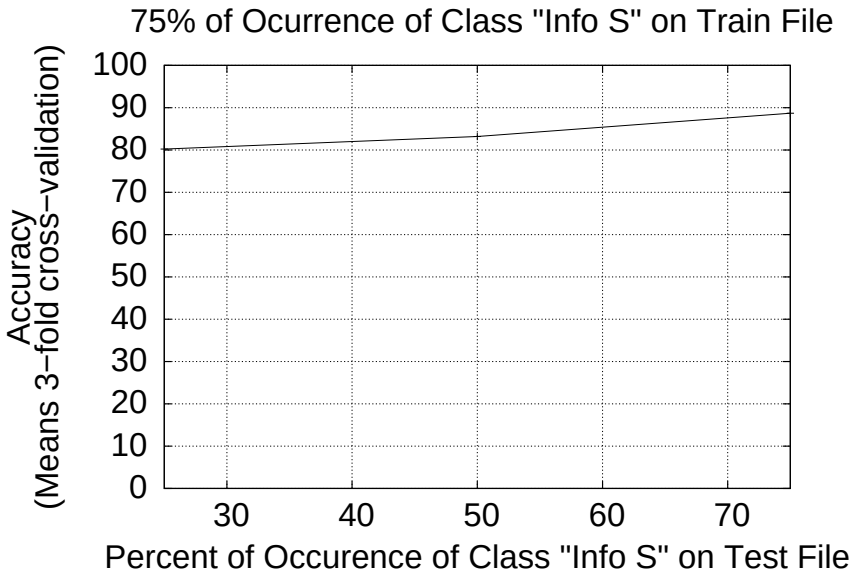
(g) ACM Collection



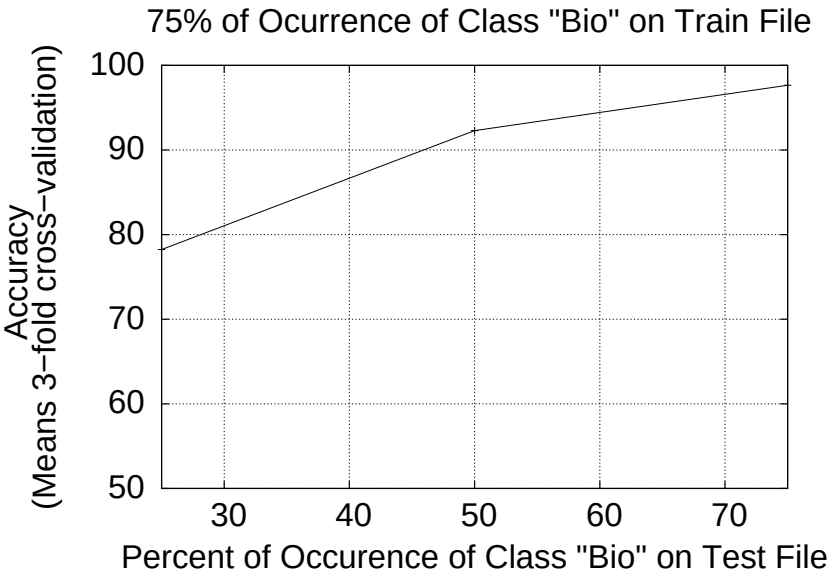
(h) MedLine Collection

Characterizing the Temporal Effects

Class Distribution



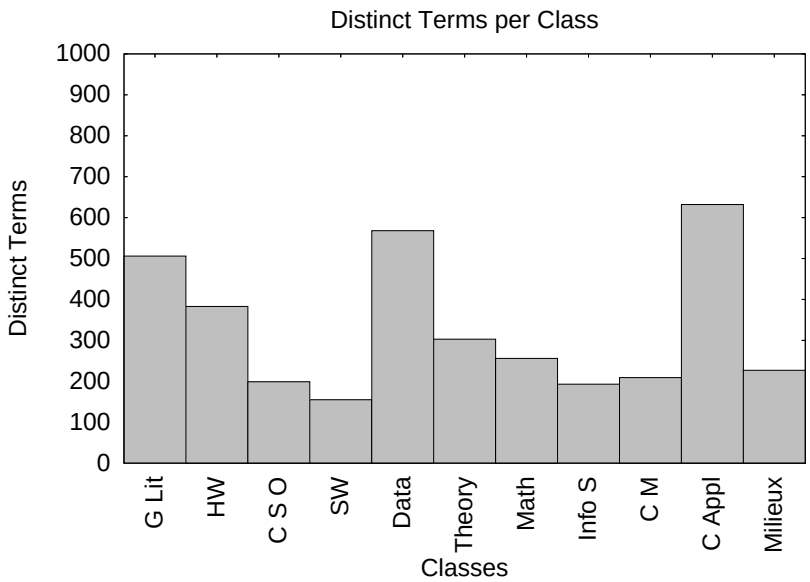
(i) ACM Collection



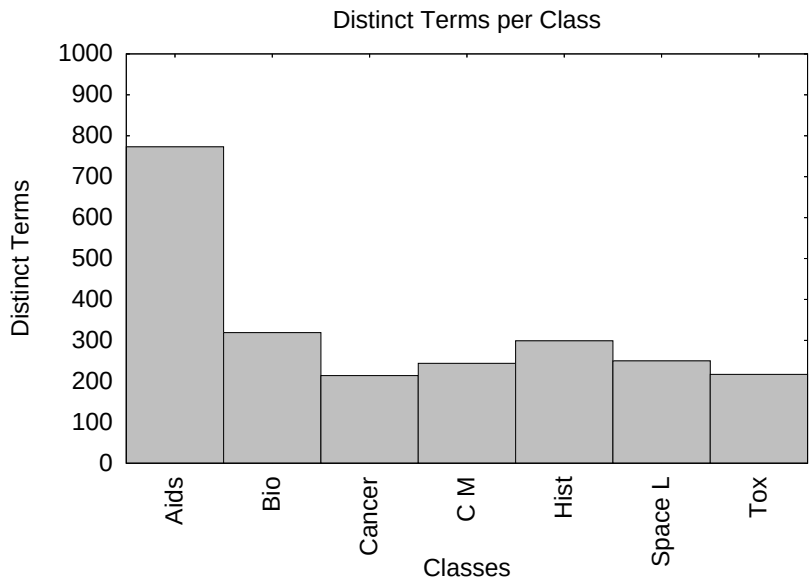
(j) MedLine Collection

Characterizing the Temporal Effects

Term Distribution



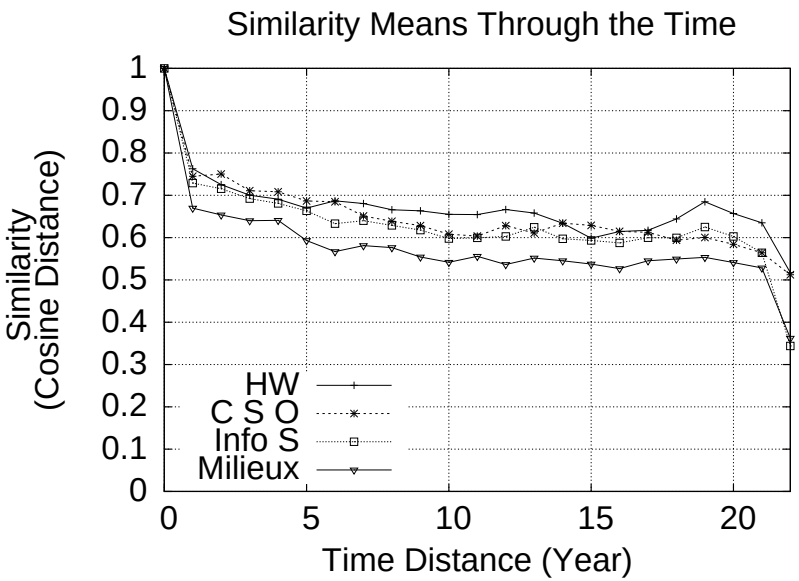
(k) ACM Collection



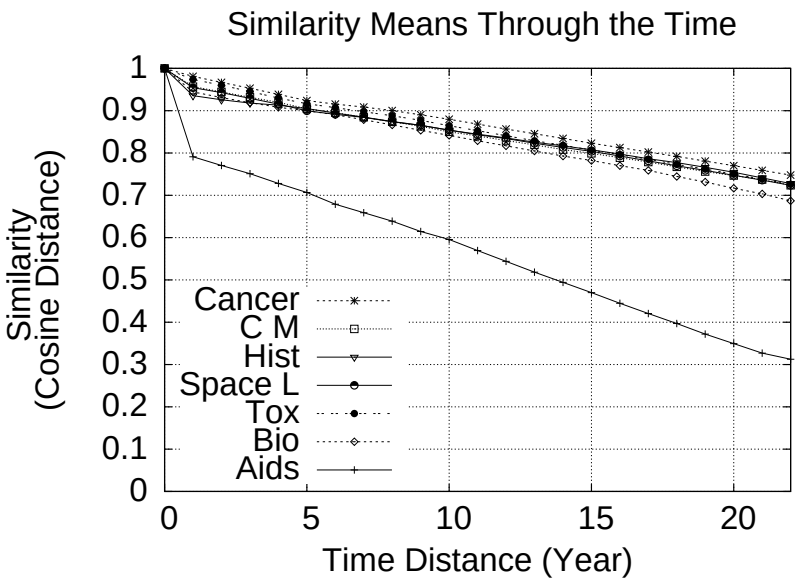
(l) MedLine Collection

Characterizing the Temporal Effects

Term Distribution



(m) ACM Collection



(n) MedLine Collection

Characterizing the Temporal Effects

Class Similarity

	Lit	HW	C S O	SW	Data	Theory	Math	Info S	C M	C App	Milieux
Lit	0	0.14	0.12	0.12	0.12	0.29	0.14	0.13	0.14	0.12	0.29
HW	-	0	0.08	0.13	0.11	0.12	0.11	0.12	0.10	0.10	0.13
C S O	-	-	0	0.10	0.09	0.10	0.08	0.07	0.08	0.10	0.13
SW	-	-	-	0	0.09	0.06	0.09	0.10	0.11	0.12	0.13
Data	-	-	-	-	0	0.05	0.08	0.09	0.10	0.13	0.13
Theory	-	-	-	-	-	0	0.14	0.13	0.07	0.06	0.29
Math	-	-	-	-	-	-	0	0.13	0.10	0.09	0.15
Info S	-	-	-	-	-	-	-	0	0.10	0.08	0.15
C M	-	-	-	-	-	-	-	-	0	0.11	0.13
C App	-	-	-	-	-	-	-	-	-	0	0.12
Milieux	-	-	-	-	-	-	-	-	-	-	0

Table: Similarity Std_Dev Matrix - ACM-DL

	Aids	Bio	Cancer	C M	Hist	Space L	Tox
Aids	0	0.19	0.16	0.18	0.19	0.18	0.19
Bio	-	0	0.04	0.20	0.17	0.19	0.12
Cancer	-	-	0	0.04	0.03	0.04	0.05
C M	-	-	-	0	0.21	0.08	0.05
Hist	-	-	-	-	0	0.20	0.11
SpaceL	-	-	-	-	-	0	0.05
Tox	-	-	-	-	-	-	0

Table: Similarity Std_Dev Matrix - MedLine

Exploiting the Temporal Effects

- May we improve classifier's accuracy using the knowledge about temporal effects?
- Are there time windows that are both stable w.r.t. temporal effects and large enough to avoid the sampling effect?

Exploiting the Temporal Effects

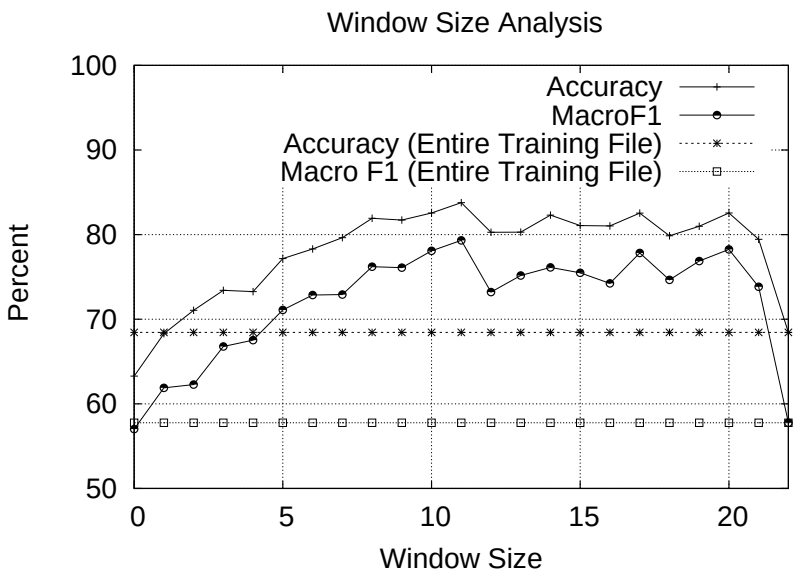
Window Size × Training Sample Size

(o) ACM Collection

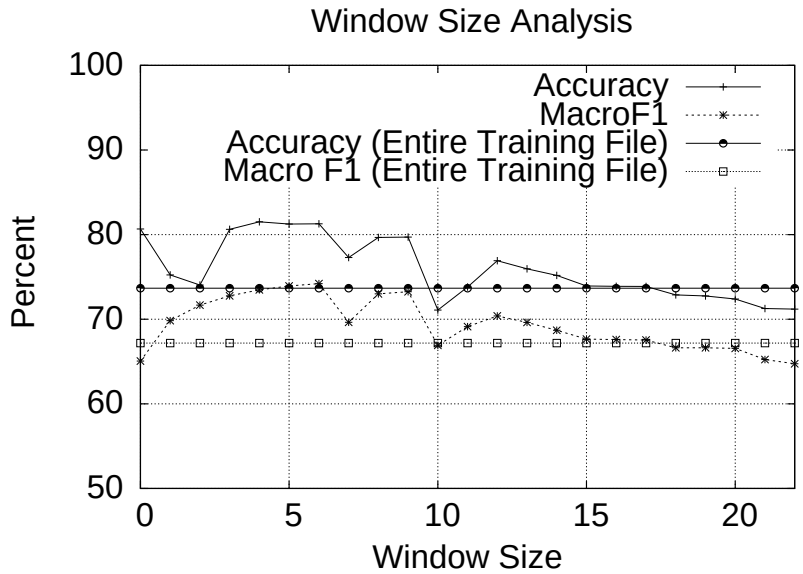
(p) Medline Collection

Exploiting the Temporal Effects

Window Size × Accuracy



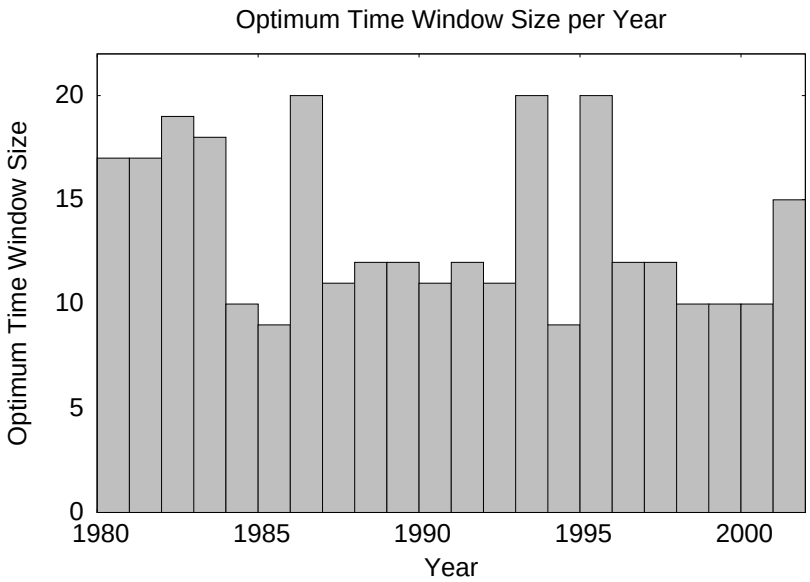
(q) ACM Collection



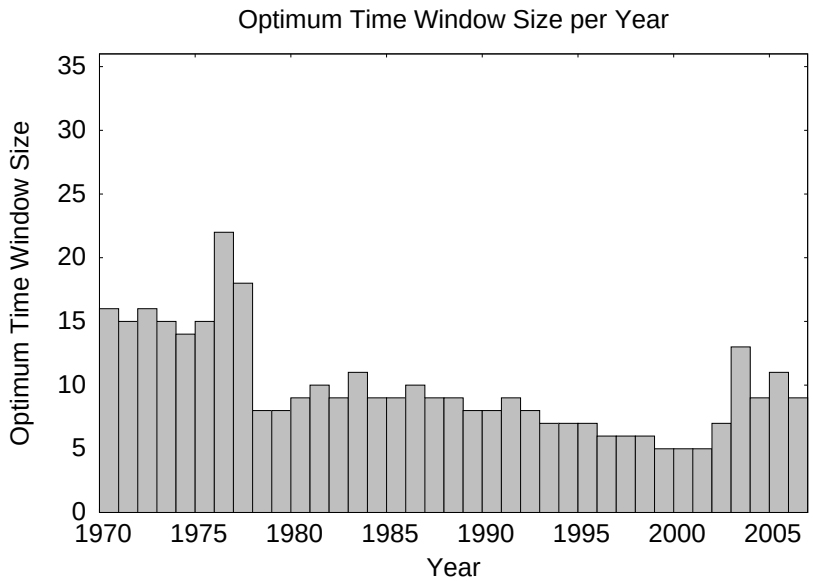
(r) MedLine Collection

Exploiting the Temporal Effects

Optimum Window Size per Year



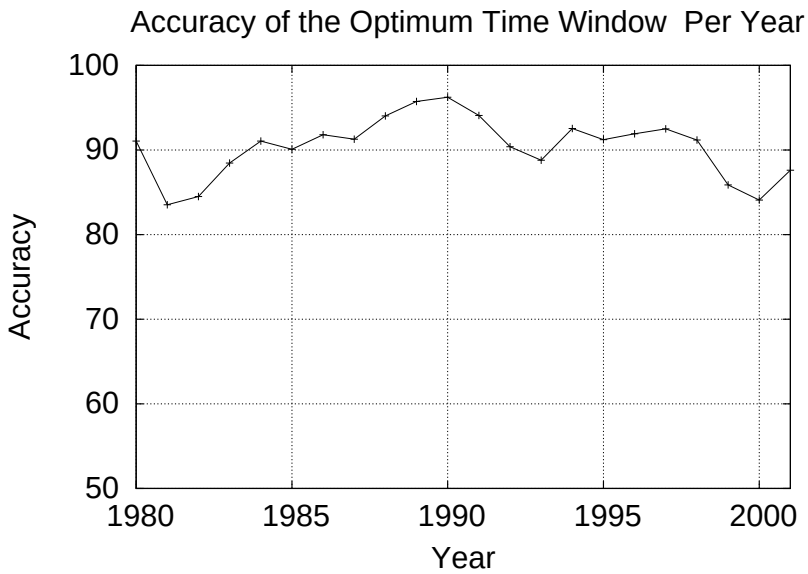
(s) ACM Collection



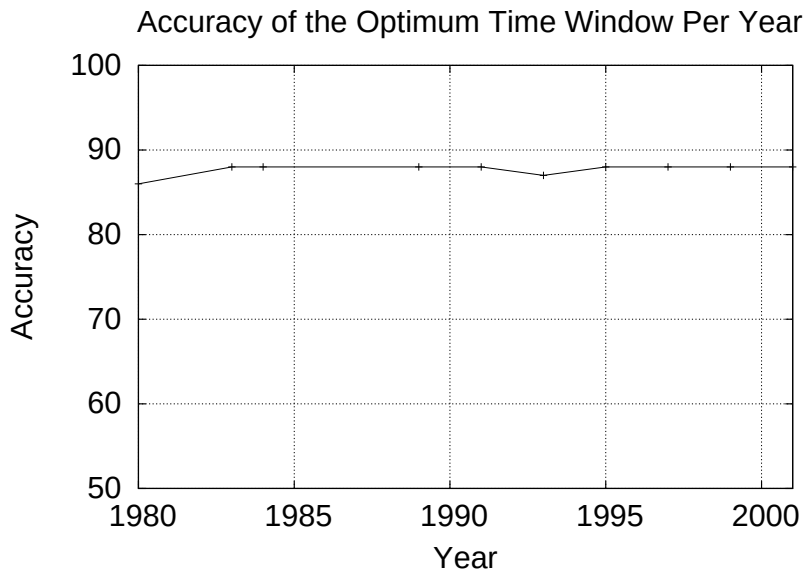
(t) Medline Collection

Exploiting the Temporal Effects

Optimum Window Size × Accuracy



(u) ACM Collection



(v) MedLine Collection

Conclusions

- Characterization methodology of three temporal effects that affect document classification
  - Class distribution
  - Terms distribution
  - Class similarity
- Assessed the trade-off between sampling and temporal effects.
- Demonstrated that exploiting the knowledge about temporal effects pays off:
  - Temporal-based selection of training dataset.
  - Accuracy Results:
    - ★ 89.76% for ACM
    - ★ 87.57% for MedLine
    - ★ up to 20% gains using significant less training data

Ongoing Work

- Characterize Web collections.
- Design novel document classification algorithms that take into account temporal effects.

Understanding Temporal Aspects in Document Classification

Fernando Mourão Leonardo Rocha Renata Araújo
Thierson Couto Marcos Gonçalves **Wagner Meira Jr.**

Universidade Federal de Minas Gerais
Brazil

February 12, 2008
WSDM'08 - California, USA