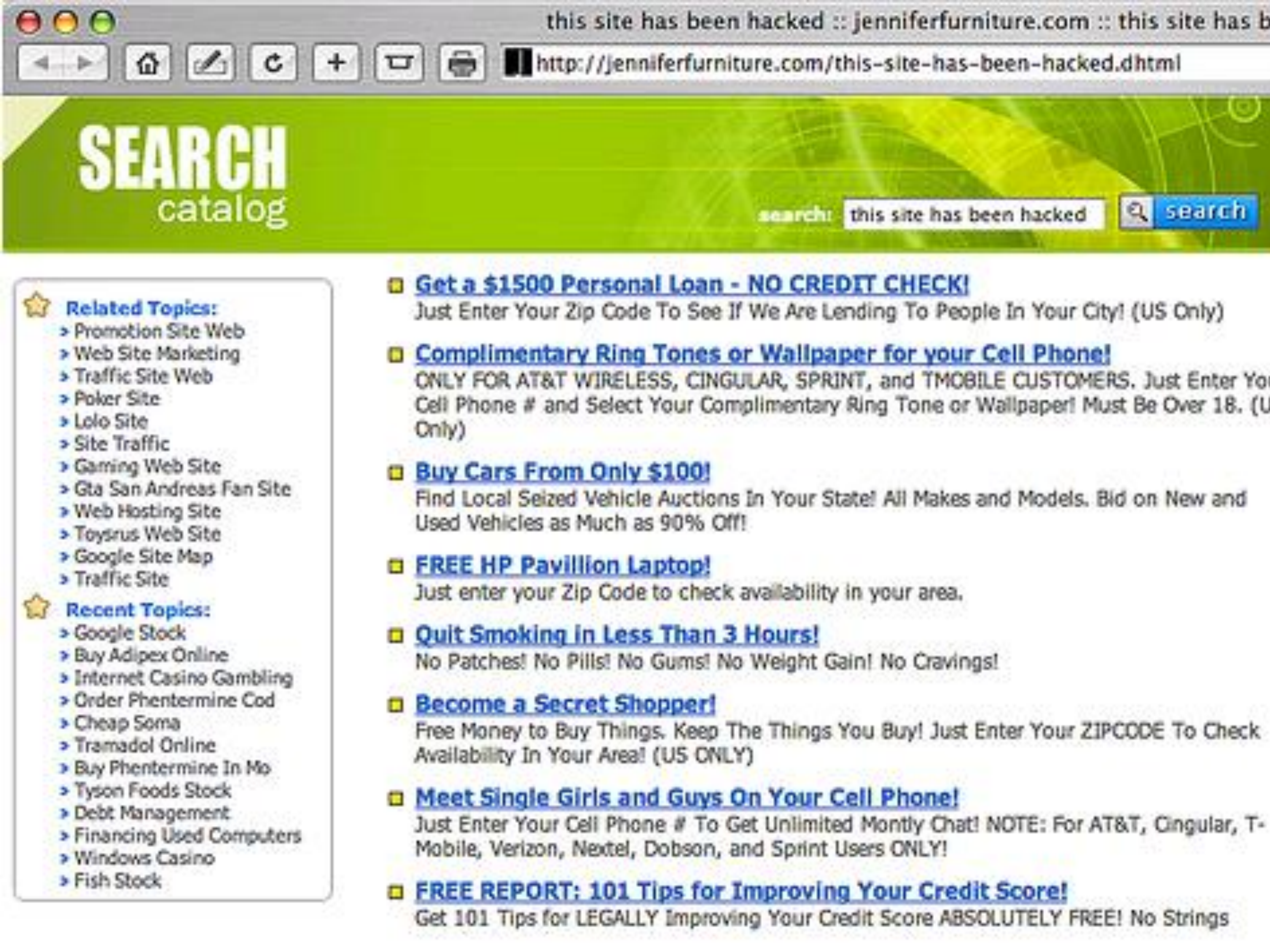
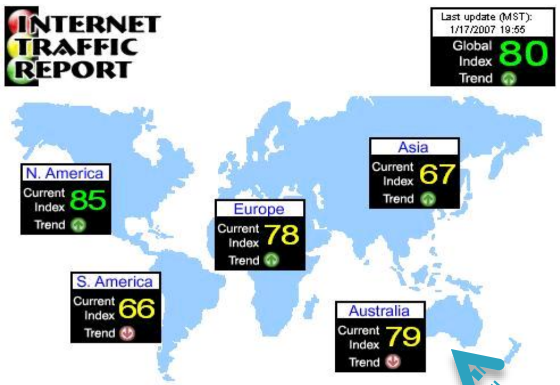
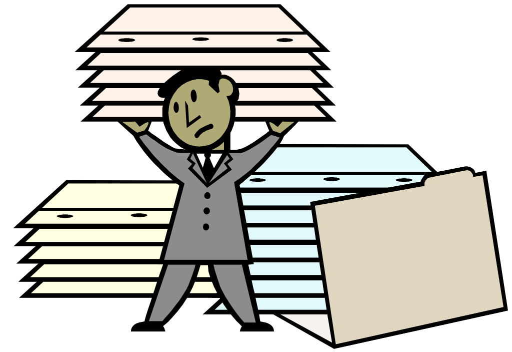
Looking into the Past to Better Classify Web Spam

Na Dai, Brian D. Davison, Xiaoguang Qi
Department of Computer Science and Engineering
Lehigh University

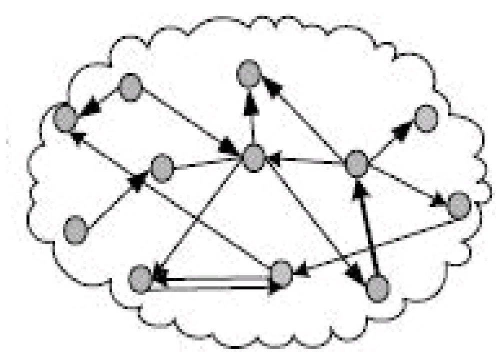

this site has been hacked :: jenniferfurniture.com :: this site has b
http://jenniferfurniture.com/this-site-has-been-hacked.dhtml
SEARCH
catalog
search:
this site has been hacked
search
Related Topics:
Promotion Site Web
Web Site Marketing
Traffic Site Web
Poker Site
Lolo Site
Site Traffic
Gaming Web Site
Gta San Andreas Fan Site
Web Hosting Site
Toysrus Web Site
Google Site Map
Traffic Site
Recent Topics:
Google Stock
Buy Adipex Online
Internet Casino Gambling
Order Phentermine Cod
Cheap Soma
Tramadol Online
Buy Phentermine In Mo
Tyson Foods Stock
Debt Management
Financing Used Computers
Windows Casino
Fish Stock
Get a $1500 Personal Loan - NO CREDIT CHECK!
Just Enter Your Zip Code To See If We Are Lending To People In Your City! (US Only)
Complimentary Ring Tones or Wallpaper for your Cell Phone!
ONLY FOR AT&T WIRELESS, CINGULAR, SPRINT, and TMOBILE CUSTOMERS. Just Enter Yo
Cell Phone # and Select Your Complimentary Ring Tone or Wallpaper! Must Be Over 18. (U
Only)
Buy Cars From Only $100!
Find Local Seized Vehicle Auctions In Your State! All Makes and Models. Bid on New and
Used Vehicles as Much as 90% Off!
FREE HP Pavillion Laptop!
Just enter your Zip Code to check availability in your area.
Quit Smoking in Less Than 3 Hours!
No Patches! No Pills! No Gums! No Weight Gain! No Cravings!
Become a Secret Shopper!
Free Money to Buy Things. Keep The Things You Buy! Just Enter Your ZIPCODE To Check
Availability In Your Area! (US ONLY)
Meet Single Girls and Guys On Your Cell Phone!
Just Enter Your Cell Phone # To Get Unlimited Montly Chat! NOTE: For AT&T, Cingular, T-
Mobile, Verizon, Nextel, Dobson, and Sprint Users ONLY!
FREE REPORT: 101 Tips for Improving Your Credit Score!
Get 101 Tips for LEGALLY Improving Your Credit Score ABSOLUTELY FREE! No Strings

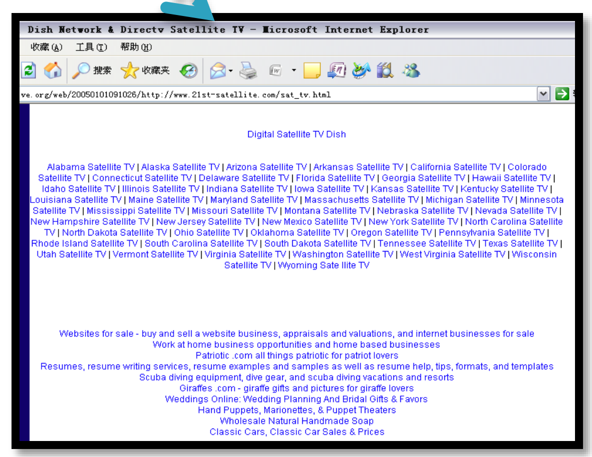




Detect Spam

Historical information about the page itself?





Introduction

- The characteristics of web pages have their own evolution patterns
- Spam pages may have distinguishable evolution patterns from normal pages

Main Questions

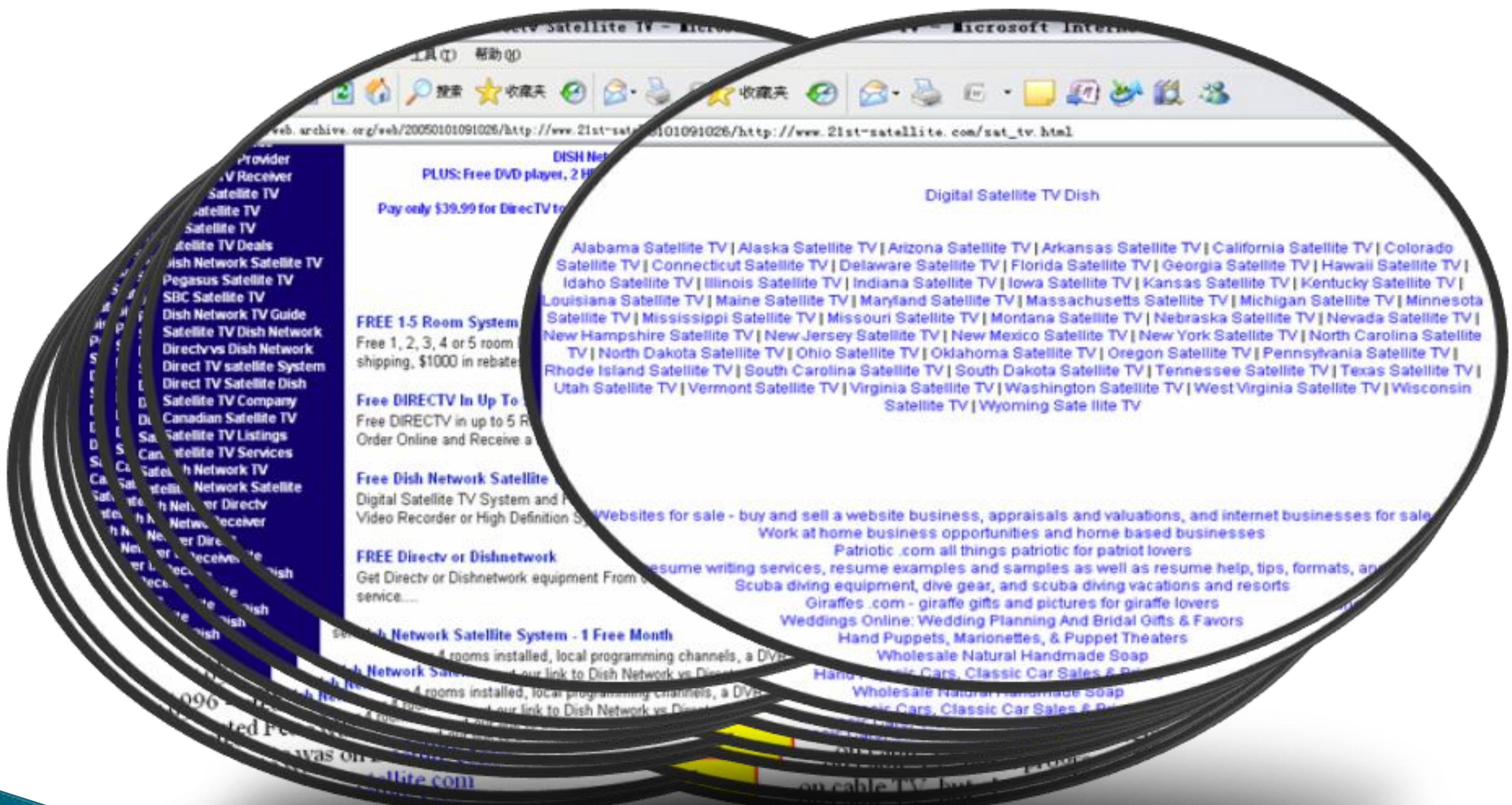
- Can we use different evolution patterns to help Web spam detection?
- Which evolution patterns will make Web pages more likely to become spam pages?
- How long should these patterns influence the decision on spam detection?

Introduction

- Our investigated characteristics
  - Variation of terms contained in web pages
  - Variation of page ownership

- Assumptions
  - Characteristics of spam pages are more likely to have some sudden changes in a previous time interval.

http://www.21st-satellite.com/sat_tv.html from 2004 to 2006



Introduction

- Our investigated characteristics
  - Variation of terms contained in web pages
  - Variation of page ownership
- Assumptions
  - Characteristics of spam pages are more likely to have some sudden changes in a previous time interval.

http://www.emrguide.com/ in 2003 and 2005

Introduction

- Our investigated characteristics
  - Variation of terms contained in web pages
  - Variation of page ownership
- Assumptions
  - Characteristics of spam pages are more likely to have some sudden changes in a previous time interval.

Principle Introduction

- Our proposed approach
  - Train separate classifiers based on multiple groups of temporal features
  - Combine the classification results to achieve the final decision on spam classification

- In our experiment, this approach can boost spam classification F-measure by 30%.

Related Work

- Google filed a patent (2005) on using historical information for scoring and spam detection.
- Lin et al. (2007) showed blog temporal characteristics with respect to splog detection.
- Shen et al. (2006) extracted temporal link features from two historical snapshots to help identify link spam.

Related Work

- Ntoulas et al. (2006) detected spam pages by combining multiple heuristics based on page content analysis.
- Gyongyi et al. (2006) proposed a concept called spam mass and successfully utilize it for link spamming detection.
- Wu and Davison (2006) detected semantic cloaking by comparing the consistency of two copies retrieved from a browser's perspective and a crawler's perspective.

Approach

- Tracking variance of term importance
  - Bucketize the time interval, and extract one snapshot in each time bucket
  - Quantify term importance and make it comparable among different snapshots (BM scores)
  - Quantify term importance change over time
    - Ave (T) – average term weight vector among the selected snapshots
    - Ave (S) – average difference (slope) between two temporally successive snapshots

Approach (Cont.)

- Dev(T) – deviation of term weight vector among the selected snapshots
- Dev(S) – deviation of difference (slope) between two temporally successive snapshots
- Decay (T) – the decayed version of accumulated term weight vectors among the selected snapshots

$$\text{Decay}(T)_i = \Sigma_j \lambda e^{\lambda(N-j)}\, t_{ij}$$

An example

	T_1	T_2	T_3			…			T_m
H_9	t_{91}	t_{92}	t_{93}			…			t_{9m}
…									
H_1	t_{11}	t_{12}	t_{13}			…			t_{1m}
C	t_{01}	t_{02}	t_{03}			…			t_{0m}

$Ave(T)_1 = 1/10 * (t_{01}+t_{11}+\ldots+t_{91})$
$Dev(T)_1 = 1/9 * ((t_{01}-Ave(T)_1)^2+(t_{11}-Ave(T)_1)^2+\ldots+(t_{91}-Ave(T)_1)^2)$
$Ave(S)_1 = 1/9 * (\|t_{01}-t_{11}\|+\|t_{11}-t_{12}\|+\ldots+\|t_{81}-t_{91}\|)$
$Dev(S)_1 = 1/8 * ((\|t_{01}-t_{11}\|-Ave(S)_1)^2+(\|t_{01}-t_{11}\|-Ave(S)_1)^2+\ldots+(\|t_{01}-t_{11}\|-Ave(S)_1)^2)$
$Decay(T)_1 = 1/10 * (\lambda t_{01}+\lambda e^{\lambda} t_{11}+\ldots+\lambda e^{9\lambda} t_{91})$

Approach (Cont.)

- Classification of page ownership change
  - Problem statement: Given a time interval, determine whether a given page has changed its ownership.
  - Extract page-level temporal features (different emphasis from previous feature groups)

Approach (Cont.)

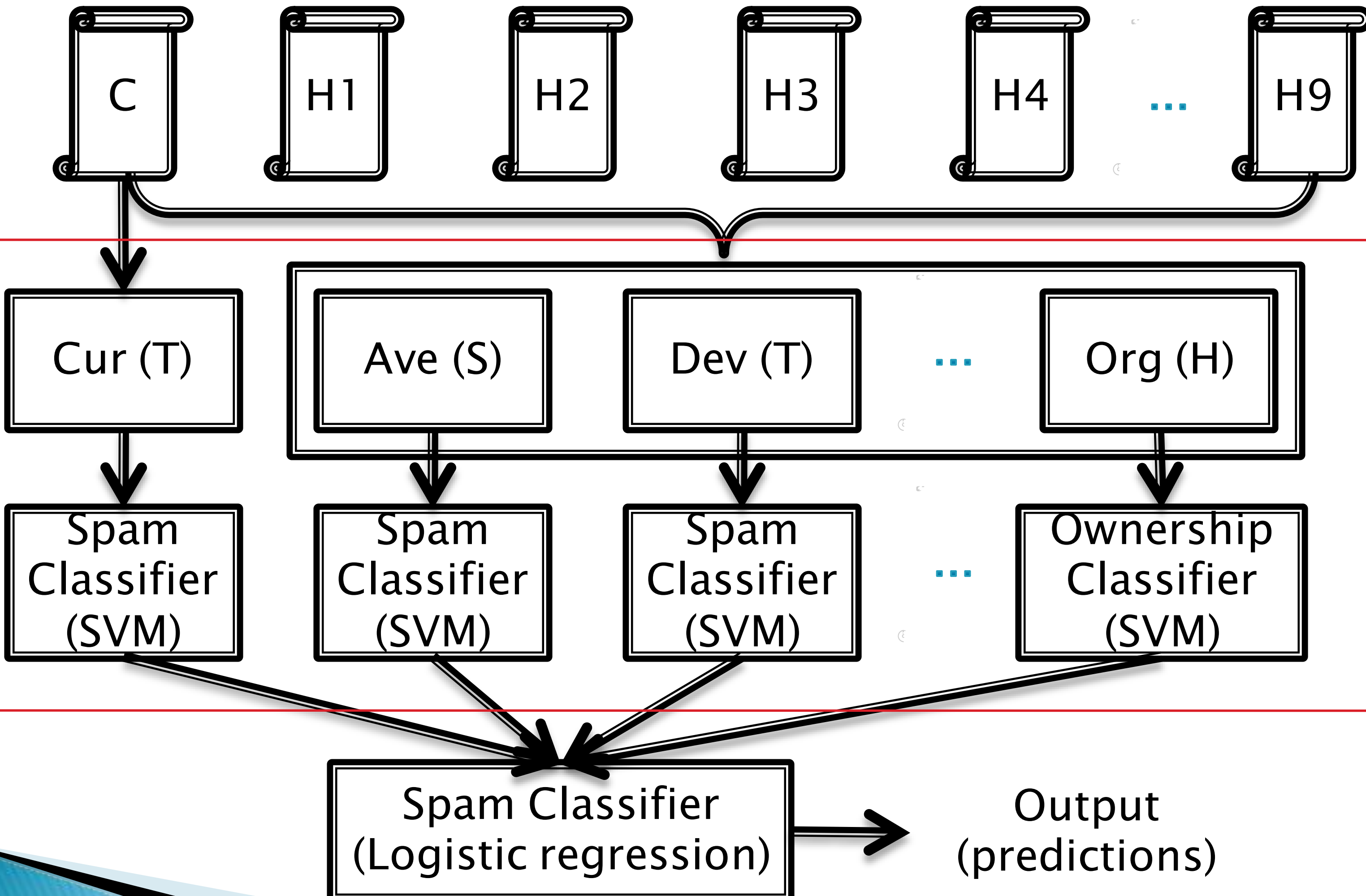
Content-based feature group(s)
➢ Features based on title information;
➢ Features based on meta information;
➢ Features based on content;
➢ Features based on time measures;
➢ Features based on the organization responsible for the target page;
➢ Features based on global bi-gram and tri-gram lists;
Category-based feature group(s)
➢ Features based on topic distribution;
Link-based feature group(s)
➢ Features based on outgoing links and anchor text;
➢ Features based on links in framesets

Approach (Cont.)

- The number of repeated terms in C(d)_i and C(d)_j
- The difference of probability of terms on a predefined list occurring in two snapshots

Content-based feature group(s)
➢ Features based on title information;
➢ Features based on meta information;
➢ Features based on content;
➢ Features based on time measures;
➢ Features based on the organization responsible for the target page;
➢ Features based on global bi-gram and tri-gram lists;
Category-based feature group(s)
➢ Features based on topic distribution;
Link-based feature group(s)
➢ Features based on outgoing links and anchor text;
➢ Features based on links in framesets

Classification architecture



Experiments

- Features' sensitivity on classification performance with respect to time-span

- The spam classification performance comparison before and after we use temporal features

Datasets

- WEBSPAM-UK2007
  - 6479 sites are labeled with about 6% spam sites
  - We select 3926 sites with 201 spam sites (5.12%).
  - Term based temporal features: 10 snapshots ranging from 2005 to 2007.
  - Use the site home page and up to 400 out-linked pages within the same site to represent the sites' content .
- ODP external pages
  - Training set for determining page ownership change.
  - Manually labeled 247 external pages within the time interval from 2005 to 2007.
  - 100 examples are labeled as positive.

Metrics

- Precision
- Recall
- F-Measure
- Confusion matrix

Features’ sensitivity on F-Measure(1)

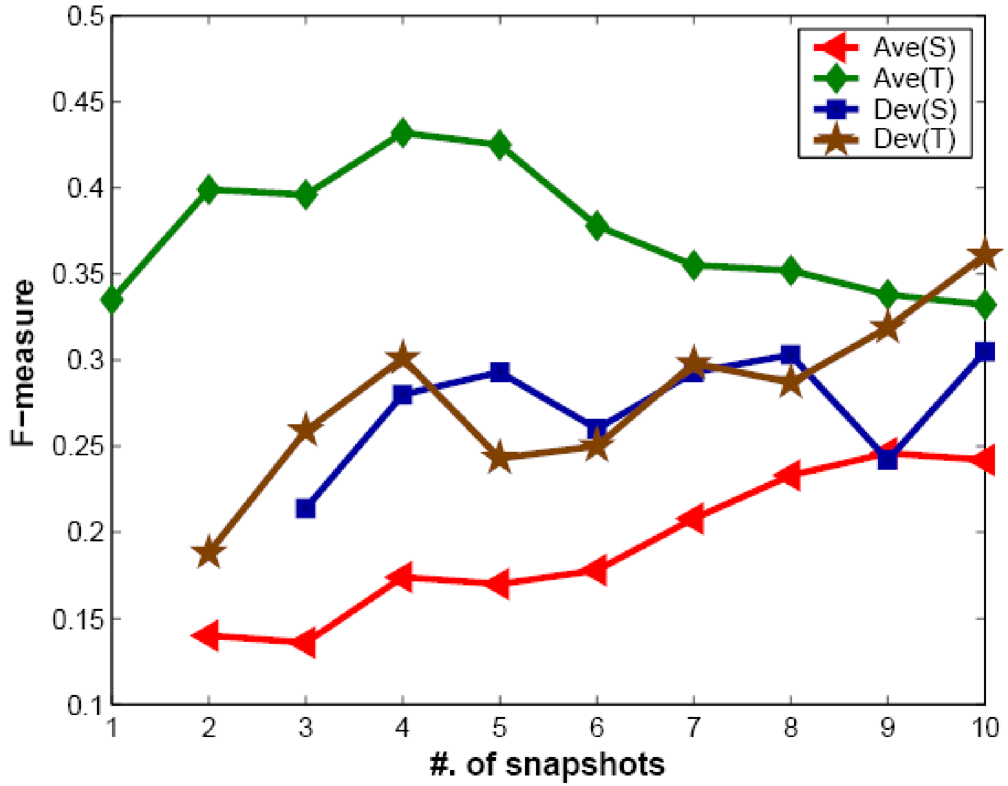


Figure 2: Features’ sensitivity on F-measure performance with respect to time-span.

Features' sensitivity on F-Measure(2)

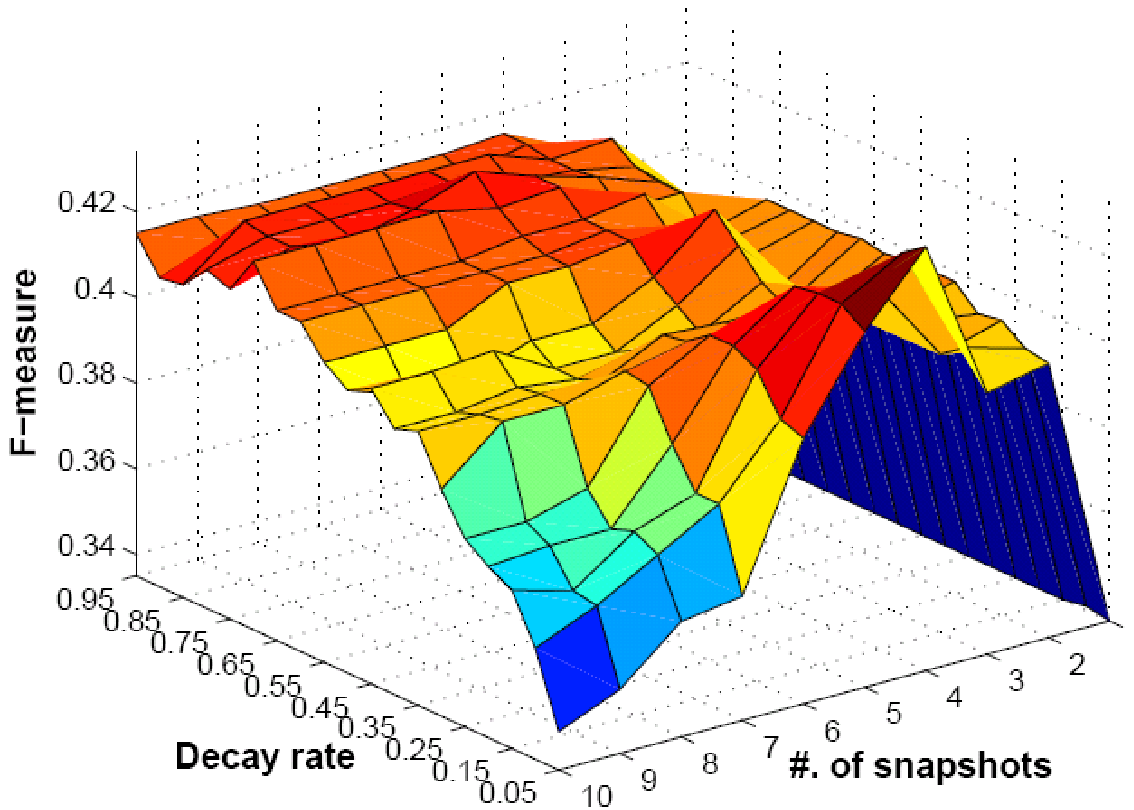


Figure 3: Feature Decay(T)'s sensitivity on F-measure performance with respect to time-span and decay rate.

Individual classification results

Combination	Precision	Recall	F-Measure
BM (baseline)	**0.674**	**0.289**	**0.404**
Dev(S)	0.530	0.214	0.304
Dev(T)	0.529	0.274	0.361
Ave(S)	**0.744**	0.144	0.242
Ave(T)	0.573	0.234	0.332
Decay(T)	0.656	0.303	**0.415**
ORG	0.120	**0.373**	0.181

Combined classification results

Combination	Precision	Recall	F-Measure
BM (baseline)	0.674	0.289	**0.404**
BM+Dev(S)+Dev(T)+ORG	0.650	0.443	**0.527**

Possible Extensions

- Tuning the number of snapshots in classification models
- Combining other temporal features
- The proposed features can be potentially used in other applications.

Conclusion

- Historical information can be a useful resource to help spam classification.
- We demonstrate its capability for spam detection in WEBSPAM-UK2007 data set, and outperform the textual baseline by 30%.

Thank you!

Packard Lab, Lehigh University

- Questions?

- Contact Info:

  - Na Dai
  - nad207(at)cse.lehigh.edu
  - WUME Laboratory
  - Department of Computer Science & Engineering
  - Lehigh University