Cleaning Search Results using Term Distance Features

Josh Attenberg, Torsten Suel

Polytechnic University

Brooklyn, NY 11201

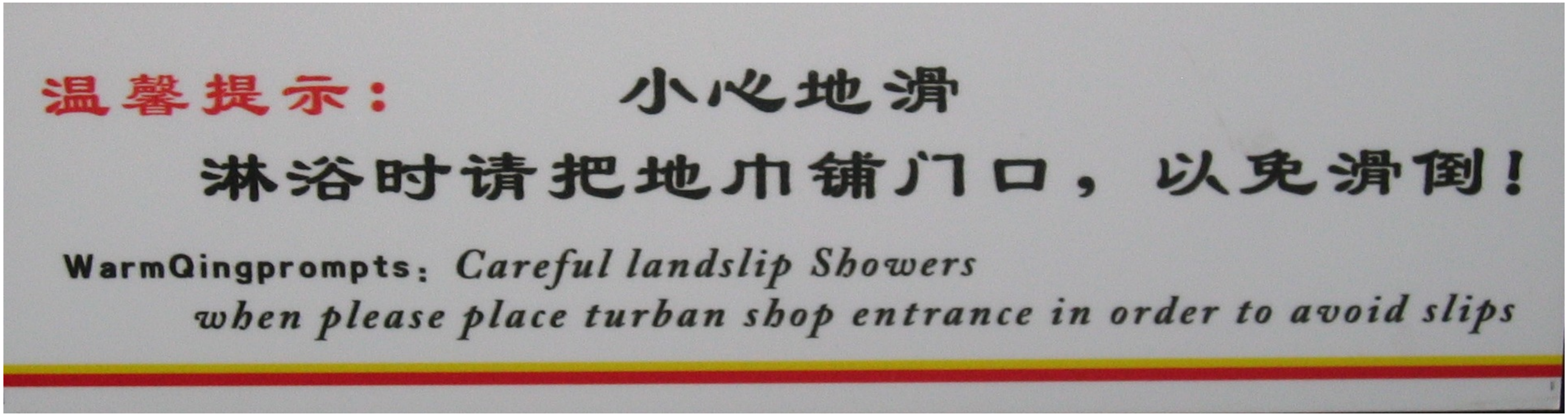
Web Exploration &
Search Technology Lab

Sophisticated Spam

- Weaving: inserting spammed terms through out an existing text
- Phrase Stitching: diverse phrases are joined together to create a new document, possibly with spam terms

…scan the table, a little old lady comes up and asks me if Id like any milk and cookies. Yes Mam I reply. She hands me a little plate with cookies and paper cup of something white. I assume its milk, but its so late. Well, I’m here to connecticut probate if connecticut probate was alone, but crowd psychology is different than individual psychology, and herd instinct connecticut probate out…

Spam or not?

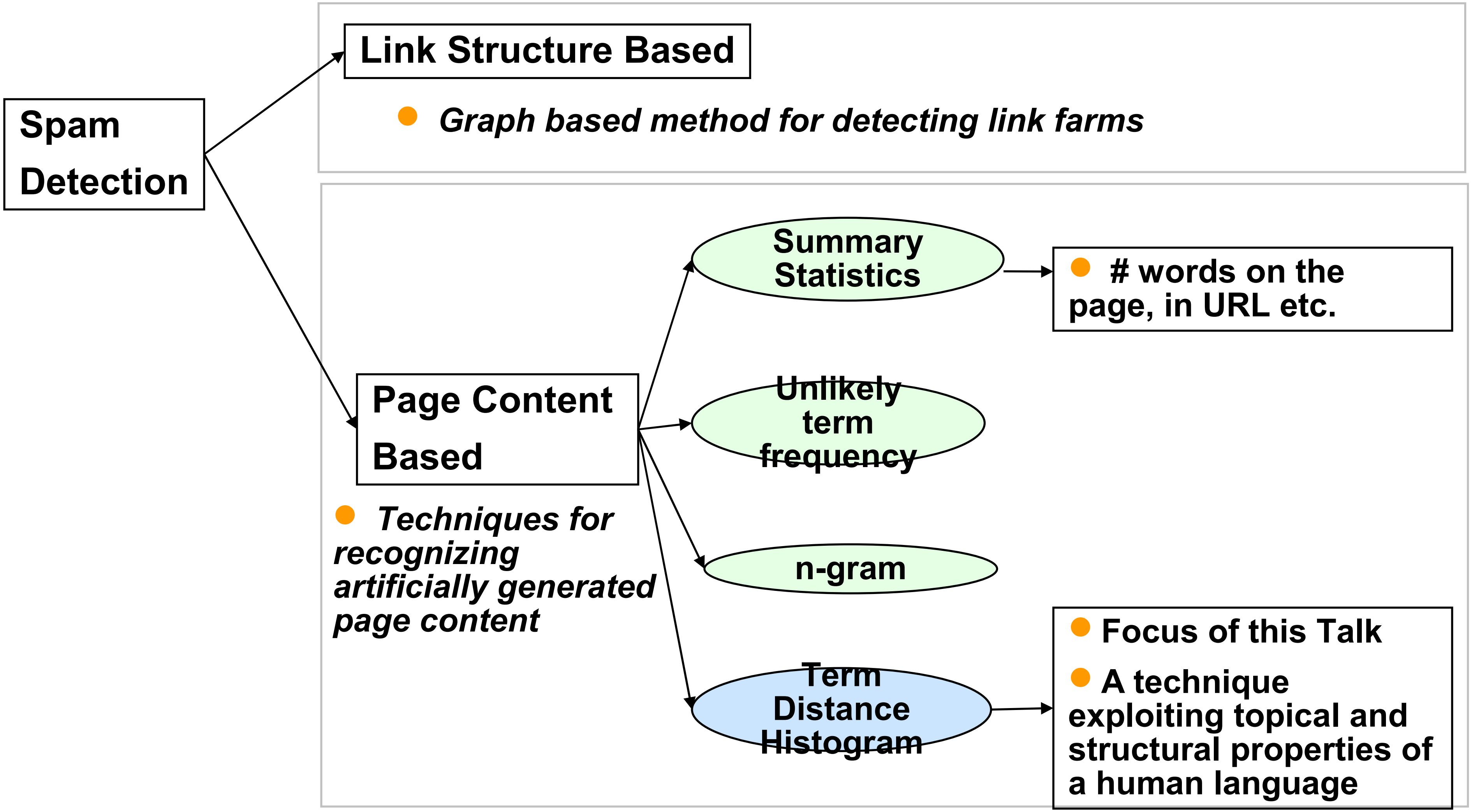


温馨提示：　　小心地滑

淋浴时请把地巾铺门口，以免滑倒！

WarmQingprompts: *Careful landslip Showers*
when please place turban shop entrance in order to avoid slips

Background



Spam Detection

Link Structure Based

- *Graph based method for detecting link farms*

Page Content Based

- *Techniques for recognizing artificially generated page content*

Summary Statistics

- # words on the page, in URL etc.

Unlikely term frequency

n-gram

Term Distance Histogram

- Focus of this Talk
- A technique exploiting topical and structural properties of a human language

Motivation: Natural Language Properties

- Features of content spam:
  - Grammatical Impossibilities
  - Unnatural word and topic patterns

- How can people identify spam?
  - We are able to recognize strange language structure and unlikely combinations of words and topics

Term Distance Histogram – Basic Idea

- We note that human text has some common pairs of words and some rare pairs of words, at varying distances.
- Our motivation is that there is a certain distribution of word-pair likelihoods across different inter-word distances: outliers from normal structure possibly spam
- We wish to create a summary data structure for a single document relating all its word-pair likelihood features: the Term Distance Histogram
- To add robustness and efficiency, we bin likelihood and distance values into a small number of classes

Term Distance Histogram – Details

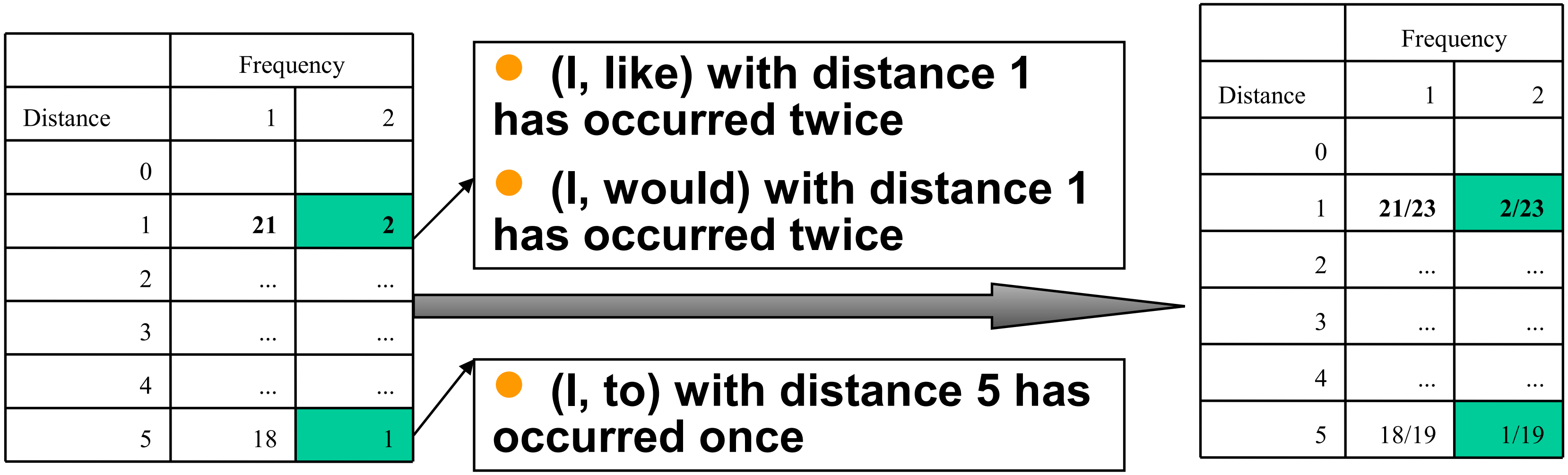
- Given a pair of words, we'd like to assign a likelihood for finding this pair, given the distance between them

- Given parameters *d*, the number of distance classes, and *c* the number of likelihood groups, we define a Term Distance Histogram, *h*, to be a *d* x *c* array of word frequency values

- For each distance class, *i*, we compute the fraction of word pairs occurring at this distance in a document. For each word pair in that distance class, we assign a likelihood class, *(i,j)*, based upon frequency of occurrence in a trusted corpus.

Term Distance Histogram – Example

- Example Text:
  - “I like Beijing. I would like to go to the great wall. I would also be happy to visit other cities in China too.”
  - There are totally 24 words. 18 unique words.
- Distance Frequency Matrix



	Frequency	
Distance	1	2
0		
1	**21**	**2**
2	...	...
3	...	...
4	...	...
5	18	1

- **(I, like) with distance 1 has occurred twice**
- **(I, would) with distance 1 has occurred twice**

- **(I, to) with distance 5 has occurred once**

	Frequency	
Distance	1	2
0		
1	**21/23**	**2/23**
2	...	...
3	...	...
4	...	...
5	18/19	1/19

Detecting Spam

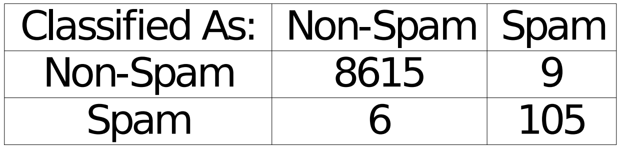
- The Term Distance Histogram for a large number of labelled pages is computed. Each is treated as *d*c* features used as input to train a C4.5 decision tree classifier.

- Term Distance Histogram features are now computed for new pages, which are classified by that decision tree.

Experimental Result

- We conducted two experiments to evaluate the performance of our algorithm.
  - 8735 pages taken from pages resulting from queries made to a major search engine
  - a sample of 50,841 pages taken from the WEBSPAM-UK2007 dataset

- Highlights of the results:
  - Ability to accurately identify content spam
  - Low rate of false positives



Classified As:	Non-Spam	Spam
Non-Spam	8615	9
Spam	6	105

Conclusion and Future Work

- Conclusions:
  - Demonstrated the utility of sentence and topic structure in spam detection
  - Presented Term Distance Histograms, a summary feature capturing structural properties of human language

- Future Work:
  - Explore other uses for Term Distance Histograms
  - Experiment with different statistical models rather than ML
  - Other techniques for spam detection utilizing structure of topics, grammar, and sentence structure.

Questions?