UNIVERSITY of DELAWARE

Machine Learning for Web Searches

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE

Machine Learning Basics

- Supervised learning
  - Need to learn how to compute a function y=f(x) based on a set of examples of the input value x and the corresponding expected output value of y (called training data).
  - Once the function is learned, the computer would take unseen values of x and compute the function f(x)
- Unsupervised learning
  - Only have the data instances x without knowing y
  - Need to learn latent properties or structures of x.

UNIVERSITY of DELAWARE

I. Web Spam

UNIVERSITY of DELAWARE

Why Web Spamming?

- Search engines direct traffic.
- Only highly ranked sites benefit from the search engine referrals.
- Ways to get search engine referrals
  - Buy advertisements — It costs money.
  - Provide better content — Not easy
  - Game the system — Spam!

UNIVERSITY of DELAWARE

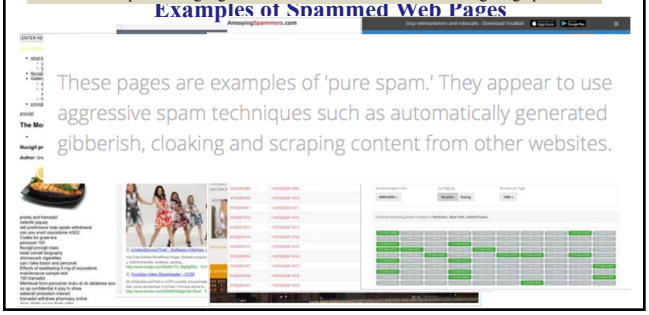
What is Web Spam?

- A page created for the sole purpose of attracting search engine referrals.
- You know when you see it.

UNIVERSITY of DELAWARE

References: https://www.google.com/insidesearch/howsearchworks/fighting-spam.html

Examples of Spammed Web Pages



Web Spam is Bad…

- Bad for users
  - Leads to frustrating search experience
- Bad for search engines
  - Pollutes the collection
  - Wastes the resources
  - Generates unsatisfying search results

Detecting Web Spam

- A classification problem
  - Given a set of features and training data, learn a model to determine whether a web page is spam
  - Use automatic classifiers
- How to identify features?
  - Need to understand spamming techniques
  - Useful features might become useless soon
    - Spammers adapt!

An example

- Given a search engine using Pivoted Normalization retrieval function to rank web pages.
- Suppose we create a useless web page and want to attract search engine traffic. How would you spam?
  - What would you do to improve the ranking of the web page for popular queries, e.g., "car rental", "pokemon go"?

Spam Strategy 1:

Keyword Stuffing

Keyword stuffing

- Observation:
  - Search engines return pages that contain query terms
- Main idea:
  - Create pages containing popular query terms
- How to do it?
  - Hand-crafted pages
  - Completely synthetic pages
  - Assembling pages from "repurposed" content

Hidden Text

- Misleading meta-tags
  - **Meta-Tags** = "… hotels, car rental, discount, booking, reservation, …"
- Hidden text with colors, etc.

Features to identify synthetic content

- Average word length
- Word frequency distribution
- N-gram frequency distribution
- Grammar correctness

Detecting content repurposing

- If the content of a spammed web page comes from a single page
  - Cluster webpages
  - If most pages on a site are very similar to pages on other sites, raise a red flag
- If the content of a spammed webpage are stitched from multiple pages
  - Test if page consists mostly of phrases that also occur somewhere else

Detour: Link-based ranking

- Most search engines use hyperlink information for ranking
- Basic idea: Peer endorsement
  - Web page authors endorse their peers by linking to them
- Link-based ranking algorithm: PageRank
  - Page is important if linked to (endorsed) by many other pages

How would you spam?

Spam Strategy 2:

Link Spam

19

Link spam

- Create more links
  - Link Farm
  - Link Exchanges
  - Links in comments on social media
- Methods:
  - Generate more links automatically
    - Use scripts to post to blogs
    - Synthesize one or many web sites
  - Making the linking page more important
    - Buy expired highly-ranked domains
    - Post links to high-quality blogs

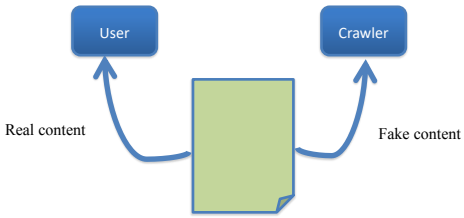
Features for link spam detection

- The number of links from low-ranked pages
- Discrepancy between the number of links and number of visitors
- The number of links from affiliated pages
  - Same domain
  - Same IP address
  - …
- Evidence that linking pages are machine-generated
- …

Spam Strategy 3:

Cloaking

- Serve fake content to search engine crawler



II. Learning To Rank

Learning to Rank

- Basic idea:
  - Define features for a pair of query and document
  - Assume the relevance score is a combination of the features.
  - Learn the feature weights by fitting the function to training data
- Two components:
  - Learning methods
  - Features
    - Content-based features, link-based features, metadata-based features, etc.

Logistic Regression

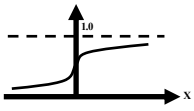
$$\log\frac{P(R=1|Q,D)}{1-P(R=1|Q,D)}=\beta_0+\sum_{i=1}^{n}\beta_i X_i$$

$$P(R=1|Q,D)=\frac{1}{1+\exp(-\beta_0-\sum_{i=1}^{n}\beta_i X_i)}$$

logit function: $x=\log\frac{y}{1-y}$

logistic (sigmoid) function:

$$y=\frac{1}{1+\exp(-x)}=\frac{\exp(x)}{1+\exp(x)}$$



An Example of Logistic Regression

	$X_1(q,d)$	$X_2(q,d)$	$X_3(q,d)$
$d_1(R=1)$	0.7	0.11	0.65
$d_2(R=0)$	0.3	0.05	0.4

$$p(\{q,d_1,R=1\},\{q,d_2,R=0\})=\frac{1}{1+\exp\{-\beta_0-0.7\beta_1-0.11\beta_2-0.65\beta_3\}}\times\left(1-\frac{1}{1+\exp\{-\beta_0-0.3\beta_1-0.05\beta_2-0.4\beta_3\}}\right).$$

Summary

- Advantages
  - It provides a general way of combining multiple features.
  - It provides more robust ranking results.
  - It can leverage all the relevance judgments.
- Challenges
  - The selection of features directly affects the performance.
  - There is no guidance on feature selection and model selection.
- In practice, they are used in all current Web search engines
  - lambdaRank
  - RankSVM
  - Deep learning