SpamChallenge 2007 : France Telecom R&D submission

AIRWEB 2007

tanguy urvoy, emmanuel chauveau, pascal filoche, thomas lavergne

france telecom R&D

may 8th, 2007
strategy overview

intuition: most amount of spam is automatically generated
▶ source code of HTML pages generated by the same script
tend to share some redundancy
▶ spreading information between pages that look alike should
improve spam classification

how to compute similarity between pages
▶ use various preprocessing strategies to extract pages footprints
▶ compute clusterings according to each HTML preprocessing

overall: a mix of supervised and unsupervised classification
▶ first classification of pages: training data, classifier output
  with most confidence, . . .
▶ smoothing and completion of classification among clusters
  according to their consistency \( \frac{Spam - Normal}{Spam + Normal} \)
sample clusterings

words similarity graph

HSS-var-space similarity graph
core components

clustering

▶ several preprocessing strategies were tried and mixed (keeps non alpha-numeric *noise*, HTML tags, words, variants with/without spaces, tags attributes . . .)

▶ LSH fingerprints were computed using *Broder* (min-hashing) and *Charikar* algorithms

▶ 2-pass clustering using connected component computation over windowed partial clusterings

classifier and features selection

▶ *MODL* discretization toolkit and selective bayes classifier [*Marc Boullé - 2004*]
two experiments

#1: enrich labels with MODL classifier, smooth using similarity clusters

#2: enrich labels with stacked graphical feature, build classifier using clusters and other input features
thanks for your attention