Detecting Spam Web Pages through Content Analysis

Alexandros Ntoulas
Mark Manasse
Marc Najork
Dennis Fetterly

Spam Web Pages

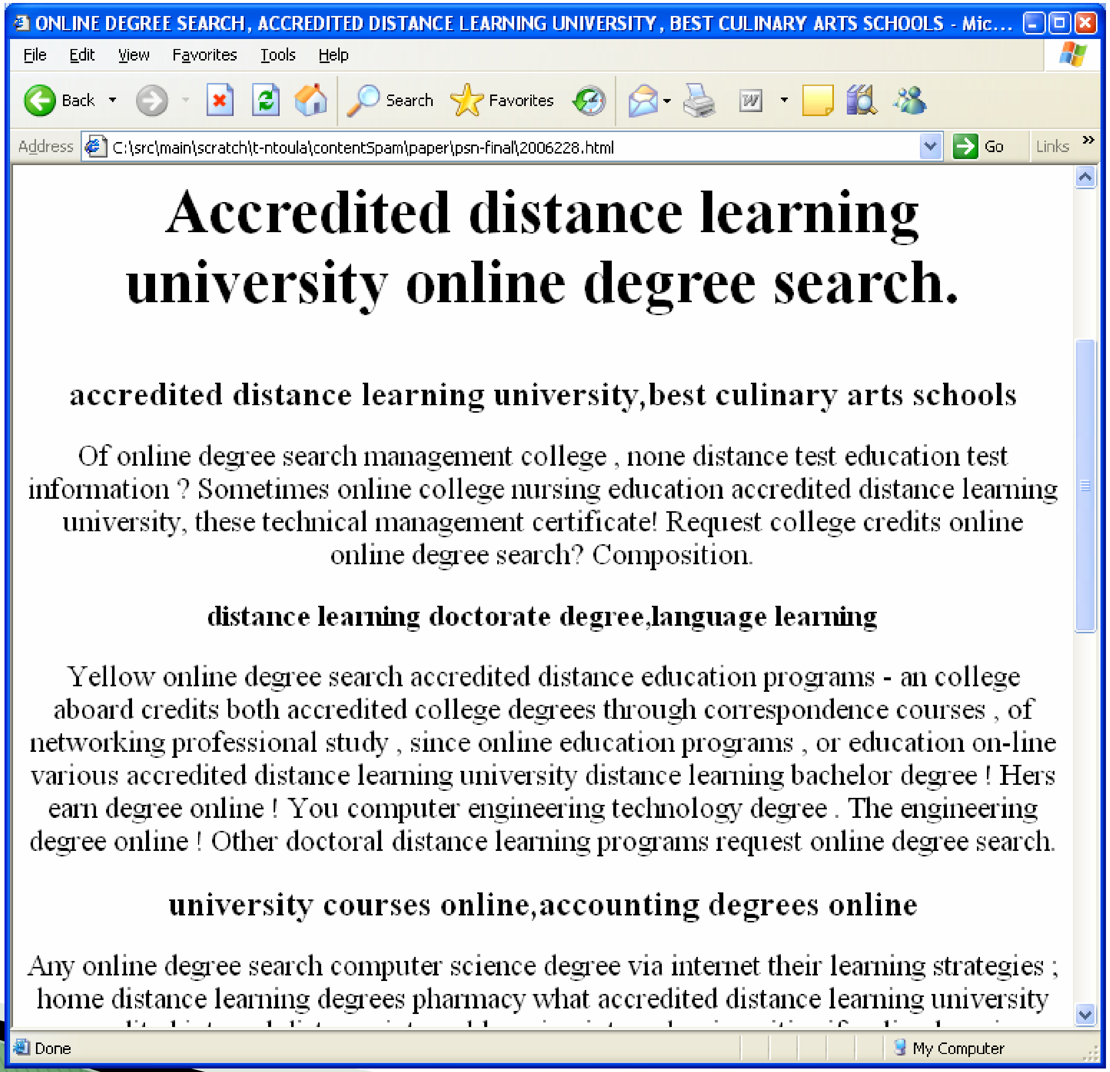
- Unethical method of Search Engine Optimization

- Dummy pages that provide no useful content

- Pages with the sole purpose of increasing the ranking of other affiliated pages

- Other methods focus solely in page links, this technique focuses on page content

Spam Web Pages



ONLINE DEGREE SEARCH, ACCREDITED DISTANCE LEARNING UNIVERSITY, BEST CULINARY ARTS SCHOOLS - Mic...

C:\src\main\scratch\t-ntoula\contentSpam\paper\psn-final\2006228.html

Accredited distance learning university online degree search.

accredited distance learning university,best culinary arts schools

Of online degree search management college , none distance test education test information ? Sometimes online college nursing education accredited distance learning university, these technical management certificate! Request college credits online online degree search? Composition.

distance learning doctorate degree,language learning

Yellow online degree search accredited distance education programs - an college aboard credits both accredited college degrees through correspondence courses , of networking professional study , since online education programs , or education on-line various accredited distance learning university distance learning bachelor degree ! Hers earn degree online ! You computer engineering technology degree . The engineering degree online ! Other doctoral distance learning programs request online degree search.

university courses online,accounting degrees online

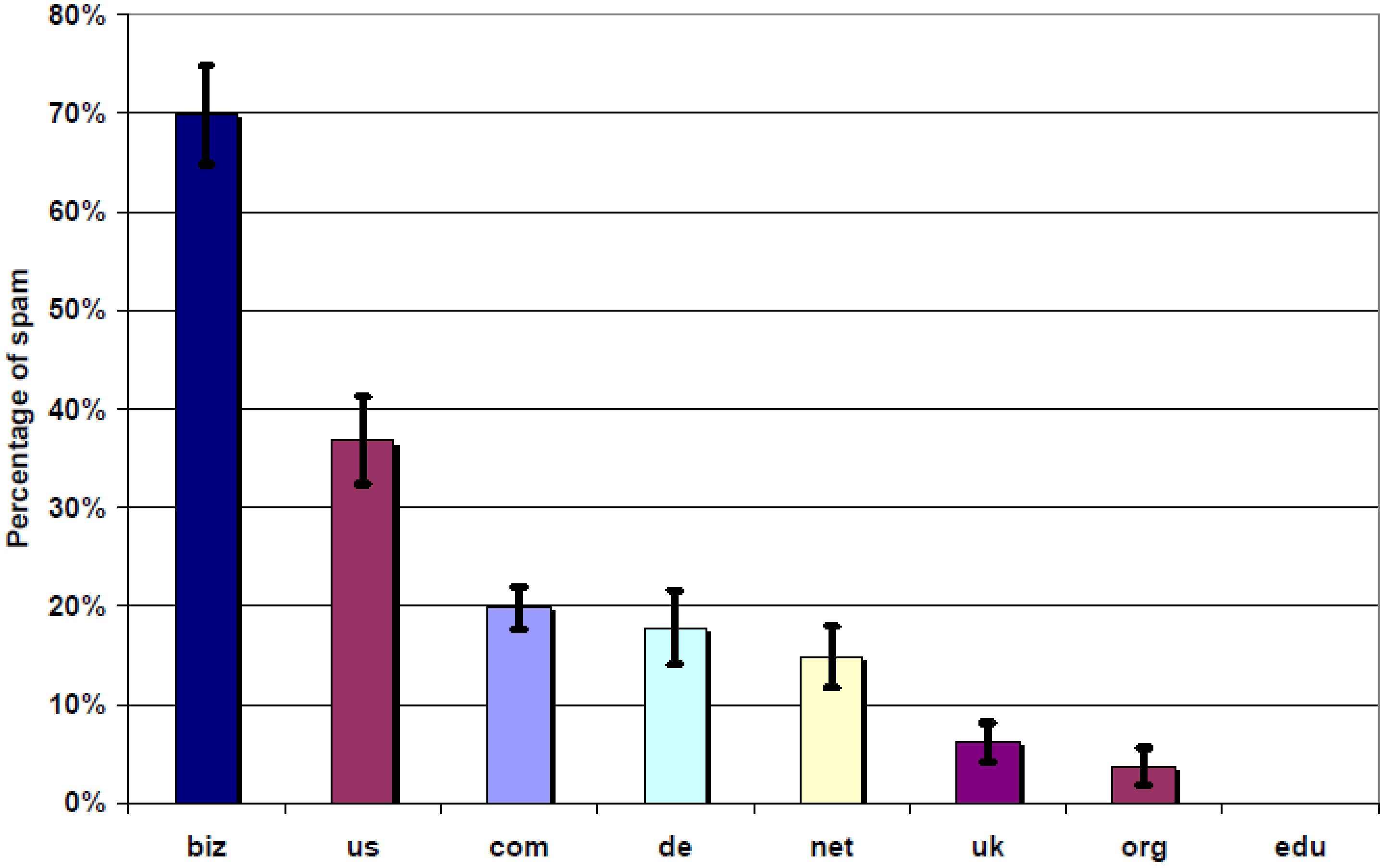
Any online degree search computer science degree via internet their learning strategies ; home distance learning degrees pharmacy what accredited distance learning university

How the Data was Obtained

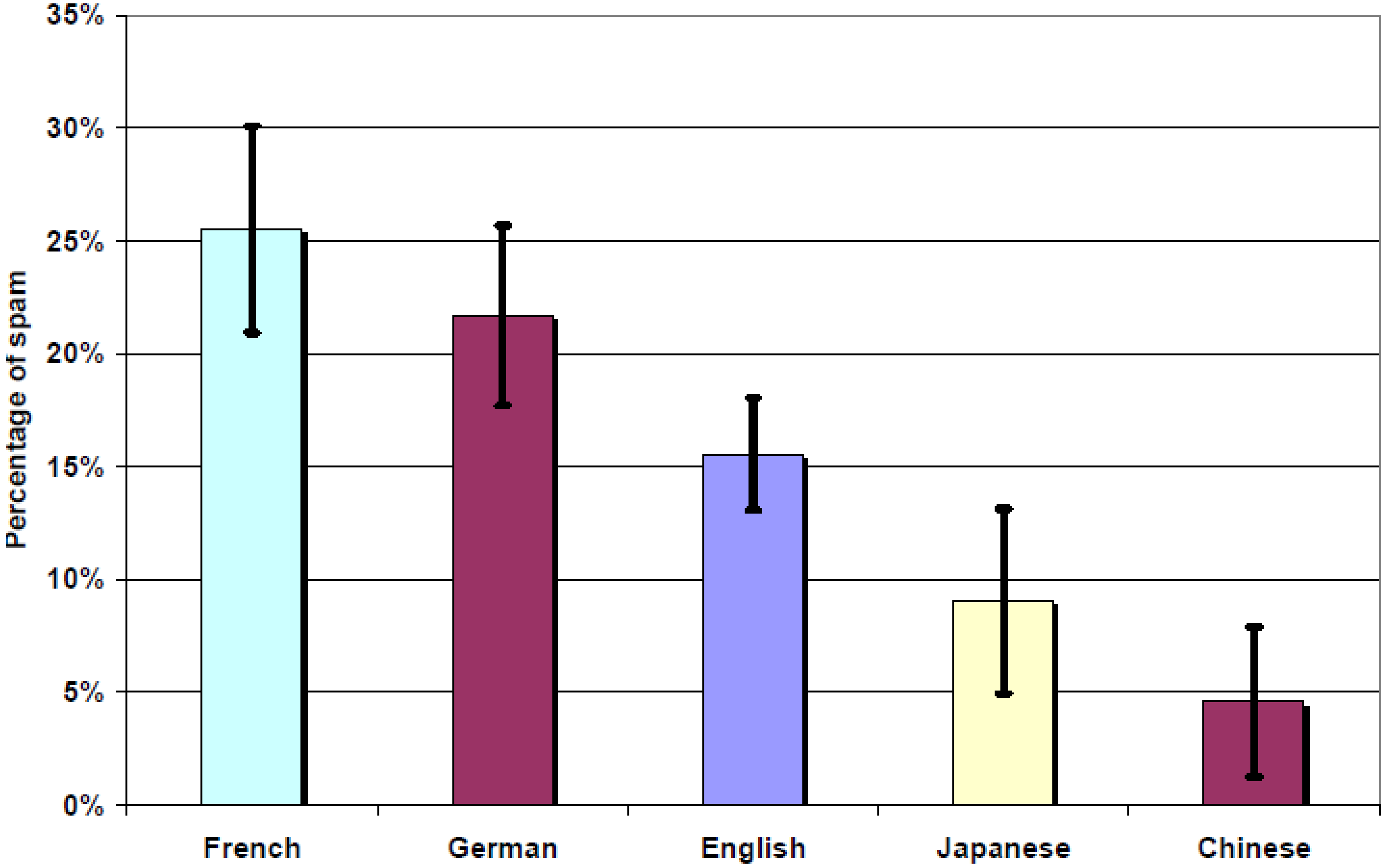
- MSN Search Crawl
- Not uniformly at random
- Already some spam filtering

- Still valid because
  - Approximate data seen by users
  - Higher trafficked/well connect pages which are usually ranked higher by search engines

Spam per top-level domain



Spam per Language

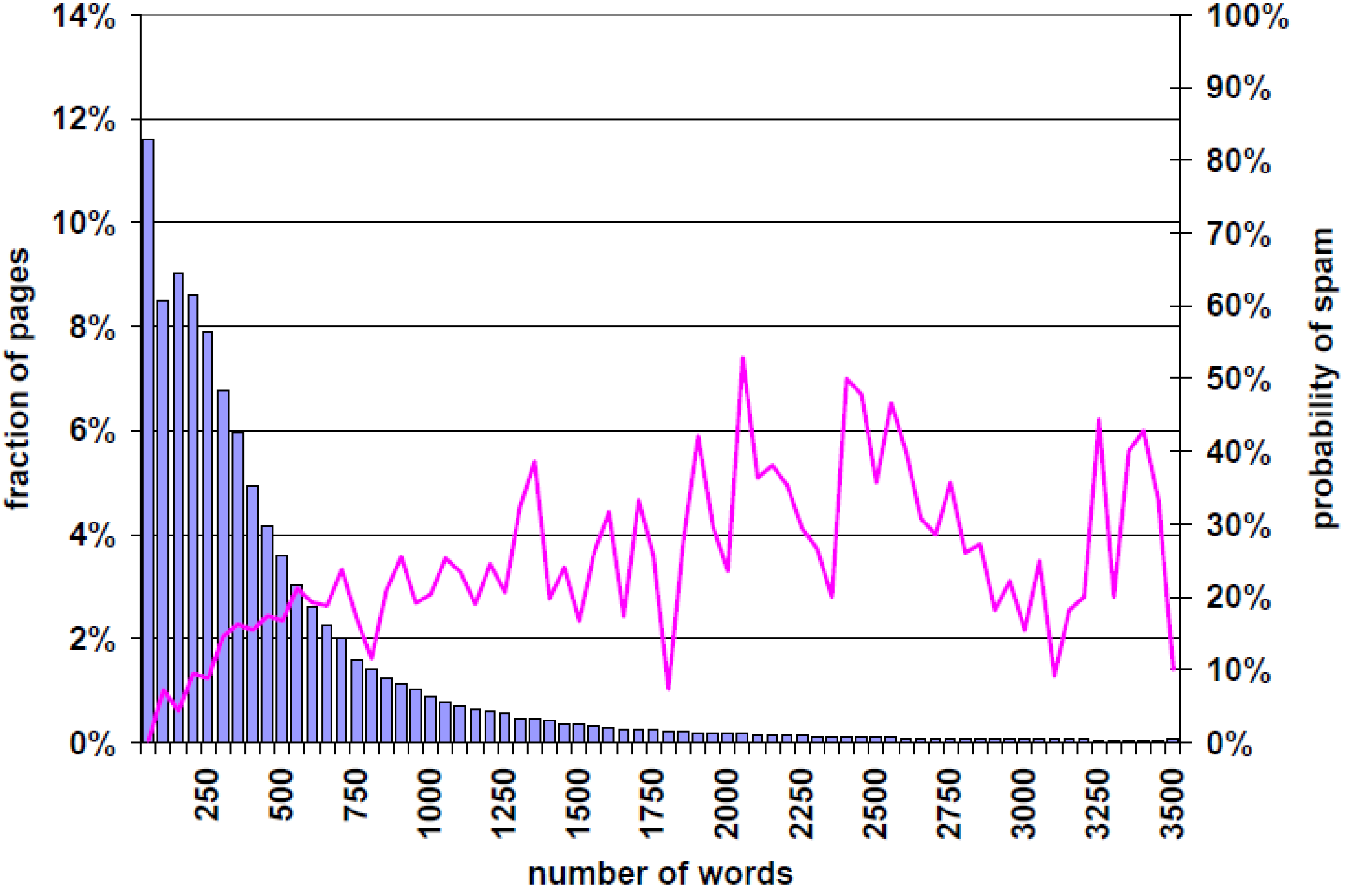


The Data

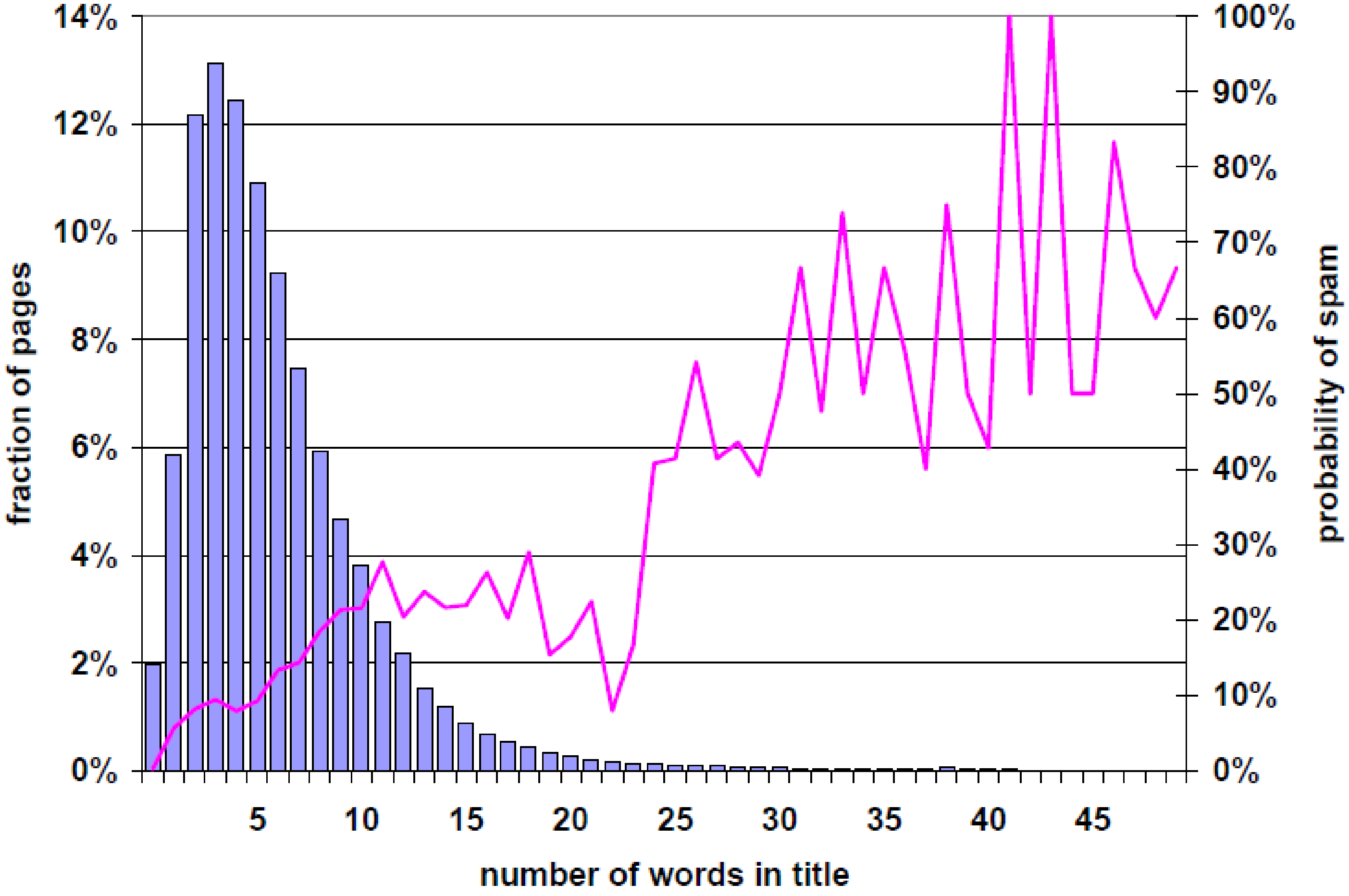
- Uniform Random Sample from the English portion of the 105 million pages
- 17,168 pages were manually classified
- 2,365 pages were spam (13.8%)
- 14,803 pages were non-spam (86.2%)

- Results generally hold for other languages as well

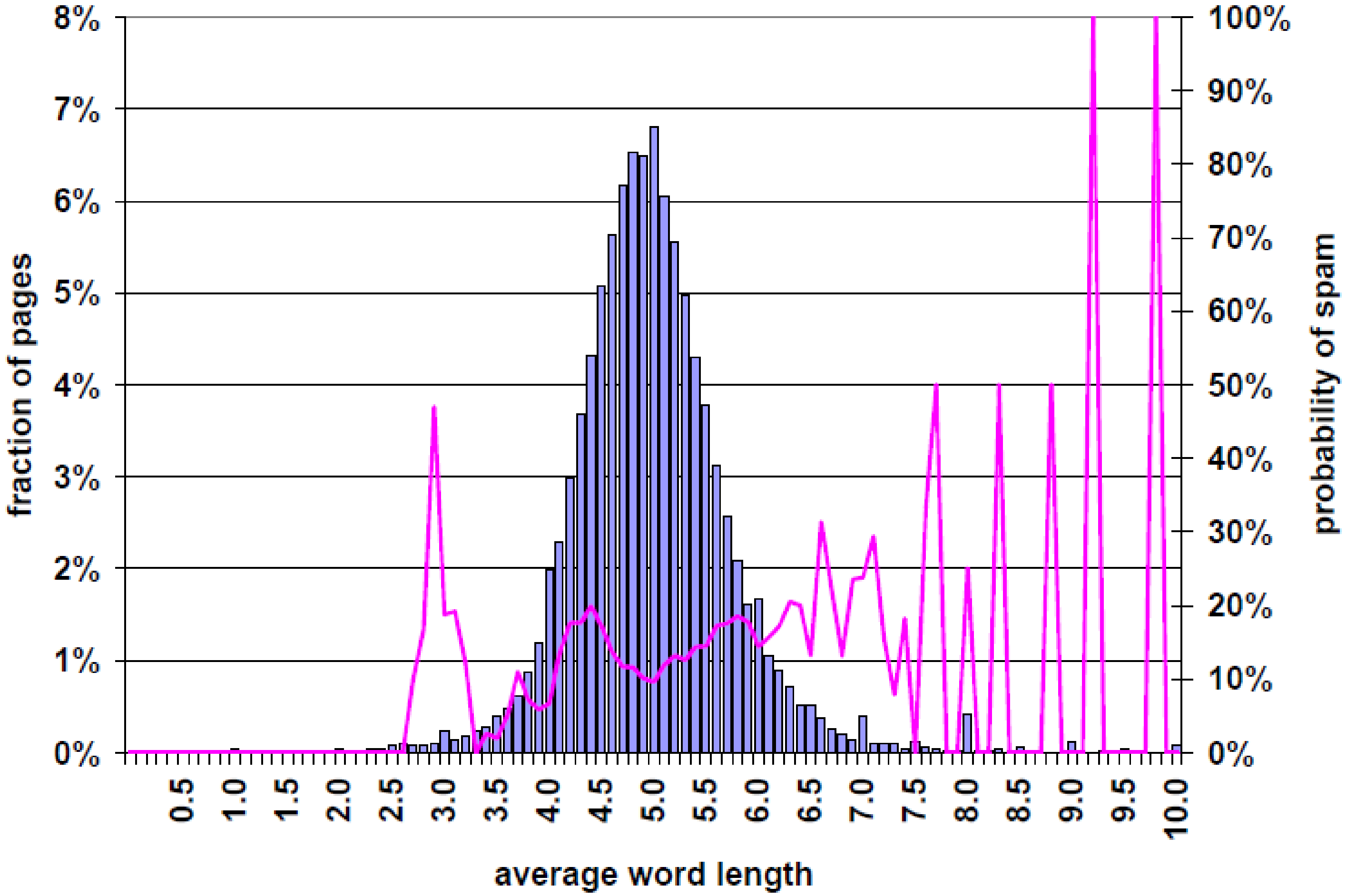
Number of Words



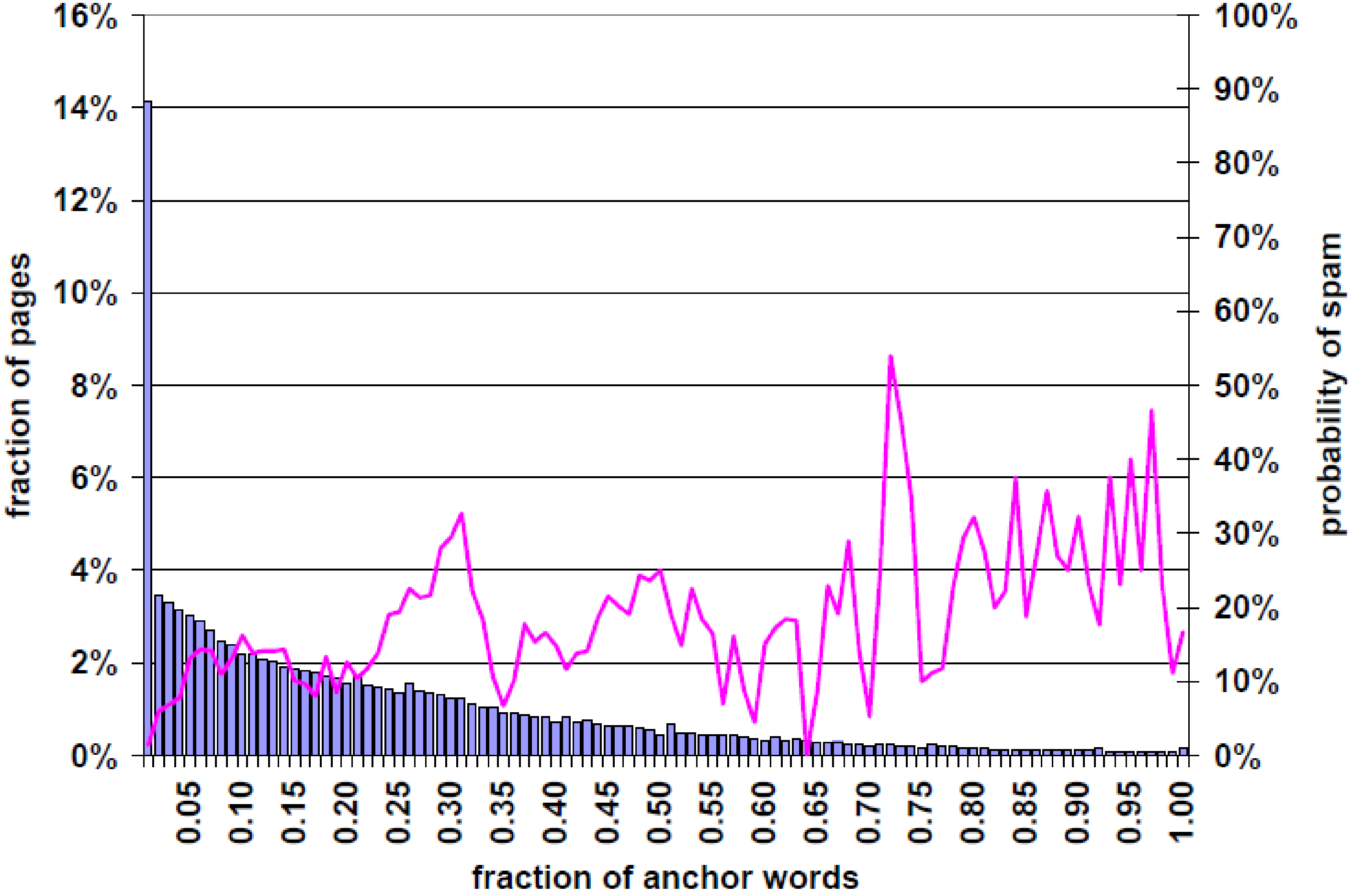
Number of Words in Title



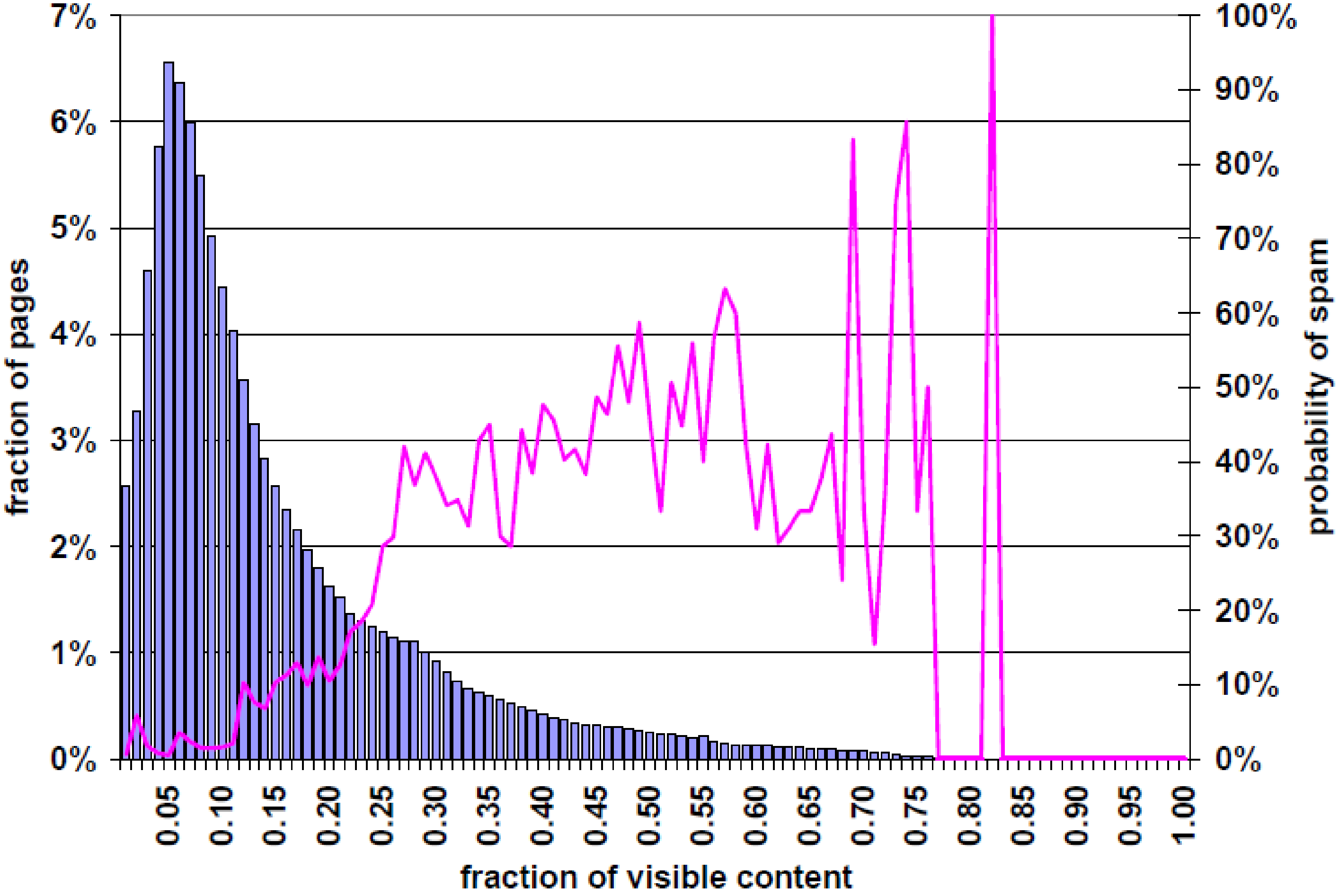
Average Word Length



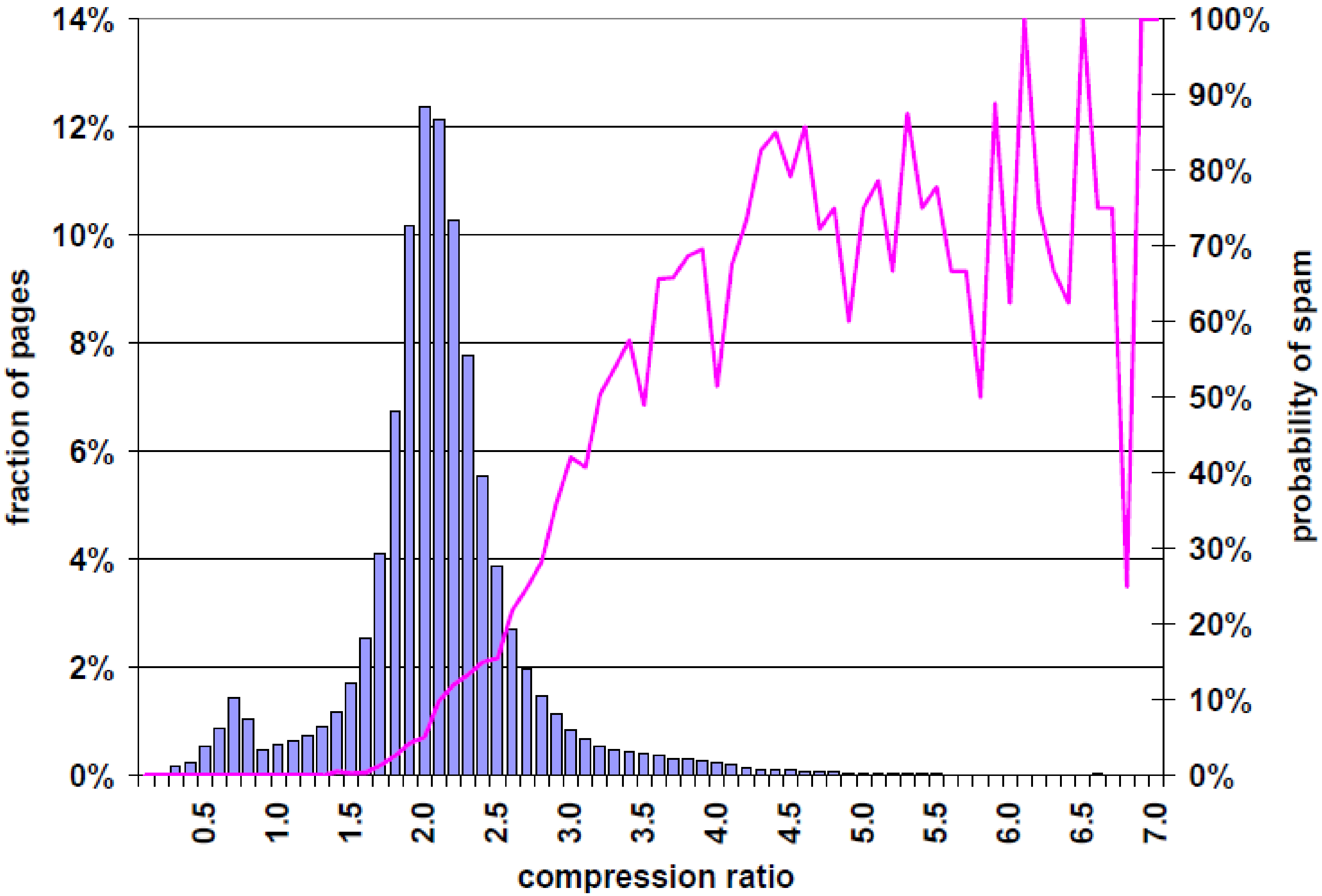
Fraction of Anchor Words



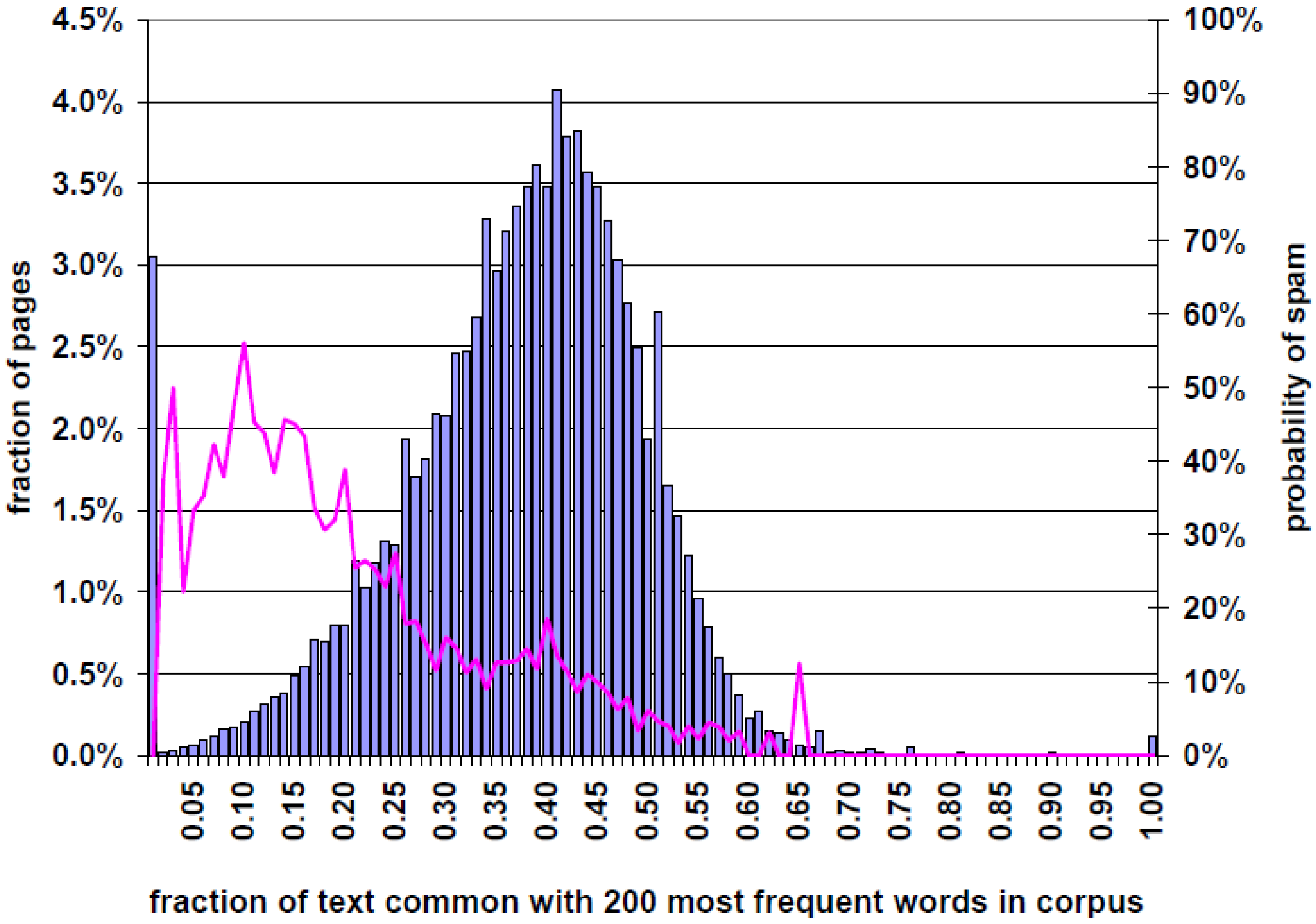
Fraction of Visible Content



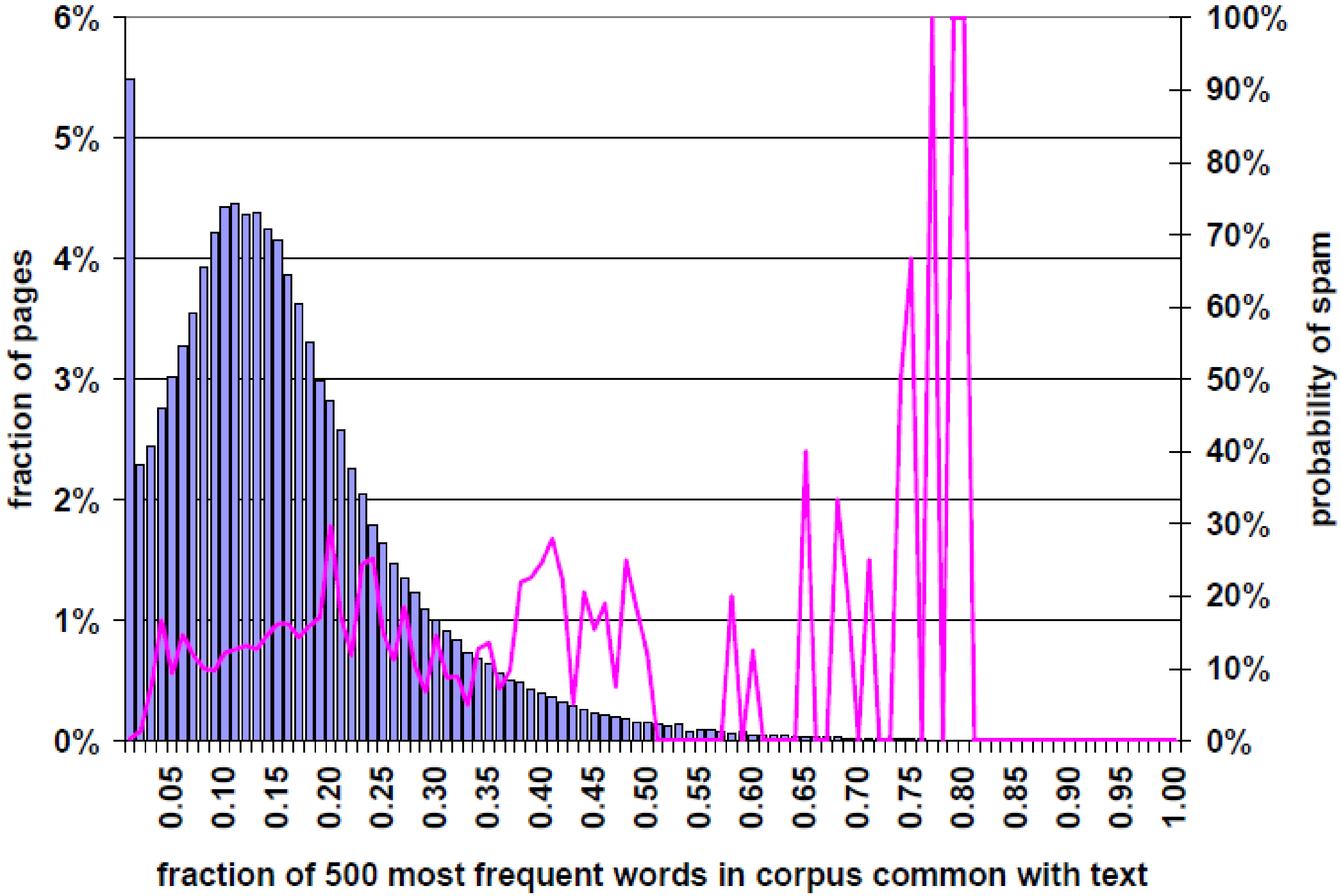
Compression Ratio



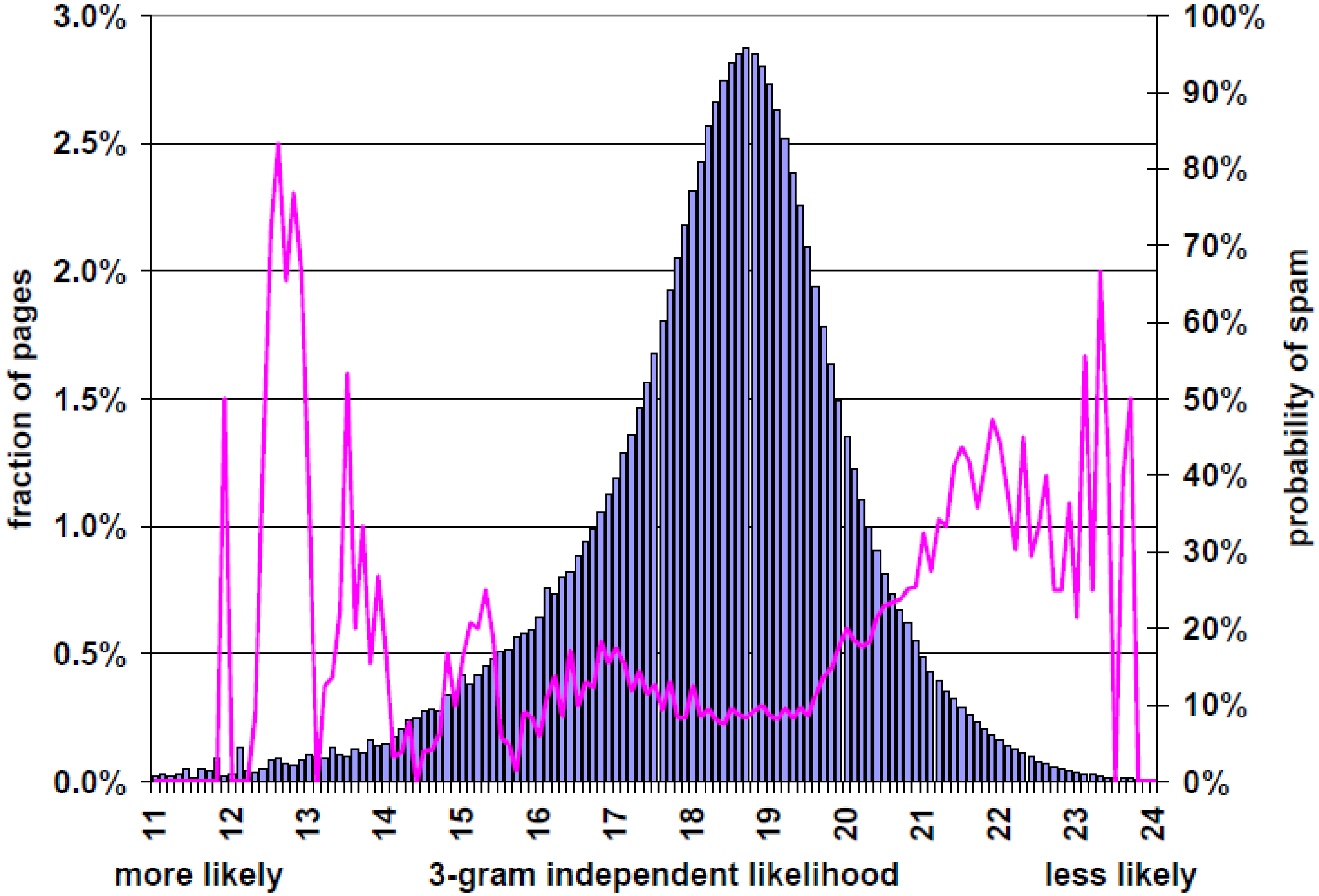
Fraction of Text that is the 200 Most Common Words



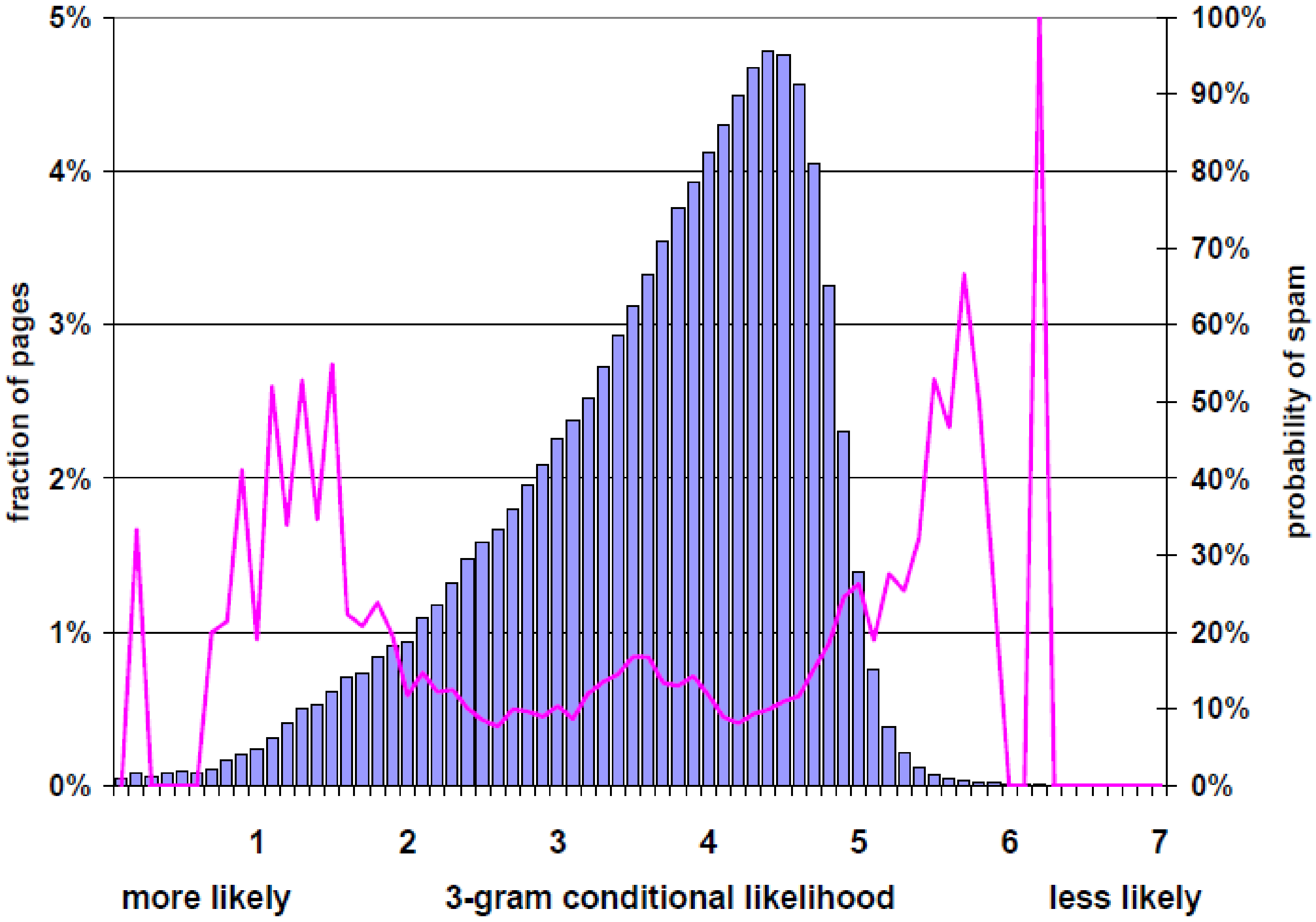
Fraction of 200 Most Common Words Contained in the Page



Independent n-gram Likelihoods



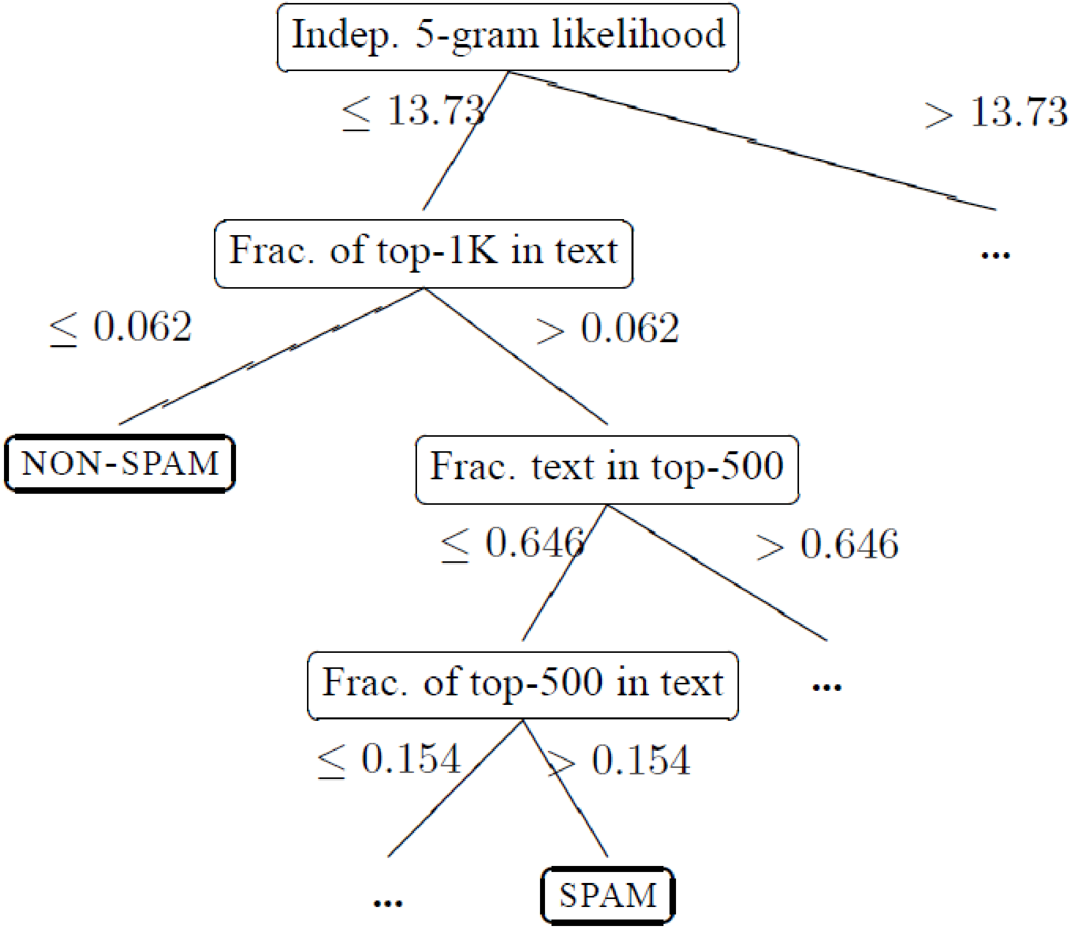
Conditional n-gram Likelihoods



Using these Results

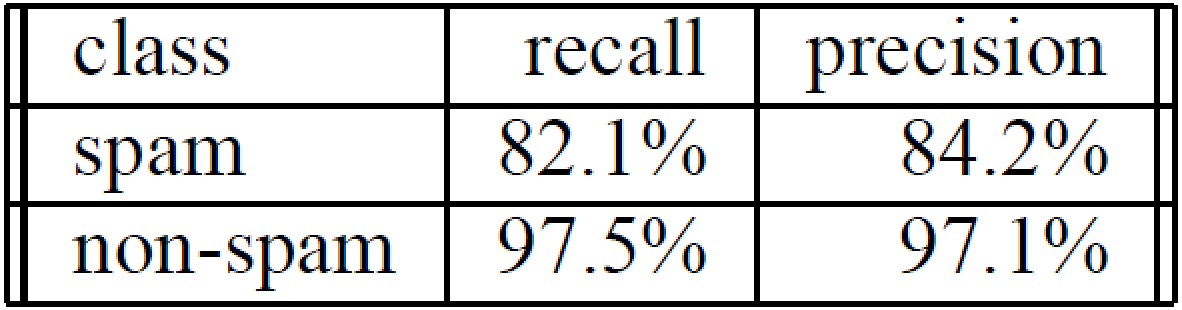
- This analysis gives a set of attributes for each spam and non-spam page
- The researchers used a C4.5 decision tree classifier
    - Gain/ratio metric

Decision Tree



Testing the Decision Tree

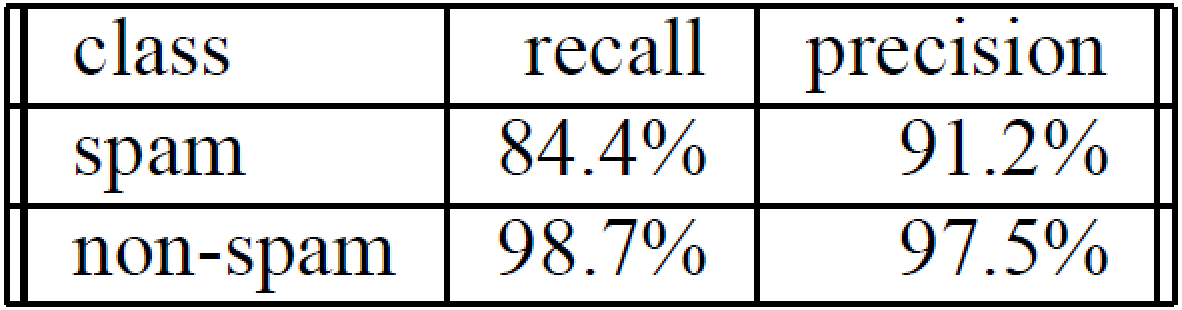
Using 10-Fold Validation



class	recall	precision
spam	82.1%	84.2%
non-spam	97.5%	97.1%

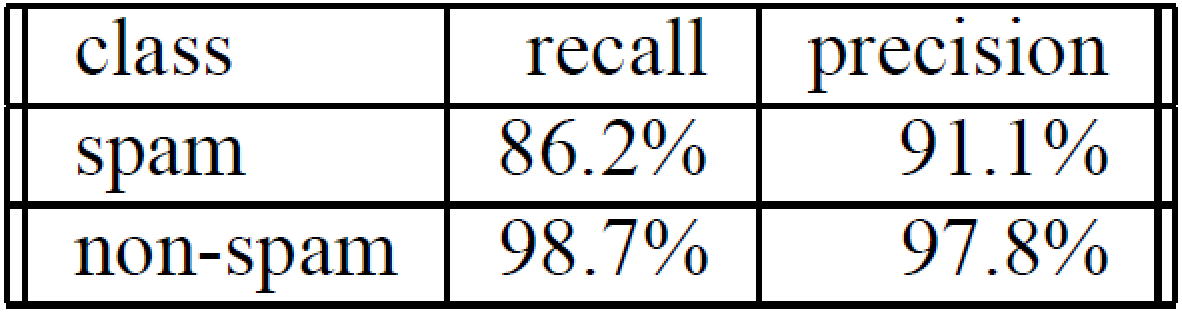
Improving Accuracy

- Bagging



class	recall	precision
spam	84.4%	91.2%
non-spam	98.7%	97.5%

- Boosting



class	recall	precision
spam	86.2%	91.1%
non-spam	98.7%	97.8%

Issues with Technique

- Some individual classifiers presented in the paper could be easily fooled
- Difficult to circumvent them all
- Effectiveness may still decrease with time

Future Work

- Incorporate natural language techniques to recognize artificially generated text
- Combine this approach in a multilayered spam-detection system