Autonomous spam detection on Wikipedia
Avantika Agrawal, Brittney Exline, & Phillip Baker
Advisors: Andrew West, Oleg Sokolsky, Insup Lee

Objective: Design and implement an intelligent routing system for Wikipedia link edits that directs human reviewers to likely spam edits for potential revocation.

The Situation

WIKI = $$$: Wikipedia is a frame-breaking innovation, providing information:
- free, after the price of an internet connection;
- aggregated, into a single, consistent collection;
- comprehensive, with over 3.5 million articles.

This compilation of collective intelligence represents a value of more than $7 billion in 2008.

SPAM ≠ $$$: Non-productive editors directly destroy value, but also harm Wikipedia’s reputation.

EASY TARGET: A homogeneous base of the underlying software package creates an attractive target for autonomous, replicated attacks, as shown by recent work. In addition:
- The open, collaborative nature of wikis makes it hard to detect a diversity of spam link types;
- Current anti-spam tools are static and simplistic;
- Human-editors are dwindling and relatively slow.

Our Solution

WHAT’S SPAM? While Wikipedia defines multiple types of spam, our system focus on a computationally attainable task: external link spam.

DATA COLLECTION: Our system collected human editor decisions and the associated metadata they likely used to make a conclusion regarding the validity of over 238,000 link additions to English Wikipedia to build a real-time classifier.

REAL TIME SYSTEM STRUCTURE: The system combines multiple information sources to present candidate spam revisions to users in a dedicated client.

Classified: Spam Ham
Spam 890 205
Ham 136 473

Tasks involved in the creation of this system

Design a corpus
Listen for link edits on Wikipedia
Collect metadata for each edit

Analyze & process data
Label as spam/ham
Filter for only edits that are labeled with certainty

Develop a predictive system
Design a decision tree using corpus
Design a real-time system that scores edits

The Results

THE SCORE: The alternating decision tree classifier attains 86.7% precision at 81.3% recall with a ROC area under the curve of 0.967.

Spam detection on Wikipedia

Objective:

REAL TIME SYSTEM STRUCTURE:
The system combines multiple information sources to present candidate spam revisions to users in a dedicated client.

Classification of spam corpus using the tree

Flow of information through our system

SPAM CORPUS: Three main classes of features were analyzed: Wiki metadata for the link; URL landing site characteristics; and third party landing site ratings.
