Paper Plot

Title

SWM: Cloaking Detection through Simhash-based Website Model

Abstract

What is cloaking?

Cloaking is severe. Traditional methods are not efficient. And cloaking is still prevalent. We propose SWM, and apply to cloaking detection, achieve high accuracy and efficiency. Systematically review the incentives behind cloaking. Then motivate and propose the new model.

1. Introduction

What is cloaking, impact and technical challenges of cloaking?

Previous work and drawbacks (Brief).

Propose our work and our contribution (Brief).

1. SWM and detection result
2. Systematic review of categories and incentives of cloaking
3. Deployment of the system, and propose a novel framework employing crowdsourcing with privacy guarantee.

Talk about structure of the paper.

2. Related-work

What is cloaking and categories of cloaking (More details)? What is SEO and SEM cloaking?

Description of past approaches in cloaking detection.

Challenges that past work haven’t been solved. Talk about our contributions.

1. Handle dynamic web page (Briefly talk about how Simhash was used in past)
2. Unrevealing cloaked content

3.Methodology

Challenges and solutions. Basic ideas are 1, 2, 3.

1. Methodology of dynamic modeling of web pages.

Simhash Models of Websites.

2. Data collection and Groundtruth

1. Collect terms

2. Query search results.

3. Crawl data and get ground truth (how many).

3.Cloaking detection

4. Results and Evaluation

1.Talk about our dataset, metrics , recall and precision.

2. Compare to previous work, show performance.

5. Measurement (Systematic review)

1. Get detected results from 60K (scale of previous research) data and analyze.

2. SEO: How severe is cloaking? How many categories and percentages of various cloaking? For each type of cloaking, what are their incentives?

3. SEM: How severe is cloaking? How many categories and percentages of various cloaking? For each type of cloaking, what are their incentives?

6. Discussion

Potential deployment

1. Server based.

Pros: 1.practical, easy deployment.

cons: IP cloaking

SEM

hard to decide crawling periods.

For different website, T should be dynamic. Key problem how to determine a crawl period based on website traffics.

Solution: Buy 100k IP.

Designing ads counting mechanisms.

2. User based (Crowd-sourcing)

pros: 1.privacy 2.Low traffic. 3.SEM 4.Distributed computation

5. Remove the need to do redirect cloaking detection, leveraging the feature that the end goal of attackers is to reach user

6. could decide crawl period passively based on user clicks, data received are based on real user’s clicks, say, website traffic

cons: user incentives.

Solution: Plugin to detect suspicious websites.

API

7. Conclusion

Current solution, multiple pass of data, doesn’t work, and hard to handle various types of cloaking.

Our model works because …

**Show theory of Simhash? EMD? Which distance measure? Which similarity?**

**Urgent!!!!!! Similarity?**

Experiments TODO:

1. Run cross validation, get parameter, detect, label and summarize
2. Run experiment on spammy search, hot search, spammy ads, hot ads
3. Label and categorize our detected dataset
4. ~~Show the ability of simhash (text, dom), separately, the number of changes maps to how many bits of change in simhash~~
5. **The speed of simhash computing weiren**
6. **Implement simhash computing in plugin? weiren**
7. Generate plots and write story