Welcome to the class of

Advanced Topics in Information Retrieval !

Min ZHANG （张敏）

z-m@tsinghua.edu.cn

Tea Time

The Evolution of Solid

Zhen Wang 王振

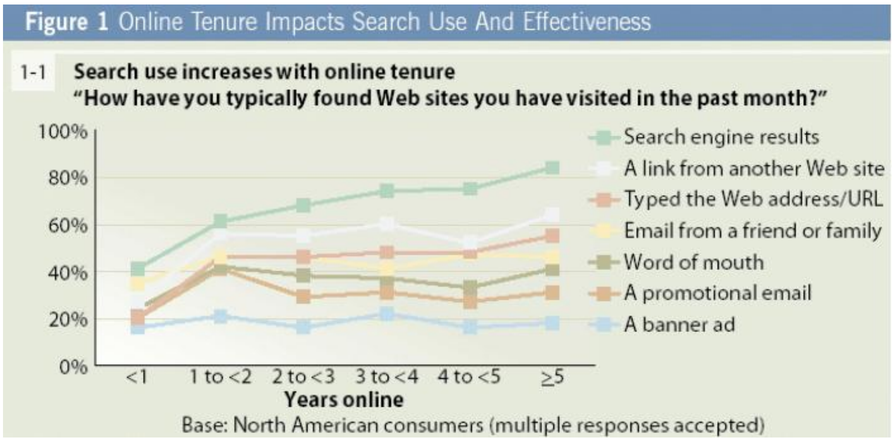
Part II. Search Engine Techniques

- Link-based analysis
- Challenging Topics
- User Modeling

CHALLENGES:

I. SPAM

Challenges I – Spam



SOURCE: Forrest Research

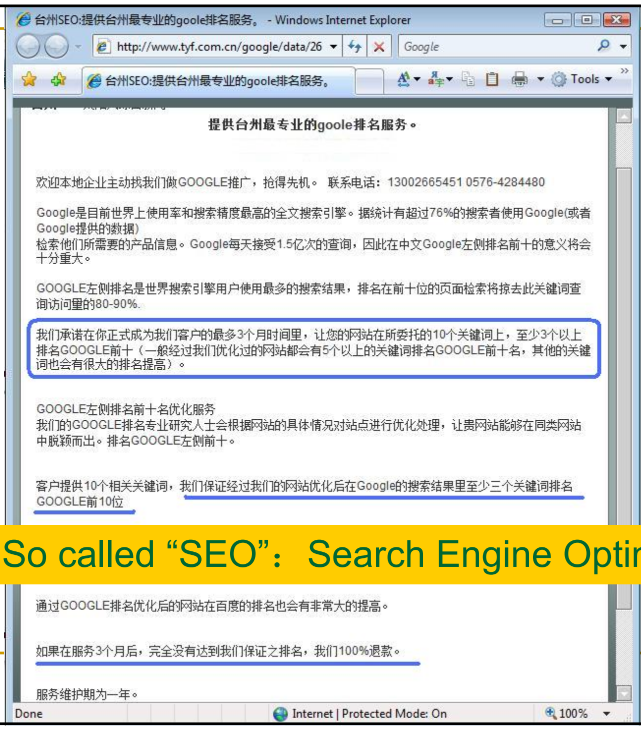
- **70-80% users use SEs to find sites**

Challenges I – spam

- Cause
  - 85% queries only request the first one/two result pages *
  - **Users follow search results**
  - **Money follow users**
  - **Spam follow money**

- Commercially-oriented web sites – be ranked in top10
- Example

* According to different study report



So called “SEO”：Search Engine Optimization

7

Challenges I – Types of Web Spams

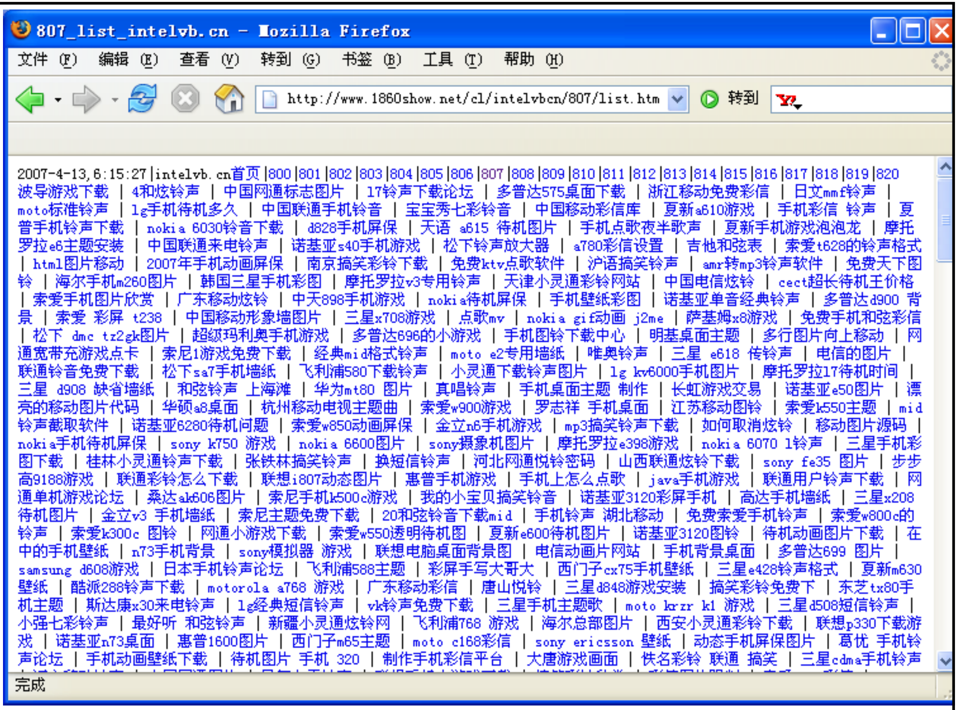
- Search engine spam

  (typical, but would be more)
  - Text-based
  - Link-based
  - Cloaking

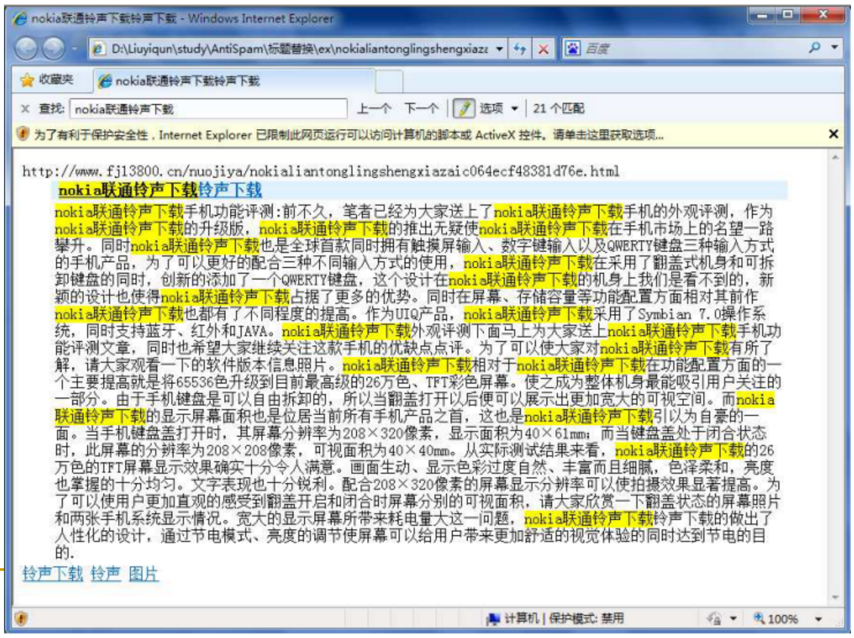
Challenges I – spam type I

- **Text-based**
  - To concentrate on a small set of keywords
    - e.g. at the bottom of the documents
    - Small font, invisible (with the background color)
  - To Increase the number of keywords
    - Include (subset of) a dictionary
    - Add text on a different topic (e.g. porn site – add famous people)

Challenges in Web IR 9



- Keyword weaving/replacing/stitching spam

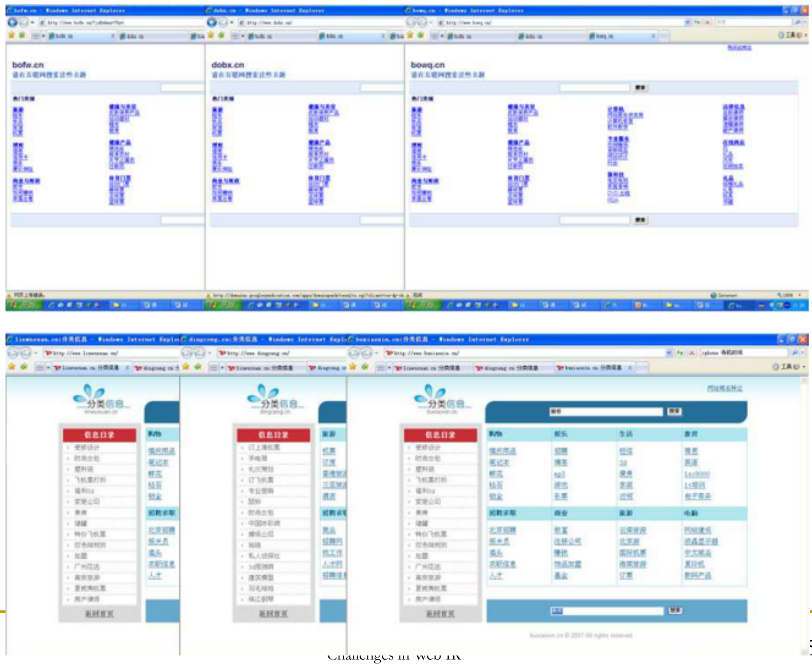


Challenges I – spam type I

- **Text-based**
  - To concentrate on a small set of keywords
    - e.g. at the bottom of the documents
    - Small font, invisible (with the background color)
  - To Increase the number of keywords
    - Include (subset of) a dictionary
    - Add text on a different topic (e.g. porn site – add famous people)
- But there're some ways to detect:
  - Use visible content
  - Detect keyword density

Challenges I – spam type II

- Link-based
  - To put a ***link farm*** at the bottom of every page
    - Have thousands of links, including multiple links to the same page
  - *Doorway pages*
    - Consist entirely of links
  - Link exchange
  - Mailing lists
  - Guestbooks
- Hurt link analysis sensitive to the absolute # of links



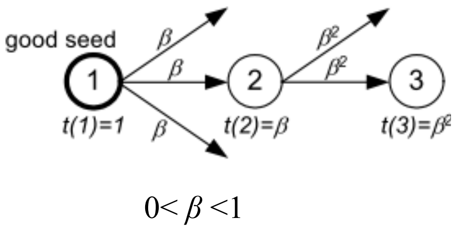
Challenges I – spam type II

- Link-based
  - To put a ***link farm*** at the bottom of every page
    - Have thousands of links, including multiple links to the same page
  - *Doorway pages*
    - Consist entirely of links
  - Link exchange
  - Mailing lists
  - Guestbooks
- Hurt link analysis sensitive to the absolute # of links
- **But**
  - You can find trusted parties and only trust links from them
  - TrustRank

Review: Trust Attenuation

Type I:

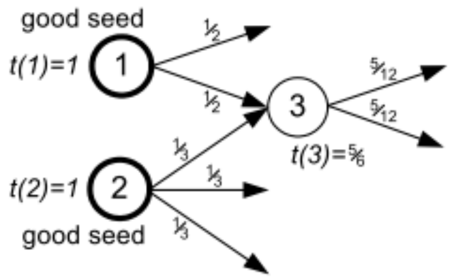
- Trust dampening



$0 < \beta < 1$

Type II:

- Trust splitting



Review: TrustRank Algorithm

function TrustRank
input
- $\mathbf{T}$ transition matrix
- N number of pages
- L limit of oracle invocations
- α_B decay factor for biased PageRank
- M_B number of biased PageRank iterations

output
- $\mathbf{t}^*$ TrustRank scores

α_B: generally set to ~0.85,

```
begin
        // evaluate seed-desirability of pages
(1)     s = SelectSeed(...)
        // generate corresponding ordering
(2)     σ = Rank({1,...,N}, s)
        // select good seeds
(3)     d = 0_N
        for i = 1 to L do
            if O(σ(i)) == 1 then
                d(σ(i)) = 1
        // normalize static score distribution vector
(4)     d = d/|d|
        // compute TrustRank scores
(5)     t* = d
        for i = 1 to M_B do
            t* = α_B · T · t* + (1 − α_B) · d
        return t*
end
```

Review: Link Based Detection

- Trust Rank Result
  - Precision and recall, PageRank vs. TrustRank

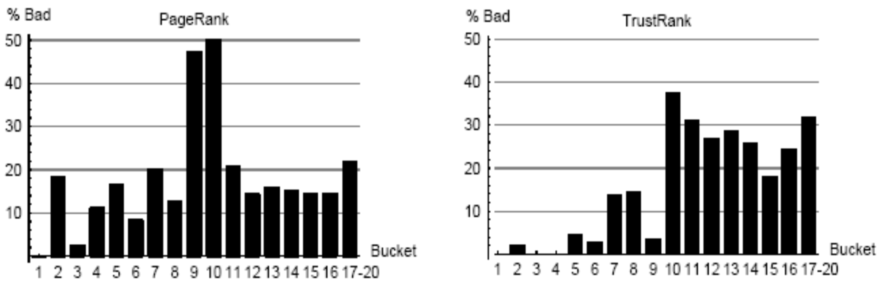


Figure 10: Bad sites in PageRank and TrustRank buckets.

Rank pages with scores. Rank 1: highest score.
Bucket setting: The sum of scores in each bucket is equal.

Z. Gyongyi, et al. Combating web spam with trustrank. In *VLDB '04, 576–587, 2004*.

Challenges I – spam type III

- Cloaking, Honey Pot
  - Serving SE crawlers different content of the page to general users
  - Some is used with the intent to "*help*" SE
    - Giving them an easily digestible, text-only version of a page
    - To provide link-based access to a database that normally only accessible via ***forms***

Challenges I – spam

- There are more… And will be more …
- **An ever-lasting mission**
  - Good news for anti-spam engineers!
  - Bad news for Web users / search engines
- Problems:
  - The anti-spam techniques are generally **type-specified**
  - It takes **a long time** for anti-spam engineers to notice the appearance of one kind of spam.
  - **“道高一尺，魔高一丈”** (“the villains can always outsmart.”)

A Promising Idea:
Web Spam Detection Based On User Behavior Analysis

A Promising Idea: Web Spam detection based on user behavior analysis

- Who will notice the existence of a new spam page at the first time? **—— The Users!**
- The behavior evidences/features we could use
  - How many user visits are oriented from search engine?
  - How many users will follow links on the page?
  - How many users will not visit the site in the future?
  - How many user visits are oriented by hot keyword searches?
  - How many pages does a certain user visit in the site?
  - How many users visit the site?
  - …

Ref: Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru. Identifying Web Spam with User Behavior Analysis. **The Fourth International Workshop on Adversarial Information Retrieval on the Web**.

Challenges I – spam: some progress

- Identifying Web Spam based on User Behavior Analysis
  - Direct information
  - Quick response
  - Ability to find new spam types
- Propose the behavior features
  - How many user visits are oriented from search engine?
  - How many users will follow links on the page?
  - How many users will not visit the site in the future?
  - How many user visits are oriented by hot keyword searches?
  - ……

Ref: Yiqun Liu, Rongwei Cen, Min Zhang, Shaoping Ma, Liyun Ru. Identifying Web Spam with User Behavior Analysis. **The Fourth International Workshop on Adversarial Information Retrieval on the Web**. 2008.4.

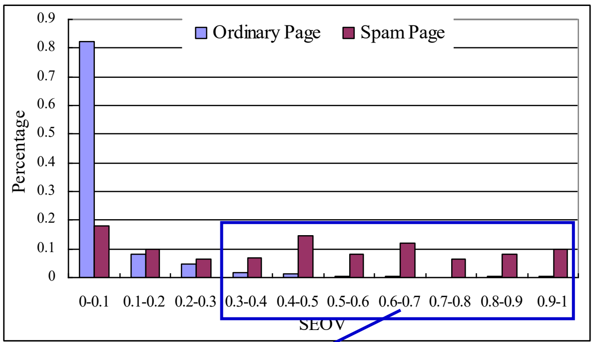
Web Spam detection based on user behavior analysis

- Who will notice the existence of a new spam page at the first time? **—— The Users!**
  - The wisdom of crowds
  - Social annotation?
    - noisy, lack of long-term interest, quality control, anti-(anti-spam)-spam
  - → Web access logs
    - Collected by a commercial search engine
    - sampled log data of 57 days
    - 2.74 billion user clicks in 800 million Web pages

Challenges in Web IR 25

User-behavior Features

- *SEOV* distribution (Search engine oriented visiting rate)

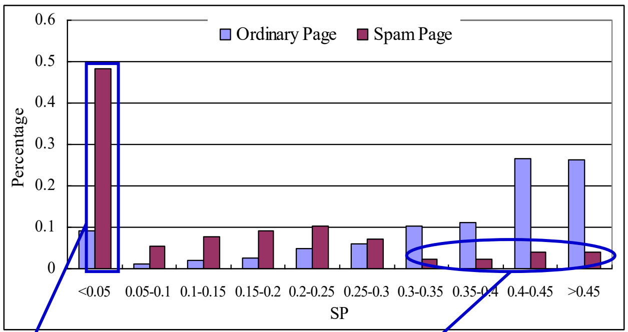


Most spam pages' visits by the users are mainly from search engines

Challenges in Web IR 27

User-behavior Features

- *SP* (Source Page) rate distribution

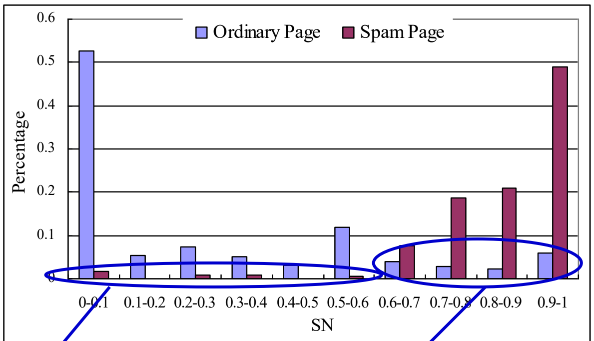


User clicks hyperlink on some spam page, too. (users may be cheated by anchor texts)

Half of spam pages have very small *SP* values

User-behavior Features

- *SN* (Short-time Navigation) rate distribution (*N* = 3)

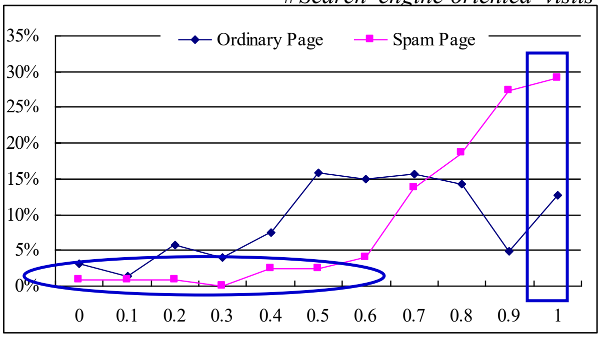


A number of ordinary pages also receive few UVs in a session. (redirection sites, low-quality sites, …)

Few spam pages are visited more than twice in a session

User-behavior Features

- Hot key word oriented visiting rate (*HKOV* rate)
  - Observation: many spam pages are lead by hot key word
  - Definition: $HKOV(p) = \frac{\#\,Hot\,keyword\ \ oriented\ \ visits}{\#\,Search\ \ engine\ \ oriented\ \ visits}$



Sparse data problem: some page have few UVs

Less than 15% spam pages have low HKOV rate → Most spam pages have high HKOV rate

Detection algorithm

- Problem:
  - Uniform sampling of negative examples (pages which are not spam) is difficult
- Solution:
  - Learning from positive examples (Web spam) and unlabeled data (Web corpus)
  - Calculate the possibility of a page *p* being Web spam using user behavior features

$$P(p \in Spam \mid SEOV(p), SP(p), SN(p))$$

Experimental Results

- Experiment setup
  - Training set:
    - 802 spam sites
    - Collected from the hottest search queries' result lists
  - Test set:
    - 1564 Web sites annotated with whether it is spam or not
    - 345 spam, 1060 non-spam, 159 cannot tell
    - Percentage of spam is higher than the estimation given by *Fetterly et al* and *Gyöngyi et al* .
      - we only retain the sites which are visited >=10 times

Web Spam detection based on user behavior analysis – Exp. Results

- How to evaluate the performance
  - Focus: find the recently-appeared spam types

  (especially for those that have passed the SEs' filtering.)

  1: Whether the spam candidates identified by this algorithm are really Web spam. (effectiveness)

  2: Whether this algorithm detect spam types in a timely manner (compare with SE's detection procedure). (timeliness)

  3: Whether the approach is dependent to spam types (type-specific or type-free)?

Experimental Results

- Detection performance (effectiveness)
  - Whether the top-ranked candidates are Web spam
  - 300 Pages with the highest *P(Spam)* values
    - Spam detection precision: 94.0%
  - Many spam types can be identified. (type independent)

Page Type	Percentage
Non-spam pages	6.00%
Web spam pages (Content spamming)	21.67%
Web spam pages (Link spamming)	23.33%
Web spam pages (Other spamming)	10.67%
Pages that cannot be accessed	38.33%

Experimental Results

- Detection performance (timeliness)
  - Check the result spam list (723 spam sites) in commercial search engine results
  - Top-ranked spam candidate sites
    - SE indexes 34 million pages from these 723 sites in March 06, 2008
    - 59 million pages are indexed in March 26, 2008

These spam are not detected by the search engine. And the search engine spent lots of resources on these useless pages

Progresses on web spam detection

- User behavior based spam page discovery
- Search Engine Click Spam Detection

More progress: Search Engine Click Spam Detection Based on Bipartite Graph Propagation

ip	query	isclick	clicked url
1323188204	9999pp.com	0	
1323188204	9999pp.com	1	http://369ii.com/
1323188204	9999pp.com	1	http://369ii.com/
1323188204	9999pp.com	1	http://369ii.com/
1323188204	9999pp.com		
1323188204	9999pp.com		
1323188204	9999pp.com		
1323188204	9999pp.com		
1323188204	9999pp.com		
1323188204	9999pp.com		

ip	query	isclick	clicked url
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	
1323188327	DNF gaming community	0	

Xin Li, Min Zhang, Yiqun Liu, Shaoping Ma, Yijiang Jin, Liyun Ru, Search Engine Click Spam Detection Based on Bipartite Graph Propagation, WSDM2014

Click spam detection

- Main idea: Bipartite Graph Propagation on User Session Behaviors

- Define 6 kinds of user actions
  - *Qi:* Submit a query, *i* is used to distinguish different queries
  - *Wi:* Click on web results, *i* is used to distinguish different results
  - *Oi:* Click on sponsored results, *i* is used to distinguish diff. res.
  - *N:* Load a new page, including click on next page, previous page and turning to a specific page number
  - *T:* Scroll the page
  - *Ai:* Other clicks, including click on tabs like "Video", "Music" and so on

Cheating modes

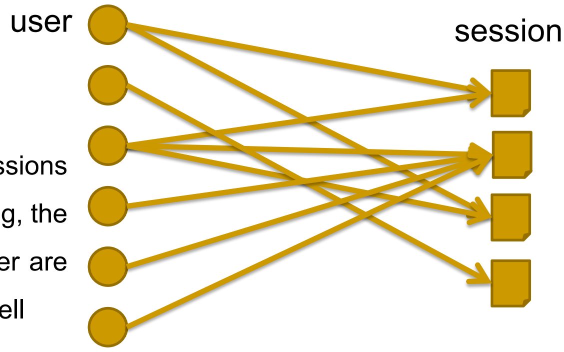
- 5 cheating modes are discovered
  - (Q Ai)*, (Qi T)*, (Qi)*, Q(Wi)*, Q(Ai)*

(Q Ai)* Example: (Q0,0) (A0,0) (Q1,0) (A0,0) (Q2,1) (A0,0) (Q3,1) (A0,0) (Q4,1) (A0,0)

time	query	isclick	_tag	_clickedurl
1323262382	China	0	--	--
1323262382	China	1	--	http://www.zzyzzy.cn/html/322.html
1323262383	Shanghai	0	--	--
1323262383	Shanghai	1	--	http://www.zzyzzy.cn/html/151.html
1323262386	Software	0	--	--
1323262386	Software	1	--	http://www.zzyzzy.cn/html/188.html
1323262389	Summit	0	--	--
1323262389	Summit	1	--	http://www.zzyzzy.cn/html/56.html
1323262391	Industry	0	--	--
1323262391	Industry	1	--	http://www.zzyzzy.cn/html/220.html

Basic assumption

If a certain number of sessions a user makes are cheating, the other sessions of this user are likely to be cheating as well



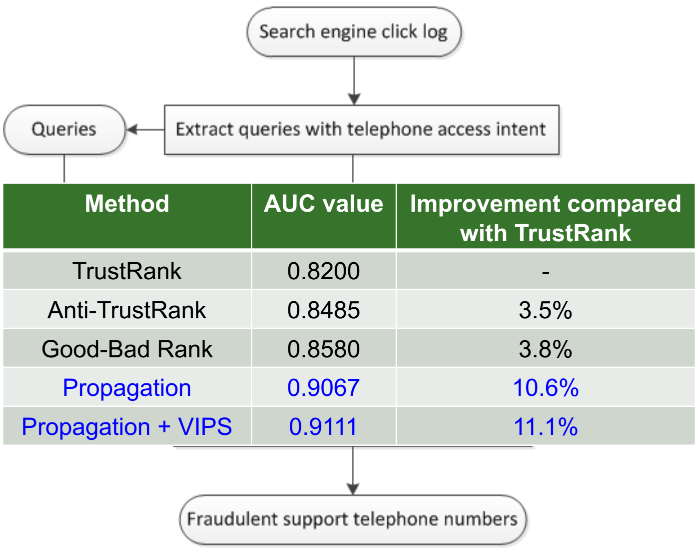
	click spam ratio	precision
Baseline	1.7%	90%
User-session	2.1%	97%
Pattern-session	2.6%	97%
User-session + Pattern-session	2.8%	97%

Progresses on web spam detection

- User behavior based spam page discovery
- Search Engine Click Spam Detection
- Fraudulent Support Telephone Num. Identification

More progress: Fraudulent Support Telephone Num. Identification Based on Co-occurrence Info. on the Web

Xin Li, Yiqun Liu, Min Zhang, Shaoping Ma, Fraudulent Support Telephone Number Identification Based on Co-occurrence Information on the Web, AAAI2014

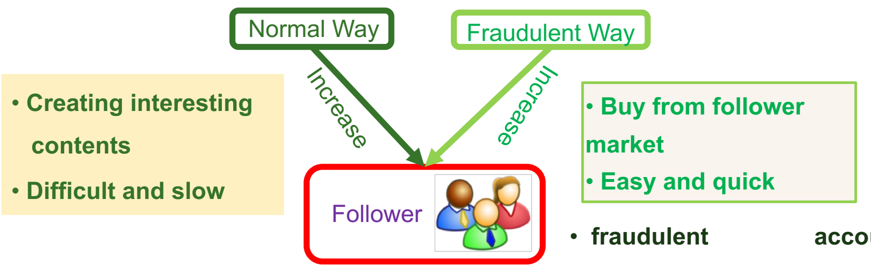


Method	AUC value	Improvement compared with TrustRank
TrustRank	0.8200	-
Anti-TrustRank	0.8485	3.5%
Good-Bad Rank	0.8580	3.8%
Propagation	0.9067	10.6%
Propagation + VIPS	0.9111	11.1%

Progresses on web spam detection

- User behavior based spam page discovery
- Search Engine Click Spam Detection
- Fraudulent Support Telephone Num. Identification
- Detecting Spams on Crowdturfing Following Activities in Microblog

Background



- A new rise of malicious following activity —
  - Oriented from normal users
  - Willing to join the follower market and get rewards
- fraudulent accounts (Almaatouq 2014)
- compromised accounts (Stringhini 2012)

Voluntary Follower Properties

	U_v (volower)	U_n (normal)
#Days since registration	882.4	934.4
#Message	519.1	588.2
#Original message	363.3	353.0
#Follower	251.1	288.6
#Followee	908.6	317.1
#Interaction per message	2.33	1.43

We registered 3 accounts on crowdsourcing platform ZhuBaJie, and have bought in total 3000 volowers.

DetectVC Algorithm

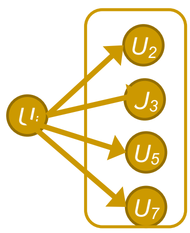
- Seed set $U_s \subset U_v$
  - Randomly selected from labeled volowers
  - With initial spam score 1
- Each user u will receive two scores $P_v(u)$ and $P_c(u)$
  - Indicate u's possibilities of being a volower and a customer

- Volower Possibilities

$$P_v^{(k)}(u_i) = \sum_{u_j:(u_i,u_j)\in E} P_c^{(k)}(u_j)$$

- Matrix form

$$P_v^{(k)} = W P_c^{(k)}$$



- Customer Possibilities

$$P_c^{(k)}(u_i) = \sum_{u_j:(u_j,u_i)\in E} P_v^{(k-1)}(u_j)$$

- Matrix form

$$P_c^{(k)} = W P_v^{(k-1)}$$

[IJCAI 2016] Pay Me and I'll Follow You: Detection of Crowdturfing Following Activities in Microblog Environment

Experimental results (F-measure)

- Volower detection

	Original	With $P_v(u)$
DetectVC	0.844	–
[Yang et al., 2012]	0.715	0.850 (+13.5%)
[Egele et al., 2013]	0.807	0.863 (+5.6%)
[Lee et al., 2014]	0.832	0.895 (+6.3%)
[Aggarwal et al., 2015]	0.825	0.868 (+4.3%)

- Customer detection

	Original	With $P_c(u)$
DetectVC	0.860	–
[Stringhini et al., 2013]	0.805	0.864 (+5.9%)
[Aggarwal et al., 2015]	0.837	0.907 (+7.0%)

Summary

- Three types of Spams
  - Text, Link, Cloaking
- Classical link-based anti-spam technique: TrustRank
- A promising approach
  - User behavior based spam page discovery
  - Search Engine Click Spam Detection
  - Fraudulent Support Telephone Num. Identification
  - Detecting Spams on Crowdturfing Following Activities in Microblog

Increasingly complex spam techniques
vs.
growing complex models
→ *Long lasting battle!*