Detecting Semantic Cloaking on the Web

Baoning Wu and Brian D. Davison

Lehigh University, USA
WWW 2006

LEHIGH UNIVERSITY

Outline

- Motivation
- Proposed Solution
- Evaluation
- Conclusion

How search engine works

- Crawler downloads pages from the web.
- Indexer puts the content of the downloaded pages into index.
- For a given query, a relevance score of the query and each page that contains the query is calculated.
- Response list is generated based on the relevance scores.

Motivation

- Cloaking occurs when, for a given URL, different content is sent to browsers versus that sent to search engine crawlers.
- Some cloaking behavior is acceptable.
- Semantic cloaking (malicious cloaking) is the type of cloaking with the effect of deceiving search engines' ranking algorithms.

Yahoo! Weather - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://we Go

Getting Started Latest Headlines http://oascentral.we...

Google yahoo weath Search PageRank Check AutoLink »

Yahoo! My Yahoo! Mail Make Yahoo! your home page

YAHOO! NEWS Weather

Sign In
New User? Sign Up

Hello, Guest

Yahoo! Weather

ADVERTISEMENT

Your weather and...

Check Pollen Levels
- Mosquito Activity
- Weather & Your Pet
- Gardening Forecast
- It's Grilling Out Season!
- Download Desktop Weather

The Weather

Find Weather

Enter city or zip code:

Search

(e.g. 95136 or San Francisco or San Francisco, CA)

Browse by location:

- United States
- Africa
- Antarctica
- Arctic
- Asia
- Caribbean
- Central America
- Europe
- Latin America
- Mediterranean

Weather Forecasts now delivered by em

choose your city or location

News and Features

Driving Rain Eases in Floode
Driving rains that caused the v
England since the 1930s final
Tuesday, but washed-out road
dam breaks prevented many p
to their homes.

Transferring data from adt.m7z.net...

Yahoo! Weather - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://cc.r Go

Getting Started Latest Headlines http://oascentral.we...

Google yahoo weath Search PageRank Check AutoLink »

page in the results to your recent query. This is not necessarily the most recent version of the page - to see the most recent version of this page, visit the page on the web.

msn MSN is not affiliated with the content nor parties responsible for the page displayed below.

Hello, Guest

Yahoo! Weather

Your weather and...

Weather Video

Henry's Severe Weather Outlook - (5/12/2006 4:46 PM)

Friday's Weather Briefing - (5/12/2006 2:21 PM)

More Video...

My Weather Edit

Chicago, IL 42...45 F

Find Weather

Enter city or zip code:

Search

(e.g. 95136 or San Francisco or San Francisco, CA)

Browse by location:

- United States
- Africa
- Antarctica
- Arctic
- Asia
- Caribbean
- Central America
- Europe
- Latin America
- Mediterranean

Weather Forecasts now delivered by em

choose your city or location

News and Features

Typhoon toll rises to 23 in Philippines
A typhoon whipped through the northw
on Saturday, killing at least 23 people
leaving more than 10,000 people stran
services were suspended.

Done

Semantic cloaking example: keywords only sent to crawler

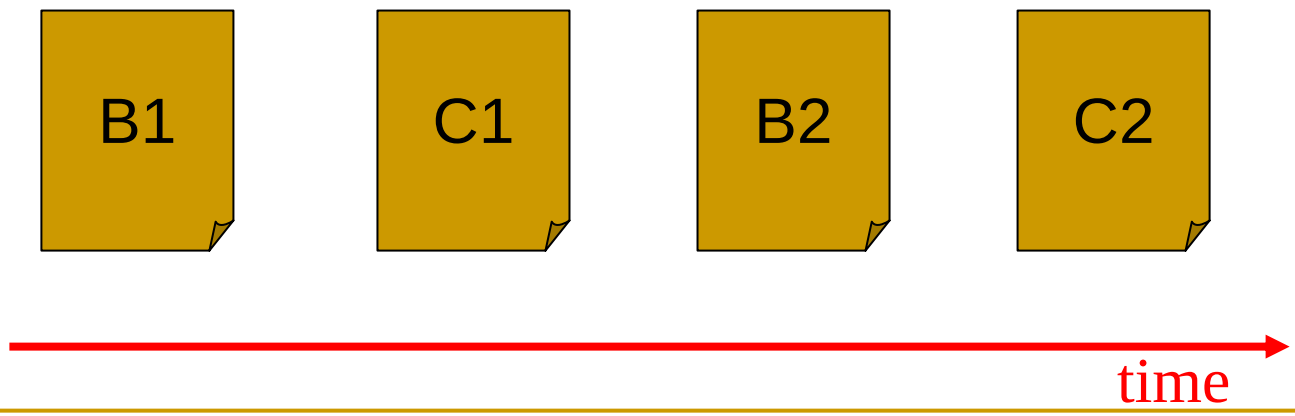
- game info, reviews, game reviews, previews, game previews, interviews, features, articles, feature articles, game developers, developers, developer diaries, strategy guides, game strategy, screenshots, screen shots, game screenshots, game screen shots, screens, forums, message boards, game forums, cheats, game cheats, cheat codes, playstation, playstation, dreamcast, Xbox, GameCube, game cube, gba, game, advance, software, game software, gaming software, files, game files, demos, game demos, play games, play games online, game release dates, Fargo, Daily Victim, Dork Tower, classics games, rpg, …….

Task

- To build an automated system to detect semantic cloaking
  - based on the several copies of a same URL from both browsers' and crawlers' perspectives

How to collect data: UserAgent

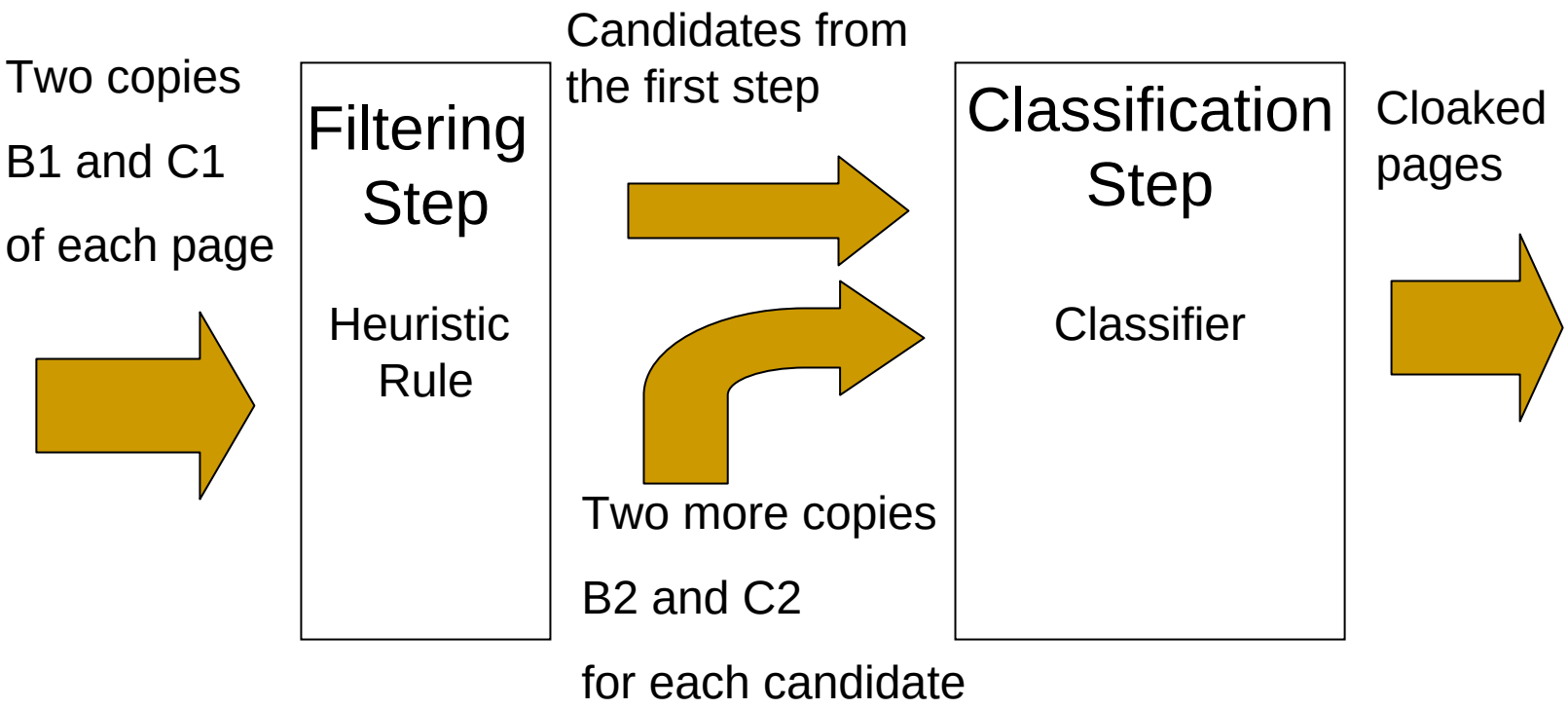
- Browser:
  - Mozilla/4.0 (compatible; MSIE 5.5; Windows 98)
- Crawler:
  - Googlebot/2.1 (+http://www.googlebot.com/bot.html)



Outline

- Motivation
- Proposed Solution
- Evaluation
- Conclusion

Architecture



Filtering Step

- To eliminate pages that do not employ semantic cloaking.

- Heuristic rules are used.
- For example, a rule might be:
  - to mark any page as long as the copy sent to the crawler contains a number of dictionary terms that don’t exist in the copy sent to the browser.

Classification Step

- A classifier is used.
  - E.g., Support Vector Machines, decision trees
- Operating on features including those from
  - Individual copies
  - Comparison of corresponding copies.

Features from individual copies

- Content-based
  - Number of terms in the page
  - Number of terms in the title field
  - Whether frame tag exists
  - ……
- Link-based
  - Number of links in the page
  - Number of links to a different site
  - Ratio of number of absolute links to the number of relative links.
  - ……

Features for corresponding copies

- Whether the number of terms in the keyword field of C1 is bigger than the one of B1
- Whether the number of links in C2 is bigger than the one in B2
- Number of common terms in C1 and B1
- Number of links appearing only in B2, not in C2
- ……

Building the classifier

- Joachims' SVM^{light} is used.
- 162 features extracted for each URL.

- Data set:
  - 47,170 unique pages (top 200 responses for popular queries).
  - We manually labeled 1,285 URLs, among which 539 are positive (semantic cloaking) and 746 are negative.

Training the classifier

- 60% of positive and 60% of negative examples are randomly selected for training and the rest are used for testing.

- Performance (average of five runs)
  - Accuracy: 91.3%
  - Precision: 93%
  - Recall: 85%

Discriminative features

- Whether the number of terms in the keyword field of the HTTP response header for C1 is bigger than the one for B1
- Whether the number of unique terms in C1 is bigger than the one in B1
- Whether C1 has the same number of relative links as B1
- ……

Outline

- Motivation
- Proposed Solution
- Evaluation
- Conclusion

Detecting semantic cloaking

- We used pages listed in dmoz Open Directory Project to demonstrate the value of our two-step architecture of detecting semantic cloaking.
- ODP 2004 gives us 4.3M URLs
  - Two copies of each of these URLs are downloaded for the filtering step.

Filtering step

- Rule: if the copy sent to crawler has more than three unique terms that do not exist in the copy sent to browser, or vice versa, the URL will be marked as a candidate.
- The filtering step marked 364,993 pages (4.3M pages in total) as candidates.
- All semantic cloaking of significance is marked.

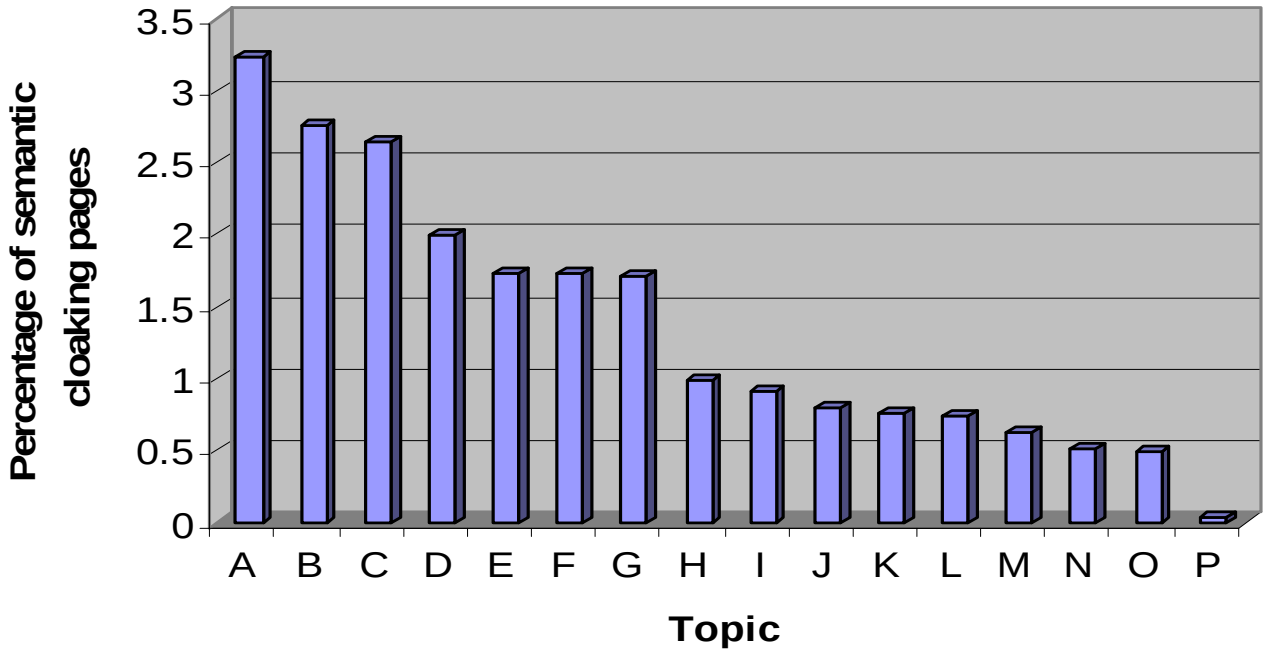
Classification results

- For each of these 364,993 pages, two more copies are downloaded.
- The classifier (trained on the earlier data set) marked 46,806 pages as utilizing semantic cloaking.
- 400 random pages are selected from the 364,993 pages for manual evaluation.
  - Accuracy 96.8%
  - Precision 91.5%
  - Recall 82.7%

Semantic cloaking pages in DMOZ

- 46,806 * 0.915 / 0.827 = 51,786
- 4.3M pages in total
- So, more than 1% of all pages within ODP are expected to utilize semantic cloaking

Semantic cloaking pages in ODP



A. Arts
B. Games
C. Recreation
D. Sports
E. Home
F. Society
G. Kids&Teens
H. Computers
I. Health
J. Science
K. Regional
L. World
M. Shopping
N. Reference
O. Business
P. News

Outline

- Motivation
- Proposed Solution
- Evaluation
- Conclusion

Discussion & Conclusion

- An automated system to detect semantic cloaking is possible!
- What if the spammers read this paper?
  - Need to be less ambitious to bypass the filtering step
  - Difficult to avoid all the features used in the classification step
- Future work
  - Better heuristic rules for the filtering step
  - More features to improve recall
  - IP-based semantic cloaking

Thank You!

- Baoning Wu
- baw4@cse.lehigh.edu
- http://wume.cse.lehigh.edu/