Improving Cloaking Detection Using Search Query Popularity and Monetizability

Kumar Chellapilla and David M Chickering
Live Labs, Microsoft
Weight Loss & Slimming Made Simple

"Lose Weight Now! -- the Easy Way...

"My Friends Laughed When I Bought Your Diet System... But When I Lost 10 Sizes...

Before SunSil Diet After SunSil Diet

Now 10 Sizes Smaller

I Lost 10 Sizes and I never stopped eating. Besides going from size 24 to size 14 (and still losing) I've found a safe, inexpensive, easy way of losing weight. I never feel hungry or deprived.

I also threw away all the Tums® and Maalox®. Diet fads come and go, but your SunSil Diet Program changed my life forever. This is definitely not a fad. I feel great.

Ms. Charley Miller

And Yes! --- It's True...

Thousands Who Never Thought They Could... Now Lose Weight ...and Stay Slim -- with this Physician's Proven Diet System

* Safe! -- A Drug-Free Diet System
* Proven! -- Hundreds of Success Stories

• Want to lose weight?

• Browser View
Weight Loss & Slimming Made Simple

"Lose Weight Now! -- the Easy Way...

"My Friends Laughed When I Bought Your Diet System... But When I Lost 10 Sizes...

Before Swilson Diet  After Swilson Diet

Now 10 Sizes Smaller

I lost 10 sizes and I never stopped eating. Besides going from size 24 to size 14 (and still losing) I've found a safe, inexpensive, easy way of losing weight. I never feel hungry or deprived.

I also threw away all the Tums® and Metamucil®. Diet fads come and go, but your Swilson Diet Program changed my life forever. This is definitely not a fad. I feel great.

Ms. Charley Miller

And Yes! -- It’s True...

'Thousands Who Never Thought They Could... Now Lose Weight...and Stay Slim -- with this Physician’s Proven Diet System"

• Safe! -- A Drug-Free Diet System

Turbo charge your weight loss with this physicians proven weightloss diet.

Regular exercise, changing food choices, and regular check-ups with your doctor are great ways to increase your weight loss success.

If's True! Holiday Weight Gain. Price deterioration includes being very orderly, thoughtful, very impatient of hot weather, easy to persuade, having restless, silly, or blank head, scoldom, insatiable of egg foods, very self-critical and critical of others, determined and impetuous.

The 5 best foods for healthy weight loss and fast slimming.

Low-density lipoproteins (LDL) carry cholesterol from the liver to the rest of the body. When there is too much in the body, it is deposited in the coronary arteries. This is not good. A build-up of cholesterol in our arteries could prevent blood from getting to part of our heart. That means that our heart won’t get the oxygen and nutrients it needs, which could result in heart attack, stroke, or sudden death. So, if your LDL is higher than your HDL, you’re at a greater risk for developing heart disease. For weight loss success, keep your LDL at 100 or below.

That’s why professional athletes and bodybuilders seem overweight according to the BMI body mass index, even though their percentage of body fat is low.

Obesity Speeds Heart Attacks.

Why You Must Be Careful: * Eat plenty of vegetables, fruits, and grains products like bread and rice. Weight Loss Best: * Eat plenty of vegetables, fruits, and grains products like bread and rice. Weight Loss Best: * Eat plenty of vegetables, fruits, and grains products like bread and rice. Weight Loss Best: * Eat plenty of vegetables, fruits, and grains products like bread and rice. Weight Loss Best: "No one is ever too thin!" The following is a weight loss plan that is designed to help you lose weight safely and effectively.

Good News:

Without knowing specific limits of high energy intake on days of pregnant women makes losing weight extra easy. Weight loss is healthy. Women should aim for 22% to 25% body fat. If you have a weight problem, consult your doctor. The fast few days of meals can be expensive. These individuals participating in the Omaha plan who were able to stick with their diet for the entire year achieved a weight loss rate of about 8%, the highest of all groups. Weight Loss: * Weight loss is self sufficient. Any kind of food is fine if you made a meal plan that works for you. * Weight loss is self sufficient. Any kind of food is fine if you made a meal plan that works for you. * Weight loss is self sufficient. Any kind of food is fine if you made a meal plan that works for you. * Weight loss is self sufficient. Any kind of food is fine if you made a meal plan that works for you.
• Want to buy blinds for your windows?

• Browser View
Welcome to your online store for beautiful window treatments from Bali, Levolor, Kirsch, Graber, Nank, Hunter Douglas, Comfortex, Skandia, our own National Blind Store brand and many others, all at the lowest prices on the Internet—guaranteed!

How and why are the questions sometimes asked our customer satisfaction policy includes the procedures necessary铝制百叶窗：deliver to our clients the best possible aluminum blinds.

Wood Blinds | wood blinds | wood blind discount

Automatic Blinds

Top quality at the best possible price treat our customers right by always providing the best Bali, how and why are the questions sometimes asked Bali.

Remote Control Blinds | hunter douglas window treatment | hunter douglas window treatment

Bali Cellular Shades

Our customer satisfaction policy includes the procedures necessary aluminum cellular shades. Our customer satisfaction policy includes the procedures necessary aluminum cellular shades.

Vertical Blinds | window shades | wood blinds

Bali Shades

No matter how many times our customer satisfaction policy includes the procedures necessary aluminum cellular shades. Treat our customers right by always providing the best Bali.

Wood Window Blinds | window shades | woven wood shades
We are recovering addicts with over 17 years of experience. We understand.

1-800-559-9503

1-800-559-9503

1-800-559-9503

1-800-559-9503

• Browser View

• Crawler View
We are recovering addicts with over 17 years of experience. We understand!

- The Holistic Drug and Alcohol Rehab Center offers a unique approach to addiction recovery at an affordable price. Our approach to drug and alcohol addiction treatment has 3 components: addressing the mind, body, and spirit of the individual. We are committed to helping each client achieve their specific goals. We combine the best of all traditional forms of therapy and treatment with the most effective holistic therapies available today.

- Our treatment team deals not only with drug and alcohol abuse and addiction, but with the underlying issues that are the contributing factors to the disease, including dual diagnosis treatment. The holistic foundation is built upon the 12 step recovery program. We treat the individual as a whole: mentally, physically, emotionally, and spiritually. Our holistic drug rehab center is tailored to meet the individual's specific needs by treating the person as well as their families. All this is accomplished in a loving and caring professional manner which we feel is the hallmark of quality treatment. Those who are addicted can and do recover. Holistic counseling helps continue the process, offering a one year aftercare program that provides a support system, with a focus on a sober living plan and problem solving. We give people their lives back everyday.

- Call us toll free anytime day or night and get help for yourself or a loved one before it's too late. 24 hour support and assistance is always available. We are here to help and have a lot of experience!

1-800-559-9503

- The Holistic Foundation offers a unique approach to addiction recovery at an affordable price. Our approach to drug and alcohol addiction treatment has 3 components: addressing the mind, body, and spirit of the individual. We are a client-centered facility based on respect and compassion. At the holistic drug rehab center, each program area is individualized to help each client achieve their specific goals. We combine the best of all traditional forms of therapy and treatment with the most effective holistic therapies available today.

- Our treatment team deals not only with drug and alcohol abuse and addiction, but with the underlying issues that are the contributing factors to the disease, including dual diagnosis treatment. The holistic foundation is built upon the 12 step recovery program. We treat the individual as a whole: mentally, physically, emotionally, and spiritually. Our holistic drug rehab center is tailored to meet the individual's specific needs by treating the person as well as their families. All this is accomplished in a loving and caring professional manner which we feel is the hallmark of quality treatment. Those who are addicted can and do recover. Holistic counseling helps continue the process, offering a one year aftercare program that provides a support system, with a focus on a sober living plan and problem solving. We give people their lives back everyday.

- Call us toll free anytime day or night and get help for yourself or a loved one before it's too late. 24 hour support and assistance is always available. We are here to help and have a lot of experience!

1-800-559-9503

- The Holistic Foundation offers a unique approach to addiction recovery at an affordable price. Our approach to drug and alcohol addiction treatment has 3 components: addressing the mind, body, and spirit of the individual. We are a client-centered facility based on respect and compassion. At the holistic drug rehab center, each program area is individualized to help each client achieve their specific goals. We combine the best of all traditional forms of therapy and treatment with the most effective holistic therapies available today.

- Our treatment team deals not only with drug and alcohol abuse and addiction, but with the underlying issues that are the contributing factors to the disease, including dual diagnosis treatment. The holistic foundation is built upon the 12 step recovery program. We treat the individual as a whole: mentally, physically, emotionally, and spiritually. Our holistic drug rehab center is tailored to meet the individual's specific needs by treating the person as well as their families. All this is accomplished in a loving and caring professional manner which we feel is the hallmark of quality treatment. Those who are addicted can and do recover. Holistic counseling helps continue the process, offering a one year aftercare program that provides a support system, with a focus on a sober living plan and problem solving. We give people their lives back everyday.

- Call us toll free anytime day or night and get help for yourself or a loved one before it's too late. 24 hour support and assistance is always available. We are here to help and have a lot of experience!

1-800-559-9503

- The Holistic Foundation offers a unique approach to addiction recovery at an affordable price. Our approach to drug and alcohol addiction treatment has 3 components: addressing the mind, body, and spirit of the individual. We are a client-centered facility based on respect and compassion. At the holistic drug rehab center, each program area is individualized to help each client achieve their specific goals. We combine the best of all traditional forms of therapy and treatment with the most effective holistic therapies available today.

- Our treatment team deals not only with drug and alcohol abuse and addiction, but with the underlying issues that are the contributing factors to the disease, including dual diagnosis treatment. The holistic foundation is built upon the 12 step recovery program. We treat the individual as a whole: mentally, physically, emotionally, and spiritually. Our holistic drug rehab center is tailored to meet the individual's specific needs by treating the person as well as their families. All this is accomplished in a loving and caring professional manner which we feel is the hallmark of quality treatment. Those who are addicted can and do recover. Holistic counseling helps continue the process, offering a one year aftercare program that provides a support system, with a focus on a sober living plan and problem solving. We give people their lives back everyday.

- Call us toll free anytime day or night and get help for yourself or a loved one before it's too late. 24 hour support and assistance is always available. We are here to help and have a lot of experience!

1-800-559-9503

- The Holistic Foundation offers a unique approach to addiction recovery at an affordable price. Our approach to drug and alcohol addiction treatment has 3 components: addressing the mind, body, and spirit of the individual. We are a client-centered facility based on respect and compassion. At the holistic drug rehab center, each program area is individualized to help each client achieve their specific goals. We combine the best of all traditional forms of therapy and treatment with the most effective holistic therapies available today.

- Our treatment team deals not only with drug and alcohol abuse and addiction, but with the underlying issues that are the contributing factors to the disease, including dual diagnosis treatment. The holistic foundation is built upon the 12 step recovery program. We treat the individual as a whole: mentally, physically, emotionally, and spiritually. Our holistic drug rehab center is tailored to meet the individual's specific needs by treating the person as well as their families. All this is accomplished in a loving and caring professional manner which we feel is the hallmark of quality treatment. Those who are addicted can and do recover. Holistic counseling helps continue the process, offering a one year aftercare program that provides a support system, with a focus on a sober living plan and problem solving. We give people their lives back everyday.

- Call us toll free anytime day or night and get help for yourself or a loved one before it's too late. 24 hour support and assistance is always available. We are here to help and have a lot of experience!

1-800-559-9503
• A hiding technique
  - Browser: Serve *true* intended content
  - Crawler: Serve content that will rank the page high on search engine

• Web spam
  - Actions intended to *mislead* search engines to rank certain pages higher than they deserve

• Cloaking reduces information reliability, as a result search engines take strict measures against sites that cloak
How do servers cloak?

• Cloaking techniques
  – User-Agent string
    • Crawlers
      – msnbot/1.0 (+http://search.msn.com/ msnbot.htm)
      – Mozilla/5.0 (compatible; Googlebot/2.1;
        +http://www.google.com/ bot.html)
    • Browsers
      – Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)
  – IP based
    • Easily available lists of crawler IPs/ ranges

• IP techniques are quite successful
We study the distribution of cloaking spam over two sets of queries
- Popular Queries
- Monetizable Queries

**Assumption**: Spammers are economically motivated

**Hypothesis**: the more monetizable a query is the more likely that it will be spammed
Motivation behind Web Spam

• Profitability of online businesses
  – Conversion ratios
    • Impression-to-click
    • Click-to-sale
  – Quality of website (features, usefulness, etc)
• Usually increasing raw traffic to site increases revenue and improves profitability
  – Search engine optimization (White hat & Black hat)
  – Web Spam
• Online advertising
  – Advertising keywords
  – Sponsored links are presented separately from organic results
  – Web spam results are inter-mixed with organic results
• Other (non-economic) motivations do exist
  – Google bombs (Negritude-Ultramarine)
• Economic motivations are well known for e-mail spam
Query classes

- **Popular Queries**
  - Search engine query logs
  - Frequency $\propto$ Popularity

- **Monetizable Queries**
  - Search engine ad logs (sponsored links)
  - Frequency of clicks $\propto$ Monetizability
  - Revenue generated $\propto$ Monetizability

- **Not disjoint sets!!**

- Top-5000 queries for study
Popular Query Set

- **Queries**
  - List of top-5000 queries that generated the most traffic
  - MSN Search user query logs
  - Only query ranks are used, their frequencies were discarded

- **Urls**
  - Top-200 search results from MSN Search, Google, and Ask.com
  - $5000 \times 200 \times 3 = 3$ million URLs (not unique)
Monetizable Query Set (5000 Queries)

• Queries
  - List of top-5000 queries that generated the most revenue (PPC) from sponsored ads on a single day
  - MSN Search advertisement logs
  - Only query ranks are used, their raw monetization values were discarded

• Urls
  - Top-200 search results from MSN Search, Google, and Ask.com
  - $5000 \times 200 \times 3 = 3$ million Urls (not unique)
Data sets

- Queries
  - 5000 popular, 5000 monetizable
  - Overlap between the two sets
    - 826 queries (17%)
- Popular URLs
  - 3 million produced 1.49 million unique URLs
- Monetizable URLs
  - 3 million produced 1.28 million unique URLs
- Each URL was processed once for cloaking
- Assumption: Search engines apply anti-spam and URL editing techniques uniformly over the set of queries and URLs
Cloaking Detection

• Extension of technique proposed by Wu and Davison (2005;2006)

• Download **up to** 4 copies of each Url
  - Browser
    • IE user-agent string
    • Up to **2 copies** (B1, B2)
  - Crawler
    • msnbot user-agent string
    • Up to **2 copies** (C1,C2)

• Urls crawled in random order

• Over 2 days
Cloaking Score

- Comparing a pair of documents
- Normalized term frequency difference

\[
D(T_1, T_2) = \frac{|(T_1 \setminus T_2) \cup (T_2 \setminus T_1)|}{|(T_1 \cup T_2)|} = 1 - 2 \frac{|(T_1 \cap T_2)|}{|(T_1 \cup T_2)|}
\]

- \( T_1 \) and \( T_2 \) are sets of terms
- \((T_1 \setminus T_2) = \) set of terms in 1 but not in 2
- Sets can contain repeats
- Normalization by \((T_1 \cup T_2)\) reduces any bias that stems from the size of the web page
**Cloaking Test Procedure**

- **URL**
  - Download URL user-agent: MsnBot (C1)
  - Download URL user-agent: IE (B1)

- **74.7%, 73.1%**
  - Same HTML?
    - YES: Not Cloaked
    - NO: HTML=>Txt

- **13.6%, 13.4%**
  - Same Txt?
    - YES: Not Cloaked
    - NO

- **0.46%, 0.67%**
  - Same Terms? D(C1,B1) = 0
    - YES: Not Cloaked
    - NO

- **8.2%, 9.8%**
  - Download URL again user-agent: MsnBot (C2)
  - Download URL again user-agent: IE (B2)

- **Overall: 3% Failed to download**
  - Cloaked
  - S >= Thld
  - S < Thld: Not Cloaked
• **Processing stages (popular, monetizable)**
  - (C1,B1) Resolved as not cloaking (91.8%, 90.2%)
    - 74.7%, 73.1% resolved (not cloaking) – same HTML
    - 13.6%, 13.4% resolved (not cloaking) – same Txt
    - 0.46%, 0.67% resolved (not cloaking) – same words (incl. freq)
  - 8.2%, 9.8% remain for which (B2,C2) downloaded

• **Normalized term frequency differences**
  - Cloaking: D(C1,B1), D(C2,B2)
    \[ \Delta_D = \min(D(C_1, B_1), D(C_2, B_2)) \]
  - Dynamic: D(C1,C2), D(B1,B2)
    \[ \Delta_S = \max(D(C_1, C_2), D(B_1, B_2)) \]

• **Simple measure of cloaking (threshold \( t \))**
  \[ S = \frac{\Delta_D}{\Delta_S} \]
  \[ 0 < S < \infty \Rightarrow \text{dynamic URLs} \]
  \[ 0 < t < S \Rightarrow \text{cloaking spam} \]
• Dynamic urls
  - 8.2% of popular urls = 122,180 urls
  - 9.8% of monetizable urls = 125,440 urls

• 4000 URLs were randomly chosen
  - 2000 from Popular set (8.2%)
  - 2000 from Monetizable set (9.8%)

• Manually labeled for cloaking spam
Precision and Recall

![Graph showing precision and recall metrics for popular and monetizable categories.]

- **Precision**: 98.5%
- **Recall**: 74.0%
• $F_1$, $F_{0.5}$, and $F_2$ give best $t = 0$ (100% recall)
• Cloaking detection algorithm
  – 98.5% precision (Monetizable)
  – 74.0% precision (Popular)
• % Cloaked urls
  – 9.7% (Monetizable)
  – 6.0% (Popular)
• It is much easier to detect cloaking in monetizable query results
• Monetizable queries are 62% more likely to produce cloaking spam results
Distribution of Cloaked URLs

- Monetizable Queries
- Popular Queries

- X-axis: Sorted Query Rank
- Y-axis: Percentage of total cloaked pages
Distribution of Cloaked URLs

Monetizable Queries
Popular Queries

Independently Sorted Queries
Distribution of Cloaked URLs

Percentage of total cloaked pages vs. Sorted Query Rank

- Blue line: Monetizable Queries
- Red line: Popular Queries

- 2% Queries
- 98% Queries
Distribution of cloaking

• Top 100 (2%) most cloaked queries
  - have 10x as many cloaking URLs in comparison with bottom 4900 queries (98%)

• Very skewed distribution

• An effective way of monitoring and detecting cloaked URLs
  - Start with most cloaked queries (found in this study) and work towards the least cloaked queries
  - True for both Popular and Monetizable Queries
Summary

• Amount of cloaking in search results depends on query properties such as popularity and monetizability
• Improved cloaking detection algorithm
  – High precision for monetizable queries
  – Moderate precision for popular queries
• Focusing on most popular and monetizable queries can produce significant reduction in cloaking spam with minimal effort
Questions?