Multi Level Filtering Undesirable Explicit Material in Website

ICESTI 2015 – Sept, 10-12, 2015

Patra Jasa Bali Resort & Villas
Gunadarma University

• S3, S2, S1 and Professional Program
• Faculties
  1. Computer Science and Information Technology
  2. Industrial Technology
  3. Economic
  4. Civil Engineering and Plan
  5. Psikology
  6. Literature
• Research Organizations
  – Research Organization University and for every Faculty
  – Special Science Group Discussion
  – Lecturer Group Research: Foshema & Scimed
Background
Background

• In Internet many types of websites that provide information of various categories. Now we can access it from any computer or gadget any where and any time.
• Internet users get information such as text, images, sound and video from their visit to a particular website address.
• For our kids or student : Ease of access to the website can cause a negative impact because the number of websites that provide pornographic content.
Fact about Pornography attack the kids

Pornographics Effect to Kids:

1. **Long-term damage**: affects to brain function can cause addictions, immoral attitude, illegal sex behavior, cyberbullying (Eric Owens, West Chester University, 2012).

2. **Aggression**: Violent media content causes child aggression (Anderson 2007).


20% - 30% of Internet pornography consumption is by children of ages 12 - 17.
Fact about Pornography attack the kids

Leisure time on an average day

- Relaxing and thinking (17 minutes)
- Playing games; using computer for leisure (25 minutes)
- Participating in sports, exercise, recreation (19 minutes)
- Reading (18 minutes)
- Socializing and communicating (38 minutes)
- Other leisure activities (17 minutes)
- Watching TV (2.7 hours)

Total leisure and sports time = 5.0 hours

Fact about Pornography attack the kids

Technology Use Guidelines for Children and Youth

<table>
<thead>
<tr>
<th>Developmental Age</th>
<th>How Much?</th>
<th>Non-violent TV</th>
<th>Handheld devices</th>
<th>Non-violent video games</th>
<th>Violent video games</th>
<th>Online violent video games and or pornography</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-2 years</td>
<td>none</td>
<td>never</td>
<td>never</td>
<td>never</td>
<td>never</td>
<td>never</td>
</tr>
<tr>
<td>3-5 years</td>
<td>1 hour/day</td>
<td>check</td>
<td>never</td>
<td>never</td>
<td>never</td>
<td>never</td>
</tr>
<tr>
<td>6-12 years</td>
<td>2 hours/day</td>
<td>check</td>
<td>never</td>
<td>never</td>
<td>never</td>
<td>never</td>
</tr>
<tr>
<td>13-18 years</td>
<td>2 hours/day</td>
<td>check</td>
<td>limit to 30 minutes/day</td>
<td>never</td>
<td>never</td>
<td>never</td>
</tr>
</tbody>
</table>

Please contact Cris Rowan at info@zonein.ca for additional information. © Zone’in February 22, 2014

What’s parents do to facing that problem?

Source: movingtolearn.ca/2014/ten-reasons-why-hand-held-devices-should-be-banned-for-children-under-the-age-of-12

Source: http://www.pewinternet.org/
Fact about Pornography attack the kids

Weekday time use of married women living with young children, by employment status

<table>
<thead>
<tr>
<th>Activity</th>
<th>Not employed</th>
<th>Employed part-time</th>
<th>Employed full-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sleeping</td>
<td>9.0</td>
<td>7.7</td>
<td>8.0</td>
</tr>
<tr>
<td>Household activities</td>
<td>3.5</td>
<td>2.7</td>
<td>1.3</td>
</tr>
<tr>
<td>Caring for household children</td>
<td>3.5</td>
<td>3.2</td>
<td>1.9</td>
</tr>
<tr>
<td>Working &amp; related activities</td>
<td>3.4</td>
<td>-0.1</td>
<td>3.7</td>
</tr>
<tr>
<td>Leisure &amp; sports</td>
<td>3.7</td>
<td>2.8</td>
<td>2.1</td>
</tr>
</tbody>
</table>

Time use on an average work day for employed persons ages 25 to 54 with children

- Sleeping (7.7 hours)
- Eating and drinking (1.6 hours)
- Household activities (1.1 hours)
- Caring for others (1.3 hours)
- Leisure and sports (2.5 hours)
- Other (1.7 hours)
- Working and related activities (8.7 hours)

Total = 24.0 hours

~0 Estimate is approximately zero.

NOTE: Data include all married women, ages 25 to 54, with a child under 6 present in the household. Data include non-holiday weekdays and are annual averages for 2013.


So, actually Our children is almost always alone to facing pornography attack

Problem Identification

• We need automatics application integrated into a web browser as the main gateway to access the information to filter an explicit content in internet for our children.
• The application can prevent users who are not old enough to access pornographic content.
Research Purposes

Images

Classification pornographic content

Crawling URL, meta tag, link, content web homepage

Block access to Website with pornography content
Problem Formulation and Scope

1. How to design an application to classify the pornographic content in websites by applying **tokenization filtering** method for **text mining**, and **face detection and skin detection** for **image analysis**?

2. Is the N-gram algorithm, haar algorithm and skin chromaticity detection can be useful to support the classification process of pornographic content websites such as text and images?
RESEARCH METHODOLOGY
RESEARCH METHODOLOGY

1. Conventional / mechanical filtering
2. Text-based filtering
3. Image-based filtering

Study Literature
1. Journal
2. Proceeding
3. Internet filtering application and Websites

Descriptive Analysis

Analysis and Design Prototype

Codification and System Implementation

Testing

Knowledge Base: user & content type
USER TYPE: Use Case Diagram

Admin

- Login
- Seeing home page
- Fill out the form URL address
- Seeing comparison between tokenization and filtering of meta tag, URL website, text and images content result
- Log out

Seeing meta tag and URL website process result

seeing text content process results

seeing image content process result

General User

- Seeing home page
- Fill out the search form words
- Seeing results of the search word
ANALYSIS
AND
DISCUSSION
The Big Structures of Filtering System

- User
  - Text Database Test
  - Keyword Filtering
    - HTML Analysis
      - Keyword=OK
        - YES
          - Hide Images
          - Extract Images Phase
            - Skin Detection Phase
              - NO
                - Skin Decision = Porno
                - Display the web page
              - YES
                - Hide Images From web page
          - NO
            - Update the Black List Automatically
  - Black List Filtering
    - Domain Database Test
    - URL Database Test
Text & Images Detection

1. Start
   - URL Embed
   - URL Crawling
   - URL Analyze
     - Check Domain Name
       - Table Domain
     - URL Segmentation
     - URL Tokenization
     - URL Filtering

2. N-gram Algorithm
   - N-gram Algorithm Result
     - URL Forward Link
       - Forward Link?
         - Yes
           - Forward Link Crawling
         - No
           - N-gram Algorithm Result
             - Text Mining
               - Text Mining Result
                 - HAAR Algorithm
                   - Website Classification Result
                     - Stop
The Mining Texts and Images Algorithm

- Implement N-Gram Algorithm to URL name, meta-tag, links and text in homepage
- Weighting Using tf-idf
- Extract Images found in the website homepage.
  - Execute the Haar-like to detect a face [Viola and Jones].
  - Execute the Skin Detection Algorithm: Locate skin regions based on the detected skin pixels based on research [Girgis].
  - Analyze the skin regions for clues of Pornography Decision as follows:
    - If skin color pixels are up to the level of between 5% and 20% it indicates a human being is most probably in the image.
    - If the percentage of skin pixels is between 20% and 25%, it indicates more than likely Pornography image.
    - If the percentage of skin pixels is more than 25%, it indicates a Pornography image.


## Application Interface

### Inscheme Object Debugger

Input URL to be debugged

http://localhost/thesis/poom

Submit the URL

### These are the ordinary raw tags that we found

11. **Meta tag**
    
    `<meta name="description" content="Poom.com is the best and the first weporn in US and it have 1000+ free porn to be downloaded by you as our user">`

12. **Meta tag**
    
    `<meta name="content-type" content="siegel">`

13. **Meta tag**
    
    `<meta name="author" content="Poom Inc.">`

14. **Meta tag**
    
    `<meta name="keywords" content="Porn, Mif, Buckake, Gay, Teen, 17+, Brunete, Amateur, Young boy, Anal and etc">`

15. **Meta tag**
    
    `<meta charset="UTF-8">`

### These are the open graph based raw tags that we found

1. **Meta tag**
    
    `<meta property="og:title" content="Poom.com the 1st weporn in US">`

2. **Meta tag**
    
    `<meta property="og:description" content="Poom.com is the best and the first weporn in US and it have 1000+ free porn to be downloaded by you as our user">`

3. **Meta tag**
    
    `<meta property="og:url" content="http://poorn.com">`

4. **Meta tag**
    
    `<meta property="og:site_name" content="poorn.com">`
1-URL N-Gram Result

### Output Report

<table>
<thead>
<tr>
<th>No.</th>
<th>URL</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.hoddamateurs.com">http://www.hoddamateurs.com</a></td>
<td>1</td>
</tr>
</tbody>
</table>

### N-Grams Result

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.hoddamateurs.com">http://www.hoddamateurs.com</a></td>
<td>n = 3: hoo ood oda dam ama mak aleu eur urs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 4: hoo ood oda dam ama mak aleu eur urs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 5: hoo oda oda dam ama mak aleu eur urs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 6: hoo dama oda dam ama mak aleu eur urs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 7: hoo dama oda dam ama mak aleu eur urs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 8: hoo dama oda dam ama mak aleu eur urs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>n = 9: hoo dama oda dam ama mak aleu eur urs</td>
</tr>
</tbody>
</table>

### Multiplied result from URL Segmentation

**For keyword ‘Business’**

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
<th>Count</th>
<th>Index</th>
<th>Multiplied</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**For keyword ‘Healthy’**

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
<th>Count</th>
<th>Index</th>
<th>Multiplied</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2-URL Process Result

<table>
<thead>
<tr>
<th>Keywords</th>
<th>Count</th>
<th>Index</th>
<th>Multiplied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nothing Result Exist for Any Keywords</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For keyword 'Government' |

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For keyword 'Organization' |

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For keyword 'None' |

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Hasil : [http://www.hoodamateurs.com](http://www.hoodamateurs.com) merupakan website Pornography dengan akumulasi index sebesar 12
3-Forward Link N-Gram Result

<table>
<thead>
<tr>
<th>No.</th>
<th>URL</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://www.hoodamateurslive.com/">http://www.hoodamateurslive.com/</a></td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td><a href="http://www.trannysextapes.com">http://www.trannysextapes.com</a></td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td><a href="http://www.hoodthots.com">http://www.hoodthots.com</a></td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td><a href="http://hoodamateurslive.com/">http://hoodamateurslive.com/</a></td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td><a href="http://trannysextapes.com">http://trannysextapes.com</a></td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td><a href="http://www.trannysextapes.com/index2.html">http://www.trannysextapes.com/index2.html</a></td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td><a href="http://www.shegotass.info/">http://www.shegotass.info/</a></td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td><a href="http://Ebonyamateurxxx.com/forum">http://Ebonyamateurxxx.com/forum</a></td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td><a href="http://www.HOODAMATEURS.com">http://www.HOODAMATEURS.com</a></td>
<td>1</td>
</tr>
</tbody>
</table>

N-Grams Result

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
<th>Count:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>n = 3: hoo ood oda dam ama mat ate teu eur urs rsli ilv ive</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n = 4: hood ooda odam dama amat mate au teu eurs ursi rsl sivl ilve</td>
<td></td>
</tr>
<tr>
<td></td>
<td>n = 5: hood ooda odam oda amma mate au teu eursi eursi rsl sivl ilve</td>
<td></td>
</tr>
<tr>
<td>No.</td>
<td>Inner Text</td>
<td>Category</td>
</tr>
<tr>
<td>-----</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>---------------</td>
</tr>
<tr>
<td>1.</td>
<td>HOOD AMATEURS FREE BLACK EBONY AMATEUR PORN Hoodamateurs Cams Tranny sextapes Hoodamateurs HOOD THOTS Signup &amp;middot Login &amp;middot Help  English (United Kingdom)  English (United States)  Videos  Photos Pornstars Premium Users Forum Bookmark us</td>
<td>Pornography</td>
</tr>
</tbody>
</table>

Hasil: www.hoodamateurs.com merupakan website Pornography dengan akumulasi index sebesar 0.08484848484848485
5- Application Interface – non pornography

Output Report

<table>
<thead>
<tr>
<th>No.</th>
<th>URL</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://ilmubisnis.co.id/ilmu-bisnis-kontak-kami/">http://ilmubisnis.co.id/ilmu-bisnis-kontak-kami/</a></td>
<td>1</td>
</tr>
</tbody>
</table>

N-Grams Result

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><a href="http://ilmubisnis.co.id/ilmu-bisnis-kontak-kami/">http://ilmubisnis.co.id/ilmu-bisnis-kontak-kami/</a></td>
<td>n = 3 : Ilmu mub ubi bis isni nis is s i ilim ilmu mu u b bi bis isni nis is s k ko kon konta nta tak ak k k ka kam kali n = 4 : Ilmu mub ubi ubi bien isni snis nis is s i ilim ilmu mu b u b bi bis bien isni snis nis is k s ko kon konta npta nta tak ak k k ka kam kali n = 5 : Ilmu mub ubiubis bisni isni snis nis is s i ilim ilmu ilmu b mu b u b bisn bisni isni snis nis k s kon ko kon ko kam kali n = 6 : Ilmu mub ubiubis bisni isni snis nis is s i ilim ilmu ilmu ilmu b mu b u b bisn bisni isni snis nis k s kon ko kon ko kam kali n = 7 : Ilmu mub ubiubis bisni isni snis nis is s i ilim ilmu ilmu ilmu b mu b u b bisn bisni isni snis nis k s kon ko kon ko kam kali n = 8 : Ilmu mub ubiubis bisni isni snis nis is s i ilim ilmu ilmu ilmu b mu b u b bisn bisni isni snis nis k s kon ko kon ko kam kali n = 9 : Ilmu mub ubiubis bisni isni snis nis is s i ilim ilmu ilmu ilmu b mu b u b bisn bisni isni snis nis k s kon ko kon ko kam kali</td>
</tr>
</tbody>
</table>

Multiplied result from URL Segmentation

For keyword "Bisnis"
5 - Application Interface - non pornography

### Multiplied result from URL Segmentation

For keyword ‘Bisnis’

<table>
<thead>
<tr>
<th>No.</th>
<th>Keywords</th>
<th>Count</th>
<th>Index</th>
<th>Multiplied</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Ilmu</td>
<td>2</td>
<td>0.122</td>
<td>0.244</td>
</tr>
<tr>
<td>2</td>
<td>Bisnis</td>
<td>2</td>
<td>0.069</td>
<td>0.138</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.191</td>
<td>0.382</td>
</tr>
</tbody>
</table>

Hasil: http://ilmubisnis.co.id/ilmu-bisnis-kontak-kami/ merupakan website Bisnis dengan akumulasi index sebesar 0.382
Problem in skin detection

LEFT: Image classified as adult content, skin detector (with color-based detector) can detect more than 80% of skin in one person.

RIGHT: Image mis-classified as adult content, skin detector (with color-based detector) also can detect more than 80% of skin in face.

CONCLUSION
Conclusion

• This research is presented an online client-side filtering application that allows the user to perform multi level combination of pornographics filtering techniques such as domain and URL filtering, meta tag / keyword filtering for text-based detection and skin detection filtering for image-based detection.

• For skin detection filtering, we implemented some established skin detection techniques for color images and combine it with haar-like features for face detection as reference to find skin area, as well as markers of the object in the image in the mode close up / macro or in the scene modes.

• After testing the program either in groups or users in the admin group, it can be said that the application can run well and useful for the classification process for filtering pornographic content websites such as text and images.
Future Work

- the system can be developed to discover text content more details into internal links that exist in a website, because in this study, keyword search perform on the first level of a website only (in homepage).
- Moreover, it is still possible to improve run time efficiency in face recognition and skin detection algorithms.
Discussion?