



Automatic User Comment Detection in Flat Internet Fora

Mathias Bank

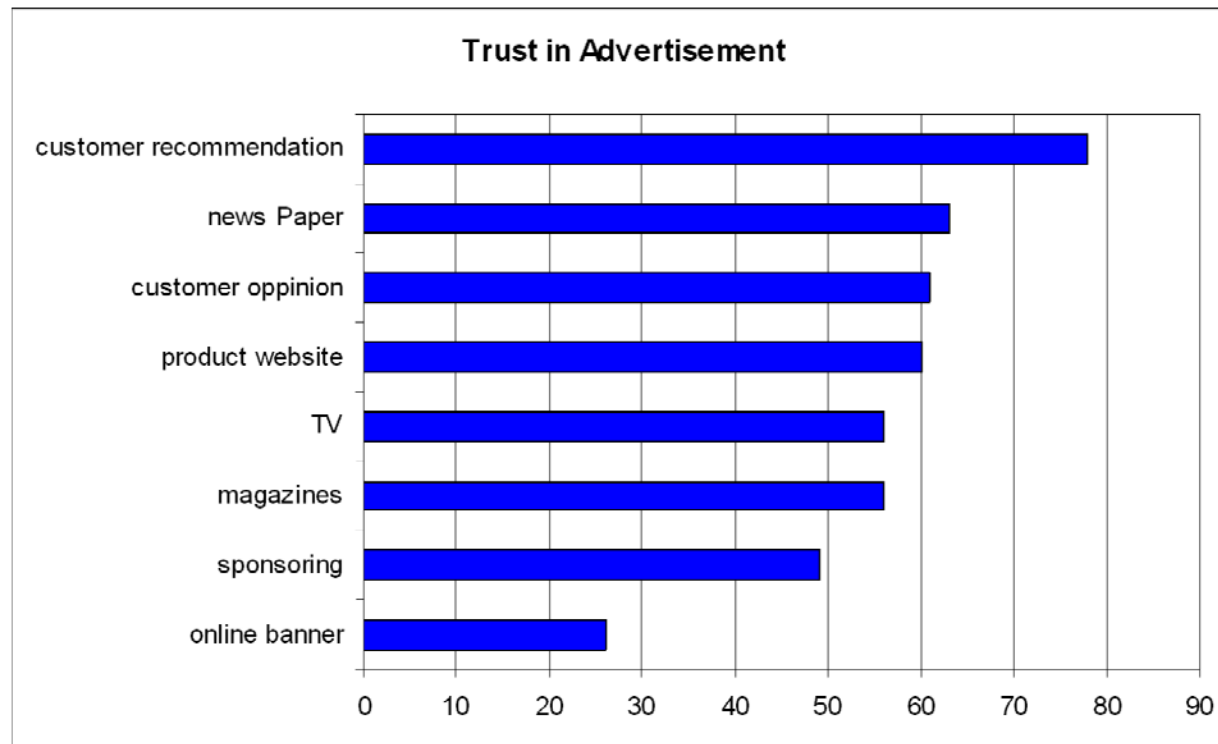
mathias.bank@uni-ulm.de

Faculty for Mathematics and Economics
University of Ulm

Community Analysis



...



User Content Detection

The screenshot shows the vBulletin forum homepage. At the top, there is a navigation bar with links for Register, Profile, Blog, FAQ, Community, Calendar, and Today's Posts. Below this, there is a search bar and a list of sub-forums. The main content area displays a list of threads under the heading "Threads in Forum - vBulletin 3.6 Questions, Problems and Troubleshooting". The threads are listed in a table with columns for Thread/Thread Starter, Rating, Last Post, Replies, and Views.

Thread / Thread Starter	Rating	Last Post	Replies	Views
Sticky: Information on the Gumbiar Virus Wayne Luke		Sat 23rd May 09 10:45am by Wayne Luke	6	3,129
Sticky: Forbidden and Not acceptable errors? Zachery		Sat 31st Jan 09 3:04pm by Steve Machol	1	22,414
Sticky: How to integrate Google AdSense in 3.6 Wayne Luke		Fri 9th Jun 09 9:20am by Wayne Luke	3	5,929
Sticky: Quick Reply not working? Zachery		Fri 9th Jun 09 8:05am by Zachery	0	3,707
Sticky: Evaluating fatal errors received during vBulletin Operations. Wayne Luke		Fri 27th Jun 09 5:56am by Wayne Luke	1	4,686
Sticky: Error: Security token was missing or mismatched Steve Machol		Sun 22nd Jun 08 3:04pm by Steve Machol	4	18,066
Sub-Forums 1000000		Today 7:23am by jib	0	1
MySQL Error - The table 'attachment' is full jib		Today 7:18am by jib	3	16
IP Banning Question vbulletin		Today 6:03am by vbulletin	0	30
Minimum Post Count for Posting Links? Andri Kello		Today 6:00am by Andri Kello	2	36
Moving threads: change default redirect period. Gronch		Today 5:18am by Gronch	0	10
After converting to UTF-8, Big problem Pharmacist		Today 4:55am by Pharmacist	3	32
moving users not working Gronch		Today 4:18am by Gronch	0	11
How do I transfer vbulletin to a new website Liam		Today 3:24am by Liam	1	74

The screenshot shows a forum thread titled "AdSense Integration". The thread is posted by a user named "Davros" (New Member) on Saturday, April 18, 2009, at 9:24am. The thread content discusses the user's experience with AdSense integration on vBulletin 3.6, mentioning that they were banned by Google for invalid click activity after integrating AdSense. The user expresses frustration and asks for help, as they have signed up with AdBright but cannot help with the issue.

AdSense Integration

I recently bought vBulletin and downloaded the version with my AdSense integrated, but since doing this I have now been banned by Google for invalid click activity. I have never and will never click ads that I publish on my web sites.

However I have been a part of the AdSense program since 2006 and have never had a problem. I run 5 websites only my latest site is vBulletin and I have made a reasonable amount of money from AdSense in the past so to say I am unhappy at losing the \$300 march payment is an understatement. I have appealed against the decision but Google have said they have checked my account and the decision will stand.

Not only have they blocked my AdSense account but they have stopped Google Bots crawling my pages, or at least I haven't seen any since this has happened. I'm just wondering if anybody else has had the same issues?

I've now signed up with AdBright, but can't help but think this is somehow related to vBulletin AdSense integration.

Digicom (New Member) posted on Saturday, April 25, 2009, at 9:21am:

This also happened to me, I had been with AdSense since 2005 and in february this year I download AdSense integrated in vBulletin, then when I was due to be paid in march I was suddently banned from AdSense, I then appealed but was told:

Quote:
However, after thoroughly reviewing your account data and taking your feedback into consideration, we have re-confirmed we are unable to reinstate your account.

Since this happened Google Bots have also stopped crawling my site, it seems strange that my problem is identical to the problem Davros has and only happened after downloading AdSense integrated in vBulletin.

User Content Detection (1)

The screenshot shows a forum thread on vBulletin 3.8.3. The thread title is "AdSense Intergration" and it is located in the "vBulletin 3.8 Questions, Problems and Troubleshooting" section. The thread contains three posts, all dated Saturday, April 4, 2009, at 9:24pm, 9:31pm, and 9:33pm respectively. The first post is by user "Davros" (New Member, Join Date: Feb 2009, Posts: 9) and discusses being banned by Google for invalid click activity after integrating AdSense into vBulletin. The second post is by user "Digicom" (New Member, Join Date: Apr 2007, Posts: 5) and shares a similar experience, including a quote from Google: "However, after thoroughly reviewing your account data and taking your feedback into consideration, we have re-confirmed we are unable to reinstate your account." The third post is by user "Nick" (Senior Member, Join Date: Feb 2008, Location: Tampa, FL, Age: 29, Posts: 3,192) and is partially visible. The forum interface includes a navigation bar with links for Register, Projects, Blogs, FAQ, Community, Calendar, Today's Posts, and Search. A user login box is visible in the top right corner. The bottom of the page shows a security warning: "Skripte sind momentan verboten | <SCRIPT>: 36 | <OBJECT>: 0" and a link for "Einstellungen...".

vBulletin 3.8.3
www.vbulletin.com

vBulletin Community Forum > vBulletin 3.8 > vBulletin 3.8 Questions, Problems and Troubleshooting

AdSense Intergration

Register Projects Blogs FAQ Community Calendar Today's Posts Search

User Name Remember Me?
Password

Post Reply

Page 1 of 2 1 2 >

View First Unread Thread Tools Search this Thread Rate Thread Display Modes

Sat 4th Apr '09, 9:24pm #1

Davros
New Member
Join Date: Feb 2009
Posts: 9

AdSense Intergration

I recently bought vBulletin and downloaded the version with my AdSense integrated, but since doing this I have now been banned by Google for invalid click activity.

I have never and will never click ads that I publish on my web sites.

However I have been a part of the AdSense program since 2006 and have never had a problem. I run 5 websites only my latest site is vBulletin and I have made a reasonable amount of money from AdSense in the past so to say I am unhappy at losing the \$300 march payment is an understatement. I have appealed against the decision but Google have said they have checked my account and the decision will stand.

Not only have they blocked my AdSense account but they have stopped Google Bots crawling my pages, or at least I havn't seen any since this has happened.

I'm just wondering if anybody else has had the same issues?

I've now signed up with Adbright, but can't help but think this is somehow related to vBulletin AdSense integration.

Sat 4th Apr '09, 9:31pm #2

Digicom
New Member
Join Date: Apr 2007
Posts: 5

This also happened to me, i had been with adsense since 2005 and in february this year i download AdSense integrated in vBulletin,then when i was due to be paid in march i was suddenly banned from adsense,i then appealed but was told

Quote:

Since this happened Google Bots have also stopped crawling my site,it seems strange that my problem is identical to the problem Davros has and only happened after downloading AdSense integrated in vBulletin

Sat 4th Apr '09, 9:33pm #3

Nick
Senior Member
Join Date: Feb 2008
Location: Tampa, FL
Age: 29
Posts: 3,192

Skripte sind momentan verboten | <SCRIPT>: 36 | <OBJECT>: 0

User Content Detection (2)



Main Content Detection

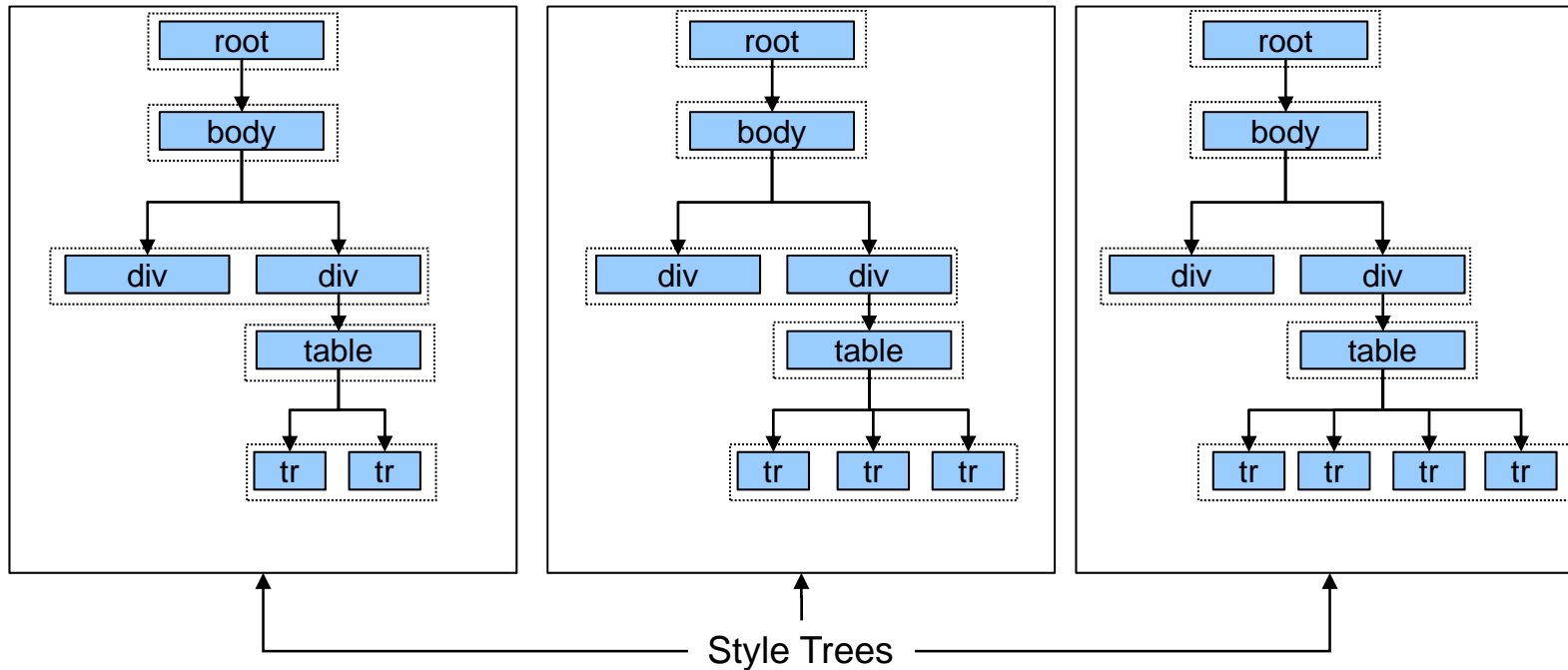


Post Segmentation

Main Content Detection - SST



Template based Systems



Main Content Detection – SST (1)



Node importance:

- number of presentation styles
- Number of content diversity
- Based on Shannon's Entropy

Composite Importance

- Propagating child importance to its parent nodes

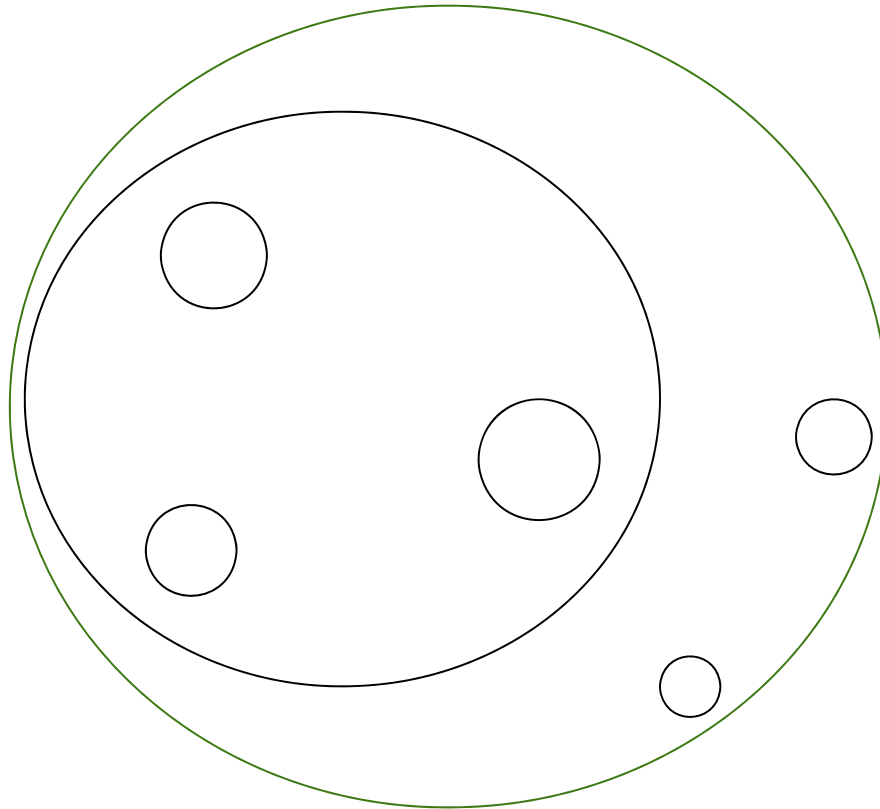
→ Noise detection by searching nodes with low Composite Importance

→ Main Content detection by searching nodes with high Composite Importance

Main Content Detection – Content Detection

Main Assumption:

Node with Main Content Section will get the highest importance value.



Body node

Discussion node

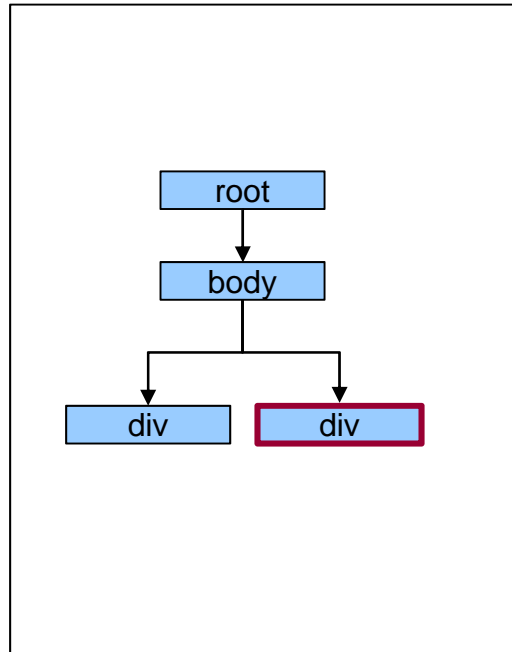
Post node

Wrong node

Node comparison
over multiple pages

```
/root/body/div[@class="content"][contains(@id,"thread")/
```

Main Content Detection – XPath Creation



Default XPath generation:

- /html/body/div/
- /html/body/div[2]/

Improved XPath generation:

- Use all possible tag attributes
- Don't use position counter
- Replace numerical attributes with "contains"-XPath expression

```
/root/body/div[@class="content"][contains(@id,"thread")/
```

Main Content Detection – Evaluation

Evaluation data:

- 51 real internet fora
- Based on 14 different fora systems
- Different template structures (not only color)
- Not used except for evaluation

Evaluation results:

- perfect: 82.4%
- Correct: 100%

For 17.6% (9 fora) the extracted XPath expression could be more specific.

User Content Detection (3)

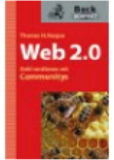


Main Content Detection



Post Segmentation

Post Segmentation – Related Work

1.  **Online-Communities im Web 2.0: So funktioniert** und Marco Ripanti (Broschiert - 1. Juni 2008)
 Neu kaufen: **EUR 34,90** [28 Angebote](#) ab EUR 28,
 Lieferung bis **Mittwoch, 26. August**: Bestellen Sie inner
Kostenlose Lieferung möglich.
 ★★★★★ (7)
Bücher: Alle 27.123 Artikel ansehen
2.  **Community Marketing Management. Wie man**
 von Frank Mühlenbeck und Klemens Skibicki (Broschiert - 1. Juni 2008)
 Neu kaufen: **EUR 29,90** [53 Angebote](#) ab EUR 22,
 Lieferung bis **Mittwoch, 26. August**: Bestellen Sie inner
Kostenlose Lieferung möglich.
 ★★★★★ (16)
Auszug - Seite 10: "... gerade die Schwester-Community
 hängeschildern schießen jeden Tag neue Community-Kc
Bücher: Alle 27.123 Artikel ansehen
3.  **Web 2.0 - Geld verdienen mit Communities**
 von Frank Mühlenbeck (Broschiert - 1. Juni 2008)
 Neu kaufen: **EUR 6,80** [59 Angebote](#) ab EUR 2,36
 Lieferung bis **Mittwoch, 26. August**: Bestellen Sie inner
Kostenlose Lieferung möglich.
 ★★★★★ (4)
Bücher: Alle 27.123 Artikel ansehen
4.  **Community Building on the Web: Secret Strategies**
 Addison-Wesley Longman, Amsterdam (Taschenbuch - 1. Juni 2008)
[4 Angebote](#) ab EUR 36,17
 ★★★★★ (12)
Englische Bücher: Alle 488.948 Artikel ansehen

B. Liu et al:

- Data regions are presented in continuous regions with similar tags
 - Data regions can be found in one subtree
- ➔ Data list by searching „identical“ subtrees
- Post-Order traversing for nested structures

Detecting „identical“ subtrees in community systems very difficult!

Post Segmentation - Observations

Analyzing

- 3,500 posts
- 13 different fora
- 8 different fora systems

All Postings

- 320 characters
- 4.3 line breaks

No line breaks

- 25% of all posts
- 94 characters

Observations

1. User comments mainly consist of textual data
2. Complete comment in one subtree (Liu), splitted by further elements
3. No similarity between user comments except base structure
4. Nested structures are possible (e.g. Drupal)

Post Segmentation – Algorithm

Post Candidate Search

Wrapper Generation

Post Segmentation – Algorithm (1)

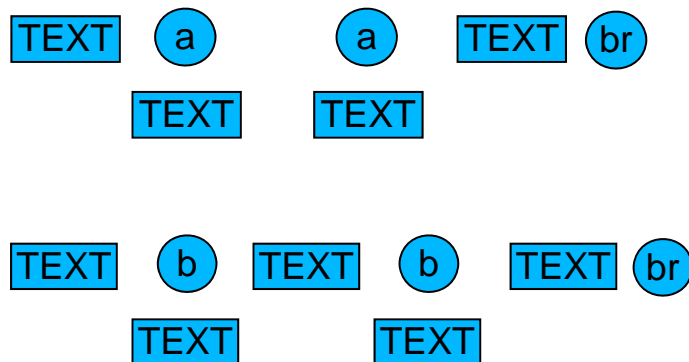
Post Candidate Search

Wrapper Generation

Merge inline elements

Ich habe die Suche bemüht und bin dabei auf zwei Threads gestoßen, die aber nicht genau mein Anliegen behandeln.

Mir geht's darum, die Elemente `<blockquote>`, `<q>` und `<cite>` richtig zu verwenden (mehr gibt's doch nicht für Zitate/Zitationen, oder?).



Post Segmentation – Algorithm (2)

Post Candidate Search

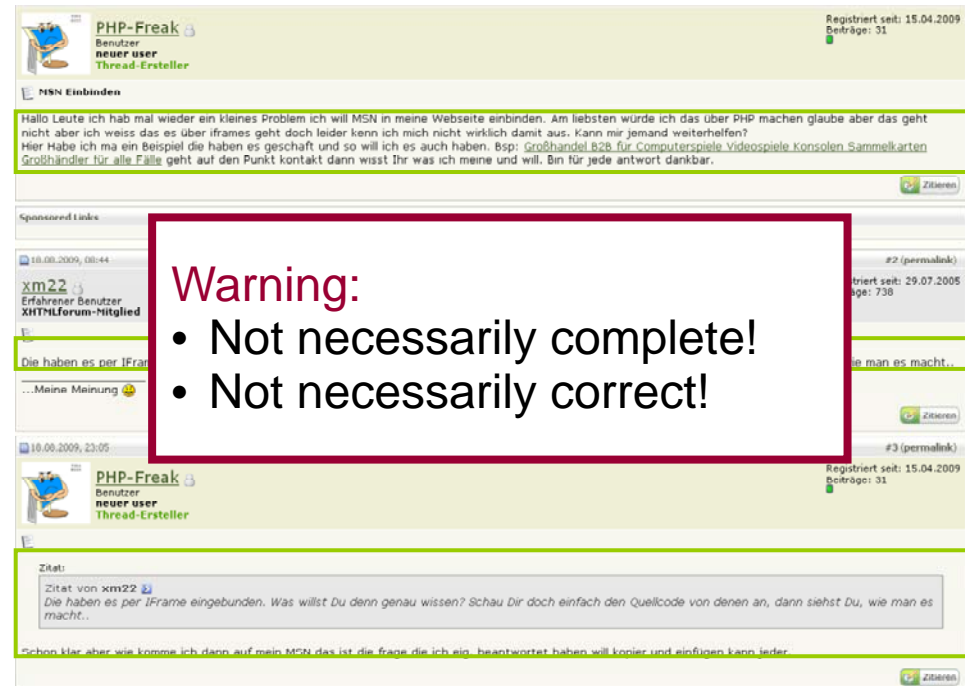
Merge inline elements

Text node identification

3 rules:

- > 150 characters
- > 50 characters & new lines
- Not beginning with „_____“

Wrapper Generation



The screenshot shows a forum post from the user 'PHP-Freak' (neuer user, Thread-Ersteller) with the title 'MSN Einbinden'. The post content is: 'Hallo Leute ich hab mal wieder ein kleines Problem ich will MSN in meine Webseite einbinden. Am liebsten würde ich das über PHP machen glaube aber das geht nicht aber ich weiss das es über iframes geht doch leider kenn ich mich nicht wirklich damit aus. Kann mir jemand weiterhelfen? Hier Habe ich ma ein Beispiel die haben es geschafft und so will ich es auch haben. Bsp: [Großhandel.B2B für Computerspiele Videospiele Konsolen Sammelkarten Großhändler für alle Fälle](#) geht auf den Punkt kontakt dann wisst Ihr was ich meine und will. Bin für jede antwort dankbar.'

A red-bordered warning box is overlaid on the post content, containing the text: 'Warning: Not necessarily complete! Not necessarily correct!'

Below the main post, there is a quote from user 'xm22' (Erfahrener Benutzer, XHTMLForum-Mitglied) with the text: 'Die haben es per IFrame eingebunden. Was willst Du denn genau wissen? Schau Dir doch einfach den Quellcode von denen an, dann siehst Du, wie man es macht..'

Post Segmentation – Algorithm (3)

Post Candidate Search

Merge inline elements

Text node identification

3 rules:

- > 150 characters
- > 50 characters & new lines
- Not beginning with „____“

Wrapper Generation

Ich bin sicher nicht auf dem Niveau, um das alles wirklich zu verstehen, auch, wenn ich schon einige Erfä
habe.

Dennoch erschließt sich mir der Sinn des Ganzen nicht so recht. Was soll dieses Standardmodul für Vorte

HTML Code:

```
<div class="mod">
  <div class="inner">
    <div class="hd">Block Head</div>
    <div class="bd">Block Body</div>
    <div class="ft">Block Foot</div>
  </div>
</div>
```

} div

Post Segmentation – Algorithm (4)

Post Candidate Search

Merge inline elements

Text node identification

Text node cleaning

Wrapper Generation

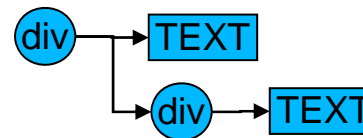
Ich bin sicher nicht auf dem Niveau, um das alles wirklich zu verstehen, auch, wenn ich schon einige Erfä
habe.

Dennoch erschließt sich mir der Sinn des Ganzen nicht so recht. Was soll dieses Standardmodul für Vorte

HTML Code:

```
<div class="mod">  
  <div class="inner">  
    <div class="hd">Block Head</div>  
    <div class="bd">Block Body</div>  
    <div class="ft">Block Foot</div>  
  </div>  
</div>
```

} div



Complete comment in
one subtree

Post Segmentation – Algorithm (5)

Post Candidate Search

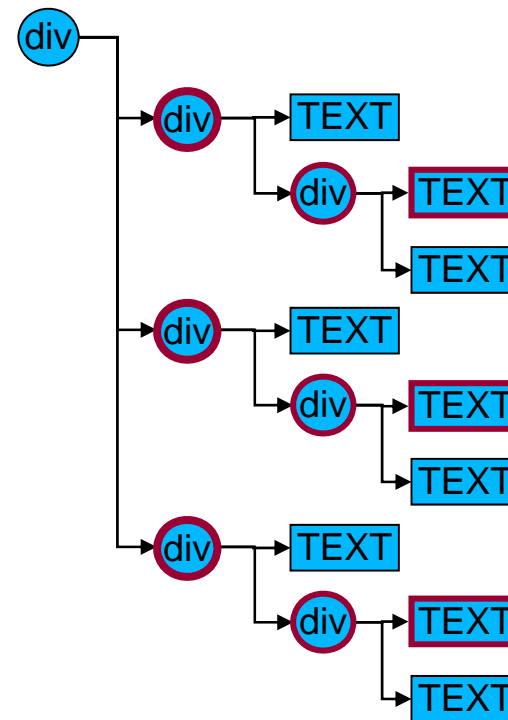
Merge inline elements

Text node identification

Text node cleaning

Node generalization

Wrapper Generation

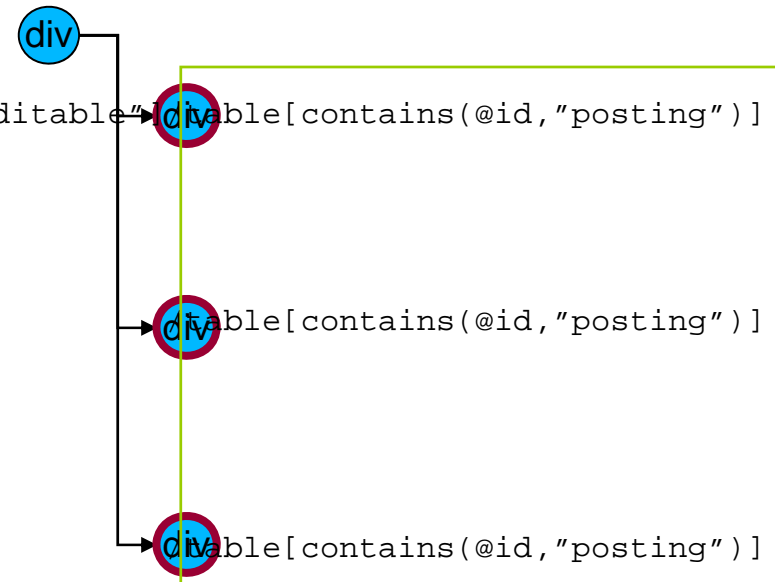


Post Segmentation – Algorithm (6)

Post Candidate Search

```
/div[class="entities"] /div[class="editable"] /div  
  
/div[class="entities"] /div  
  
/div[class="entities"] /div
```

Wrapper Generation



Post Segmentation – Evaluation

Evaluation data:

Same data as for Main Content
Detection

Main Content already selected (all
fora used!)

Evaluation results:

- perfect: 64%
- Correct: 90%

system	correct	perfect
Burning Board	100,0%	66,7%
drupal	83,3%	0,0%
IPB	75,0%	50,0%
myBB	66,7%	66,7%
phpBB	100,0%	57,1%
SMF	100,0%	33,3%
Unclassified NewsBoard	100,0%	100,0%
Vanilla	100,0%	100,0%
vBulletin	92,3%	84,6%
miscellaneous	60,0%	60,0%