Content Code Blurring: A New Approach to Content Extraction

Thomas Gottron

Institute of Computing Science
Johannes Gutenberg-University Mainz, Germany

DEXA Workshop Text-based Information Retrieval, 2008

Outline

1. Introduction
   - Motivation
   - Related Work
2. Content Code Blurring
   - General Idea
   - Code and Content
   - Content Code Ratio
   - Versions of the Algorithm
3. Evaluation
   - Evaluation Method
   - Evaluation Data
   - Evaluation Results
4. Conclusion and Future Work

Introduction

Motivation: Finding the Main Content

BBC Home News Sport Radio TV Weather Languages Search
UK version International version | About the versions Low graphics | Accessibility help
BBC NEWS One-Minute World News
News services Your news when you want it

News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
Business
Market Data
Economy
Companies
Health
Science/Nature
Technology
Entertainment
Also in the news
Video and Audio
Have Your Say
In Pictures
Country Profiles
Special Reports
RELATED BBC SITES
SPORT
WEATHER
ON THIS DAY
EDITORS' BLOG

Last Updated: Wednesday, 20 February 2008, 11:47 GMT
E-mail this to a friend Printable version

Microsoft steps up Yahoo campaign

Microsoft has hired a firm that specialises in proxy battles in a move which suggests it could try to oust the current board of Yahoo directors.

There is currently a stalemate between Yahoo and Microsoft

The Yahoo board rejected a takeover offer from Microsoft worth more than $40bn (£20.6bn) saying it was too low.

A proxy fight would see Microsoft nominate a group of directors sympathetic to a deal for shareholders to vote on at Yahoo's annual meeting.

The move would be cheaper than raising its $31-a-share offer, analysts say.

Microsoft's offer was 62% above the level at which Yahoo stock was trading when the bid approach was made on 1 February, and Microsoft has called the price "full and fair".

However, one of Yahoo's biggest shareholders, Bill Miller, an asset manager at Legg Mason, recently said a fair price would be nearer $40 a share.

> **"Microsoft is doing the smart thing. It's giving both the carrot and the stick"**
>
> Toan Tran, Morningstar analyst

But it is thought that even offering an extra $1 a share would cost Microsoft an additional $1.4bn, while waging a proxy fight could cost between $20m and $30m.

"Microsoft is doing the smart thing. It's giving both the carrot and the stick," said Morningstar analyst Toan Tran.

"The carrot was the big premium on Yahoo stock and now the stick is the threat of a proxy fight."

MICROSOFT'S BID FOR YAHOO
Latest News
- AOL 'contemplates Yahoo deal'
- Yahoo investor urges higher bid
- Murdoch rules out bid for Yahoo
- Microsoft move troubles Google
- Yahoo bid 'should get approval'

ANALYSIS

Shotgun marriage
Microsoft wants to buy internet legend Yahoo, but who will benefit from the deal?

- Searching for the next big thing
- Peston: It won't move Wall Street
- Blog: Match made in heaven?
- Profiles: Microsoft and Yahoo
- Google's might drove the bid

VIDEO ON DEMAND
- BT on Microsoft bid for Yahoo

HAVE YOUR SAY
- Your views on the Deal

SHARE PRICE CHECK
- Microsoft
- Yahoo
- Google

RELATED INTERNET LINKS
- Yahoo
- Microsoft
- New York Times
- Innisfree M&A

The BBC is not responsible for the content of external internet sites

TOP BUSINESS STORIES
- Microsoft set to open up software
- Starbucks cutting HQ jobs

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Motivation: Finding the Main Content

BBC Home News Sport Radio TV Weather Languages Search
UK version | International version | About the versions — Low graphics | Accessibility help
BBC NEWS — One-Minute World News — News services: Your news when you want it

News Front Page
Africa
Americas
Asia-Pacific
Europe
Middle East
South Asia
UK
Business
Market Data
Economy
Companies
Health
Science/Nature
Technology
Entertainment
Also in the news
Video and Audio
Have Your Say
In Pictures
Country Profiles
Special Reports
RELATED BBC SITES
SPORT
WEATHER
ON THIS DAY
EDITORS' BLOG

Last Updated: Wednesday, 20 February 2008, 11:47 GMT
E-mail this to a friend — Printable version

Microsoft steps up Yahoo campaign

Microsoft has hired a firm that specialises in proxy battles in a move which suggests it could try to oust the current board of Yahoo directors.

There is currently a stalemate between Yahoo and Microsoft

The Yahoo board rejected a takeover offer from Microsoft worth more than $40bn (£20.6bn) saying it was too low.

A proxy fight would see Microsoft nominate a group of directors sympathetic to a deal for shareholders to vote on at Yahoo's annual meeting.

The move would be cheaper than raising its $31-a-share offer, analysts say.

Microsoft's offer was 62% above the level at which Yahoo stock was trading when the bid approach was made on 1 February, and Microsoft has called the price "full and fair".

However, one of Yahoo's biggest shareholders, Bill Miller, an asset manager at Legg Mason, recently said a fair price would be nearer $40 a share.

> **"Microsoft is doing the smart thing. It's giving both the carrot and the stick"**
>
> Toan Tran, Morningstar analyst

But it is thought that even offering an extra $1 a share would cost Microsoft an additional $1.4bn, while waging a proxy fight could cost between $20m and $30m.

"Microsoft is doing the smart thing. It's giving both the carrot and the stick," said Morningstar analyst Toan Tran.

"The carrot was the big premium on Yahoo stock and now the stick is the threat of a proxy fight."

MICROSOFT'S BID FOR YAHOO
Latest News
- AOL 'contemplates Yahoo deal'
- Yahoo investor urges higher bid
- Murdoch rules out bid for Yahoo
- Microsoft move troubles Google
- Yahoo bid 'should get approval'

ANALYSIS

Shotgun marriage
Microsoft wants to buy internet legend Yahoo, but who will benefit from the deal?

- Searching for the next big thing
- Peston: It won't move Wall Street
- Blog: Match made in heaven?
- Profiles: Microsoft and Yahoo
- Google's might drove the bid

VIDEO ON DEMAND
- BT on Microsoft bid for Yahoo

HAVE YOUR SAY
- Your views on the Deal

SHARE PRICE CHECK
- Microsoft
- Yahoo
- Google

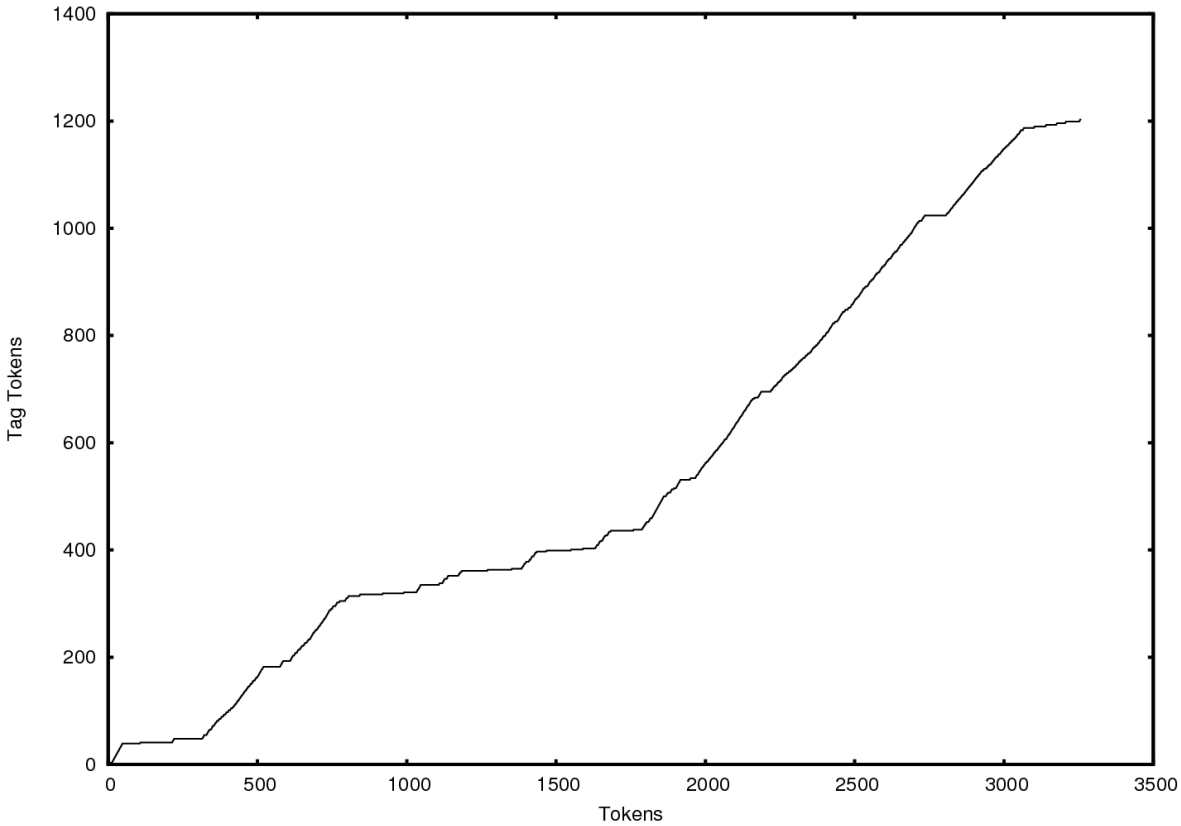
RELATED INTERNET LINKS
- Yahoo
- Microsoft
- New York Times
- Innisfree M&A

The BBC is not responsible for the content of external internet sites

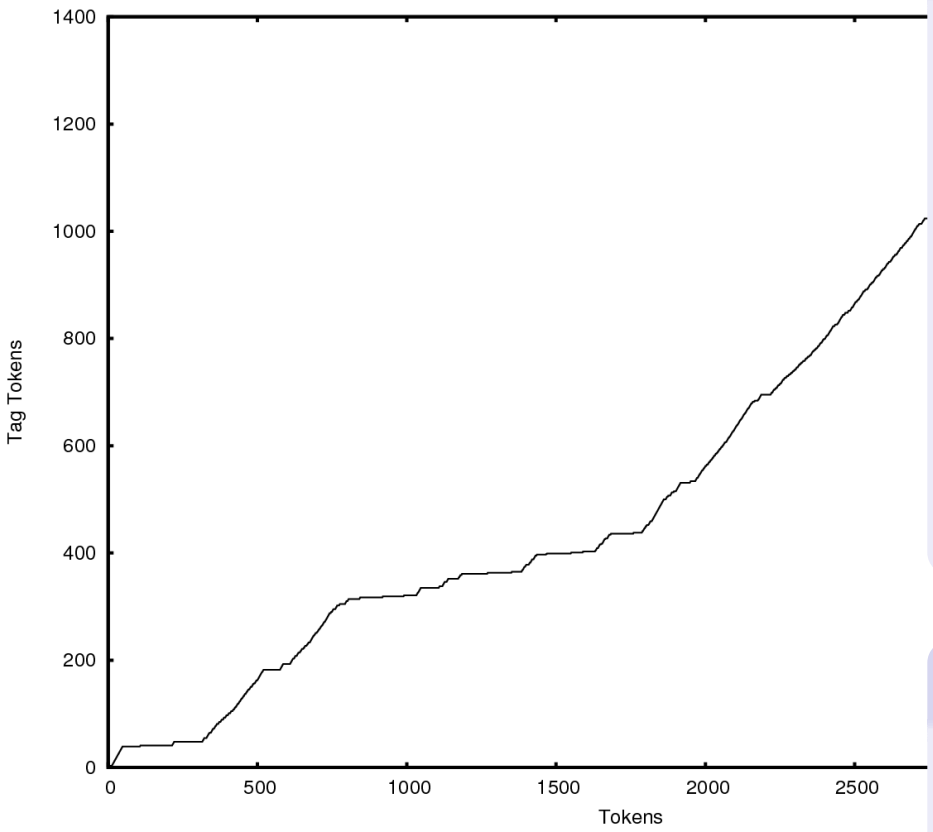
TOP BUSINESS STORIES
- Microsoft set to open up software
- Starbucks cutting HQ jobs

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Related Work - Document Slope Curve (DSC)



Related Work - Document Slope Curve (DSC)



Document Slope Curve

- Tags and words as tokens B_i
- Tag token have value 1
- Word token have value 0

$$d(i) = \sum_{n=0}^{i} B_n, \ for\ 0 \leq i \leq N-1$$

CE Idea

- Find low *slope regions*
- Sliding window technique

JCDL'02: Pinto, Branstein, Coleman, Croft, King, Li, Wei. QuASM: A System for Question Answering Using Semi-Structured Data

Related Work - Link Quota Filter (LQF)

Algorithm 3.3: Linkquota function.

Input: n: DOM node
Output: q: quota of links to overall text
begin
  $C \leftarrow \texttt{descendants}(n)$;
  $t_{tot} \leftarrow 0$;
  $t_{link} \leftarrow 0$;
  foreach $m \in C$ **do**
    if $\neg\texttt{isBlockNode}(m)$ **then**
      if $\texttt{isTextNode}(m)$ **then**
        $t_{tot} \leftarrow t_{tot} + \texttt{length}(\texttt{getText}(m))$;
      else if $\texttt{isLinkNode}(m)$ **then**
        $t_{tot} \leftarrow t_{tot} + \texttt{length}(\texttt{getText}(m))$;
        $t_{link} \leftarrow t_{link} + \texttt{length}(\texttt{getText}(m))$;
      else
        $t_{tot} \leftarrow t_{tot} + \texttt{length}(\texttt{getText}(m))$;
        $t_{link} \leftarrow t_{link} + \texttt{Linkquota}(m) \cdot \texttt{length}(\texttt{getText}(m))$;
    else
      $C \leftarrow C \setminus \texttt{descendants}(m)$;
  $q \leftarrow t_{link}/t_{tot}$;
  return q
end

Related Work - Link Quota Filter (LQF)

Algorithm 3.3: Linkquota function.

Input: n: DOM node
Output: q: quota of links to overall text
begin
  $C \leftarrow \texttt{descendants}(n)$;
  $t_{tot} \leftarrow 0$;
  $t_{link} \leftarrow 0$;
  foreach $m \in C$ **do**
    if $\neg\texttt{isBlockNode}(m)$ **then**
      if $\texttt{isTextNode}(m)$ **then**
        $t_{tot} \leftarrow t_{tot} + \texttt{length}(\texttt{getText}(m))$;
      else if $\texttt{isLinkNode}(m)$ **then**
        $t_{tot} \leftarrow t_{tot} + \texttt{length}(\texttt{getText}(m))$;
        $t_{link} \leftarrow t_{link} + \texttt{length}(\texttt{getText}(m))$;
      else
        $t_{tot} \leftarrow t_{tot} + \texttt{length}(\texttt{getText}(m))$;
        $t_{link} \leftarrow t_{link} + \texttt{Linkquota}(m) \cdot \texttt{length}(\texttt{getText}(m)$
    else
      $C \leftarrow C \setminus \texttt{descendants}(m)$;
  $q \leftarrow t_{link}/t_{tot}$;
  return q
end

Link Quota Function

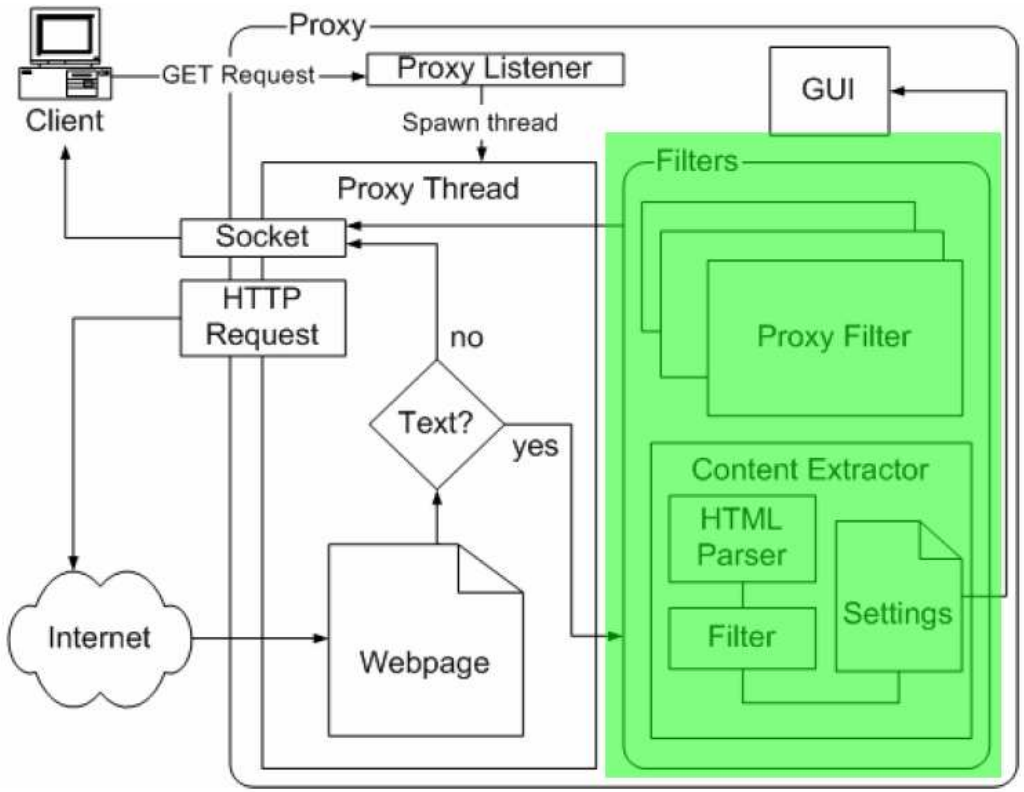
- Link rich areas are not main content
- Determine ratio of text in links
- Result is the link quota

CE Idea

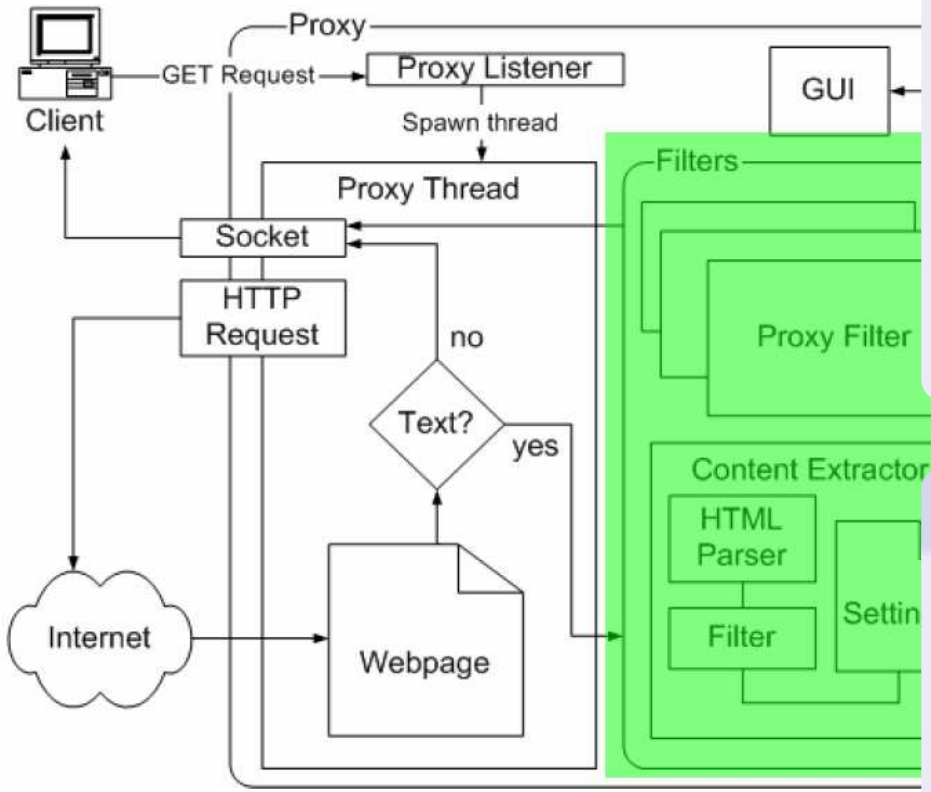
- Removes blocks with high link quota
- Affects only navigation and link-lists

HYPERTEXT'05: Mantratzis, Orgun, Cassidy. Separating XHTML Content from Navigation Clutter using DOM-structure Block Analysis

Related Work - Crunch Framework



Related Work - Crunch Framework



Crunch

- Improve web access for screen readers and small screen devices
- Proxy layout
- DOM analysis

CE Idea

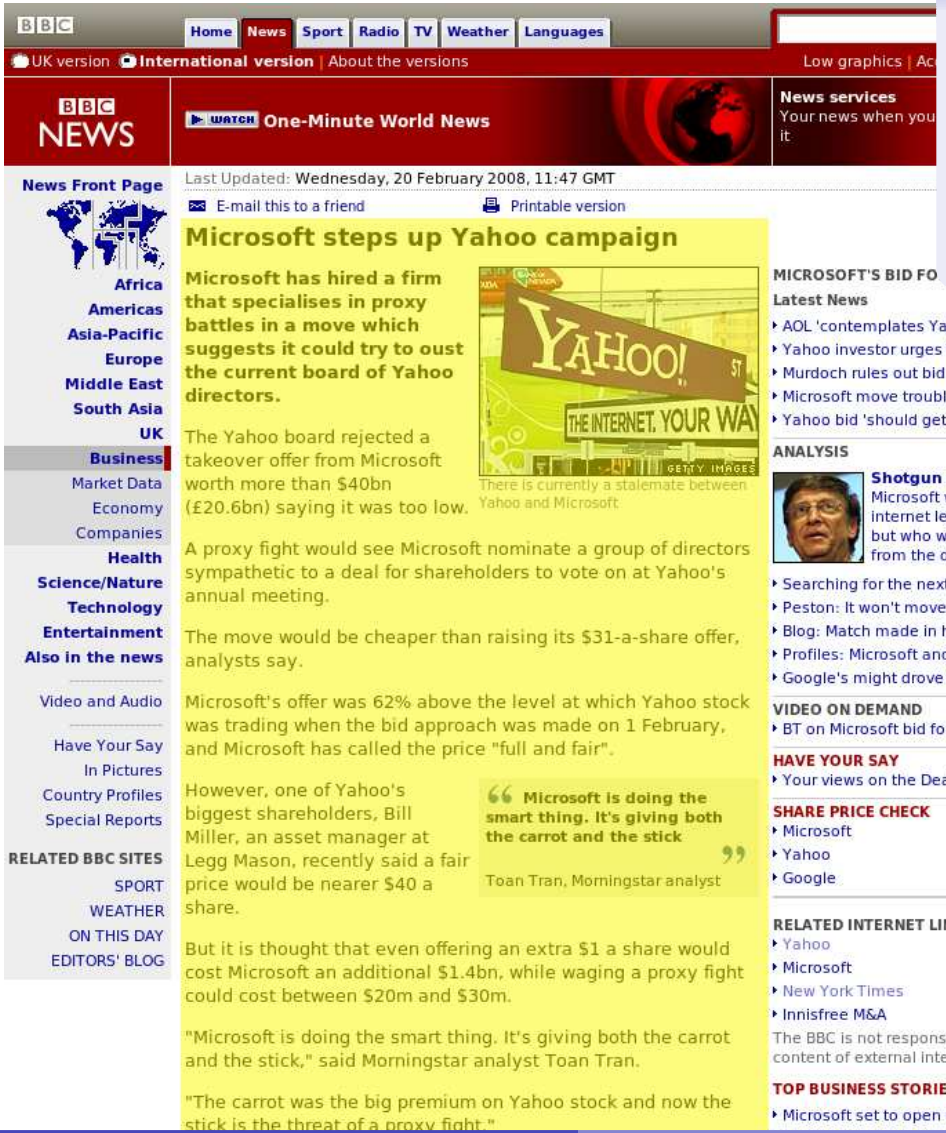
- Combine several heuristics (e.g. LQF)
- Very conservative

WWW'03: Gupta, Kaiser, Neistadt, Grimm. DOM-based Content Extraction of HTML Documents

Content Code Blurring

Content Code Blurring

Content Code Blurring



Typical observation

Main (text) content is:

- long text
- homogenously formatted

Content Code Blurring

Typical observation

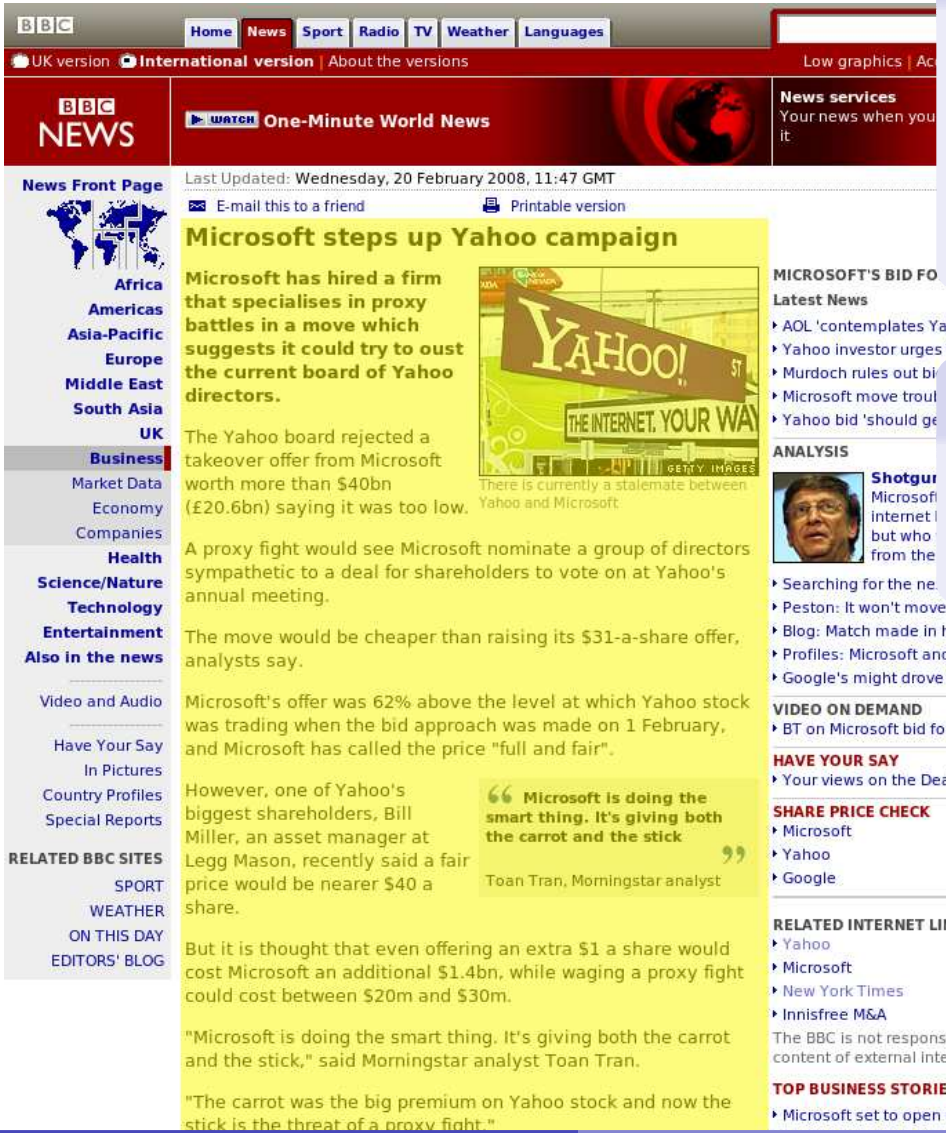
Main (text) content is:

- long text
- homogenously formatted

CE Idea:

Try to find long texts with a homogenous format.

Content Code Blurring



Typical observation

Main (text) content is:

- long text
- homogenously formatted

CE Idea:

Try to find *regions* with a *lot of content and little code*.

Code and Content

```
<p>The Silicon Valley company looks after
 two of the net's 13 DNS root servers. It
 also controls the computers that contain
 the master list of domain name suffixes
 such as .com and .net
<!-- S IIMA -->
<table border="0" cellspacing="0" align=
"right" width="226" cellpadding="0">
  <tr><td>
    <div><img
src="http://newsimg.bbc.co.uk/media/images/
44899000/jpg/_44899470_ken-silva-body.jpg"
width="226" height="170" alt="Ken Silva,
CTO VeriSign" border="0" vspace="0"
hspace="0">
      <div class="cap">"If there is a
      silver lining in all of this, it's
      that users will become more aware
      and more consious of who they do
      business with."</div>
    </div>
  </td></tr>
</table>
<!-- E IIMA -->
<p>Ken Silva, chief technology officer at
 Verisign, said: "We have anticipated
 these flaws in DNS for many years and we
 have  basically engineered around them."
```

Code

All tags are code.

Content

Everything else.

Code and Content

```
<p>The Silicon Valley company looks after
 two of the net's 13 DNS root servers. It
 also controls the computers that contain
 the master list of domain name suffixes
 such as .com and .net
<!-- S IIMA -->
<table border="0" cellspacing="0" align=
"right" width="226" cellpadding="0">
  <tr><td>
    <div><img
src="http://newsimg.bbc.co.uk/media/images/
44899000/jpg/_44899470_ken-silva-body.jpg"
width="226" height="170" alt="Ken Silva,
CTO VeriSign" border="0" vspace="0"
hspace="0">
      <div class="cap">"If there is a
      silver lining in all of this, it's
      that users will become more aware
      and more consious of who they do
      business with."</div>
    </div>
  </td></tr>
</table>
<!-- E IIMA -->
<p>Ken Silva, chief technology officer at
 Verisign, said: "We have anticipated
 these flaws in DNS for many years and we
 have  basically engineered around them."
```

Code

All tags are code.

Content

Everything else.

Special Cases

- comments
- contents of script- and style elements
- superfluous white space
- HTML entities

Code and Content

```
<p>The Silicon Valley company looks after
 two of the net's 13 DNS root servers. It
 also controls the computers that contain
 the master list of domain name suffixes
 such as .com and .net
<!-- S IIMA -->
<table border="0" cellspacing="0" align=
"right" width="226" cellpadding="0">
  <tr><td>
    <div><img
src="http://newsimg.bbc.co.uk/media/images/
44899000/jpg/_44899470_ken-silva-body.jpg"
width="226" height="170" alt="Ken Silva,
CTO VeriSign" border="0" vspace="0"
hspace="0">
      <div class="cap">"If there is a
      silver lining in all of this, it's
      that users will become more aware
      and more consious of who they do
      business with."</div>
    </div>
  </td></tr>
</table>
<!-- E IIMA -->
<p>Ken Silva, chief technology officer at
 Verisign, said: "We have anticipated
 these flaws in DNS for many years and we
 have  basically engineered around them."
```

```
<p>The Silicon Valley company looks after
 two of the net's 13 DNS root servers. It
 also controls the computers that contain
 the master list of domain name suffixes
 such as .com and .net

<table border="0" cellspacing="0" align=
"right" width="226" cellpadding="0">
  <tr><td>
    <div><img
src="http://newsimg.bbc.co.uk/media/images/
44899000/jpg/_44899470_ken-silva-body.jpg"
width="226" height="170" alt="Ken Silva,
CTO VeriSign" border="0" vspace="0"
hspace="0">
      <div class="cap">"If there is a
      silver lining in all of this, it's
      that users will become more aware
      and more consious of who they do
      business with."</div>
    </div>
  </td></tr>
</table>

<p>Ken Silva, chief technology officer at
 Verisign, said: "We have anticipated
 these flaws in DNS for many years and we
 have  basically engineered around them."
```

Code and Content

```
<p>The Silicon Valley company looks after
 two of the net's 13 DNS root servers. It
 also controls the computers that contain
 the master list of domain name suffixes
 such as .com and .net
<!-- S IIMA -->
<table border="0" cellspacing="0" align=
"right" width="226" cellpadding="0">
  <tr><td>
    <div><img
src="http://newsimg.bbc.co.uk/media/images/
44899000/jpg/_44899470_ken-silva-body.jpg"
width="226" height="170" alt="Ken Silva,
CTO VeriSign" border="0" vspace="0"
hspace="0">
      <div class="cap">"If there is a
      silver lining in all of this, it's
      that users will become more aware
      and more consious of who they do
      business with."</div>
    </div>
  </td></tr>
</table>
<!-- E IIMA -->
<p>Ken Silva, chief technology officer at
 Verisign, said: "We have anticipated
 these flaws in DNS for many years and we
 have  basically engineered around them."
```

```
<p>The Silicon Valley company looks after
two of the net's 13 DNS root servers. It
also controls the computers that contain
the master list of domain name suffixes
such as .com and .net <table border="0"
cellspacing="0" align= "right" width="226"
cellpadding="0"><tr><td><div><img
src="http://newsimg.bbc.co.uk/media/images/
44899000/jpg/_44899470_ken-silva-body.jpg"
width="226" height="170" alt="Ken Silva,
CTO VeriSign" border="0" vspace="0"
hspace="0"><div class="cap">"If there is a
silver lining in all of this, it's that
users will become more aware and more
consious of who they do business with."
</div></div></td></tr></table><p>Ken Silva,
chief technology officer at Verisign, said:
"We have anticipated these flaws in DNS for
many years and we have  basically
engineered around them."
```

Code and Content

```
<p>The Silicon Valley company looks after
 two of the net's 13 DNS root servers. It
 also controls the computers that contain
 the master list of domain name suffixes
 such as .com and .net
<!-- S IIMA -->
<table border="0" cellspacing="0" align=
"right" width="226" cellpadding="0">
  <tr><td>
    <div><img
src="http://newsimg.bbc.co.uk/media/images/
44899000/jpg/_44899470_ken-silva-body.jpg"
width="226" height="170" alt="Ken Silva,
CTO VeriSign" border="0" vspace="0"
hspace="0">
      <div class="cap">"If there is a
      silver lining in all of this, it's
      that users will become more aware
      and more consious of who they do
      business with."</div>
    </div>
  </td></tr>
</table>
<!-- E IIMA -->
<p>Ken Silva, chief technology officer at
 Verisign, said: "We have anticipated
 these flaws in DNS for many years and we
 have  basically engineered around them."
```

```
XXX-------------------------------------
----------------------------------------
----------------------------------------
----------------------------------------
----------------------XXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
XXXXXXXXXXXXXXXXXXXXXXXXXXXXX-------------
----------------------------------------
----------------------------------------
----------------------------------------
XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX----------
----------------------------------------
----------------------------------------
--------------------------------
------------------------
```

Content Code Vector (CCV)

- Clear separation of code or content.
- Data structure for further processing: Vector
- Vector entry represent atomic code or content elements

Character Based CCV

Each entry represents a single character

Token Based CCV

Each entry represents a tag of a word (as for DSC)

Determining Content Code Ratio (CCR)

- Measure how much each entry is surrounded by content and code (in a local neighbourhood)
- Calculate the CCR
- Question:
  How to calculate the CCR?

Determining Content Code Ratio (CCR)

- Measure how much each entry is surrounded by content and code (in a local neighbourhood)
- Calculate the CCR
- Question:
  How to calculate the CCR?

Desired Effects

- strong influence of near entries
- weak influence of far away entries
- convergence to uniform regions

Blurring the CCV

Our approach: local, weighted averages

weights according to a Gauss distribution

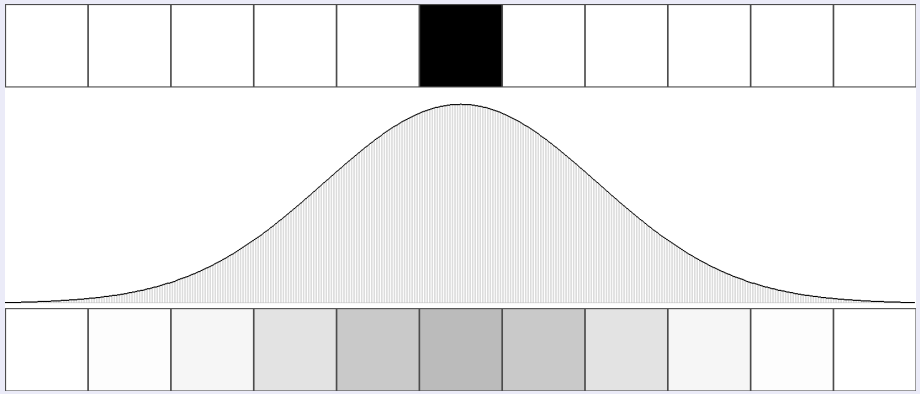
Blurring the CCV

Our approach: local, weighted averages

weights according to a Gauss distribution

Visual interpretation

- Interpret the CCV as 1-D image
- Black is code, white is content
- Apply Gaussian blurring fliter
- Shade of grey corresponds to CCR



Blurring the CCV

On a document level

Blurring the CCV

On a document level

Blurring the CCV

On a document level

Blurring the CCV

On a document level

Blurring the CCV

On a document level

Blurring the CCV

On a document level

Overall process

- Normalise document
- Convert into CCV (1 content, 0 code)
- Repeat blurring until convergence
- Delete contents with CCR below given threshold

Finetuning

Maintain text blocks

- Extract entire text block if one entry is above threshold
- Maintain semantic cohesian of text blocks

Finetuning

Maintain text blocks

- Extract entire text block if one entry is above threshold
- Maintain semantic cohesian of text blocks

Ignore anchor tags in CCV

- Consider anchors like whitespace
- Resolve problems with in-text links
- Links have no *intentional* influence on the format.

Versions of the Algorithm

CCB	character based (basic version)
ACCB	character based, ignoring anchor tags
TCCB	token based, ignoring anchor tags

Parameters

- Threshold, size of local neighbourhood
- Behaviour on a few documents.

Evaluation

Evaluation Method

- Provide Gold Standard for main content
- Use LCS on word sequence as overlap between gold standard and extract
- Determine Recall, Precision and F1
- “Plain” method as baseline

ITA’07: Gottron. Evaluating Content Extraction on HTML Documents

Evaluation Data

Package	URL	Size
bbc	http://news.bbc.co.uk	1000
chip	http://www.chip.de	361
economist	http://www.economist.com	250
espresso	http://espresso.repubblica.it	139
golem	http://golem.de	1000
heise	http://www.heise.de	1000
manual	several	65
repubblica	http://www.repubblica.it	1000
slashdot	http://slashdot.org	364
spiegel	http://www.spiegel.de	1000
telepolis	http://www.telepolis.de	1000
wiki	http://de.wikipedia.org	1000
yahoo	http://news.yahoo.com	1000
zdf	http://www.heute.de	422

Results (F1)

	ACCB	CCB	TCCB	DSC	Crunch	LQF	Plain
bbc	0.924	0.923	0.914	**0.937**	0.756	0.826	0.595
chip	0.703	0.716	**0.842**	0.708	0.342	0.502	0.173
economist	0.890	**0.914**	0.903	0.881	0.815	0.720	0.613
espresso	**0.875**	**0.876**	0.871	0.862	0.810	0.666	0.624
golem	0.959	0.939	0.947	0.958	0.837	0.806	0.502
heise	**0.916**	0.841	0.821	0.877	0.810	0.787	0.575
manual	**0.419**	**0.420**	0.404	0.403	0.382	0.381	0.371
repubblica	**0.968**	0.964	0.918	0.925	0.887	0.816	0.704
slashdot	0.177	0.160	**0.269**	0.252	0.123	0.127	0.106
spiegel	0.861	0.858	**0.910**	0.902	0.706	0.775	0.549
telepolis	0.908	**0.913**	0.902	0.859	0.910	0.906	0.858
wiki	0.682	0.403	0.660	0.594	0.725	0.752	**0.823**
yahoo	0.732	0.742	0.758	**0.780**	0.738	0.670	0.582
zdf	**0.929**	**0.929**	0.745	0.847	0.772	0.578	0.514

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Results (F1)

	ACCB	CCB	TCCB	DSC	Crunch	LQF	Plain
bbc	0.924	0.923	0.914	**0.937**	0.756	0.826	0.595
chip	0.703	0.716	**0.842**	0.708	0.342	0.502	0.173
economist	0.890	**0.914**	0.903	0.881	0.815	0.720	0.613
espresso	**0.875**	**0.876**	0.871	0.862	0.810	0.666	0.624
golem	0.959	0.939	0.947	0.958	0.837	0.806	0.502
heise	**0.916**	0.841	0.821	0.877	0.810	0.787	0.575
manual	**0.419**	**0.420**	0.404	0.403	0.382	0.381	0.371
repubblica	**0.968**	0.964	0.918	0.925	0.887	0.816	0.704
slashdot	0.177	0.160	**0.269**	0.252	0.123	0.127	0.106
spiegel	0.861	0.858	**0.910**	0.902	0.706	0.775	0.549
telepolis	0.908	**0.913**	0.902	0.859	0.910	0.906	0.858
wiki	0.682	0.403	0.660	0.594	0.725	0.752	**0.823**
yahoo	0.732	0.742	0.758	**0.780**	0.738	0.670	0.582
zdf	**0.929**	**0.929**	0.745	0.847	0.772	0.578	0.514

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Results (F1)

	ACCB	CCB	TCCB	DSC	Crunch	LQF	Plain
bbc	0.924	0.923	0.914	**0.937**	0.756	0.826	0.595
chip	0.703	0.716	**0.842**	0.708	0.342	0.502	0.173
economist	0.890	**0.914**	0.903	0.881	0.815	0.720	0.613
espresso	**0.875**	**0.876**	0.871	0.862	0.810	0.666	0.624
golem	0.959	0.939	0.947	0.958	0.837	0.806	0.502
heise	**0.916**	0.841	0.821	0.877	0.810	0.787	0.575
manual	**0.419**	**0.420**	0.404	0.403	0.382	0.381	0.371
repubblica	**0.968**	0.964	0.918	0.925	0.887	0.816	0.704
slashdot	0.177	0.160	**0.269**	0.252	0.123	0.127	0.106
spiegel	0.861	0.858	**0.910**	0.902	0.706	0.775	0.549
telepolis	0.908	**0.913**	0.902	0.859	0.910	0.906	0.858
wiki	0.682	0.403	0.660	0.594	0.725	0.752	**0.823**
yahoo	0.732	0.742	0.758	**0.780**	0.738	0.670	0.582
zdf	**0.929**	**0.929**	0.745	0.847	0.772	0.578	0.514

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Results (F1)

	ACCB	CCB	TCCB	DSC	Crunch	LQF	Plain
bbc	0.924	0.923	0.914	**0.937**	0.756	0.826	0.595
chip	0.703	0.716	**0.842**	0.708	0.342	0.502	0.173
economist	0.890	**0.914**	0.903	0.881	0.815	0.720	0.613
espresso	**0.875**	**0.876**	0.871	0.862	0.810	0.666	0.624
golem	0.959	0.939	0.947	0.958	0.837	0.806	0.502
heise	**0.916**	0.841	0.821	0.877	0.810	0.787	0.575
manual	**0.419**	**0.420**	0.404	0.403	0.382	0.381	0.371
repubblica	**0.968**	0.964	0.918	0.925	0.887	0.816	0.704
slashdot	0.177	0.160	**0.269**	0.252	0.123	0.127	0.106
spiegel	0.861	0.858	**0.910**	0.902	0.706	0.775	0.549
telepolis	0.908	**0.913**	0.902	0.859	0.910	0.906	0.858
wiki	0.682	0.403	0.660	0.594	0.725	0.752	**0.823**
yahoo	0.732	0.742	0.758	**0.780**	0.738	0.670	0.582
zdf	**0.929**	**0.929**	0.745	0.847	0.772	0.578	0.514

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Results (F1)

	ACCB	CCB	TCCB	DSC	Crunch	LQF	Plain
bbc	0.924	0.923	0.914	**0.937**	0.756	0.826	0.595
chip	0.703	0.716	**0.842**	0.708	0.342	0.502	0.173
economist	0.890	**0.914**	0.903	0.881	0.815	0.720	0.613
espresso	**0.875**	**0.876**	0.871	0.862	0.810	0.666	0.624
golem	0.959	0.939	0.947	0.958	0.837	0.806	0.502
heise	**0.916**	0.841	0.821	0.877	0.810	0.787	0.575
manual	**0.419**	**0.420**	0.404	0.403	0.382	0.381	0.371
repubblica	**0.968**	0.964	0.918	0.925	0.887	0.816	0.704
slashdot	0.177	0.160	**0.269**	0.252	0.123	0.127	0.106
spiegel	0.861	0.858	**0.910**	0.902	0.706	0.775	0.549
telepolis	0.908	**0.913**	0.902	0.859	0.910	0.906	0.858
wiki	0.682	0.403	0.660	0.594	0.725	0.752	**0.823**
yahoo	0.732	0.742	0.758	**0.780**	0.738	0.670	0.582
zdf	**0.929**	**0.929**	0.745	0.847	0.772	0.578	0.514

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Results (F1) – Details

	ACCB	TCCB	DSC
bbc	0.924	0.914	**0.937**
chip	0.703	0.842	*0.708*
economist	**0.890**	0.903	0.881
espresso	*0.875*	0.871	0.862
golem	0.959	0.947	0.958
heise	**0.916**	0.821	0.877
manual	*0.419*	0.404	0.403
repubblica	**0.968**	0.918	0.925
slashdot	0.177	0.269	**0.252**
spiegel	0.861	0.910	**0.902**
telepolis	**0.908**	0.902	0.859
wiki	**0.682**	0.660	0.594
yahoo	0.732	0.758	**0.780**
zdf	**0.929**	0.745	0.847

Observations

Results (F1) – Details

	ACCB	TCCB	DSC
bbc	0.924	0.914	**0.937**
chip	0.703	0.842	*0.708*
economist	**0.890**	0.903	0.881
espresso	*0.875*	0.871	0.862
golem	0.959	0.947	0.958
heise	**0.916**	0.821	0.877
manual	*0.419*	0.404	0.403
repubblica	**0.968**	0.918	0.925
slashdot	0.177	0.269	**0.252**
spiegel	0.861	0.910	**0.902**
telepolis	**0.908**	0.902	0.859
wiki	**0.682**	0.660	0.594
yahoo	0.732	0.758	**0.780**
zdf	**0.929**	0.745	0.847

Observations

- ACCB generally better than DSC

Results (F1) – Details

	ACCB	TCCB	DSC
bbc	0.924	0.914	**0.937**
chip	0.703	0.842	*0.708*
economist	**0.890**	0.903	0.881
espresso	*0.875*	0.871	0.862
golem	0.959	0.947	0.958
heise	**0.916**	0.821	0.877
manual	*0.419*	0.404	0.403
repubblica	**0.968**	0.918	0.925
slashdot	0.177	0.269	**0.252**
spiegel	0.861	0.910	**0.902**
telepolis	**0.908**	0.902	0.859
wiki	**0.682**	0.660	0.594
yahoo	0.732	0.758	**0.780**
zdf	**0.929**	0.745	0.847

Observations

- ACCB generally better than DSC
- Sometimes tag based CCV better?

Results (F1) – Details

	ACCB	TCCB	DSC
bbc	0.924	0.914	**0.937**
chip	0.703	0.842	*0.708*
economist	**0.890**	0.903	0.881
espresso	*0.875*	0.871	0.862
golem	0.959	0.947	0.958
heise	**0.916**	0.821	0.877
manual	*0.419*	0.404	0.403
repubblica	**0.968**	0.918	0.925
slashdot	0.177	*0.269*	**0.252**
spiegel	0.861	0.910	**0.902**
telepolis	**0.908**	0.902	0.859
wiki	**0.682**	0.660	0.594
yahoo	0.732	0.758	**0.780**
zdf	**0.929**	0.745	0.847

Observations

- ACCB generally better than DSC
- Sometimes tag based CCV better?
- Short main content problematic

Conclusion and Future Work

Conclusions and Future Work

Conclusions

- ACCB does not have drawbacks compared to CCB
- ACCB better than DSC

Future Work

- New models for CCV
- Incorporate DOM structure
- Explore parameter space
- Combination of heuristics
- Use “external” knowledge (TD)

Questions?