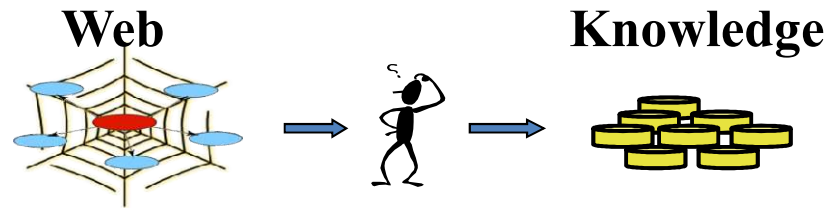


# WEBIST 2016

12<sup>th</sup> International Conference on Web Information Systems and Technologies

Rome, Italy // 23 - 25 April, 2016



## Content-based Title Extraction from Web Page

Najlah Gali and **Pasi Fränti**



UNIVERSITY OF  
EASTERN FINLAND

24.4.2016

# Motivation

# Application: search engine

**MOPSI** See what's around

bar Search

Mopsi service Photo Web search

1 **Bar Play Joensuu S-**  
**kanava\***  
Kauppakatu 23 80100  
Joensuu  
882 m  
Check route

2 Jet Set Sport Bar\*  
Kauppakatu 35 80100  
Joensuu  
1 km 192 m  
Check route

3 Super Smoothie  
Joensuu Cafe & Salad  
Bar smoothie bar\*  
Torikatu 31 Joensuu  
1 km 243 m  
Check route

**SINUN ETUSI**

**Jet Set**

Bar Play Joensuu ...

**SINUN ETUSI**

Osastojuhlat  
Nallerock  
25-26.7.2014

Kaupunkikatu 23 80100 Joensuu

882 m Street View

**Title** **Address** **Image**

**Calculating distance**

# Summary Extraction

- **Title** *Rosso restaurant*, *“City pharmacy”*
- **Keywords** *“restaurant, food, lunch, dinner”*
- **Representative Image**
- **Short description**

ma-pe: 16.00-22.00

la: 12.00-22.00

puh. 013 227 874



What we deal with

# Content of Web Page

Hypertext Markup Language (HTML, XHTML)

Logo image



Navigation bar

Title

Raspberry Pi 3 adds wi-fi and Bluetooth

Keywords



UK astronaut Tim Peake took a Raspberry Pi to the International Space Station

**The Raspberry Pi** has become the most popular British computer yet made.

The title was formerly held by the Amstrad PCW which is believed to have sold a total of eight million units.


Sales of the Raspberry Pi will surpass that figure this month, said the Raspberry Pi project founder Eben Upton.

Text


Top Stories

- Oscars 2016: DiCaprio finally wins  
1 hour ago
- UN to expand Syria aid as truce holds  
30 minutes ago
- Pakistan hangs killer of state governor  
4 hours ago

Features & Analysis



He said yes!  
Eight women who proposed to their partners



Images

# Web Page Title

The title can be in three different places:

Title Tag

`<title>Wentworth House Hotel Bath Hotels - Cheap Hotels in Bath, Somerset, UK</title>`

Logo image

Web Page body



Segmenting title tags

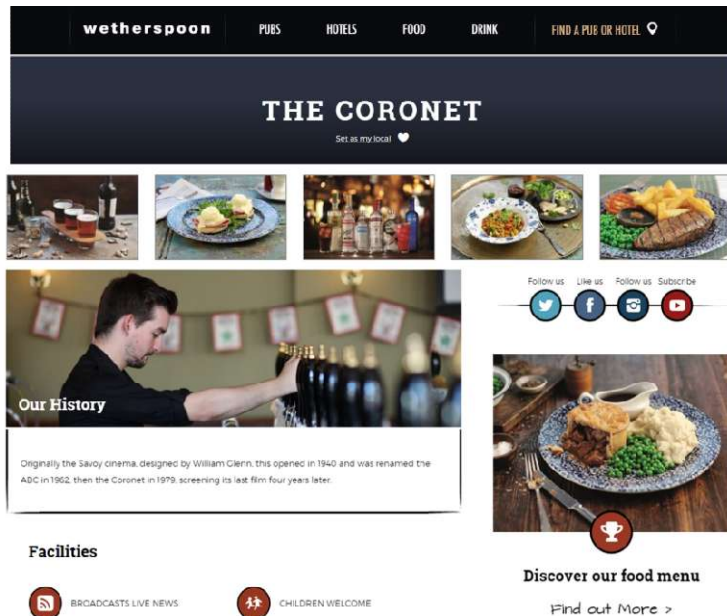


# Title and Meta Tags

- The obvious source
- But includes also additional information
  - `<title>` *Piato Restaurant* – 123 Blues Point Road, McMahons Point, Sydney | Visit Piato and experience the life & flavour of Europe. North Sydney Functions. North Sydney Restaurants.`</title>`
  - `<title>` Joensuu Keskusta | *Intersport* - Sport to the people `</title>`
- Segmentation is needed!
  - Joensuu Keskusta
  - **Intersport**
  - Sport to the people

# Work flow

<https://www.jdwetherspoon.com/pubs/all-pubs/england/london/the-coronet-holloway>



Web page

Extract title & meta tags from the page

Segment content by delimiters

Construct candidate list

Score candidate segments

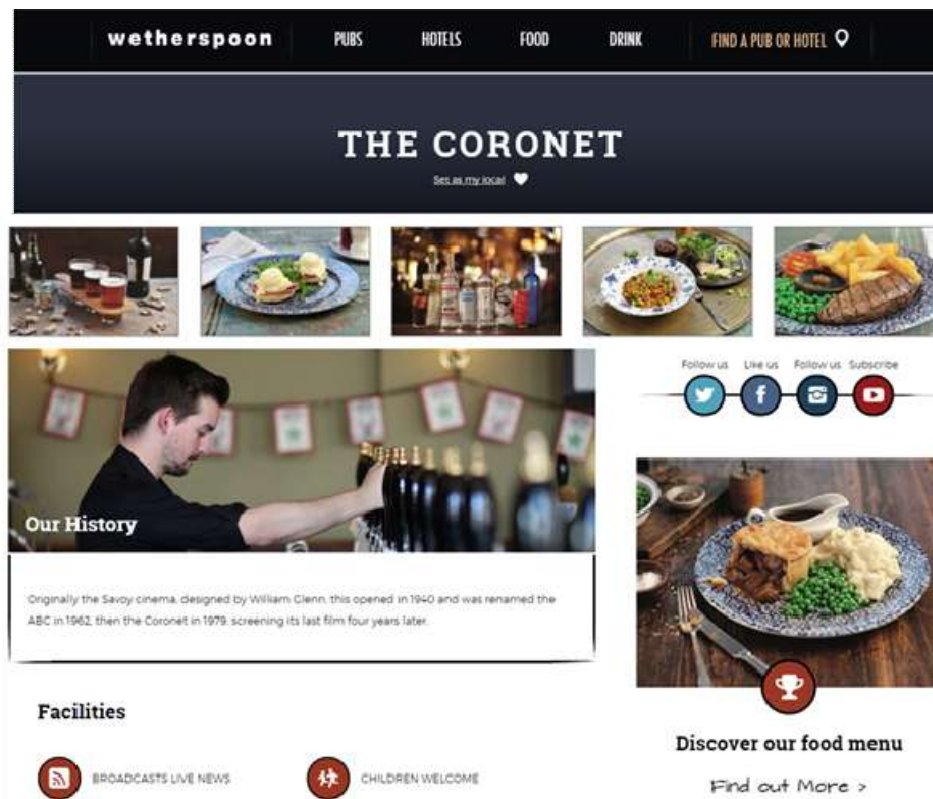
1. Placement in title & meta tags
2. Popularity in header tags
3. Position in the web link

**Title** The coronet

# Content of Title and Meta tags

<title>*The Coronet*, Holloway | Our Pubs | J D Wetherspoon </title>

<meta name="keywords" content="*The Coronet*" />



The screenshot shows the website for 'The Coronet' pub, part of the J D Wetherspoon chain. The top navigation bar includes 'wetherspoon', 'PUBS', 'HOTELS', 'FOOD', 'DRINK', and a search button 'FIND A PUB OR HOTEL'. The main heading is 'THE CORONET' with a sub-link 'See all my local'. Below this is a row of five small images: a bar with drinks, a plate of food, a bar with bottles, a plate of food, and a plate of food. To the right of these images are social media icons for Twitter, Facebook, Instagram, and YouTube, with labels 'Follow us', 'Like us', 'Follow us', and 'Subscribe'. Below the social media icons is a large image of a man pouring beer, with the heading 'Our History' and a text box stating: 'Originally the Savoy cinema, designed by William Clenn, this opened in 1940 and was renamed the ABC in 1962, then the Coronet in 1979, screening its last film four years later.' Below the history section is a 'Facilities' section with icons for 'BROADCASTS LIVE NEWS' and 'CHILDREN WELCOME'. To the right of the facilities section is a large image of a plate of food, with a heading 'Discover our food menu' and a link 'Find out More >'. A small trophy icon is visible below the food image.

# Segmentation by delimiters

<title>Sydney Waterfront Restaurant | Restaurant Milsons Point - Aqua Dining</title>

<title>SIGNORELLI GASTRONOMIA - Pyrmont Italian Restaurant - EAT • DRINK • SHOP • COOK Italian Restaurant Pyrmont Sydney – Signorelli Gastronomia</title>

<title>Neutral Bay Club | Tennis, Bowls, Bistro & Functions | Sydney</title>

<title>The Coronet, Holloway | Our Pubs | J D Wetherspoon</title>

## Pre-defined delimiter patterns

space – space	space / space	space . space
space : space	, space	space -
: space	space :	space
space >	space «	space »
? ,	- ,	space ::
Space /	-	space <

# Candidate Segments

<**title**>The Coronet, Holloway | Our Pubs | J D Wetherspoon</**title**>

<**meta** name="keywords" content="The Coronet" />

## Candidates

- The Coronet
- Holloway
- Our Pubs
- J D Wetherspoon

The screenshot shows the Wetherspoon website for The Coronet. The navigation bar includes 'wetherspoon', 'PUBS', 'HOTELS', 'FOOD', 'DRINK', and 'FIND A PUB OR HOTEL'. The main heading is 'THE CORONET' with a 'Set as my local' link. Below the heading are five small images: a tray of drinks, a plate of food, a bar with bottles, a plate of food, and a plate of food. A large image shows a man pouring beer, with the text 'Our History' and a paragraph: 'Originally the Savoy cinema, designed by William Clenn, this opened in 1940 and was renamed the ABC in 1962, then the Coronet in 1979, screening its last film four years later.' Below this is the 'Facilities' section with icons for 'BROADCASTS LIVE NEWS' and 'CHILDREN WELCOME'. On the right, there are social media icons for Twitter, Facebook, Instagram, and YouTube, and a 'Discover our food menu' section with a 'Find out More >' link.

# Scoring the candidates

# 1. Position in Title and Meta Tags

- Appear first or last either in Title or Meta gets 0.1

**0.1**                      **0.0**                      **0.0**                      **0.1**  
<**title**>The Coronet, Holloway | Our Pubs | J D Wetherspoon </**title**>  
Or  
<**meta** name="keywords" content="The Coronet" /> **0.1**

## Candidates

- The Coronet                      **0.1**
- Holloway                      **0.0**
- Our Pubs                      **0.0**
- J D Wetherspoon                      **0.1**

# 2. Popularity Among Header Tags

`<h1 class="banner-inner__title">The Coronet</h1>`

`<h2 class="venue-finder__title-text">Find a pub or hotel</h2>`

`<h2 class="venue-finder__title-text">Our Pubs</h2>`

`<h2 class="venue-finder__title-text" ng-hide="isPubName">Check out your nearest pub or hotel</h2>`

`<h3 class="feature-panel__title">Discover our food menu</h3>`

`<h3 class="feature-panel__title">Our drinks selection</h3>`

`<h4 class="tab__title">Nearby J D Wetherspoons</h4>`

## Candidates

- The Coronet  $1 \times 6 = 6$
- Holloway  $0$
- Our Pubs  $1 \times 5 = 5$
- J D Wetherspoon  $1 \times 3 = 3$

Frequency      Weight

`<h1 > = 6`

`<h2 > = 5`

`<h3 > = 4`

`<h4 > = 3`

`<h5 > = 2`

`<h6 > = 1`



# 3. Position in Web Link

**Domain**                      **Path**                      **File name**

https:// **www.jdwetherspoon.com/** pubs/all-pubs/england/london/ **the-coronet-holloway**

**1**                                      **1.5**                                      **3**

Dice similarity measure

## Candidates

- The Coronet                       $3 \times 0.70$                       = 2.10
- Holloway                       $3 \times 0.58$                       = 1.74
- Our Pubs                       $1.5 \times 0.00$                       = 0.00
- J D Wetherspoon                       $1 \times 1.00$                       = 1.00

# Rank Segments

---

Candidates	Position in tag	Popularity among header tags	Position in web link
The Coronet	0.1	6	2.10
Holloway	0.0	0	1.74
Our Pubs	0.0	5	0.00
J D Wetherspoon	0.1	3	1.00

---

Normalizing



---

Candidates	Position in tag	Popularity among header tags	Position in web link	Total
<b>The Coronet</b>	0.1	1.00	1.00	<b>2.10</b>
J D Wetherspoon	0.1	0.50	0.48	1.08
Holloway	0.0	0.00	0.83	0.83
Our Pubs	0.0	0.83	0.00	0.83

---

# Experiments

# Data Set

- Websites: 1245 <http://cs.uef.fi/mopsi/titleextraction/data/> (in this paper)  
<http://cs.uef.fi/mopsi/titler/> (extended)
- Categories: 8  
food & drink, home & garden, accommodation & hotels, shopping, arts & entertainment, hobbies & leisure, sport, and health & social care
- Sources: Google and Google maps
- Collected: 18–31.7.2014 + 19-23.4.2015
- Ground truth titles are manually extracted according to the specifications in [Hu et al. 2005]

# Impact of criteria

**Criterion 1:** lowest impact (0.65)

- Generic words (*home, welcome*) often at the beginning
- Slogan or address often at the end.

**Criterion 2:** slightly higher impact (0.68)

- Heading tags not always used, not always correct title there.

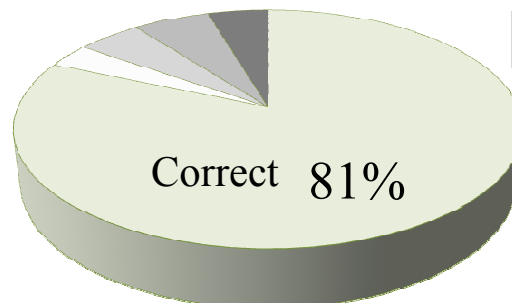
**Criterion 3:** Best (0.84).

Criteria	Average similarity
(1) Position in tag	0.65
(2) Popularity among <i>hx</i> tags	0.68
(3) Position in web link	<b>0.84</b>
1 + 2	0.70
1 + 3	<b>0.85</b>
2 + 3	<b>0.82</b>
1 + 2 + 3	<b>0.84</b>

# Qualitative Analysis of TTA

Title	Ground truth	Content of Title tag	Selected string
Correct	3 Weeds Hotel	<b>3 Weeds Hotel</b>   Unique Pub   Bars   Restaurant   Party Venue   Inner West Sydney	3 Weeds Hotel
Short	Irish Channel Restaurant & Pub	<b>Irish Channel - Restaurant &amp; Pub</b>   500 H St NW DC (202) 216-0046	Irish Channel
Long	Secret Garden Bed & Breakfast	<b>Secret Garden Bed &amp; Breakfast</b> (formerly Whitegates Guest House), near Keynsham, Bristol: Rooms, Prices and Guest Information	Secret Garden Bed & Breakfast (formerly Whitegates Guest House)
No title	Rio Pool	Hot Tubs, hot tub hire, swimming pools, Bristol, Gloucester	swimming pools
Incorrect	Slice and Dice	Home   Prepared Food   Swansea   <b>Slice and Dice</b> UK	Swansea

Long (5 %)  
[CATEGORY  
NAME] (3 %)



No title (6 %)

[CATEGORY  
NAME] (5 %)

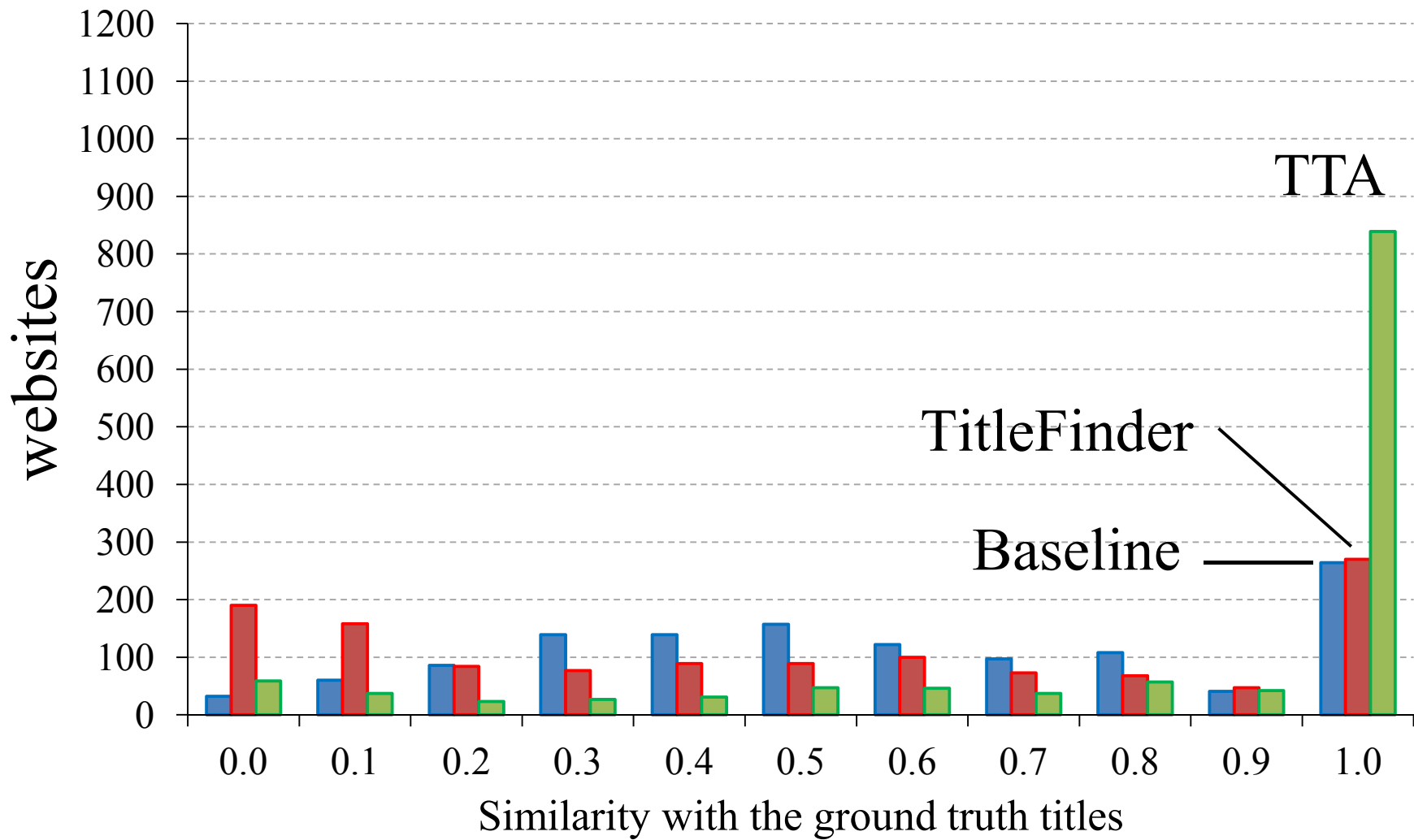
# Comparative Results

---

<b>Method</b>	<b>Average similarity</b>
Title tag (baseline)	0.62
TitleFinder (Mohammadzadeh et al. 2012)	0.52
TTA (proposed)	<b>0.84</b>

---

# Comparative Results





# Results with *Mopsi Services*



Annotated titles

Method	Rouge-1			Jaccard	Dice
	Precision	Recall	F-score		
Baseline (Title Tag)	<b>0.71</b>	0.33	0.41	0.44	0.54
TitleFinder (Moham.et al. 2012)	0.35	0.47	0.37	0.37	0.43
Styling (Changuel et al. 2009)	0.14	0.21	0.15	0.22	0.28
TTA (Gali and Fränti 2016)	0.52	<b>0.59</b>	<b>0.52</b>	<b>0.54</b>	<b>0.62</b>

# WEBIST 2016

12<sup>th</sup> International Conference on Web Information Systems and Technologies

Rome, Italy // 23 - 25 April, 2016

[PRIMORIS](#) [Contacts](#) [FAQs](#) [INSTICC Portal](#)



WEBIST 2016 will be held in conjunction with [CSEDU 2016](#), [SMARTGREENS 2016](#), [CLOSER 2016](#), [VEHITS 2016](#), [IoTBD 2016](#) and [COMPLEXIS 2016](#).

Registration to WEBIST allows free access to the CSEDU, SMARTGREENS, CLOSER, VEHITS, IoTBD and COMPLEXIS conferences (as a non-speaker).

## Actions

### On-line Registration

[Registration Fees](#)

[Deadlines and Policies](#)

### Submit Paper

[Author's Kit](#)

### Author's Login

### Reviewer's Login

## Information

### Conference Details

[Important Dates](#)

[Technical Program](#)

[Social Event](#)

[Call for Papers](#)

[Program Committee](#)

[Event Chairs](#)

[Keynote Lectures](#)

[Best Paper Awards](#)

### Satellite Events

[Workshops](#)

[Special Sessions](#)

[Tutorials](#)

[Demos](#)

[Panels](#)

[Doctoral Consortium](#)

[Open Communications](#)

**NEW REGISTRATIONS ARE NOW ONLY AVAILABLE AT THE CONFERENCE WELCOME DESK**

Please visit the [WEBIST 2017 website](#)



The purpose of the 12th International Conference on Web Information Systems and Technologies (WEBIST) is to bring together researchers, engineers and practitioners interested in the technological advances and business applications of web-based information systems. The conference has five main tracks, covering different aspects of Web Information Systems, including Internet Technology, Web Interfaces and Applications, Society, e-Communities, e-Business, Web Intelligence and Mobile Information Systems.



# WEBIST 2016

12<sup>th</sup> International Conference on Web Information Systems and Technologies

Rome, Italy // 23 - 25 April, 2016

## Title Tag:

**WEBIST 2016 - 12th International Conference on Web Information Systems and...**



### Actions

#### On-line Registration

Registration Fees

Deadlines and Policies

#### Submit Paper

Author's Kit

#### Author's Login

#### Reviewer's Login

### Information

#### Conference Details

Important Dates

Technical Program

Social Event

Call for Papers

Program Committee

Event Chairs

Keynote Lectures

Best Paper Awards

#### Satellite Events

Workshops

Special Sessions

Tutorials

Demos

Panel

Doctoral Consortium

WEBIST 2016 will be held in conjunction with [CSEDU 2016](#), [SMARTGREENS 2016](#), [CLOSER 2016](#), [VEHITS 2016](#), [IoTBD 2016](#) and [COMPLEXIS 2016](#).

Registration to WEBIST allows free access to the CSEDU, SMARTGREENS, CLOSER, VEHITS, IoTBD and COMPLEXIS conferences (as a non-speaker).

## Selected:

**WEBIST 2016**

NEW REGISTRATION INFORMATION AVAILABLE AT THE CONFERENCE WELCOME DESK

Please visit the WEBIST 2017 website

# WEBIST 2017

12<sup>th</sup> International Conference on Web Information Systems and Technologies

Porto, Portugal // 23 - 25 April, 2017

## Others:

### Mobile Information Systems

The purpose of the 12th International Conference on Web Information Systems and Technologies (WEBIST) is to bring together researchers, engineers and practitioners interested in the technological advances and business applications of web-based information systems. The conference has five main tracks, covering different aspects of Web Information Systems, including Internet Technology, Web Interfaces and Applications, Society, e-Communities, e-Business, Web Intelligence and Mobile Information Systems.



# What about logo images?

- ~89 % of web pages have their title within a logo image
- Needs to detect logo image
- Apply OCR
- Challenging !!!

LOCAL BISTRO

Santa's Reindeers

Savon Kinot



Savon Kinet

Etusivu Myynnissä Tulossa Sarjat ja Eventit Ajankohtaista Liput

Kirjaudu sisään JOENSUU

Giacomo Puccini  
**TOSCA**  
rooleissa Jonas Kaufmann,  
Karita Mattila ja Juha Uusitalo

15.3. + 17.3.  
Joensuu & Savonlinna  
12.3. + 15.3. + 17.3.  
Varkaus  
14.4.  
Iisalmi & Kitee

Etkö löydä  
etsimääsi  
esitysaikaa?  
Klikkaa ja lue lisää.

**SARJA-  
LIPPUJA  
NYT MYÖS  
NETISTÄ!**

Joensuu Tänään, 17.3.2016 Ajankohta Kaikki laityypit

Joensuun  
**PYÖRÄKELLARI**  
Keskikatu 23  
00100 Joensuu puh.013-211008

Joensuun Pyöräkellari

Parasta pyörällesi!

Polkupyörien monimerkkihuolto Joensuun keskustassa. Ammattitaitoinen huolto ja korkealaatuisten varaosien käyttö on eräs yritystoimintamme perussasioita, samoin työtilausten sujuva vastaanotto, laadukas ja nopea huoltotyö, sekä polkupyörän luovutus asiakkaalle. Meille on kertynyt vuosien saatossa tuhansia tyytyväisiä asiakkaita ja näin toimimme jatkossakin.

+++ Ketjuöljy

27.16

**LOCAL BISTRO**  
SE ASTETTA  
PAREMPI  
PAIKALLINEN

*Santa's Reindeers*  
*Laukkalan Mökit*

GSM: +358 50 5924 252 (FIN, ENG), +358 40 158 2618 (RUS)  
e-mail: info@fincottages.fi, laukkanen.svellana@gmail.com (RUS)

AKTIVITEETIT

Santa's Reindeers & Laukkalan Mökit

Tule kokemaan puhdas ja monimuotoinen luonto. Tarjoamme majoitusta kaunisä järvimaaisemassa, sekä elämyksiä ympäri vuoden, koko perheelle.

Mökit & Majoitus

Willa  
Ailexia  
Iso Mökki  
Kesk. Mökki

# Conclusions

- TTA improves baseline

62% → 84%

- Title tag works ok, but it needs to be processed

**<title>This still ok</title>** 

- Words in the page link have the highest impact

**<http://www.webist.org/>**