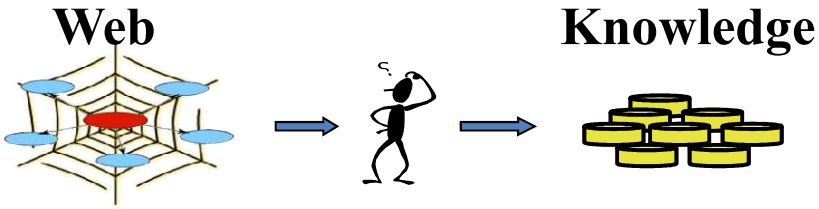
WEBIST 2016
12th International Conference on Web Information Systems and Technologies
Rome, Italy // 23 - 25 April, 2016



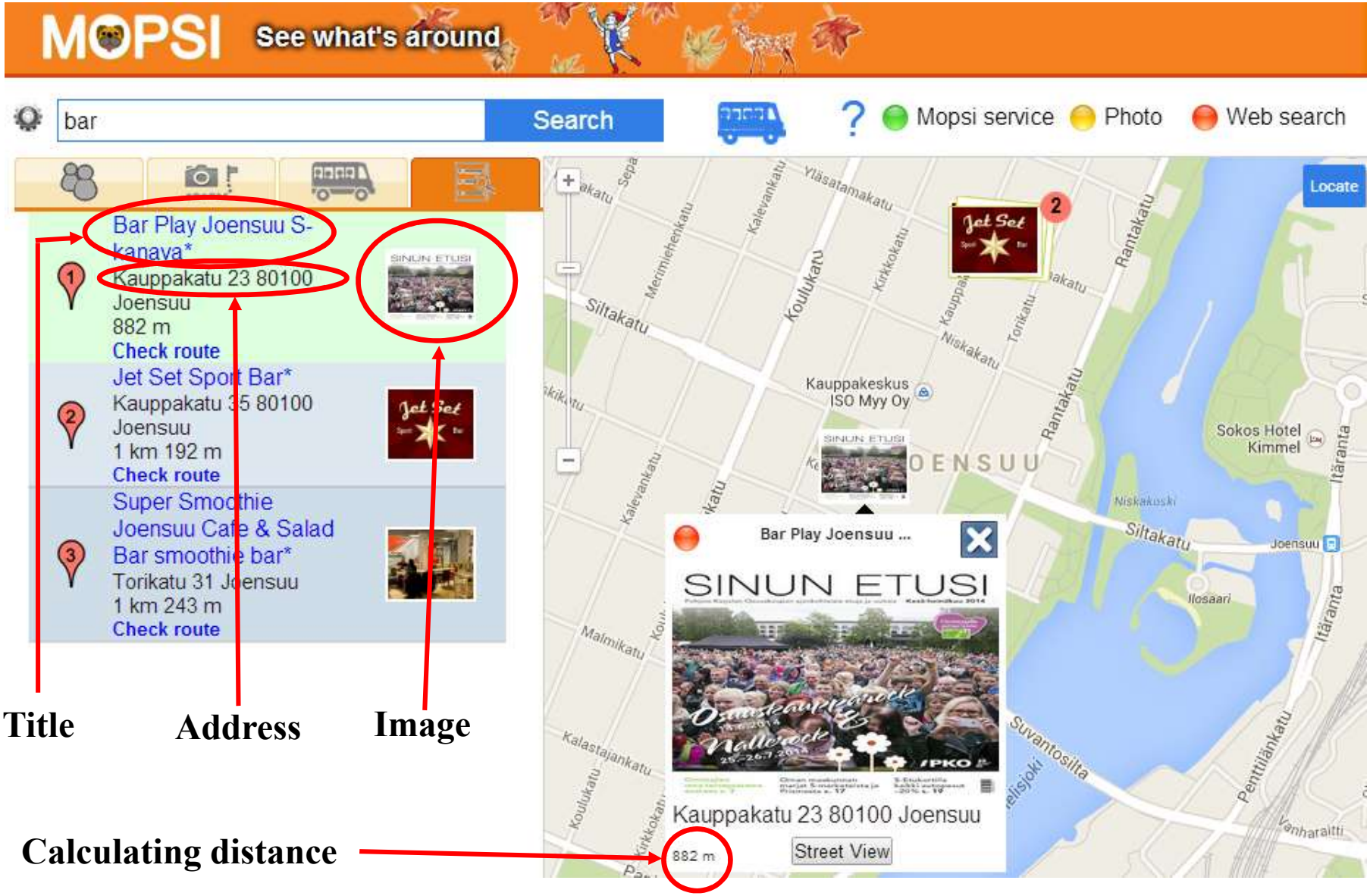
Content-based Title Extraction from Web Page

Najlah Gali and **Pasi Fränti**

UNIVERSITY OF
EASTERN FINLAND

Motivation

Application: search engine



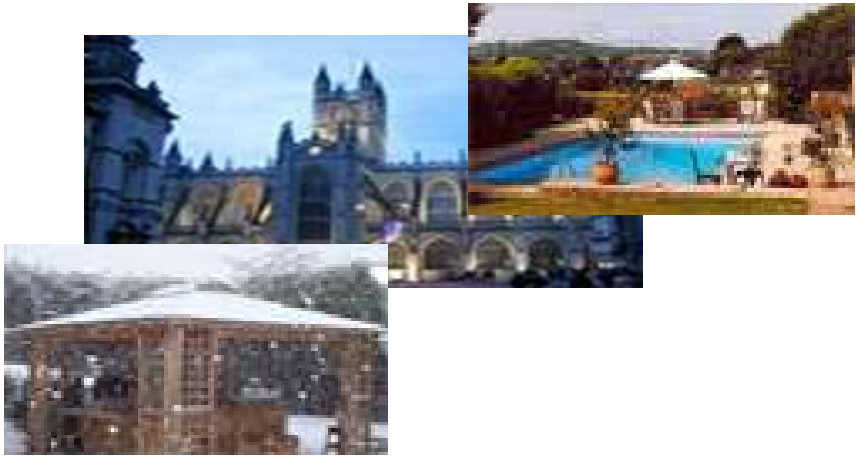
Summary Extraction

- **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

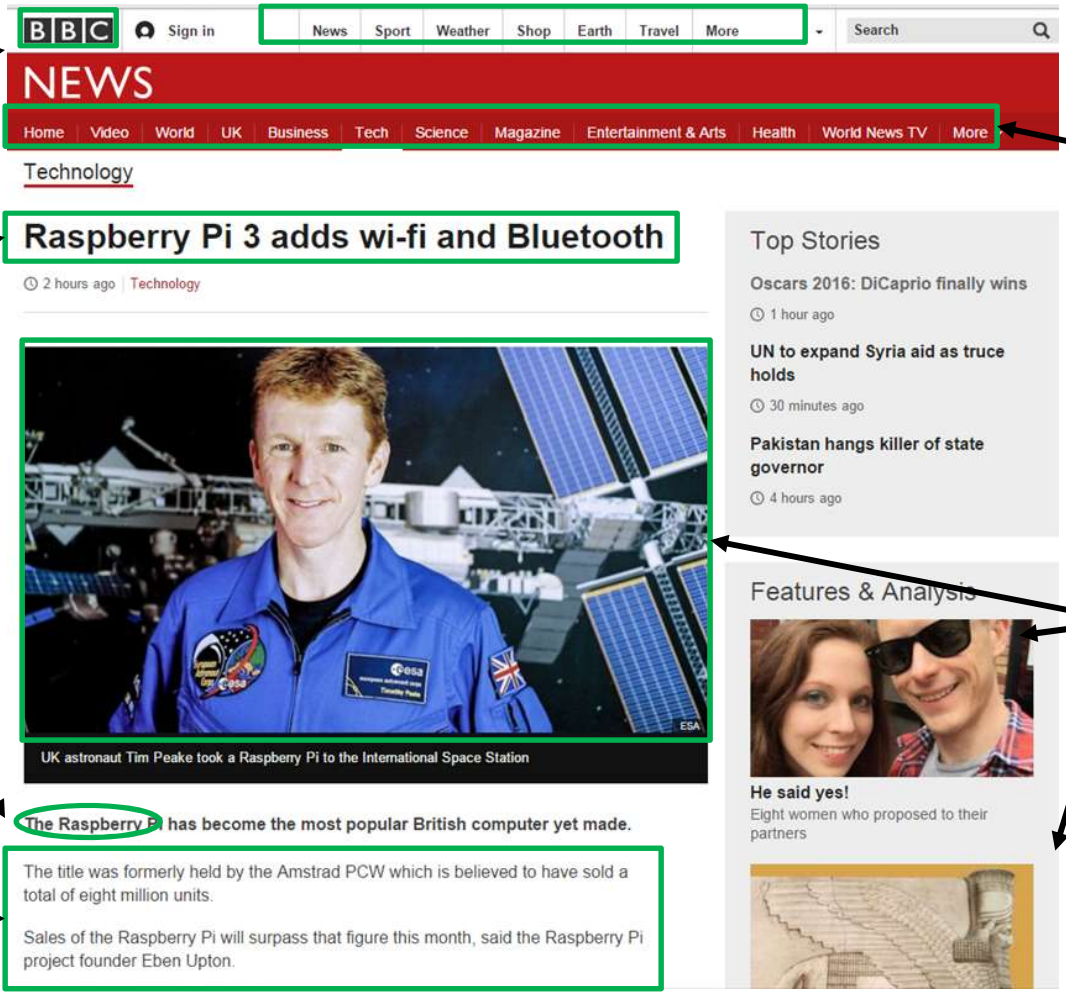
Navigation bar

Title

Keywords

Images

Text



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>

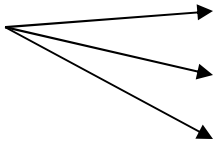
Logo image

Web Page body

Segmenting title tags

Title and Meta Tags

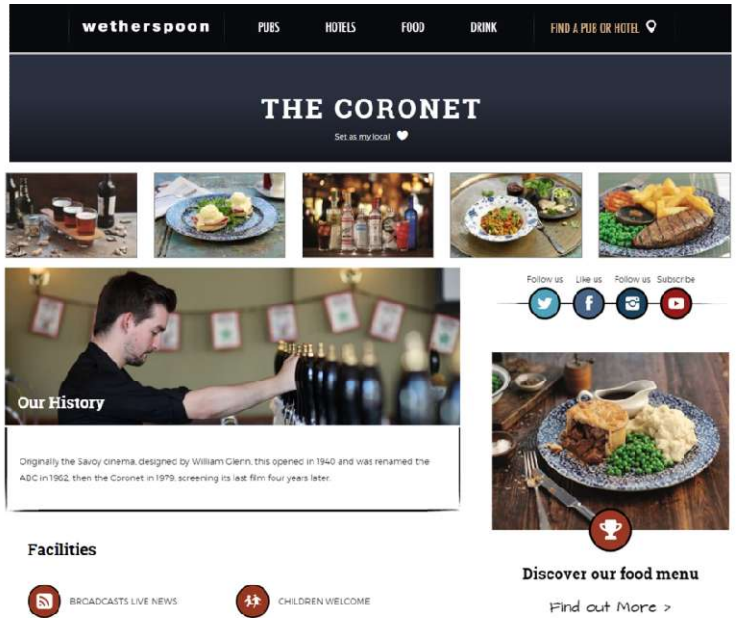
- The obvious source
- But includes also additional information
  - <**title**> ***Piato Restaurant*** – 123 Blues Point Road, McMahons Point, Sydney | Visit Piato and experience the life & flavour of Europe. North Sydney Functions. North Sydney Restaurants.</**title**>
  - <**title**> Joensuu Keskusta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

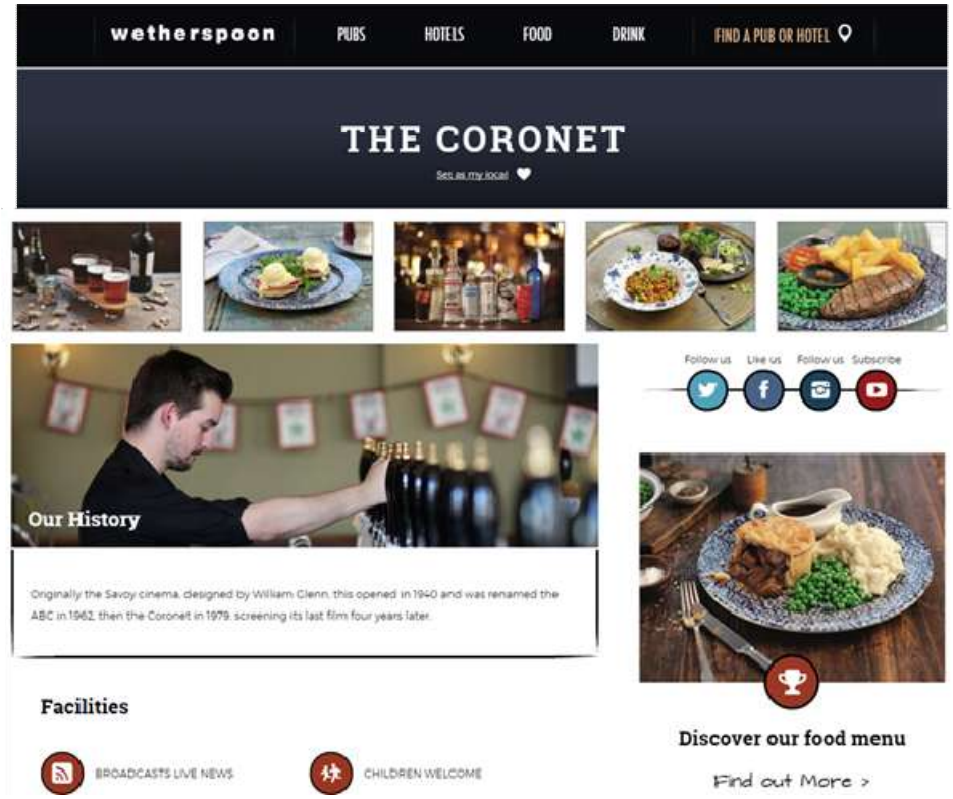
1. Placement in title & meta tags
2. Popularity in header tags
3. Position in the web link

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*" />



Segmentation by delimiters

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

<title>**SIGNORELLI GASTRONOMIA** - Pyrmont Italian Restaurant - EAT • DRINK • SHOP • COOKItalian 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 <

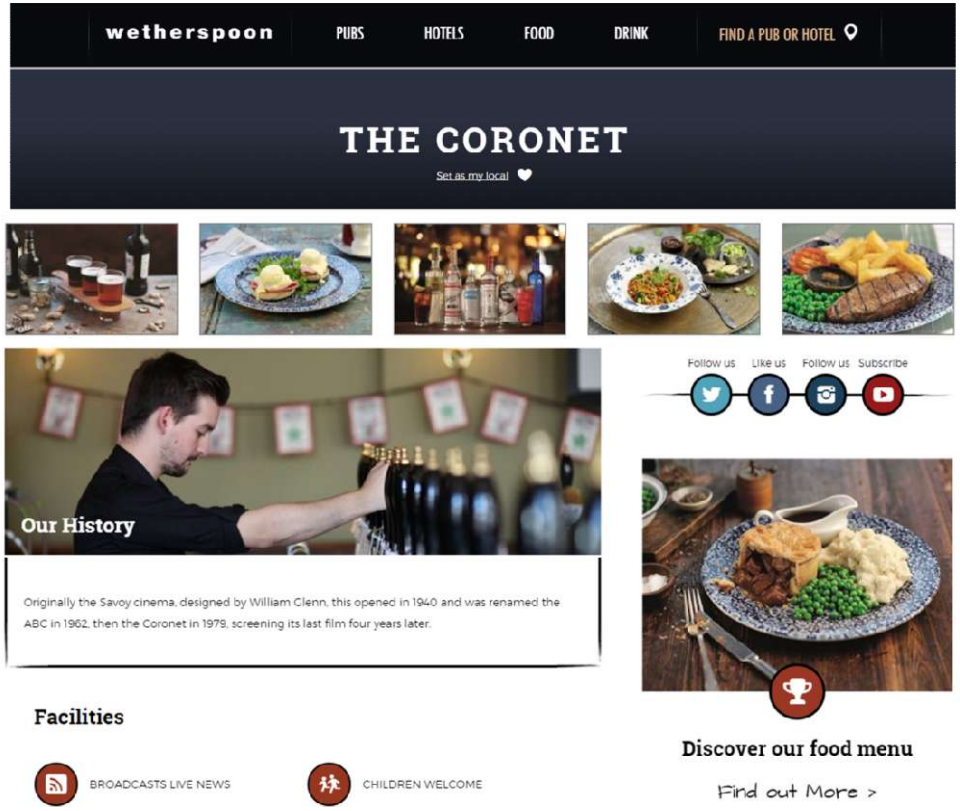
Candidate Segments

<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



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

0.1

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

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 $1 \times 6 = \mathbf{6}$
- Holloway $\mathbf{0}$
- Our Pubs $1 \times 5 = \mathbf{5}$
- J D Wetherspoon $1 \times 3 = \mathbf{3}$

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 × 0.70	= 2.10
• Holloway	3 × 0.58	= 1.74
• Our Pubs	1.5 × 0.00	= 0.00
• J D Wetherspoon	1 × 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
The Coronet	0.1	1.00	1.00	**2.10**
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)

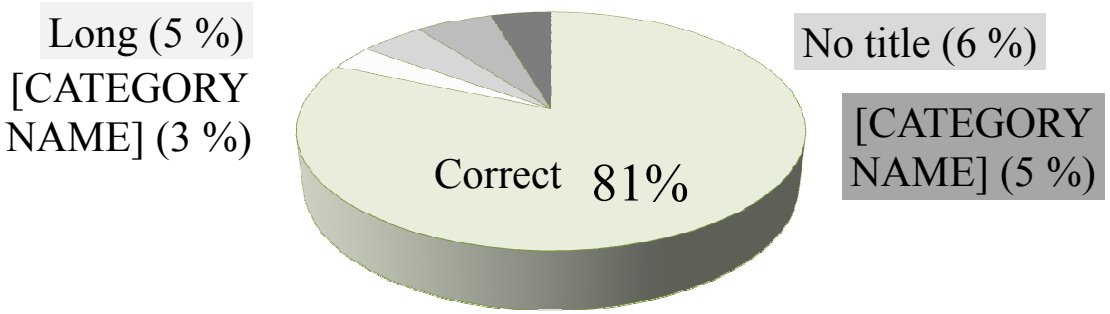
- Heading tags not always used, not always correct title there.

Criterion 3: Best (0.84).

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

Qualitative Analysis of TTA

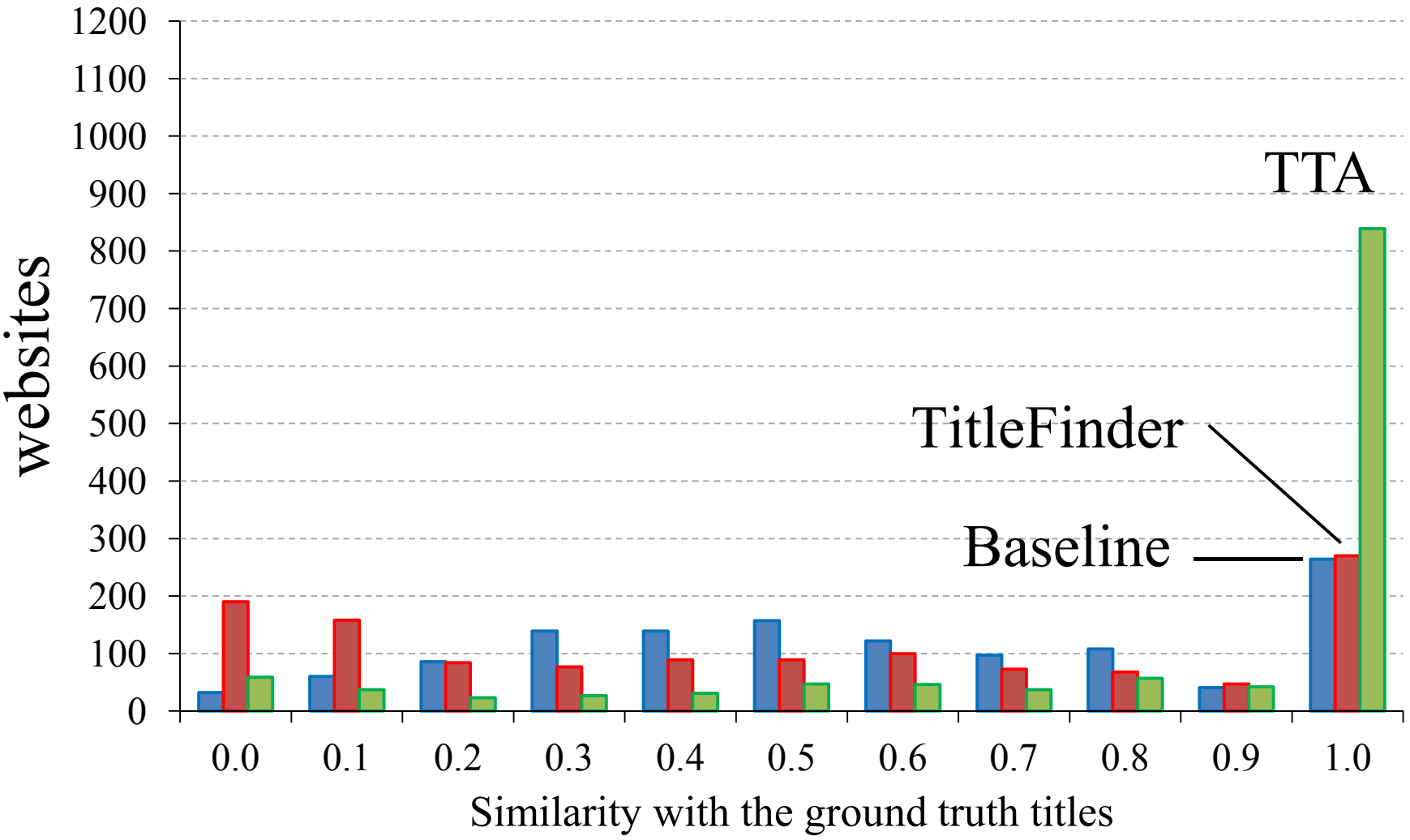
Title	Ground truth	Content of Title tag	Selected string
Correct	3 Weeds Hotel	**3 Weeds Hotel** | Unique Pub | Bars | Restaurant | Party Venue | Inner West Sydney	3 Weeds Hotel
Short	Irish Channel Restaurant & Pub	**Irish Channel - Restaurant & Pub** | 500 H St NW DC (202) 216-0046	Irish Channel
Long	Secret Garden Bed & Breakfast	**Secret Garden Bed & Breakfast** (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 | **Slice and Dice** UK	Swansea



Comparative Results

Method	**Average similarity**
Title tag (baseline)	0.62
TitleFinder (Mohammadzadeh et al. 2012)	0.52
TTA (proposed)	**0.84**

Comparative Results



Results with *Mopsi Services*

Annotated titles

Method	**Rouge-1**			**Jaccard**	**Dice**
	Precision	**Recall**	**F-score**		
Baseline (Title Tag)	**0.71**	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	**0.59**	**0.52**	**0.54**	**0.62**

WEBIST 2016

12th International Conference on Web Information Systems and Technologies

Rome, Italy // 23 - 25 April, 2016

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

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).

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.

Mobile Information Systems

Title Tag:

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

Selected:

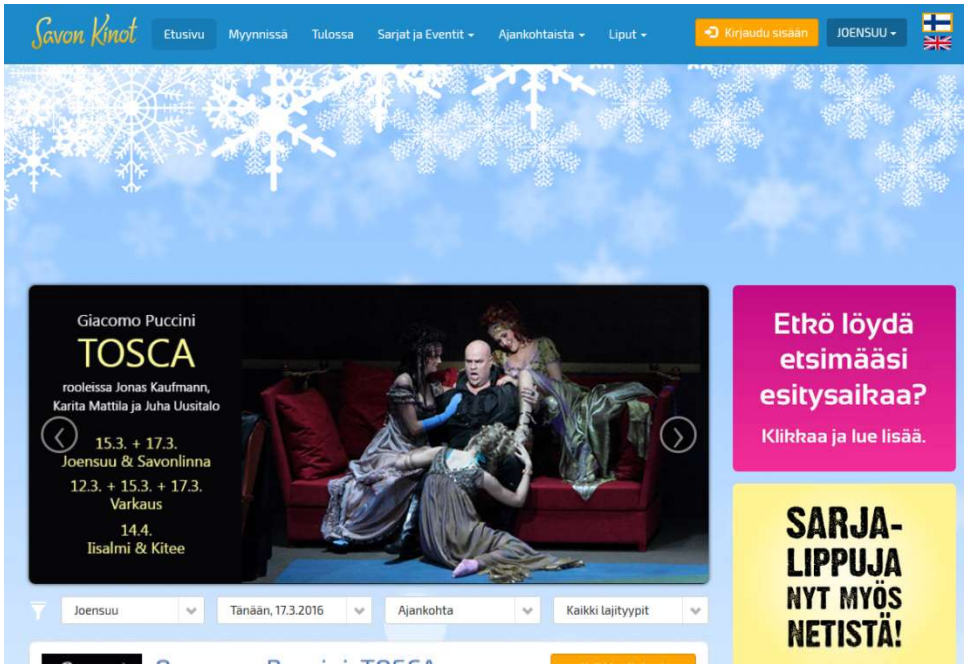
WEBIST 2016

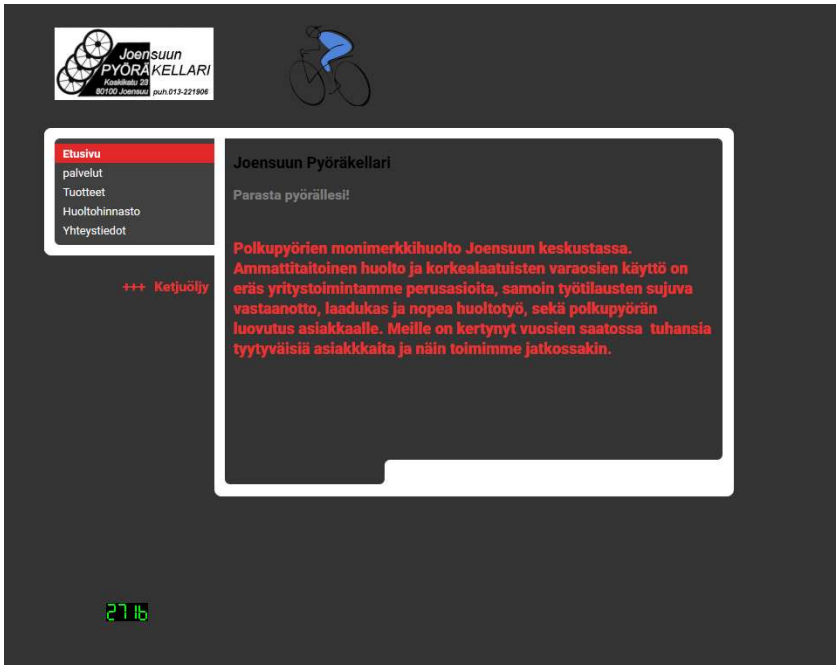
Others:

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 !!!


Savon Kinot
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 lajityypit


Joensuun PYÖRÄKELLARI
Etusivu
palvelut
Tuotteet
Huoltohinnasto
Yhteystiedot
+++ Ketjuöljy
Joensuun Pyöräkellari
Parasta pyörällesi!
Polkupyörien monimerkkihuolto Joensuun keskustassa. Ammattitaitoinen huolto ja korkealaatuisten varaosien käyttö on eräs yritystoimintamme perusasioita, samoin työtilausten sujuva vastaanotto, laadukas ja nopea huoltotyö, sekä polkupyörän luovutus asiakkaalle. Meille on kertynyt vuosien saatossa tuhansia tyytyväisiä asiakkkaita ja näin toimimme jatkossakin.


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.svetlana@gmail.com (RUS)
Aktiviteetit
Mökit & Majoitus
Santa's Reindeers & Laukkalan Mökit
Tule kokemaan puhdas ja monimuotoinen luonto. Tarjoamme majoitusta kauniissa järvimaisemissa, sekä elämyksiä ympäri vuoden, koko perheelle.
Santa's Reindeers
Poror
Villa Alexia
Iso Mökki
Keski Mökki

Conclusions

- TTA improves baseline

  62% $\rightarrow$ 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/