



Extracting Representative Image from Web page

Najlaa Gali, Andrei Tabarcea and **Pasi Fränti**



UNIVERSITY OF
EASTERN FINLAND

Motivation: summarize search result

MOPSI See what's around

bar Search

Mopsi service Photo Web search

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

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

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

Kauppakatu 23 80100 Joensuu

882 m

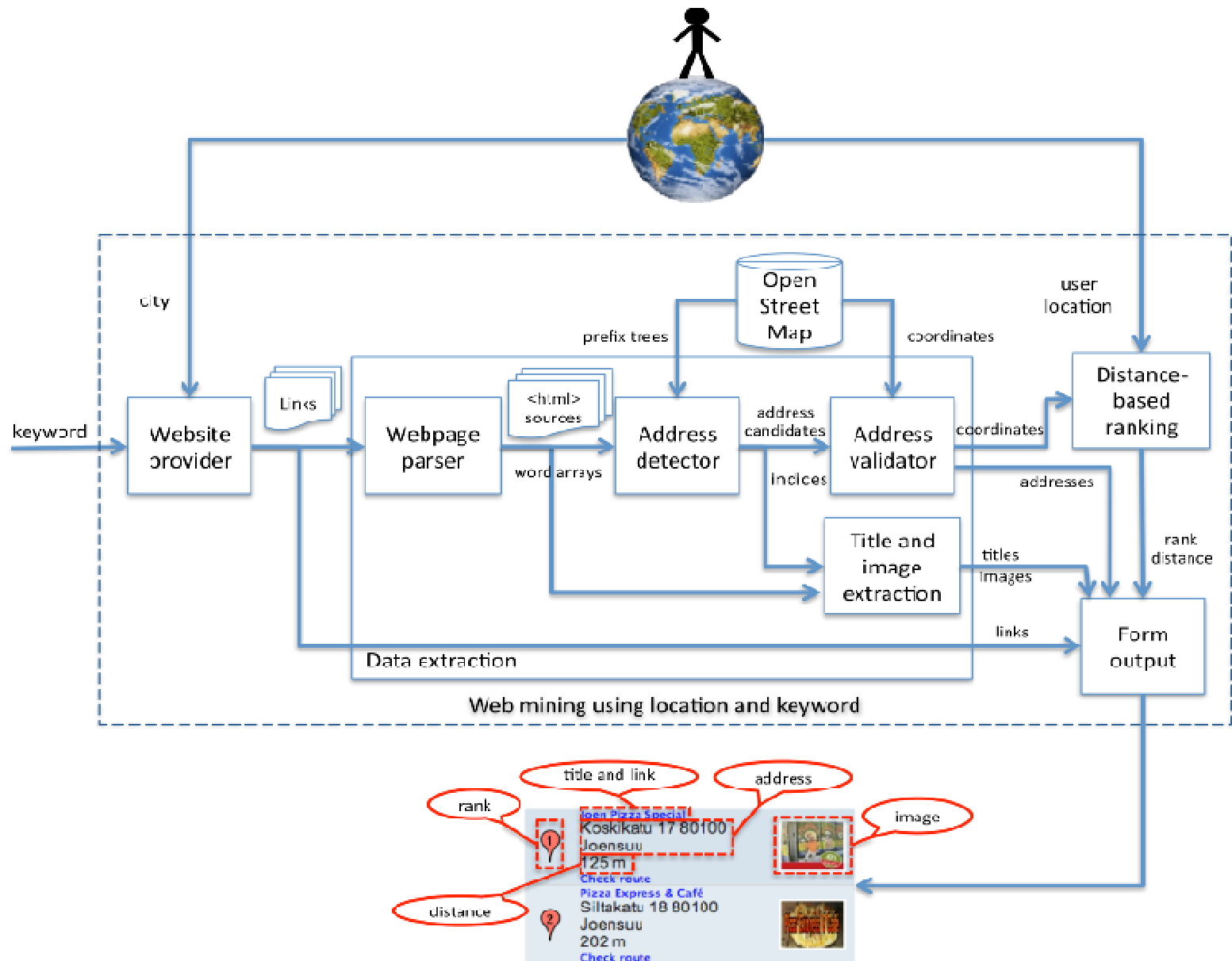
Street View

Calculating distance

Title Address Image

The screenshot displays the MOPSI application interface. At the top, there's a header with the MOPSI logo and the tagline 'See what's around'. Below this is a search bar containing the word 'bar' and a 'Search' button. To the right of the search bar are icons for Mopsi service, Photo, and Web search. The main content area is divided into two sections. The left section shows a list of search results for bars. The first result, 'Bar Play Joensuu S-kanava*', is highlighted with a red circle around its title and another red circle around its address 'Kauppakatu 23 80100'. A red arrow points from the 'Calculating distance' text at the bottom left to the distance '882 m' associated with this result. The right section shows a map of Joensuu with a red pin marking the location of 'Bar Play Joensuu S-kanava*'. A red circle highlights the 'SINUN ETUSI' image, which is a thumbnail of a music festival. A red arrow points from the 'Image' label at the bottom to this image. The map also shows other bars like 'Jet Set Sport Bar*' and 'Super Smoothie Joensuu Cafe & Salad Bar smoothie bar*'. A red arrow points from the 'Image' label to the 'SINUN ETUSI' image. The map shows the location of 'Bar Play Joensuu S-kanava*' on Kauppakatu 23, 80100 Joensuu, with a distance of 882 m from the user's location. The map also shows other streets like Siltakatu, Rantakatu, and Torikatu. The map includes a 'Locate' button in the top right corner. The map also shows a 'Street View' button at the bottom right. The map also shows a 'Calculating distance' text at the bottom left, with a red arrow pointing to the distance '882 m'.

Structure of location-based search

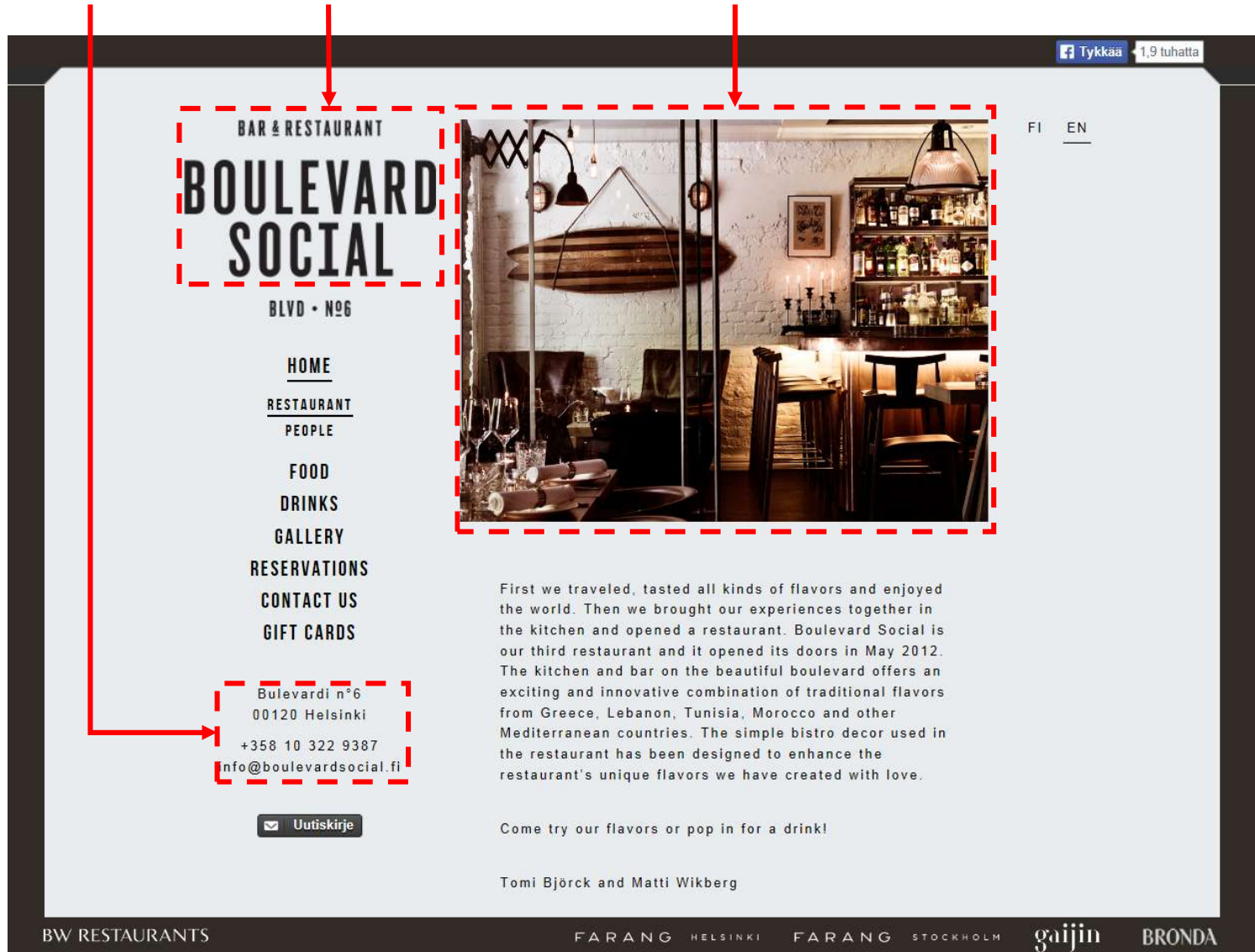


Content that we want to extract

Address

Title

Representative image

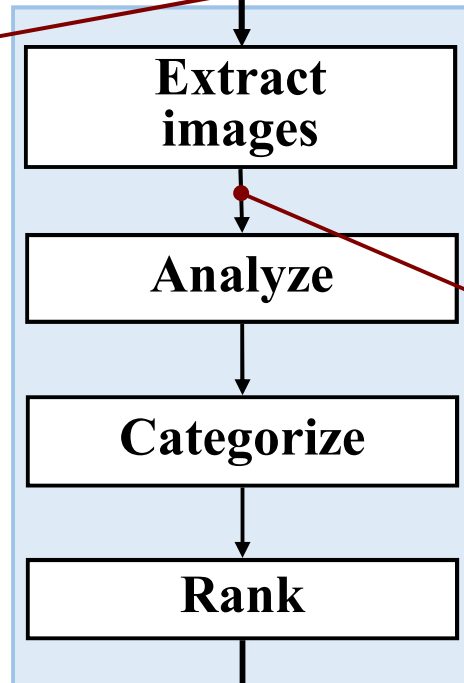


Overall extraction process

Web page



Web page link



Images found



Representative image



What to extract

Three sources: **html**, **CSS**, **JS**

```
<link rel="stylesheet" type="text/css" media="screen"
href="http://www.ompelimot.com/css/rakenne.css" title="Oletus" />
```

Representative image →

rakenne.css

```
#ylaosa {
  height: 150px;
  background:
    url("../images/2.png") no-repeat scroll
    0px 0px #EEE6C8;
  border-bottom: 2px solid #FFF;
  width: 694px;
  margin: 0px auto;
}
```

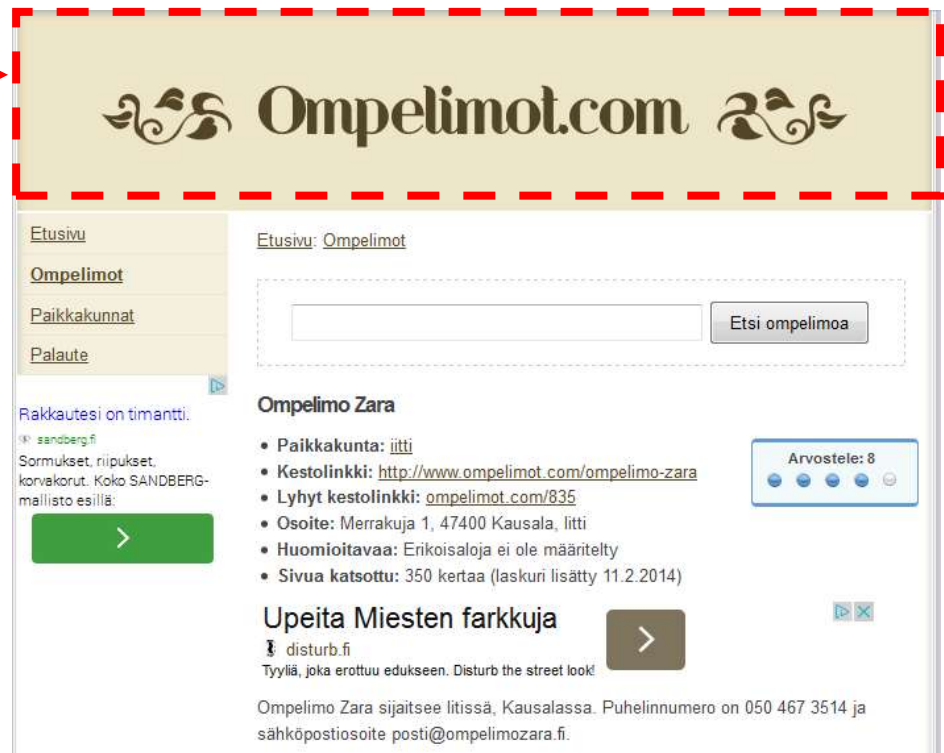


Image features used



src	http://www.ravintolakreeta.fi///images/banner.jpg
alt	--
title	--
from	css
format	jpg
width	945
height	202
size	190,890 px
aspect ratio	4.67
parent tag	<div>
class	header

Image categories

Banner

ENJOY WITH THE
ORIGINAL
GREEK WAY!



Formatting



Logo



UUSI KUOPION RAVINTOLA:
Maaherrankatu 23,
70100 Kuopio
puh. 017-2828337
varaukset(a)ravintolakreeta.fi

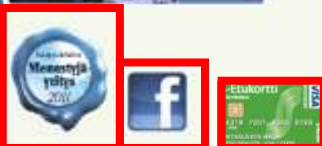
Iso Myy, Kauppakatu 28, 2.krs, 80110
Joensuu
puh. 044-260 4151
joensuu(a)ravintolakreeta.fi

AVOINNA

ma - to	11.00 - 22.00
pe	11.00 - 23.00
la	12.00 - 23.00
su	12.00 - 21.00



Icons



Representative



Advertisement

Tarjoimme Kreikkalaisen lounaan ma-pe kello 11-14.
Tule tutustumaan lounaslistaamme.

Kaikki ruokamme saat myös kätevästi kotipakettin.
Isompiin tilauksiin tarvittaessa myös kuljetus.

Mikäli sinulla on erityistoiveita ruokailun suhteen, oita Läkselle ravintolaamme.

Tule tutustumaan Kreikkalaisen keittöä aarteisiin Ravintola Kreetaan Kuopiossa,
Maaherrankatu 23,
Joensuussa, Iso Myyn 2. krs Kauppakatu 28.

Image features used



src	http://www.martina.fi/sites/martina.fi/files/styles/fiiliskuva/public/Valitse%20alikansio/Ravintolat/ravintola-martina-paakuv-a-pasta.jpg?itok=z8DMqAu2
alt	Ravintola Martina Joensuu
title	--
from	html
format	jpg
width	920
height	313
size	287.96 px
aspect ratio	2.94
parent tag	<div>
class	header_fiilis
class of parent	content clearfix

Image categories

Logo

Representative image

Banner

Advertisement

The image shows a screenshot of the Ravintola Torero website. Red arrows point from labels to specific elements: 'Logo' points to the Torero logo, 'Representative image' points to a large photo of two men in traditional Spanish attire, 'Banner' points to a navigation bar, 'Advertisement' points to a promotion for S-ryhmän makulahjat kortti, and 'Formatting' points to the overall layout. The website features a top navigation bar with links like RUOKALISTAT, LOUNAS, PÖYTÄVARAUS, AJANKOHTAISTA, LIITY POSTITUSLISTALLE, and PALAUTE. A large central image shows two men in traditional Spanish clothing, one holding a skewer of food. Text on the page includes 'VAHVOJA MAKUJA YHTEISIÄ ELÄMYKSIÄ' and 'Torero on matka espanjalaisuuteen, aitojen makujen ja rotevien viinien maailmaan. Valmiina voittamaan sinunkin sydämesi. Ole!'. A sidebar on the right promotes the S-ryhmän makulahjat kortti. The footer lists various Torero locations and contact information.

RAVINTOLA
TORERO

TORERO JOENSUU

Siltakatu 8, Joensuu
Pöytävaraukset numerosta
010-7623520 www.kattaakarjalan.fi

English | Svenska

S-Etukortti

RUOKALISTAT LOUNAS PÖYTÄVARAUS AJANKOHTAISTA LIITY POSTITUSLISTALLE PALAUTE

YHTEYSTIEDOT JA AUKIOLOAJAT

VAHVOJA MAKUJA
YHTEISIÄ ELÄMYKSIÄ

Torero on matka espanjalaisuuteen,
aitojen makujen ja rotevien viinien maailmaan.
Valmiina voittamaan sinunkin sydämesi. Ole!

ASIAKASOMISTAJAVINIT

Nyt on makulahjat kortilla.

Anna lahja hyvällä maulla.
S-ryhmän ravintoloiden
lahjakortit saatavissa
meiltä.

Amarillo | Bar & Bistro | Buffa | Coffee House | Fiesta! | Fransmanni | Grill It! | Makutarinoita-blogi | Presso | Public Corner | Rosso | Rosso Express | Sevilla
Silvukartta | Käyttöehdot © SOK | Oivahymy

Julkaisujärjestelmä: Mediasignal Communications

Category 1: Representative images

Images that are directly related to the content



Category 2: Logos

Images of logo of the company or institution

Criteria: Image **link**, **class** or **id** attribute of the `<img>` or its parent element contains text *logo*



```

```

Martina



RAVINTOLA
TORERO



TEATTERI
RAVINTOLA



```
<div class="logo">  
    
  </img>  
</div>
```

Category 3: Banners

Wide or tall images usually used as logo of the service

Criteria:

- **link, class or id** contains: *banner, header, footer, button*
- High aspect ratio (> 1.8)
- Not classified as advertisement , formatting or logo



Haluamme hurrata sinut!

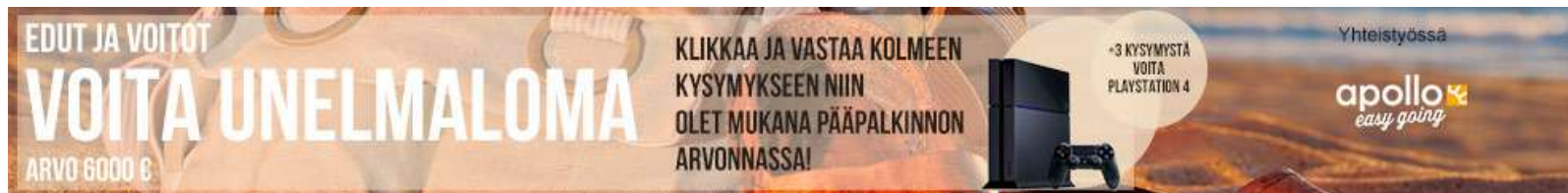


Category 4: Advertisement

Images that advertize products from other websites

Criterion:

- **Link, class or id** contains text: *free, now, buy, join, adserver, click, affiliate, adv, hits, counter*
- [Considered adding well known adv. server but not used]

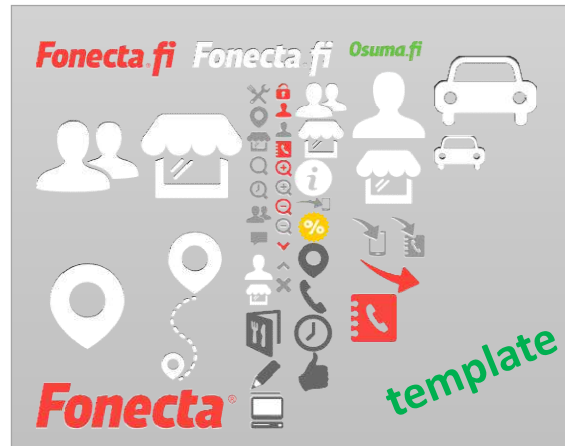


Category 5: Formatting and icons

Images used as backgrounds, decorators or icons

Criteria:

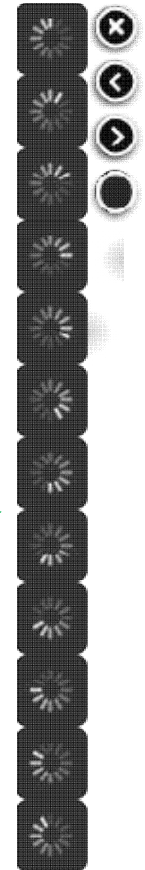
- **Link, class or id** contains text: *background, bg, sprite, template*
- **Height or width** is smaller than 100 px



VIINILISTA Список вин на русском языке



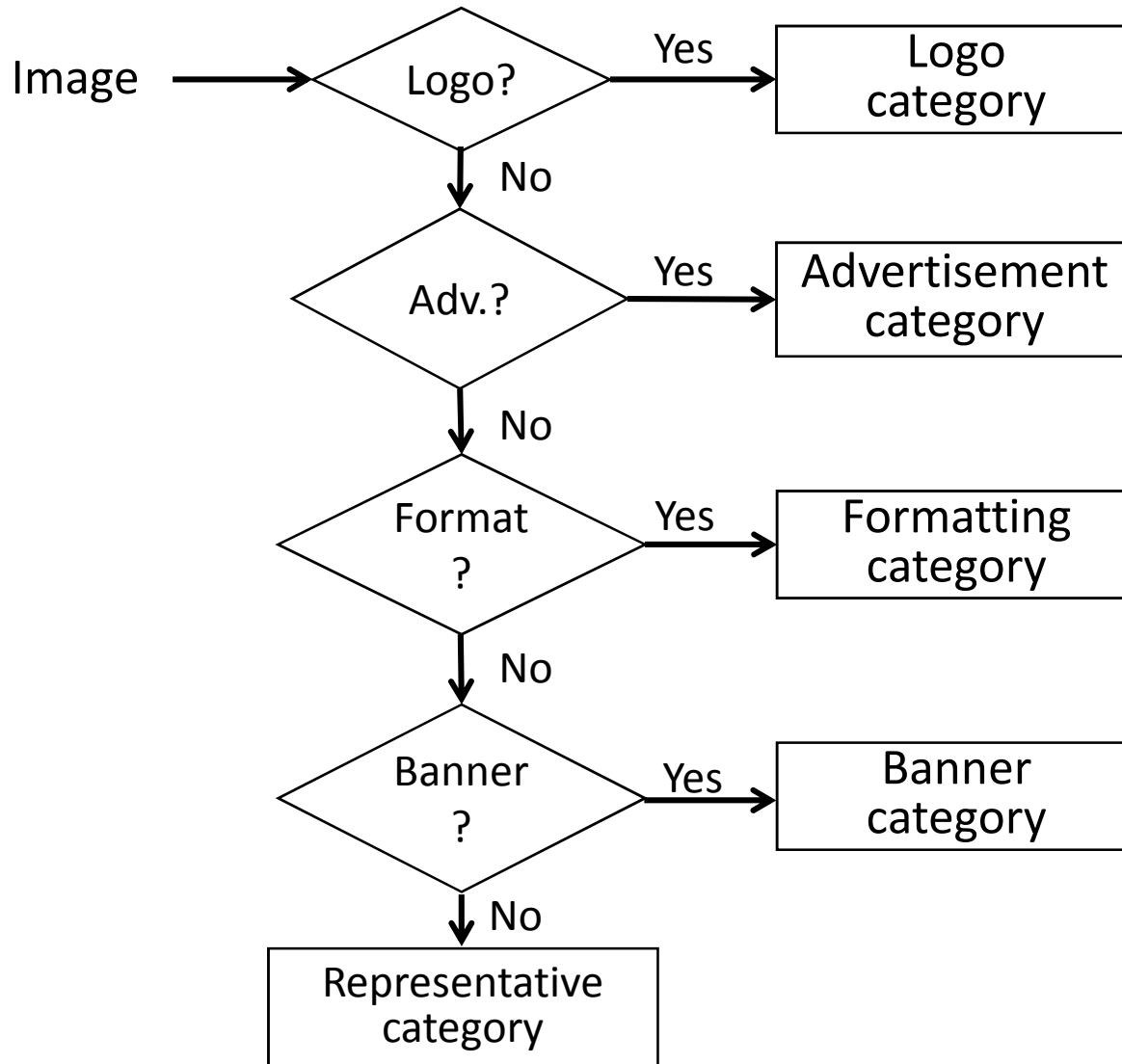
Size



Summary of rules

Category	Features	Keywords
Representative	Not in other category	
Logo		logo
Banner	Ratio > 1.8	Banner, header, Footer, button
Advertisement		Free, adserver, now, buy, join, click, affiliate, adv, hits, counter
Formatting and Icons	Width < 100 px Height < 100 px	Background, bg, spirit, templates

Decision tree for categorization



Scoring images

<http://ptiszai.com/imageext/>

Rule	Score
Image size ≥ 10.000 px	1
Aspect ratio ≤ 1.8	1
Image alt or title set a value	1
Keywords of alt or title appear also in <title> tag	1
Keywords of alt or title appear also in <h1> tag	1
Keywords of image path also in <title> or <h1> tags	1
The image is in the sub-tree of <h1> or <h2> tags	1
Format = jpg	1
Format = svg , png or gif	0.5

Mopsi WebIma dataset



<http://cs.uef.fi/mopsi/img/>

Summary of data collected:

- Websites: 1002
- Images: 2363
- Per page: Min=1, Average=2.36, Max=154

Collection details:

- Who: 117 volunteers
- When: September 2014
- What: Pages of own choice or Mopsi search
- How: Select 1-3 most representative images
- Issues: Some level of subjectivity unavoidable

Overall results

	Accuracy	Extracted Images
WebIma	64%	99%
Google+	48%	92%
Facebook	39%	90%

GOOD cases

Subset for which Weblma gives 100% accuracy

	Set 1			
	Ground truth (%)	Weblma (%)	Google+ (%)	Facebook (%)
Representative	63	63	13	20
Logo	37	37	13	10
Banner	0	0	57	60
Advertisement	0	0	0	0
Formatting	0	0	17	10

BAD cases

Subset for which Weblma gives 0% accuracy

	Set 2			
	Ground truth (%)	Weblma (%)	Google+ (%)	Facebook (%)
Representative	33	83	27	40
Logo	30	7	30	7
Banner	37	3	27	33
Advertisement	0	0	3	3
Formatting	0	7	13	17

Good enough?

Subjective Ground truth



WebIma



Facebook



Google+



Actions

On-line Registration
Submit Paper
Authors Area
Reviewers Area

Information

Conference Details
Important Dates
Technical Program
Call for Papers
Program Committee
Event Chairs

Venue

Conference Venue
Reaching the Venue

Invited Speakers

Keynote Lectures
Previous Invited Speakers

Satellite Events

Workshops
Special Sessions
Tutorials
Concours
Workshop Consortium
European Project Space
Industrial Track

Authors Kit

Guidelines and Templates
Glossary

Registration

Registration Fees
Important Information
Social Event

Travel and Accommodation

Reaching the City
Visa Information
Hotel Reservation
Local Information
About the Region
What to See and Do
Useful Information

Awards

Best Paper Award
Previous Awards

Post-Publications

Books published

Partners

Academic Partners
Industrial Partners
Institutional Partners
Media Partners
Partner Events
Publication Partners
Venues
Abstracts

Sponsored by:



INSTICC is Member of:



Logistics Partner:

WEBIST 2015 will be held in conjunction with SMARTGREENS 2015, CLOSER 2015, ICT4Ageing Well 2015 and VENITS 2015. Registration to WEBIST allows free access to the SMARTGREENS, CLOSER, ICT4Ageing Well and VENITS conferences (as a non-speaker).

UPCOMING DEADLINES

Regular and Position Paper Camera Ready and Registration: March 25, 2015

The purpose of the 11th 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, Security, e-Commerce, e-Business, Web Intelligence and Mobile Information Systems.

CONFERENCE AREAS

1. Internet Technology
2. Web Interfaces and Applications
3. Society, e-Business and e-Government
4. Web Intelligence
5. Mobile Information Systems

CONFERENCE CO-CHAIRS

Valérie Mounier, LAMM Valbonne, UMR, CNRS 5201, France
Karl-Heinz Kampel, RWTH Aachen University, Germany

PROGRAM CO-CHAIRS

Tim A. Majchrzak, University of Cologne, Germany
Zige Turk, University of Ljubljana, Slovenia / Refson Group, Slovenia

KEYNOTE LECTURES

Victor Chang, Leeds Beckett University, United Kingdom
Paolo Traverso, Center for Information Technology - IST (P5 K-ICT), Italy
Saverio Amadio, L3S - RWTH Aachen University, Germany
Rishabh Saxena, King's College London, United Kingdom
Alberto Broggi, Viterbo University & Parma, Italy
Gernot Klein, Siemens AG, Germany

SPECIAL SESSIONS

Socio-Technical Dynamics in Information Systems - STDS
Chair(s): Marius Jermak

DOCTORAL CONSORTIUM

WEBIST 2014 received 152 submissions, 42% of which were orally presented (15% as full papers).



All papers presented at the conference venue will be available at the SCITEPRESS Digital Library.



A short list of presented papers will be selected as that revised and extended versions of these papers will be published by Springer in a UICP Series book.



A short list of papers presented at the conference venue will be selected for publication of extended and revised versions in a special issue of Springer's Computing Journal.

Technically Sponsored by:



In Cooperation with:



Conclusions

- **Lightweight** method suitable for real time applications
- **Unsupervised**: No training, no user feedback needed
- Finds correct image **64%** of the cases.
Outperforms Google+ (48%) and Facebook (39%)
- In use in MOPSI: *Search and Service upgrade*



Thank you!