Without Users, no Retrieval! Overview and Future of iCLEF

Jussi Karlgren

September 16, 2008, Århus

Interactive CLEF

Julio Gonzalo
Javier Artiles
Víctor Peinado
Fernando López-Ostenero

UNED
Spain

Paul Clough
Emma Barker

U. Sheffield
United Kingdom

Jussi Karlgren

SICS
Sweden

Big question

How best to *assist users* when searching information written in unknown languages,
rather than how best an algorithm can *find information* written in languages different from the query language.

Avoid the dread MT Trap!

Big question

How best to *assist users* when searching information written in unknown languages,
rather than how best an algorithm can *find information* written in languages different from the query language.

Avoid the dread MT Trap!

Traditional evaluation

- Strict and well-established evaluation scheme.
- Test-collections, pre-assessed data.
- Clear (but debated) target notion: “relevance”.
- Great benefits to field.

So — why go interactive?

- Users introduce noise in evaluation. (Bad)
- User studies are first step towards extrinsic evaluation. (Good)
- User studies are fraught with methodological issues.

  (Interesting)

There is great interest in interactive evaluation among practicioners in the field!

So — why go interactive?

- Users introduce noise in evaluation. (Bad)
- User studies are first step towards extrinsic evaluation. (Good)
- User studies are fraught with methodological issues. (Interesting)

There is great interest in interactive evaluation among practicioners in the field!

So — why go interactive?

- Users introduce noise in evaluation. (Bad)
- User studies are first step towards extrinsic evaluation. (Good)
- User studies are fraught with methodological issues. (Interesting)

There is great interest in interactive evaluation among practicioners in the field!

So — why go interactive?

- Users introduce noise in evaluation. (Bad)
- User studies are first step towards extrinsic evaluation. (Good)
- User studies are fraught with methodological issues.

  (Interesting)

There is great interest in interactive evaluation among practicioners in the field!

So — why go interactive?

- Users introduce noise in evaluation. (Bad)
- User studies are first step towards extrinsic evaluation. (Good)
- User studies are fraught with methodological issues.

  (Interesting)

There is great interest in interactive evaluation among practicioners in the field!

Incidental factors

- System characteristics (Engineering, Design)
- Test subject ennui (Bad)
- Task realism (Good)

Incidental factors

- System characteristics (Engineering, Design)
- Test subject ennui (Bad)
- Task realism (Good)

Incidental factors

- System characteristics (Engineering, Design)
- Test subject ennui (Bad)
- Task realism (Good)

Incidental factors

- System characteristics (Engineering, Design)
- Test subject ennui (Bad)
- Task realism (Good)

Multilingual image retrieval

- Interesting and well motivated real task (Good)
- Adds complexity (Bad)
- Acts as grey filter (Good)

→ Methodology to tease out the salient factors!

Multilingual image retrieval

- Interesting and well motivated real task (Good)
- Adds complexity (Bad)
- Acts as grey filter (Good)

→ Methodology to tease out the salient factors!

Multilingual image retrieval

- Interesting and well motivated real task (Good)
- Adds complexity (Bad)
- Acts as grey filter (Good)

→ Methodology to tease out the salient factors!

Multilingual image retrieval

- Interesting and well motivated real task (Good)
- Adds complexity (Bad)
- Acts as grey filter (Good)

→ Methodology to tease out the salient factors!

Multilingual image retrieval

- Interesting and well motivated real task (Good)
- Adds complexity (Bad)
- Acts as grey filter (Good)

$\longrightarrow$ Methodology to tease out the salient factors!

High entry threshold

- Building a system for CLIR is hard work, especially if you need an interface. (Bad)
- Recruiting test subjects is hard work (Bad)
- Running experiments is hard work (Bad)
- Extraneous noise lowers generality of results (Bad)

High entry threshold

- Building a system for CLIR is hard work, especially if you need an interface. (Bad)
- Recruiting test subjects is hard work (Bad)
- Running experiments is hard work (Bad)
- Extraneous noise lowers generality of results (Bad)

High entry threshold

- Building a system for CLIR is hard work, especially if you need an interface. (Bad)
- Recruiting test subjects is hard work (Bad)
- Running experiments is hard work (Bad)
- Extraneous noise lowers generality of results (Bad)

High entry threshold

- Building a system for CLIR is hard work, especially if you need an interface. (Bad)
- Recruiting test subjects is hard work (Bad)
- Running experiments is hard work (Bad)
- Extraneous noise lowers generality of results (Bad)

Making changes

(Good!)

- Move from news collections to using images
- Move from canned information needs towards more naturalistic scenarios
- Lower threshold of entry for test subjects and experimenters alike
- Move from system design towards log analysis

Making changes

(Good!)

- Move from news collections to using images
- Move from canned information needs towards more naturalistic scenarios
- Lower threshold of entry for test subjects and experimenters alike
- Move from system design towards log analysis

Making changes

(Good!)

- Move from news collections to using images
- Move from canned information needs towards more naturalistic scenarios
- Lower threshold of entry for test subjects and experimenters alike
- Move from system design towards log analysis

Making changes

(Good!)

- Move from news collections to using images
- Move from canned information needs towards more naturalistic scenarios
- Lower threshold of entry for test subjects and experimenters alike
- Move from system design towards log analysis

Making changes

(Good!)

- Move from news collections to using images
- Move from canned information needs towards more naturalistic scenarios
- Lower threshold of entry for test subjects and experimenters alike
- Move from system design towards log analysis

Task requirements

- Simple, need no training
- Engaging and addictive
- Clear indication of success.
- Adaptive level of difficulty
- Naturally multilingual.

iCLEF 2008: search log analysis for Multilingual Image Retrieval

- Work on shared log
- Experiment on single interface
- Game-like task
- Find given image using the interface

iCLEF 2008: search log analysis for Multilingual Image Retrieval

- Work on shared log
- Experiment on single interface
- Game-like task
- Find given image using the interface

iCLEF 2008: search log analysis for Multilingual Image Retrieval

- Work on shared log
- Experiment on single interface
- Game-like task
- Find given image using the interface

iCLEF 2008: search log analysis for Multilingual Image Retrieval

- Work on shared log
- Experiment on single interface
- Game-like task
- Find given image using the interface

iCLEF 2008: search log analysis for Multilingual Image Retrieval

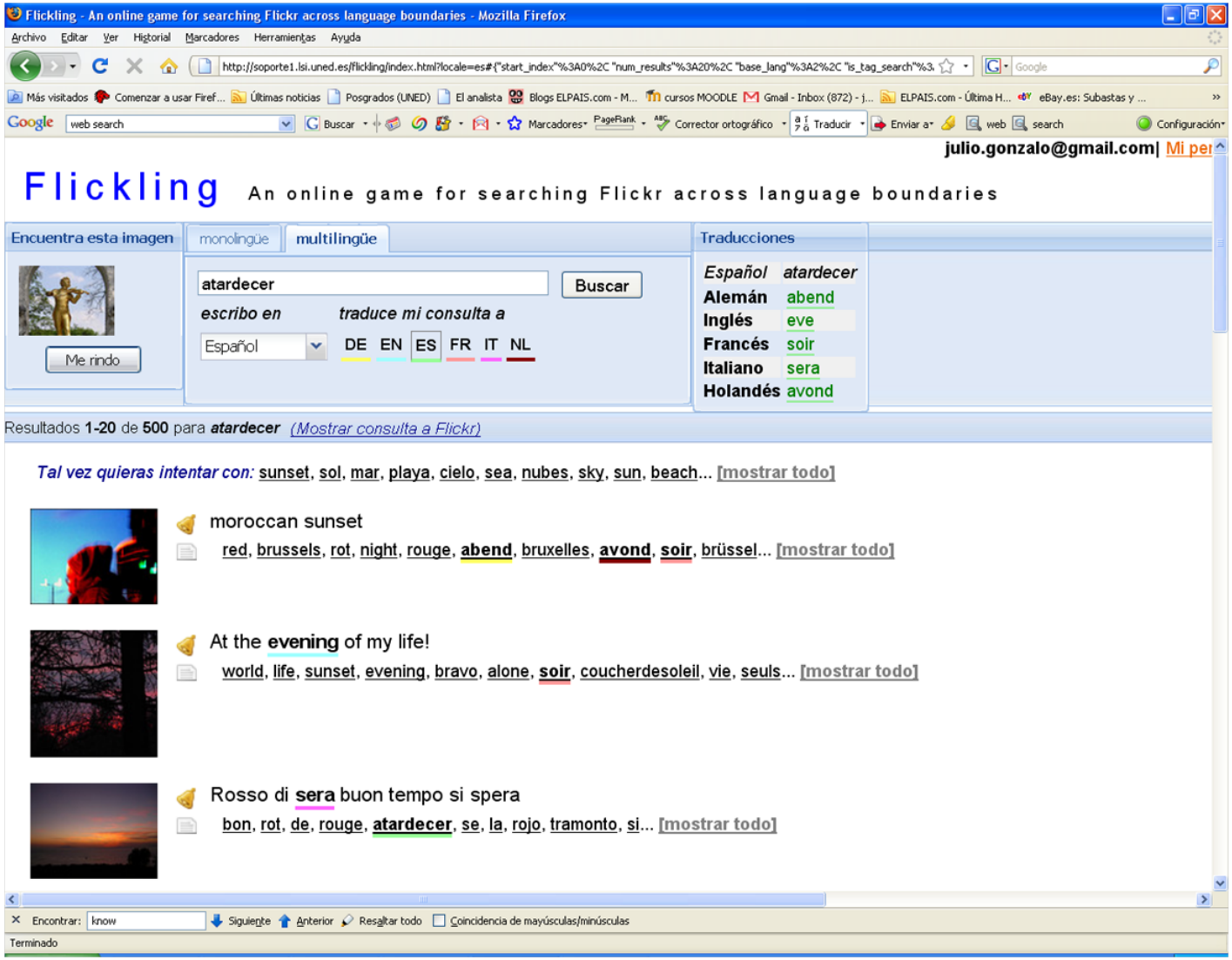
- Work on shared log
- Experiment on single interface
- Game-like task
- Find given image using the interface

Task Definition

Example images

Flickling interface

- Use dictionary, (add to it if necess) block translations
- Hints (manually fixed, first hint always target language making search mono or bilingual)
- Scores 25p/image -5 for each hint
- No time limit
- See Flickling paper


Flickling - An online game for searching Flickr across language boundaries - Mozilla Firefox
Archivo Editar Ver Historial Marcadores Herramientas Ayuda
http://soporte1.lsi.uned.es/flickling/index.html?locale=es#{"start_index"%3A0%2C "num_results"%3A20%2C "base_lang"%3A2%2C "is_tag_search"%3
julio.gonzalo@gmail.com| Mi per
Flickling An online game for searching Flickr across language boundaries
Encuentra esta imagen
Me rindo
monolingüe
multilingüe
atardecer
Buscar
escribo en
Español
traduce mi consulta a
DE EN ES FR IT NL
Traducciones
Español atardecer
Alemán abend
Inglés eve
Francés soir
Italiano sera
Holandés avond
Resultados 1-20 de 500 para atardecer (Mostrar consulta a Flickr)
Tal vez quieras intentar con: sunset, sol, mar, playa, cielo, sea, nubes, sky, sun, beach... [mostrar todo]
moroccan sunset
red, brussels, rot, night, rouge, abend, bruxelles, avond, soir, brüssel... [mostrar todo]
At the evening of my life!
world, life, sunset, evening, bravo, alone, soir, coucherdesoleil, vie, seuls... [mostrar todo]
Rosso di sera buon tempo si spera
bon, rot, de, rouge, atardecer, se, la, rojo, tramonto, si... [mostrar todo]
Encontrar: know
Siguiente Anterior Resaltar todo Coincidencia de mayúsculas/minúsculas
Terminado

Logs

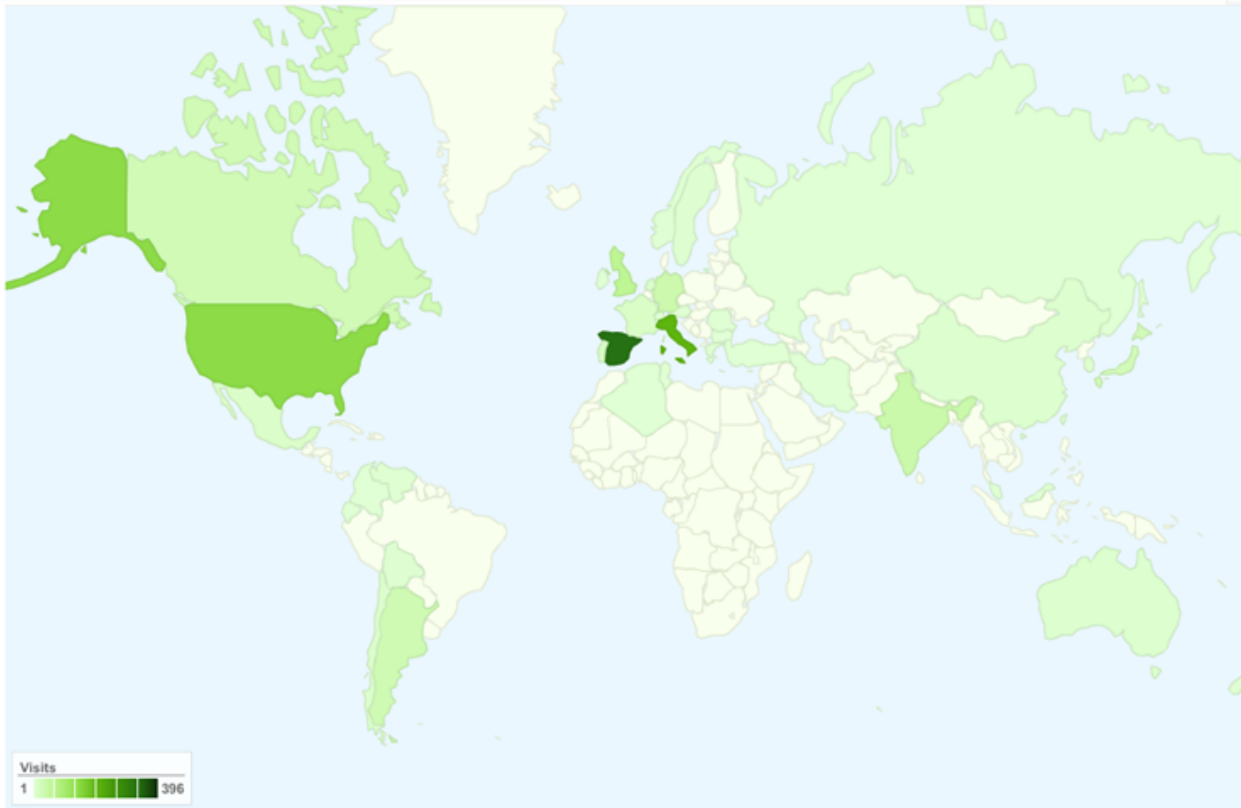
```
...
Fri Apr 25 17:50:08 2008 UNED LSI user 2 ...  search|saved QueryCache 'cangrejo arena
playa'
Fri Apr 25 17:50:08 2008 UNED LSI user 2 ...  search|68 results retrieved from Flickr
Fri Apr 25 17:50:10 2008 UNED LSI user 2 ...  playTime|time set on play mode
Fri Apr 25 17:50:28 2008 UNED LSI user 2 ...  log|click on found it 324171287:  wrong
Fri Apr 25 17:50:33 2008 UNED SCC user 1 ...  pauseTime|time was already paused
Fri Apr 25 17:50:33 2008 UNED SCC user 1 ...  search|launch query 'lobo' (0:20)
Fri Apr 25 17:50:33 2008 UNED SCC user 1 ...  search|using Flickr's API
Fri Apr 25 17:50:34 2008 UNED SCC user 1 ...  search|saved QueryCache 'lobo'
Fri Apr 25 17:50:34 2008 UNED SCC user 1 ...  search|500 results retrieved

...
```

CLEF Flickr challenge

Publicity in information access forums and photo blog sites.

Two prizes awarded by CLEF: best individual, best participating group.

Success!



1,212 visits came from 40 countries/territories

$>$ 300 participants, $\approx$ 230 active; researchers, students, photo buffs.

Sustainable resource

Truly reusable data set!

- $>$ 5000 search sessions with associated questionnaires, almost half truly cross-lingual
- $>$ 100 queries
- $>$ 200 active users, with heterogeneous language profiles

Sustainable resource

Truly reusable data set!

- $>$ 5000 search sessions with associated questionnaires, almost half truly cross-lingual
- $>$ 100 queries
- $>$ 200 active users, with heterogeneous language profiles

Experiments

Log analysis

Indian Institute of Information Technology:
success $\otimes$ reformulation

University of Westminster:
language confidence $\otimes$ personal dictionary use and modification

SICS:
user confidence $\approx$ dictionary modification, reformulation $\otimes$ success

UNED:
language competence, learning curves $\otimes$ success, end questionnaires

Observational experiments

Manchester Metropolitan University:
user perception of linguistic factors

Padua:
timing restraints for session

iCLEF 2009: what now?

- It works, don't break it!
- Tweaks likely
- New task possible
- Come to the session on Friday and make your thoughts known!
- Join in the fun, play with the logs!

Announcing The iCLEF Bender Awards


TrebleCLEF
CLEF
Interactive track for the Cross-Language Evaluation Forum
iCLEF 2008 - Flickling Challenge
An online game for searching
Flickr across language boundaries
Award for the best individual player
Nuno Cardoso


TrebleCLEF
CLEF
iCLEF 2008 - Flickling Challenge
An online game for searching
Flickr across language boundaries
Award for the top scoring group
University of Padua
Interactive track for the Cross-Language Evaluation Forum