

WebCLEF... the future

Maarten de Rijke

Three parts

- Between ad hoc and known-item
- Exploring the news
- Exploring cultural heritage

WebCLEF 2005/2006

- **Two tasks using known-item topics**
 - **Mixed Monolingual task**
 - stream of monolingual KI topics covering range of languages
 - language of topic statement is language of desired page
 - **Multilingual task**
 - stream of the English topics (translated from the monolingual)
 - the desired pages may cover any language/domain

WebCLEF 2005/2006

- Great success!
 - many known-item topics
 - CLIR effective for mixed-monolingual retrieval
 - multilingual retrieval is the hard nut to crack
- But ...
 - no informational topics yet, only navigational known-item topics
 - naturalness of the tasks?

WebCLEF 2007

- No more known-item retrieval against EuroGOV
- But then what?

Topics first, task later

Jaap Kamps



Topics First, Task Later

- Reverse standard practice of task definition and topic creation:
 - Topic creator gets access to a search box, and is left to explore the collection and may create any search topic that she thinks is natural
- Questionnaire data is collected
 - The sort of topic, the language preferences
 - Topic set will be close reflection of actual CLWR
 - Mix of topics + metadata
 - Breakdown in familiar categories (ad hoc, known-item, ...)

Limited Assessments

- Judgments as relevance assessments for informational topics, as duplicate detection for navigational topics
 - Very shallow pools: 10–50 results, or ranked pools
 - leave it to the assessor to decide when to stop judging
 - Page relevant only if judge understands language
 - Graded judgments: page may be non-relevant because it contains no useful information, or is not exact same page as visited before, or the language is not understood, ...

Proposal 1

- Reversal of standard IR practice:
 - Topics first, task later
- Retrieval task basically the same with
 - Natural CLWR topic stream
 - Broad range of topic types (including ad hoc)
- Questionnaires → contextual metadata
 - User context (e.g., languages) crucial for CLWR
 - Beyond generalizing one-size-fits-all approach
 - See K. Spärck Jones, *SIGIR Forum*, 40:10–20, 2006

In a Nutshell

- So far, WebCLEF restricted to known-item search
 - No informational ad hoc search topics
 - Current topic set may not reflect actual CLWR
- Proposal: invert traditional topic creation
 - Collect topics first, then classify according to type

Advantage

- Allows for mixed-monolingual and multilingual tasks
 - Evaluated over a (more) realistic stream of topics
 - Or over individual topic types (known-item, ad hoc, etc)

Tracking European News

Maarten de Rijke

NewsExplorer

- Publicly accessible news aggregation and analysis system
 - Navigate across languages and over time
 - 13 languages, 9 countries
 - 1,000 newsfeeds, 15,000 articles/day
- <http://press.jrc.it/NewsExplorer/>

NewsExplorer

EMM NewsBrief	EMM NewsExplorer	Name Search	Text Search
----------------------	-------------------------	-------------	-------------

News Explorer
Europe Media Monitor

News Analysis

RSS RSS feed for the latest news summary
Daily News Analysis, across languages and over time

Main Menu

[Latest News Summary](#)
[About EMM NewsExplorer](#)

News language and date

 Language or country:

 Date:

Mo	Tu	We	Th	Fr	Sa	Su
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1

Analysis over time

 Timeline
Timeline [en] for 09/20/2006

Clustered news for Wednesday, September 20, 2006

Countries

- United States (486)
- Iraq (130)
- United Kingdom (99)
- Israel (80)
- Thailand (47)
- Iran, Islamic Republic Of (46)
- Japan (42)
- Australia (38)

[View with Google Earth](#)

Thai Coup Leader Says He'll Be Acting PM [85]

de es fr it nl

Thai Coup Leader Says He'll Be Acting PM
Wednesday September 20, 2006 2:16 PM AP Photo XSL116 By DENIS D. GRAY Associated Press Writer BANGKOK, Thailand (AP) - The army commander who seized Thailand's government in a swift, bloodless coup promised Wednesday to act as prime minister for only two....
guardian 3:41:00 PM CEST

Budapest erupts for second night [27]

es fr it nl

BUDAPEST, Hungary (CNN) -- Police have clashed for a second night with protesters in Budapest amid

This Week's New Stories
[Police kill gunman after shooting](#)
September 13, 2006 - September 18, 2006
[Hurricane Lane Becomes Category 3](#)
September 14, 2006 - September 18, 2006
[150 People Injured In Hungary Protests](#)
September 18, 2006 - September 20, 2006
[Senate panel rejects Bush anti-terror plan](#)
September 13, 2006 - September 20, 2006
[E. coli outbreak traced to bagged spinach; 1 dead](#)
September 15, 2006 - September 20, 2006

This Month's New Stories
[US to probe Israeli cluster bombing](#)
August 25, 2006 - September 2, 2006
[Muslim Leaders Demand Apology From Pope](#)
September 8, 2006 - September 20, 2006
[Atlantis shuttle takes off, at last](#)
August 26, 2006 - September 20, 2006
[Hurricane on track to hit Florida](#)
August 26, 2006 - September 20, 2006
[Blair Won't Set Specific Date for Exit](#)
August 27, 2006 - September 18, 2006

Biggest Stories
[At least 40 die in Israeli attack on Qana](#)
December 5, 2005 - September 20, 2006

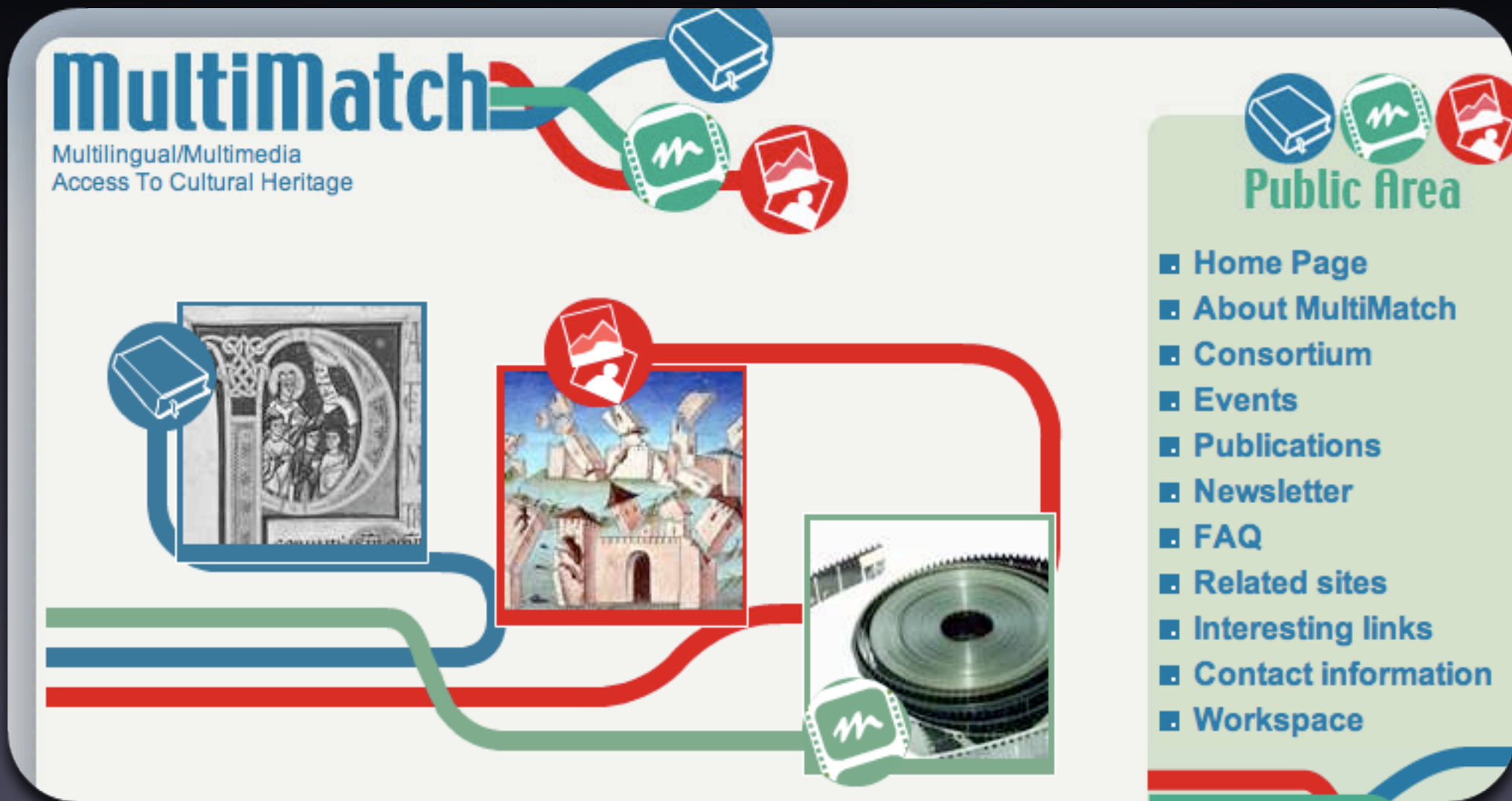
NewsExplorer

- News-related search tasks
 - E.g., track and detect topics or events across Europe
 - + Can use logs and other resources from NewsExplorer
 - CLEF would have to create a news corpus, but could use information on feeds from NewsExplorer

MultiMATCH

Maarten de Rijke

MultiMATCH



MultiMATCH

- FP6 project
- Multilingual search engine for access, organisation and personalised presentation of cultural heritage information
 - “Present the complete picture”
 - Explore and interact with online accessible cultural heritage content, across media types and language boundaries

MultiMATCH

- + “White-list crawl” of CH institutions to be available Dec 06/Jan 07
- + Log files and input on search behavior from CH institutions
- Search/access tasks
 - Faceted retrieval, CH event mining, link generation (e.g., Wikipedia), ...
 - Later: link up to user generated content, reviews, ...

Now what?

- Any thoughts at this point?
- Join discussions on the WebCLEF wiki
 - <http://ilps.science.uva.nl/WebCLEF/>