

Evaluating Cultural Heritage Information Retrieval

Avi Arampatzis[†] Jaap Kamps[†] Marijn Koolen[†] Vincent de Keijzer[‡] Nir Nussbaum[†]

[†] University of Amsterdam [‡] Gemeentemuseum Den Haag

The MuSeUM project

MuSeUM addresses the prototypical problem of a cultural heritage institution with the ambition to disclose all of its content in a single, unified system. The institution has various [legacy systems](#),

- each dealing with a small part of the collection,
- each constructed for different purposes,
- in different times,
- by different people,
- working in different traditions,
- based on different design principles,
- with different access methods, etc.

In short, the cultural heritage institution is confronted with its [own history](#).

Evaluation

Why is evaluation important?

- How do different techniques affect retrieval effectiveness (system oriented)?
- How do different techniques affect user satisfaction (user oriented)?

Why is evaluation so hard?

- We have to average over wildly diverging users, and over wildly diverging search requests.
- Is the evaluation valid and reliable?
- Making topics and doing assessments is extremely time consuming. Not every topic is suitable for evaluating certain aspects. Set up of the experiment needs careful planning. Participants need to be well-instructed.

Manual Evaluation

[System Oriented Evaluation](#) is based on a frozen collection of *documents*; a set of *search requests*; and *relevance judgments*, and can be reused.

Known-item Topics

We already had a set of 66 Known-Item topics for a heterogeneous collection of museum, library and archival descriptions. We have expanded our test suite with a set of 150 Known-Item topics and 40 Ad hoc topics on the website objects.

Ad Hoc Topic Creation

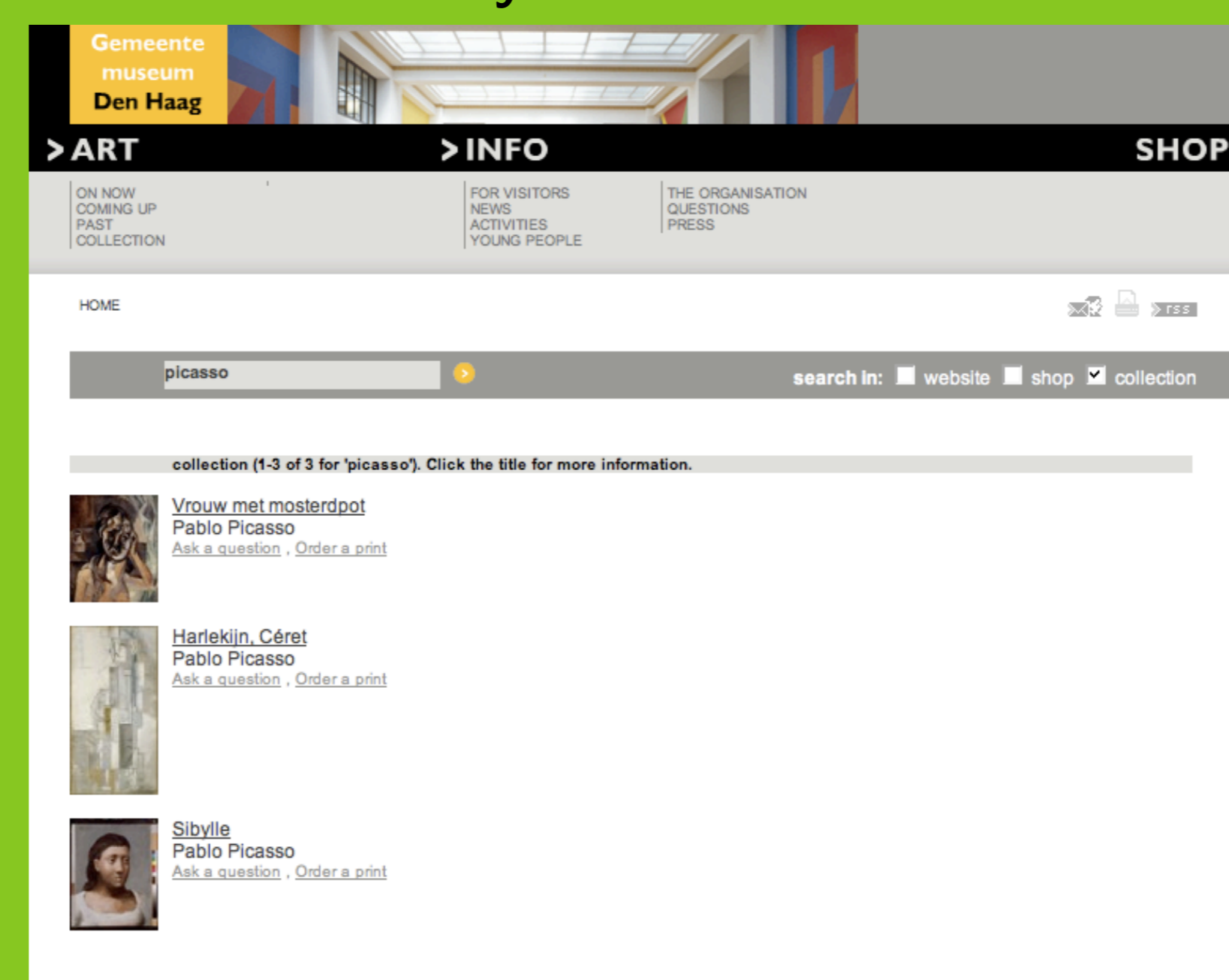
We are in the process of creating another Ad Hoc Topic set:

- Topics come from the Gemeentemuseum employees.
 - ★ This requires careful instruction, so that resulting topics will not only be realistic, but also suitable to test aspects like Precision and Recall.
- The document collection contains over 120,000 object descriptions of the Gemeentemuseum.

Automatic Evaluation

Since evaluation is very expensive, we investigate possibilities of doing *automatic* evaluation. We use the queries typed in by users on the Gemeentemuseum web site as topic to evaluate IR-systems.

The search facility on the museum's web site



- Do automatically extracted topic sets lead to the same system ranking as a set of manually constructed topics?
- We compare the extracted topics to a set of manually constructed Known-Item topics.

Extraction methods

We used 4 extraction methods to construct a test set:

1. **Bag queries:** each query appearing in the log is used, i.e. the bag of queries. Here, a topic consist of a query and the corresponding clicked results.
2. **Unique union:** All unique queries are used, i.e. the set of queries. All the results clicked by all users typing the same query are considered relevant documents.
3. **Unique intersection:** All unique queries are used, i.e. the set of queries. The intersection of the results clicked by all users typing the same query are considered relevant documents. Thus, a result is relevant only if all users who typed the query, clicked on that result.
4. **Majority Vote:** Only queries typed by two or more users are selected, and only results clicked by *more* than 50% of the users.

Statistics on the extracted topic sets.

Topic set	# Topics	Query length		# Rel. docs.	max
		mean	median		
Bag	7,531	1.22	1	2.38	206
Union	1,183	1.51	1	3.86	380
Intersec.	974	1.57	1	1.41	22
Majority	409	1.40	1	1.45	20
KI-topics	150	2.38	2	1	1
Ad hoc	40	1.80	2	13.65	52

System Ranking

We have tested 10 different systems on each of the topic sets, and compared the system rankings of each of the topic sets.

MRR scores of all systems on each topic set

	Topic set					
	Ad hoc	KI-topics	Majority	Bag	Inters.	Union
Cosine	0.6925	0.5614	0.5898	0.6218	0.4497	0.5226
Cosine BF	0.5198	0.4720	0.3843	0.4216	0.2589	0.4979
Indri	0.7131	0.5674	0.6221	0.6466	0.5606	0.5795
Indri JM	0.7140	0.5636	0.6173	0.6430	0.5786	0.5781
KL-div	0.7155	0.5687	0.6239	0.6484	0.5655	0.5815
KL-div. Jelinek	0.7113	0.5623	0.6108	0.6372	0.5464	0.5584
Okapi	0.7036	0.5600	0.6048	0.6325	0.5272	0.5521
Okapi K ₁ = 1	0.7018	0.5596	0.6030	0.6311	0.5213	0.5481
Tf.Idf	0.7122	0.5661	0.6153	0.6399	0.5654	0.5780
Tf.Idf BF	0.6584	0.4955	0.5674	0.5906	0.5196	0.6077
# Topics	409	7,531	974	1,183	150	40

We want to know if the automatically extracted topic sets have the same ability to rank IR systems as a set of manually constructed topics:

Rank correlation coefficients between topic sets

	Bag	Inters.	Union	Majority	KI-topics	Ad hoc
Bag	<u>1</u>	<u>0.87</u>	<u>0.87</u>	<u>0.82</u>	<u>0.69</u>	0.51
Intersection		<u>1</u>	<u>1</u>	<u>0.96</u>	<u>0.82</u>	<u>0.64</u>
Union			<u>1</u>	<u>0.96</u>	<u>0.82</u>	<u>0.64</u>
Majority				<u>1</u>	<u>0.87</u>	<u>0.6</u>
KI-topics					<u>1</u>	<u>0.56</u>
Ad hoc						<u>1</u>

- The system rankings of the automatically created topics sets are strongly correlated with the system ranking of known-item topics.
- Less but still significant correlation with ad hoc topics.
- The high number of topics lead to very stable topic sets.

Conclusions

Evaluation is crucial to assess the quality of, and further improve CH information access

- Manual evaluation of system performance requires an enormous investment, and is often not an option.
- Automatic evaluation can be an interesting alternative when no domain specific test collection is available.

We plan to do the following:

- Evaluate the effectiveness of using non-content features like link structure and relations between objects and documents, and images of objects.
- Evaluate techniques that combine/merge results from different collections/indexes.

Further Information

Visit the webpage of the [MuSeUM](http://www.nwo.nl/catch/museum/) project at <http://www.nwo.nl/catch/museum/>.

MuSeUM is funded by the Continuous Access To Cultural Heritage (CATCH) program of the Netherlands Organization for Scientific Research (NWO) under grant # 640.001.501.