Query Classification via Topic Models for an art image archive

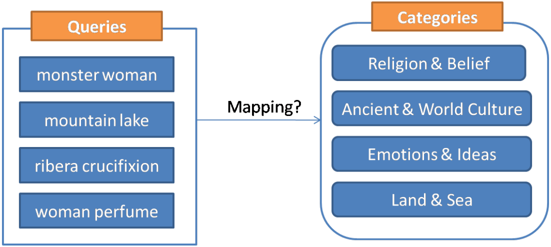
Dieu-Thu Le[1], Raffaella Bernardi[1], Edwin Vald[2]

[1]Trento University
[2]Bridgeman Art Library

September 16, 2011

Introduction

- Query Classification: map user queries to given target categories
- Query Classification in a specific domain: art, culture & history
  - Case study: Bridgeman Art Library
  - Challenges: specific vocabulary, short length of queries, lack of suitable available training data



Queries

monster woman

mountain lake

ribera crucifixion

woman perfume

Mapping?

Categories

Religion & Belief

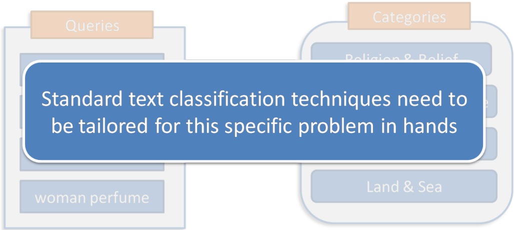
Ancient & World Culture

Emotions & Ideas

Land & Sea

Introduction

- Query Classification: map user queries to given target categories
- Query Classification in a specific domain: art, culture & history
  - Case study: Bridgeman Art Library
  - Challenges: specific vocabulary, short length of queries, lack of suitable available training data



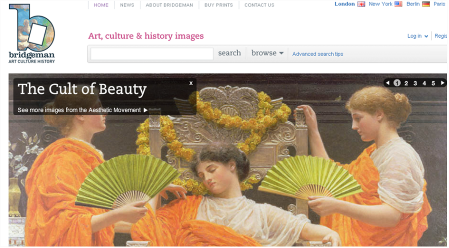
Standard text classification techniques need to be tailored for this specific problem in hands

Our proposal

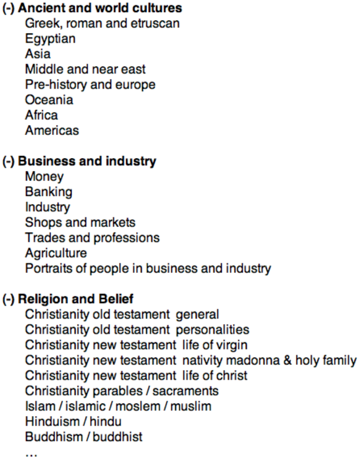
Enrichment of the query with:

- the information mined from the library via click-through links
- the information collected from external sources via Topic Models.

Bridgeman Art Library (BAL)



- one of the worlds top image libraries for art, culture and history
- contains images from over 8,000 collections; more than 29,000 artists
- Images in the library are classified to a target taxonomy



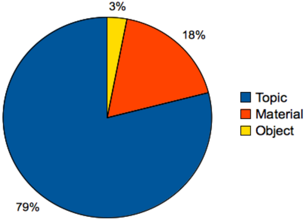
(-) Ancient and world cultures
- Greek, roman and etruscan
- Egyptian
- Asia
- Middle and near east
- Pre-history and europe
- Oceania
- Africa
- Americas

(-) Business and industry
- Money
- Banking
- Industry
- Shops and markets
- Trades and professions
- Agriculture
- Portraits of people in business and industry

(-) Religion and Belief
- Christianity old testament general
- Christianity old testament personalities
- Christianity new testament life of virgin
- Christianity new testament nativity madonna & holy family
- Christianity new testament life of christ
- Christianity parables / sacraments
- Islam / islamic / moslem / muslim
- Hinduism / hindu
- Buddhism / buddhist
- ...

Bridgeman Art Library (BAL)

Categories are divided into 3 main groups: topic, object, material



Title	Mountain Lake near Piedmont, Maryland (oil on canvas)
Keywords	American landscape, river, tranquil, atmospheric, rural, ...
Sub-Category	Peace & Relaxation
Top-Category	Emotions & Ideas

Does click-through information help in deciding categories for queries?

- Select 100 queries
- Three experts are asked to assign to each query up to 3 categories by:
  - looking only at the query (a)
  - looking at the query & the click-thru information & image (b)

Does click-through information help in deciding categories for queries?

- Select 100 queries
- Three experts are asked to assign to each query up to 3 categories by:
  - looking only at the query (a)
  - looking at the query & the click-thru information & image (b)

Does click-through information help in deciding categories for queries?

- Results:
  - agreement among 3 annotators in both cases is moderate
    → It is hard to decide a category for a given query, even for human annotators
  - (a): 20% of the queries are tagged as “Unknown”. (b): 4% of the queries are tagged as “Unknown”
    → Many queries are ambiguous and unable to be classified without looking at the click through information
  - agreement of the same annotator in 2 cases (a) and (b) is higher for categories in topic group (kappa ≈ 0.8) , lower in other 2 groups (kappa ≈ 0.57, 0.62)
    → Click-through information is important for deciding query's category, especially for those in “material” and “object” groups.

Does click-through information help in deciding categories for queries?

- Results:
  - agreement among 3 annotators in both cases is moderate
    → It is hard to decide a category for a given query, even for human annotators
  - (a): 20% of the queries are tagged as "Unknown". (b): 4% of the queries are tagged as "Unknown"
    → Many queries are ambiguous and unable to be classified without looking at the click through information
  - agreement of the same annotator in 2 cases (a) and (b) is higher for categories in topic group (kappa $\approx$ 0.8) , lower in other 2 groups (kappa $\approx$ 0.57, 0.62)
    → Click-through information is important for deciding query's category, especially for those in "material" and "object" groups.

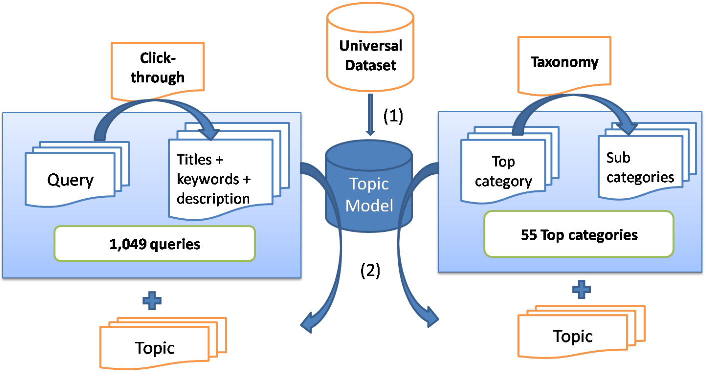
Does click-through information help in deciding categories for queries?

- Results:
  - agreement among 3 annotators in both cases is moderate
    → It is hard to decide a category for a given query, even for human annotators
  - (a): 20% of the queries are tagged as “Unknown”. (b): 4% of the queries are tagged as “Unknown”
    → Many queries are ambiguous and unable to be classified without looking at the click through information
  - agreement of the same annotator in 2 cases (a) and (b) is higher for categories in topic group (kappa $\approx$ 0.8) , lower in other 2 groups (kappa $\approx$ 0.57, 0.62)
    → Click-through information is important for deciding query's category, especially for those in “material” and “object” groups.

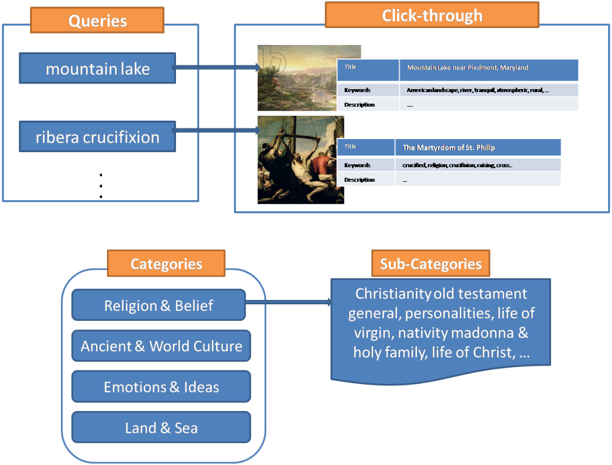
Does click-through information help in deciding categories for queries?

- Results:
  - agreement among 3 annotators in both cases is moderate
    → It is hard to decide a category for a given query, even for human annotators
  - (a): 20% of the queries are tagged as "Unknown". (b): 4% of the queries are tagged as "Unknown"
    → Many queries are ambiguous and unable to be classified without looking at the click through information
  - agreement of the same annotator in 2 cases (a) and (b) is higher for categories in topic group (kappa ≈ 0.8) , lower in other 2 groups (kappa $\approx 0.57, 0.62$)
    → Click-through information is important for deciding query's category, especially for those in "material" and "object" groups.

Data Enrichment: Our proposed framework



Query Enrichment with click-through



Title	Mountain Lake near Piedmont, Maryland
Keywords	American landscape, river, tranquil, atmospheric, rural, ...
Description	...

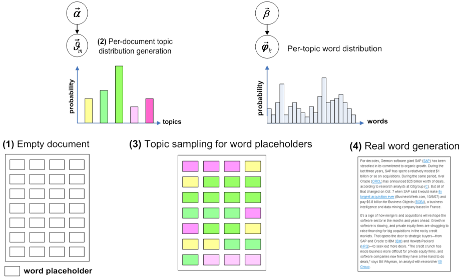
Title	The Martyrdom of St. Philip
Keywords	crucified, religion, crucifixion, raising, cross...
Description	...

Hidden Topic Models

- Documents exhibits multiple topics
- Given a set of documents, we estimate a topic model:
  - what are words for each topic?
  - what are topics for each documents?

Hidden Topic Models

- Documents exhibits multiple topics
- Given a set of documents, we estimate a topic model:
    - what are words for each topic?
    - what are topics for each documents?

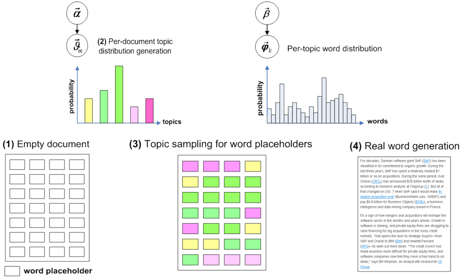

$\vec{\alpha}$
$\vec{\vartheta}_m$
(2) Per-document topic distribution generation
probability
topics
$\vec{\beta}$
$\vec{\varphi}_k$
Per-topic word distribution
probability
words
(1) Empty document
word placeholder
(3) Topic sampling for word placeholders
(4) Real word generation

For decades, German software giant SAP (SAP) has been steadfast in its commitment to organic growth. During the last three years, SAP has spent a relatively modest $1 billion or so on acquisitions. During the same period, rival Oracle (ORCL) has announced $25 billion worth of deals, according to research analysts at Citigroup (C). But all of that changed on Oct. 7 when SAP said it would make its largest acquisition ever (BusinessWeek.com, 10/8/07) and pay $6.8 billion for Business Objects (BOBJ), a business intelligence and data mining company based in France.

It's a sign of how mergers and acquisitions will reshape the software sector in the months and years ahead. Growth in software is slowing, and private equity firms are struggling to raise financing for big acquisitions in the rocky credit markets. That opens the door to strategic buyers—from SAP and Oracle to IBM (IBM) and Hewlett-Packard (HPQ)—to seek out more deals. "The credit crunch has made business more difficult for private equity firms, and software companies now feel they have a free hand to do deals," says Bill Whyman, an analyst with researcher ISI Group.

Hidden Topic Models

- Documents exhibits multiple topics
- Given a set of documents, we estimate a topic model:
  - what are words for each topic?
  - what are topics for each documents?

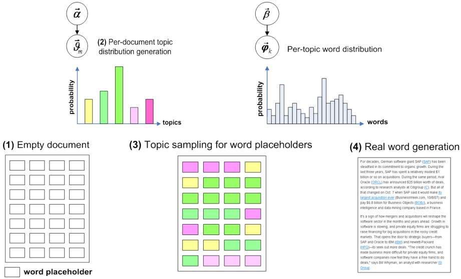


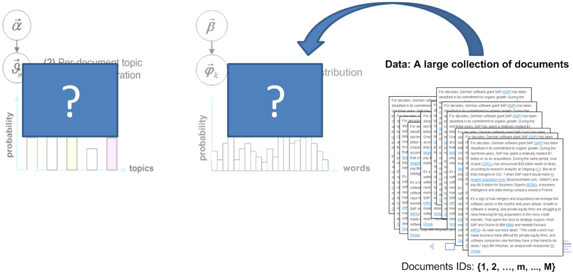
For decades, German software giant SAP (SAP) has been steadfast in its commitment to organic growth. During the last three years, SAP has spent a relatively modest $1 billion or so on acquisitions. During the same period, rival Oracle (ORCL) has announced $25 billion worth of deals, according to research analysts at Citigroup (C). But all of that changed on Oct. 7 when SAP said it would make its largest acquisition ever (BusinessWeek.com, 10/8/07) and pay $6.8 billion for Business Objects (BOBJ), a business intelligence and data mining company based in France.

It's a sign of how mergers and acquisitions will reshape the software sector in the months and years ahead. Growth in software is slowing, and private equity firms are struggling to raise financing for big acquisitions in the rocky credit markets. That opens the door to strategic buyers—from SAP and Oracle to IBM (IBM) and Hewlett-Packard (HPQ)—to seek out more deals. "The credit crunch has made business more difficult for private equity firms, and software companies now feel they have a free hand to do deals," says Bill Whyman, an analyst with researcher ISI Group.

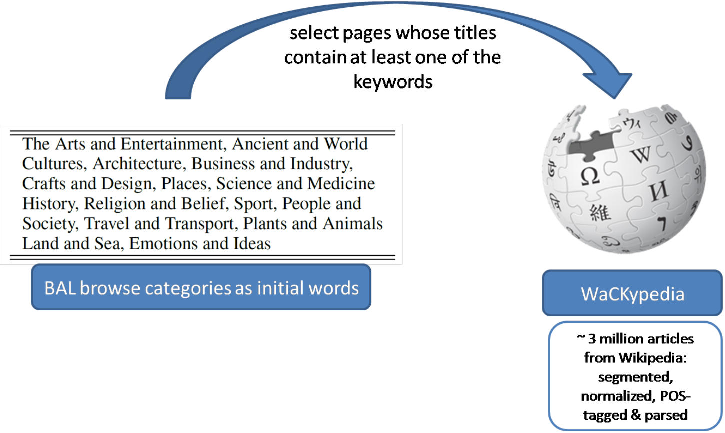
Hidden Topic Models

- Documents exhibits multiple topics
- Given a set of documents, we estimate a topic model:
  - what are words for each topic?
  - what are topics for each documents?





Selecting relevant pages from Wikipedia



select pages whose titles contain at least one of the keywords

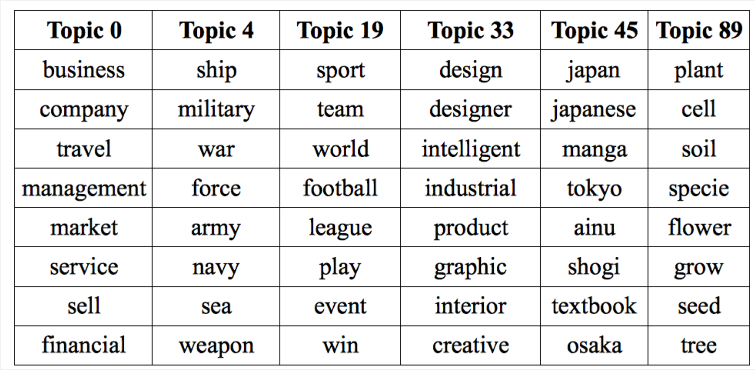
The Arts and Entertainment, Ancient and World Cultures, Architecture, Business and Industry, Crafts and Design, Places, Science and Medicine History, Religion and Belief, Sport, People and Society, Travel and Transport, Plants and Animals Land and Sea, Emotions and Ideas

BAL browse categories as initial words

WaCKypedia

~ 3 million articles from Wikipedia: segmented, normalized, POS-tagged & parsed

The Hidden Topic estimated from selected pages of Wikipedia



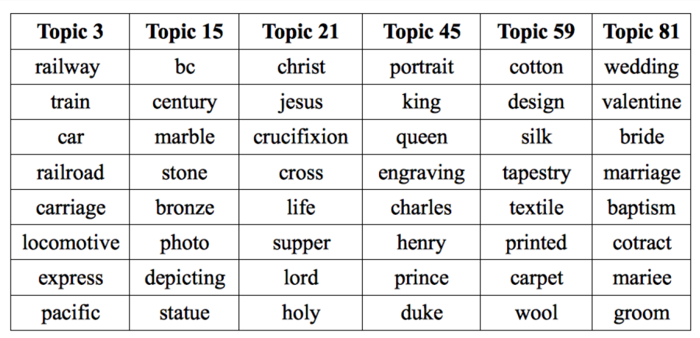
Topic 0	Topic 4	Topic 19	Topic 33	Topic 45	Topic 89
business	ship	sport	design	japan	plant
company	military	team	designer	japanese	cell
travel	war	world	intelligent	manga	soil
management	force	football	industrial	tokyo	specie
market	army	league	product	ainu	flower
service	navy	play	graphic	shogi	grow
sell	sea	event	interior	textbook	seed
financial	weapon	win	creative	osaka	tree

$\approx$ 14K documents, $\approx$200K words

100 Topics

The Hidden Topic estimated from BAL catalogue

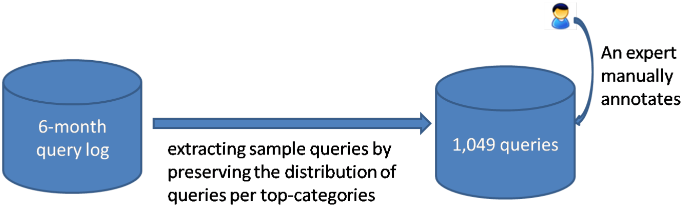
Images in the same category are grouped into one document.



Topic 3	Topic 15	Topic 21	Topic 45	Topic 59	Topic 81
railway	bc	christ	portrait	cotton	wedding
train	century	jesus	king	design	valentine
car	marble	crucifixion	queen	silk	bride
railroad	stone	cross	engraving	tapestry	marriage
carriage	bronze	life	charles	textile	baptism
locomotive	photo	supper	henry	printed	cotract
express	depicting	lord	prince	carpet	mariee
pacific	statue	holy	duke	wool	groom

732 documents, ≈ 136K words
100 Topics

Gold Standard

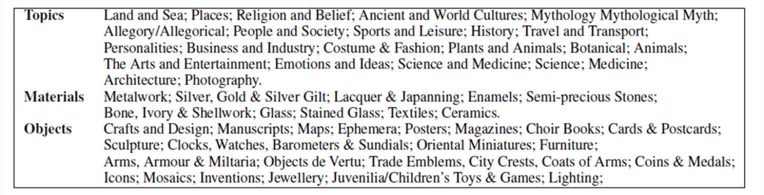


6-month query log

extracting sample queries by preserving the distribution of queries per top-categories

1,049 queries

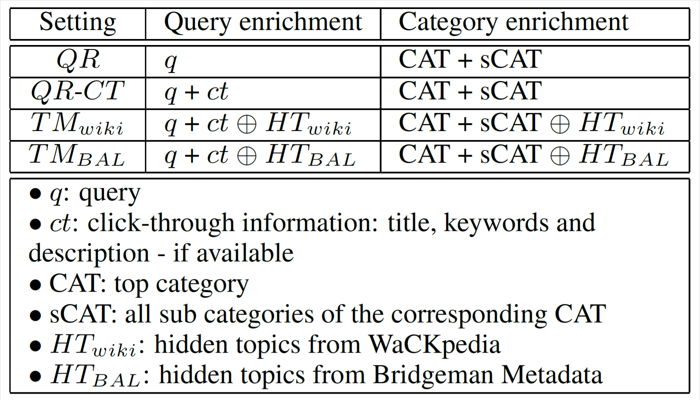
An expert manually annotates



Topics	Land and Sea; Places; Religion and Belief; Ancient and World Cultures; Mythology Mythological Myth; Allegory/Allegorical; People and Society; Sports and Leisure; History; Travel and Transport; Personalities; Business and Industry; Costume & Fashion; Plants and Animals; Botanical; Animals; The Arts and Entertainment; Emotions and Ideas; Science and Medicine; Science; Medicine; Architecture; Photography.
Materials	Metalwork; Silver, Gold & Silver Gilt; Lacquer & Japanning; Enamels; Semi-precious Stones; Bone, Ivory & Shellwork; Glass; Stained Glass; Textiles; Ceramics.
Objects	Crafts and Design; Manuscripts; Maps; Ephemera; Posters; Magazines; Choir Books; Cards & Postcards; Sculpture; Clocks, Watches, Barometers & Sundials; Oriental Miniatures; Furniture; Arms, Armour & Miltaria; Objects de Vertu; Trade Emblems, City Crests, Coats of Arms; Coins & Medals; Icons; Mosaics; Inventions; Jewellery; Juvenilia/Children's Toys & Games; Lighting;

Categories used by the annotator

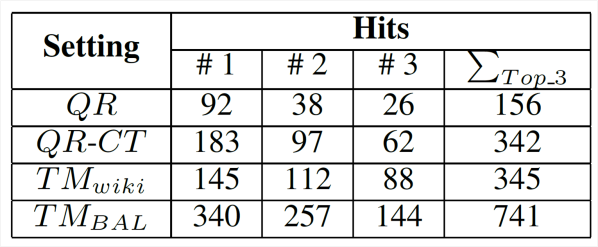
Experimental Setting



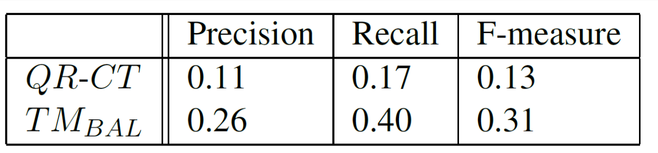
Setting	Query enrichment	Category enrichment
QR	q	CAT + sCAT
$QR\text{-}CT$	$q + ct$	CAT + sCAT
TM_{wiki}	$q + ct \oplus HT_{wiki}$	CAT + sCAT $\oplus$ HT_{wiki}
TM_{BAL}	$q + ct \oplus HT_{BAL}$	CAT + sCAT $\oplus$ HT_{BAL}

- q: query
- ct: click-through information: title, keywords and description - if available
- CAT: top category
- sCAT: all sub categories of the corresponding CAT
- HT_{wiki}: hidden topics from WaCKpedia
- HT_{BAL}: hidden topics from Bridgeman Metadata

Results

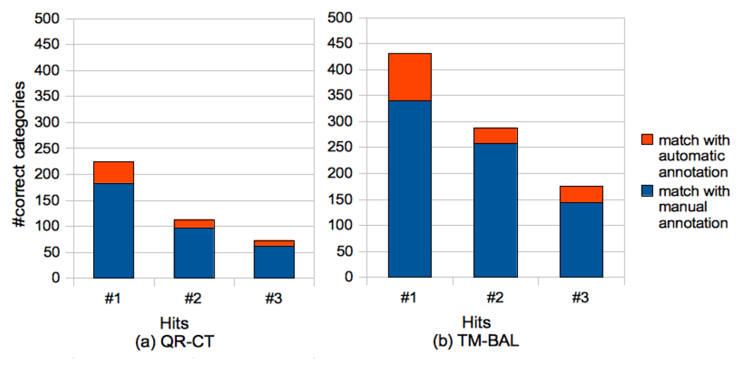


Setting	Hits			
	# 1	# 2	# 3	$\sum_{Top\_3}$
QR	92	38	26	156
$QR\text{-}CT$	183	97	62	342
TM_{wiki}	145	112	88	345
TM_{BAL}	340	257	144	741



	Precision	Recall	F-measure
$QR\text{-}CT$	0.11	0.17	0.13
TM_{BAL}	0.26	0.40	0.31

Results

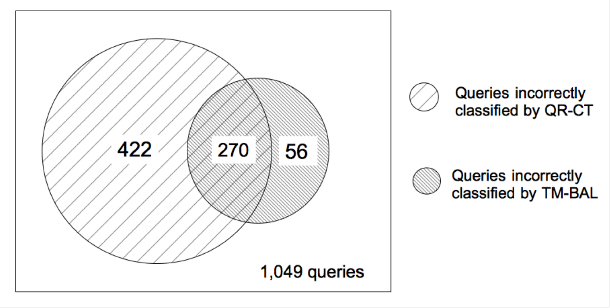


#correct categories
500
450
400
350
300
250
200
150
100
50
0
#1
#2
#3
Hits
(a) QR-CT
500
450
400
350
300
250
200
150
100
50
0
#1
#2
#3
Hits
(b) TM-BAL
🟥 match with automatic annotation
🟦 match with manual annotation

Analysis of wrong classification

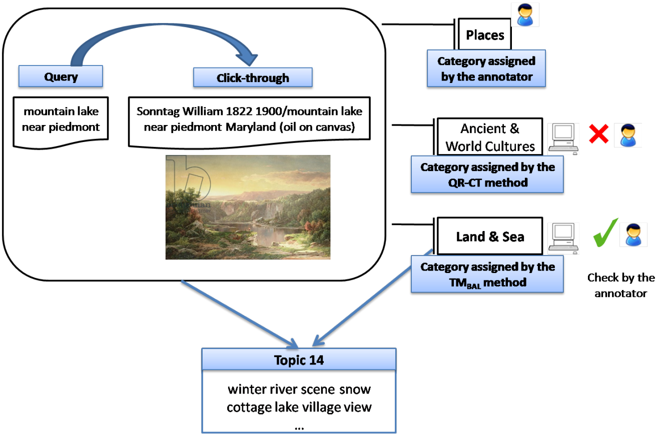
Selecting the most challenging queries:

- Queries that are not correctly classified by QR-CT and TM-BAL in any of the top 3-positions using gold standard (1) and (2)
  - GS (1): manual annotation
  - GS(2): automatic extraction from the meta-data of the clicked-image

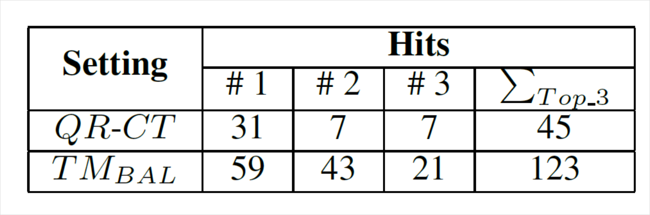


Analysis of wrong classification

Ask a domain expert to check again these "most challenging" queries:


Query
mountain lake near piedmont
Click-through
Sonntag William 1822 1900/mountain lake near piedmont Maryland (oil on canvas)
Places
Category assigned by the annotator
Ancient & World Cultures
✗
Category assigned by the QR-CT method
Land & Sea
✓
Category assigned by the TM_{BAL} method
Check by the annotator
Topic 14
winter river scene snow cottage lake village view
...

Analysis of wrong classification



Setting	Hits			
	# 1	# 2	# 3	$\sum_{Top\_3}$
$QR\text{-}CT$	31	7	7	45
TM_{BAL}	59	43	21	123

Conclusion

- Confirm the effect of the click-through information in query classification for art, history & culture closed domain
- Propose the use of metadata as a source to train topics models for query & category enrichment
- Future work: consider more click-through images per query, use this data enrichment as features for a machine learning based classifier

Bibliography

- Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta, *The WaCky wide web: a collection of very large linguistically processed web-crawled corpora*, Language Resources and Evaluation (2009).

- David M. Blei, Andrew Y. Ng, and Michael I. Jordan, *Latent dirichlet allocation*, J. Mach. Learn. Res. **3** (2003), 993–1022.

- T. L. Griffiths and M. Steyvers, *Finding scientific topics*, Proceedings of the National Academy of Sciences **101** (2004), no. Suppl. 1, 5228–5235.

- Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, and Quang-Thuy Ha, *A hidden topic-based framework towards building applications with short web documents*, IEEE Transactions on Knowledge and Data Engineering **99** (2010), no. PrePrints.