

Transkribus User Conference 2018 - Vienna (Austria)

Indexing and Searching of Manuscript Collections

Alejandro H. Toselli and Enrique Vidal

Pattern Recognition and Human Language Technology Research Center
Universitat Politècnica de València, Spain
`{ahector,evidal}@prhlt.upv.es`



February 9th, 2018

Outline

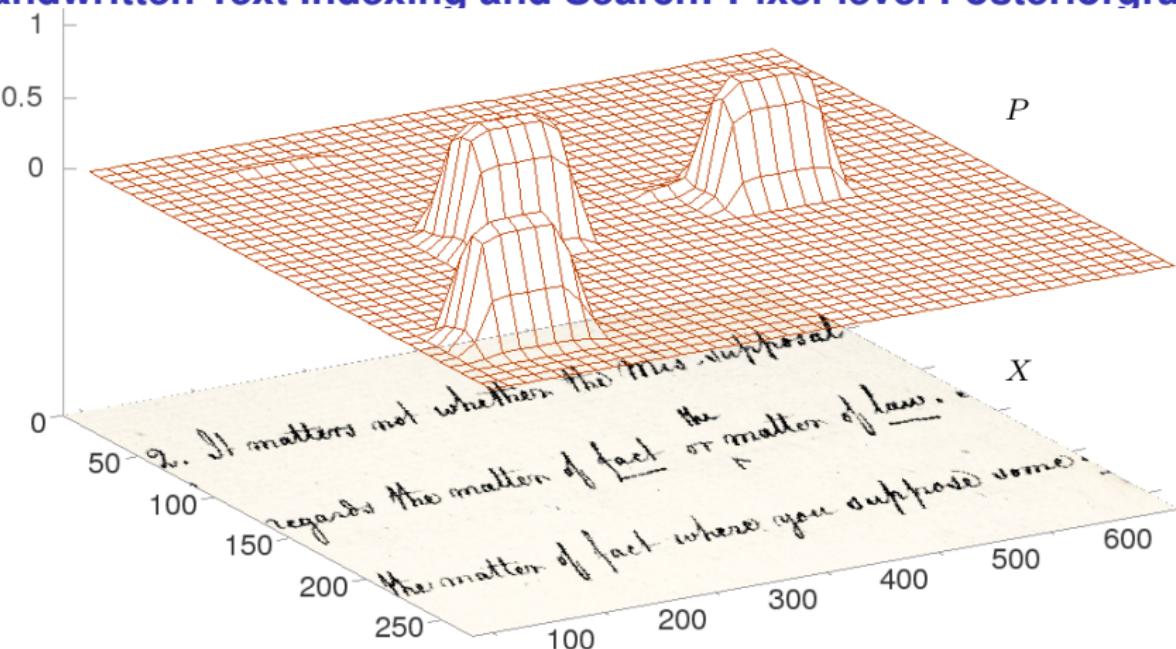
- Textual Access to Untranscribed Manuscripts ▷ 3
- Probabilistic Indexing of Text Images ▷ 4
- Performance Measures ▷ 8
- Preparatory Steps and Laboratory Results ▷ 9
- Keyword Search Demonstrators ▷ 11
- Other Indexing Tasks ▷ 15

Textual access to Untranscribed Manuscripts

- ▶ Massive text image collections have been compiled by libraries and archives all over the world, but their textual content remains practically inaccessible
- ▶ If perfect or sufficiently accurate text image transcripts were available, image textual context could be straightforwardly indexed for plaintext textual access.
- ▶ But manual or even interactive-predictive, assisted transcription is entirely prohibitive to deal with massive image collections
- ▶ And fully automatic transcription results lack the level of accuracy needed for useful text indexing and search purposes

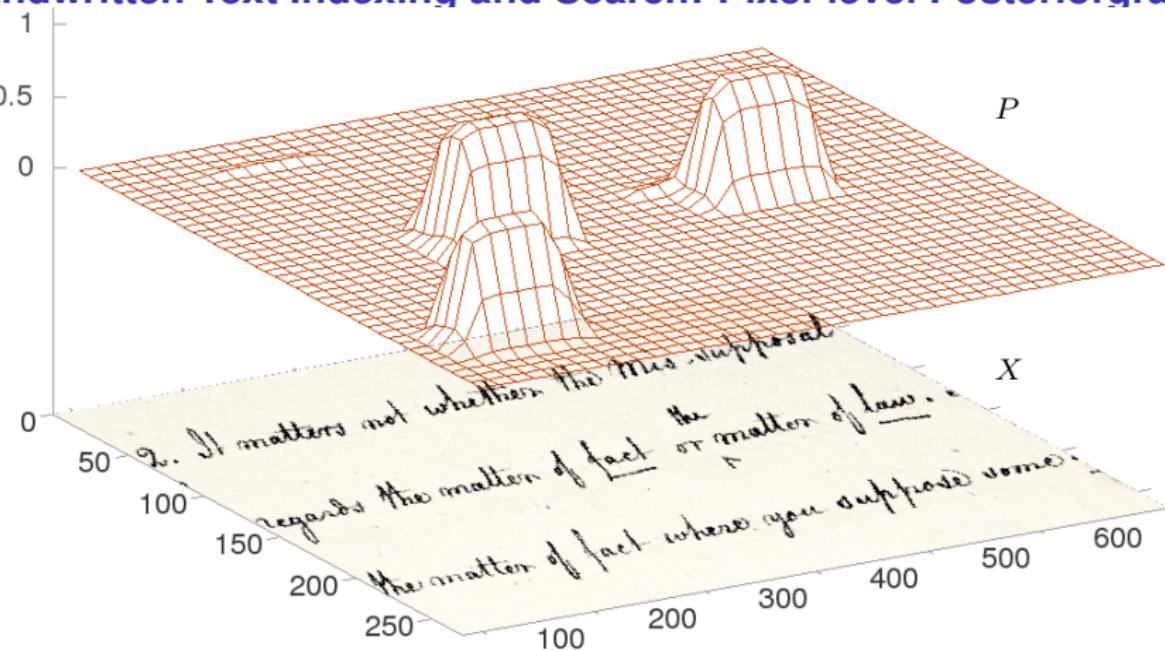
Good news: *indexing and textual search* can be directly carried out on *untranscribed images*, as we will see now.

Handwritten Text Indexing and Search: Pixel-level Posteriorogram



Pixel-level posterior probabilities (P) for a text image X and word $v = \text{"matter"}$.

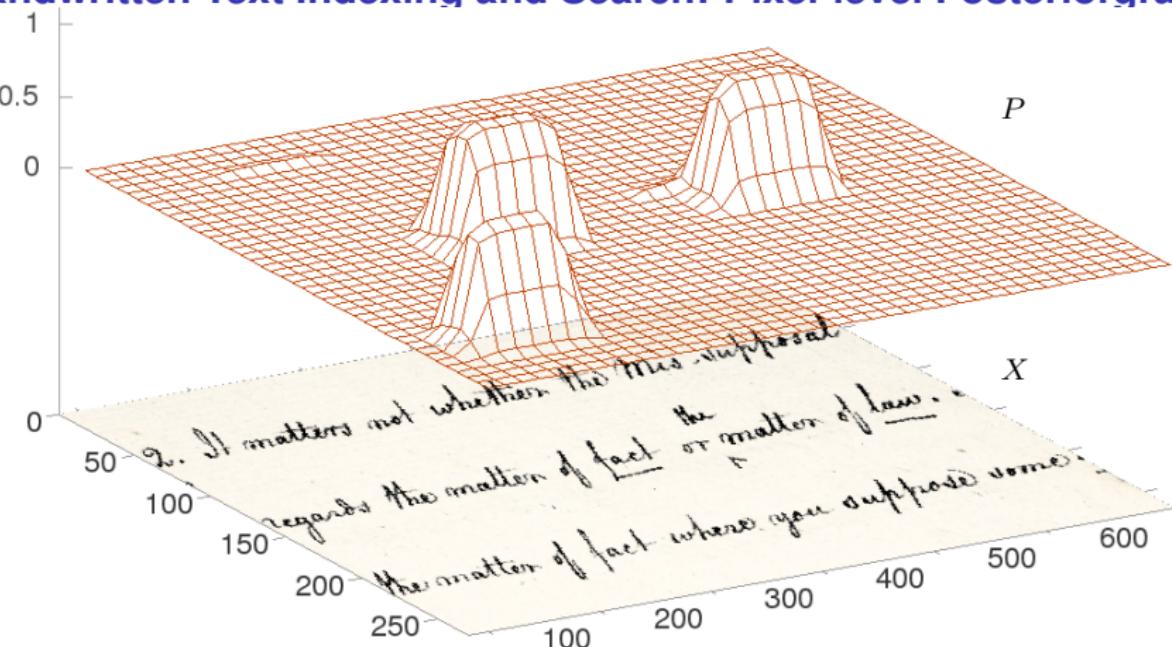
Handwritten Text Indexing and Search: Pixel-level Posteriorogram



Pixel-level posterior probabilities (P) for a text image X and word $v = \text{"matter"}$.

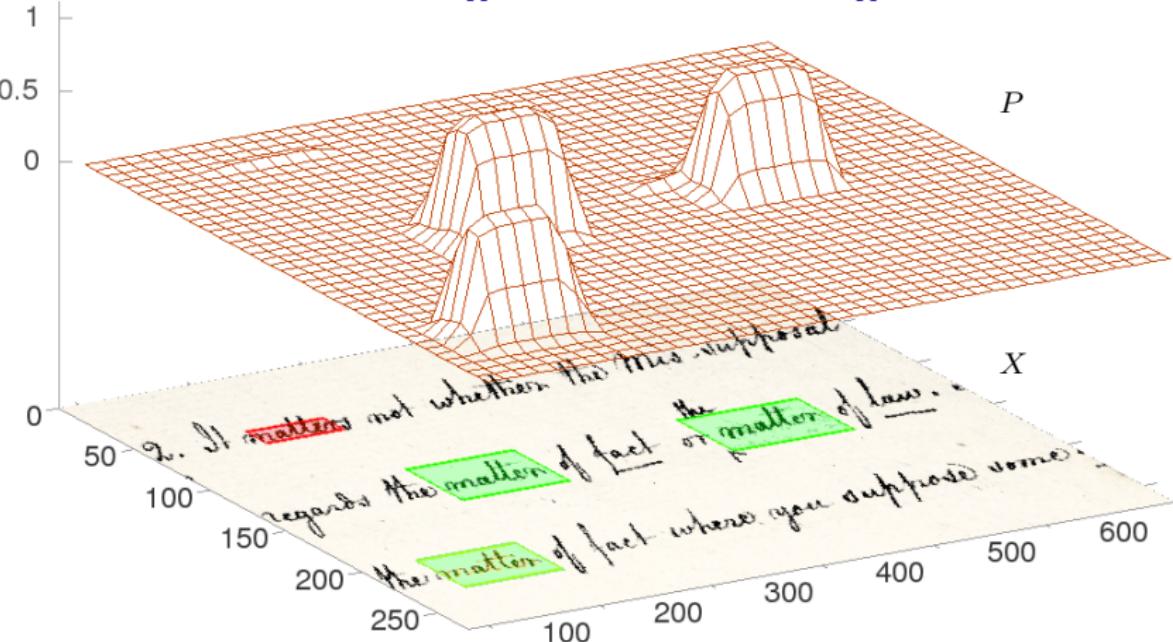
To compute P an accurate, contextual (n -gram based) [word classifier](#) can be used. In this example, this helped to achieve very low posteriors in a region of X around $(i=100, j=60)$, where a very similar (but [different](#)) word, "matters", is written.

Handwritten Text Indexing and Search: Pixel-level Posteriorogram



Directly computing and using a full pixel-level posteriorogram would entail a formidable computational load and would require prohibitive amounts of indexing storage.

Probabilistic Word Indexing from the Posteriorogram



Directly computing and using a full pixel-level posteriorogram would entail a formidable computational load and would require prohibitive amounts of indexing storage.

But, for each word, image region *relevance probabilities* and *locations* are easily derived from the Posteriorogram – and used to probabilistically index the word in an efficient way.

Lexicon-free Probabilistic Index: Example

0 100 200 300 400 500 600

50. 2. It matters not whether the mis-supposal
 100. regards the matter of fact or matter of law.
 150. The matter of fact where you suppose some.
 200.

# pageID="Bentham-071-021-002-part"		REGARDS	0.857	5	115	84	31	THE	0.990	1	198	28	31
# keyword relPrb	bounding box	UGARDS	0.138	5	115	80	31	MATTER	0.934	61	198	64	31
#		THE	0.993	110	115	43	31	OF	0.988	141	198	28	31
2	0.929	1	36	20	31			FAST	0.367	182	198	62	31
21	0.064	1	36	24	31			FAR	0.186	182	198	36	31
IT	0.982	33	36	27	31		
IF	0.012	33	36	26	31			FACT	0.017	182	198	46	31
MATTERS	0.989	77	36	99	31			AS	0.142	200	198	29	31
MATTER	0.011	77	36	93	31			HAE	0.022	200	198	29	31
NOT	0.999	216	36	7	31			WHERE	0.992	255	198	90	31
WHETHER	1.000	256	36	99	31			YOU	0.761	365	198	45	31
THE	0.997	389	36	33	31			YOW	0.030	365	198	45	31
MIS-SUPPOSAL	1.000	455	36	193	31			GOUS	0.064	372	198	47	31
				SUPPOSE	0.975	429	198	120	31
THE	0.927	430	88	30	31			SUPFROSE	0.024	429	198	125	31
LHE	0.056	434	88	25	31			SOME	0.834	570	198	78	31
...			SONER	0.016	576	198	83	31
								OME	0.109	580	198	65	31
								ME	0.022	620	198	22	31

All character strings or “pseudo-words” which are likely enough to be real words are indexed.

Lexicon-free Probabilistic Index: Example

0 100 200 300 400 500 600

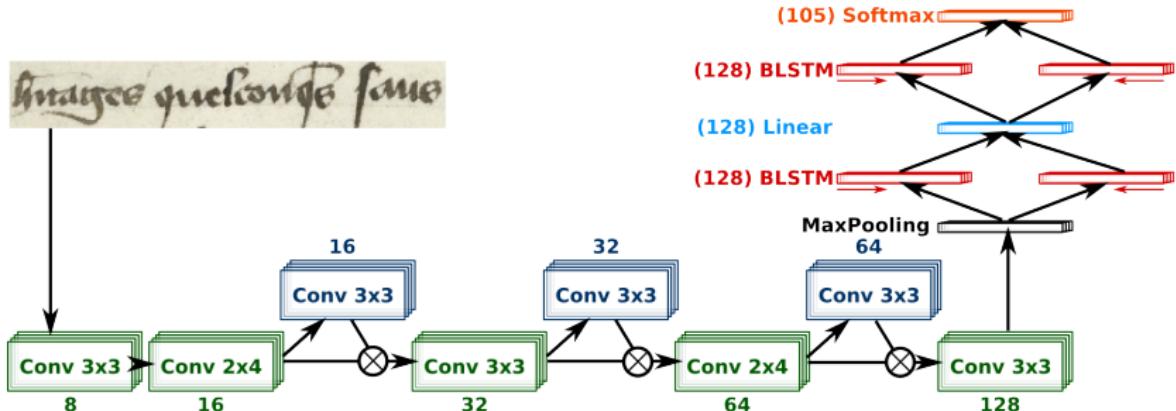
50. 2. It **matter** not whether the **Mis-supposal**
 regards the **matter** of fact or **matter** of law.
 The **matter** of fact where you suppose some-

#	pageID="Bentham-071-021-002-part"	REGARDS	0.857	5	115	84	31	THE	0.990	1	198	28	31
#	keyword relPrb	UGARDS	0.138	5	115	80	31	MATTER	0.934	61	198	64	31
#	bounding box	THE	0.993	110	115	43	31	OF	0.988	141	198	28	31
2	0.929	1	36	20	31			FAST	0.367	182	198	62	31
21	0.064	1	36	24	31			FAR	0.186	182	198	36	31
IT	0.982	33	36	27	31			FACT	0.017	182	198	46	31
IF	0.012	33	36	26	31			AS	0.142	200	198	29	31
MATTERS	0.998	160	115	93	31			HAE	0.022	200	198	29	31
MATTER	0.011	77	36	93	31			WHERE	0.992	255	198	90	31
NOT	0.999	216	36	7	31			YOU	0.761	365	198	45	31
WHETHER	1.000	256	36	99	31			YOW	0.030	365	198	45	31
THE	0.997	389	36	33	31			GOUS	0.064	372	198	47	31
MIS-SUPPOSAL	1.000	455	36	193	31			SUPPOSE	0.975	429	198	120	31
		LAW	0.032	575	115	36	31	SUPFROSE	0.024	429	198	125	31
THE	0.927	430	88	30	31			SOME	0.834	570	198	78	31
LHE	0.056	434	88	25	31			SONER	0.016	576	198	83	31
...	...	TAUE	0.031	575	115	55	31	OME	0.109	580	198	65	31
...	ME	0.022	620	198	22	31

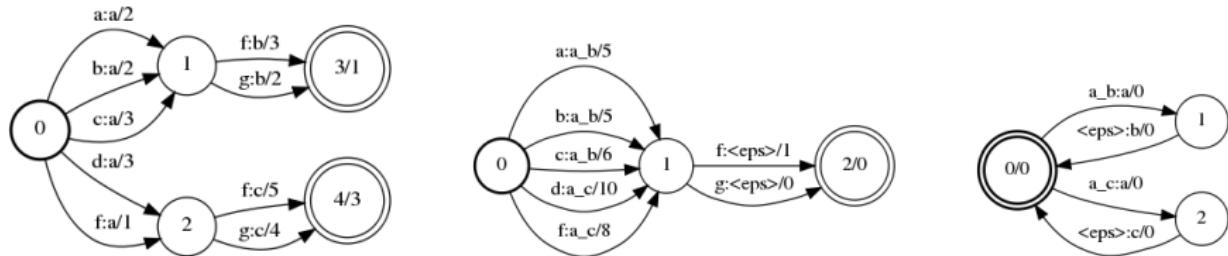
Spots for **MATTER** and **MATTERS** marked in colors according to their Relevance Probabilities.

Technologies Involved

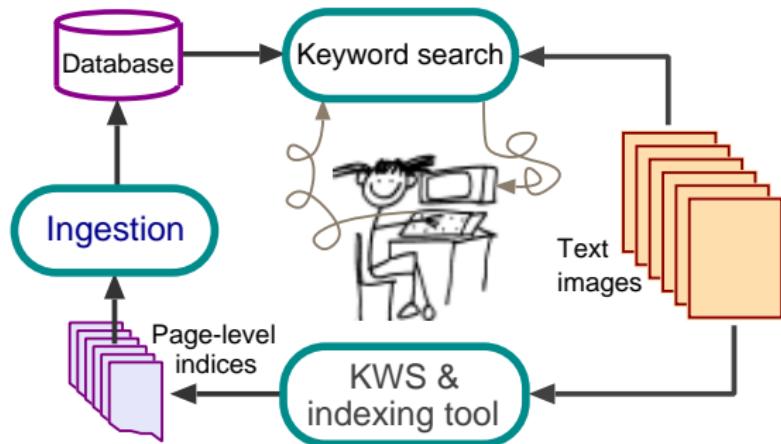
- ▶ Optical modelling: Deep CNN-RNN network:



- ▶ Textual context modelled by finite state character n -grams.



Probabilistic Text Image Indexing and Search: System Diagram



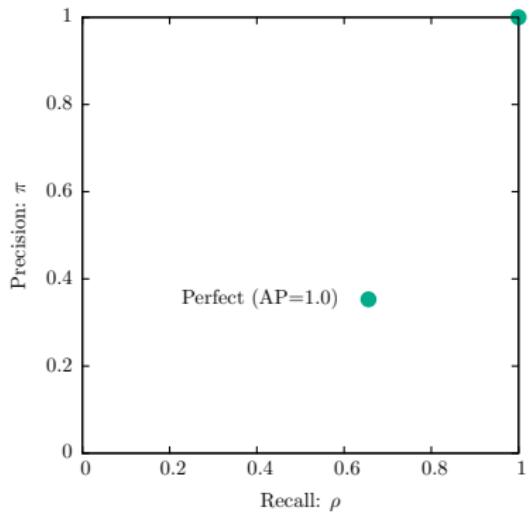
- ▶ “*KWS & indexing tool*”: Off-line pre-computation of probabilistic indices
- ▶ “*Ingestion*”: Off-line creation of the actual database. Typically a simple and computationally cheap process
- ▶ “*Keyword search*”: On-line user query analysis, find the requested information and present the retrieved images. Short response times needed.

Probabilistic Indexing & Search: Precision-Recall Tradeoff Model

Indexing and search quality can be assessed by means of *precision* (π) & *recall* (ρ) performance.

Precision is high if most of the retrieved results are correct while recall is high if most of the existing correct results are retrieved.

If perfectly correct text were indexed, you'd get a single, "ideal" point with $\rho = \pi = 1$.



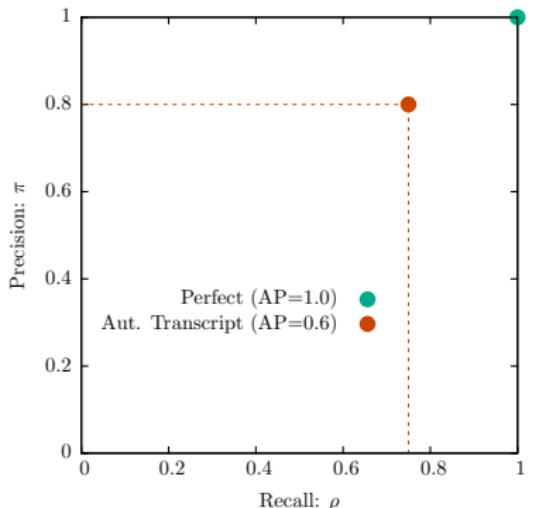
Probabilistic Indexing & Search: Precision-Recall Tradeoff Model

Indexing and search quality can be assessed by means of *precision* (π) & *recall* (ρ) performance.

Precision is high if most of the retrieved results are correct while recall is high if most of the existing correct results are retrieved.

If perfectly correct text were indexed, you'd get a single, "ideal" point with $\rho = \pi = 1$.

If automatic (typically noisy) handwritten text transcripts are naively indexed just as plaintext, precision and recall are also fixed values, albeit not "ideal" (perhaps something like, with Average Precision AP 0.6).



Probabilistic Indexing & Search: Precision-Recall Tradeoff Model

Indexing and search quality can be assessed by means of *precision* (π) & *recall* (ρ) performance.

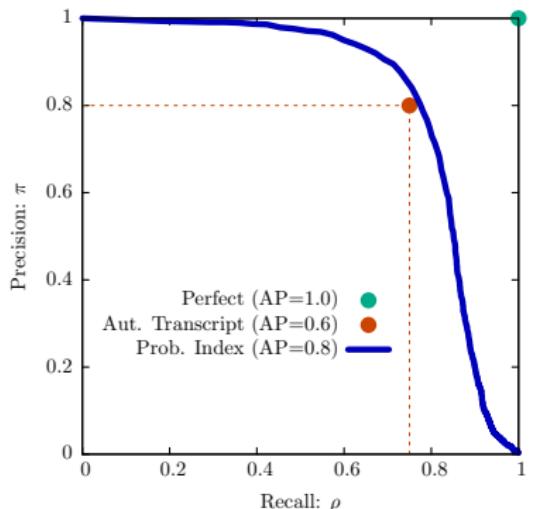
Precision is high if most of the retrieved results are correct while recall is high if most of the existing correct results are retrieved.

If perfectly correct text were indexed, you'd get a single, "ideal" point with $\rho = \pi = 1$.

If automatic (typically noisy) handwritten text transcripts are naively indexed just as plaintext, precision and recall are also fixed values, albeit not "ideal" (perhaps something like, with Average Precision AP 0.6).

In contrast, probabilistic indexing allows for arbitrary precision-recall tradeoffs by setting a threshold on the system confidence (relevance probability).

This flexible "*precision-recall tradeoff model*" obviously allows for better search and retrieval performance than naive plaintext searching on automatic noisy transcripts.



Probabilistic Indexing & Search: Precision-Recall Tradeoff Model

Indexing and search quality can be assessed by means of *precision* (π) & *recall* (ρ) performance.

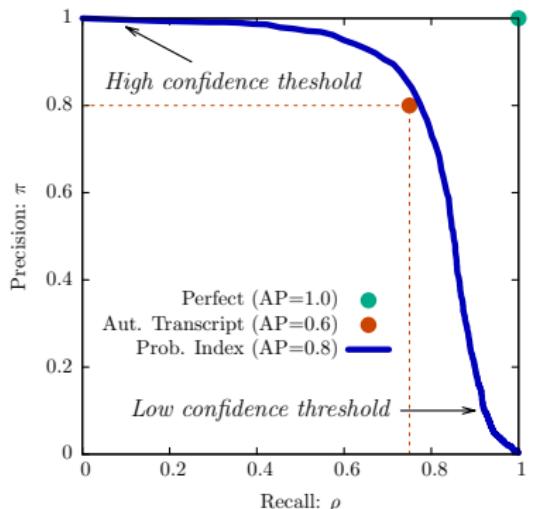
Precision is high if most of the retrieved results are correct while recall is high if most of the existing correct results are retrieved.

If perfectly correct text were indexed, you'd get a single, "ideal" point with $\rho = \pi = 1$.

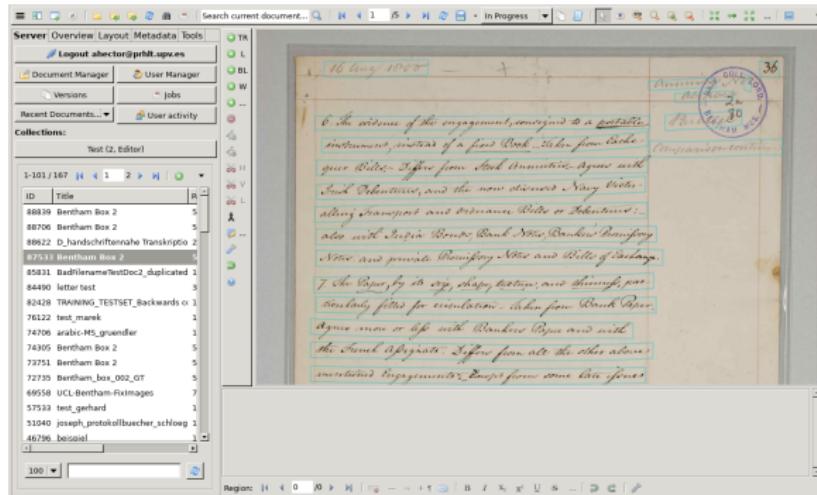
If automatic (typically noisy) handwritten text transcripts are naively indexed just as plaintext, precision and recall are also fixed values, albeit not "ideal" (perhaps something like, with Average Precision AP 0.6).

In contrast, probabilistic indexing allows for arbitrary precision-recall tradeoffs by setting a threshold on the system confidence (relevance probability).

This flexible "*precision-recall tradeoff model*" obviously allows for better search and retrieval performance than naive plaintext searching on automatic noisy transcripts.



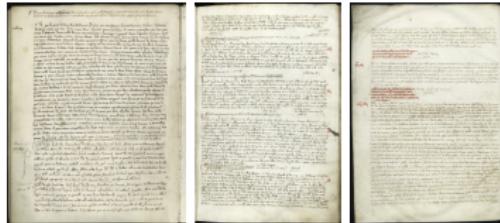
Preparatory Steps: Use of TransKribus Tools



- ▶ Line detection
- ▶ Transcribing (or page-image text aligning) a small number of training pages.
- ▶ Resizing images as required to approximately achieve uniform resolution.

Large Scale Collections: Laboratory Results

Chancery



XIV-XV century medieval registers written in old French by many hands.

Bentham

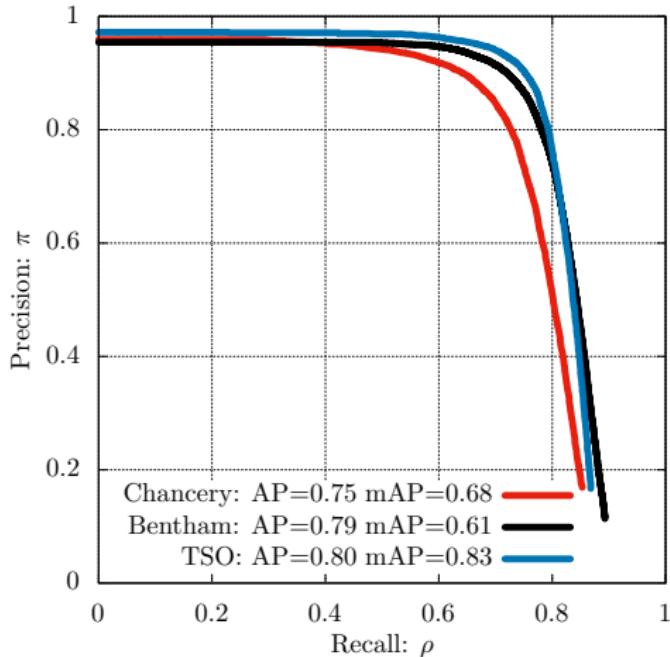


18th-19th century manuscripts written in English by many hands.

TSO



15th-16th century manuscripts written in old Spanish by many hands.



Chancery: AP=0.75 mAP=0.68

Bentham: AP=0.79 mAP=0.61

TSO: AP=0.80 mAP=0.83

Collection	train/test	Char LM	Query words
Chancery	341/95 acts	5-gram	6,506
Bentham	155/846 pgs	8-gram	3,597
TSO	cross-val 286 pgs	8-gram	5,409

Chancery: Indexing and Search Live Demonstration

PRHLT Search Interface: <http://prhl-kws.prhl/himanis>

PRHLT search engine **BETA**

Search options

Confidence: 50 Max results: 100 You are here: home

199 medieval manuscripts

© 2018 HIMANIS. See also this and this & PRHLT, IUPV. SITEMAP. CITE. All. Browse Help.

Probabilistic Index basic statistics (as of Sep-2017)

Computed:

#Volumes	167
#Page Indexed	67,413
#Spots	266,301,333
Average #Spots / Page	1,594,619

Estimated from index probabilities:

Running words	44,216,365
Running words / Page	264,769
Average #Spots / Running word	6.0

Bentham papers: Indexing and Search Large Scale Demonstrator

PRHLT Search Interface: <http://prhlt-carabela.prhlt.upv.es/bentham>

PRHLT READ UCL

Help Confidence: 50 Max. results: 50 You are here: home

How to measure Pain and Pleasure.

In small sums the quantity of pleasure can not go beyond which is nearly as the quantity of pain, as he goes on writing of Bentham papers, and carry to this interest.

But where bodies upon the part of a hundred thousand could not carry it further. Here then is the quantity of money increased a thousand fold, and that of pleasure not at all. For all this it is true enough.

© 2018 Bentham project, UCL Special Collections, British Library, PRHLT, READ – see also this and this

Probabilistic Index basic statistics (as of Nov-2018)

Computed:

#Boxes	173
#Page images / Indexed	95,247 / 89,911
#Spots	197,651,336
Average #Spots / Page	2,198

Estimated from index probabilities:

Running words	25,487,932
Running words / Page	283
Average #Spots / Running word	7.8

TSO: Indexing and Search Large-Scale Demonstrator

PRHLT Search Interface: <http://prhlt-carabela.prhlt.upv.es/tso>

The screenshot shows a search result for a manuscript page. At the top, there is a header for "TEATRO DEL SIGLO DE ORO ESPAÑOL" and a sub-header "Búsqueda de texto en imágenes – sistema PRHLT". Below the header, there are search parameters like "Categoría: SG" and "Max. resultados: 10". A "Buscar" button and links for "Acerca y ejemplos" and "Info. de imágenes indexadas" are also present. The main content area displays a photograph of a handwritten manuscript page. Two specific sections of text are highlighted with black boxes. The first highlighted section contains the lyrics:

mi gracio Amor te soy.
y ay me expusiste tan bien
Autores Conocidos
182 manuscritos

que tan diferente soy.

The second highlighted section contains the lyrics:

A pura imaginación
de la que el Vndeffer
en los palacios me llevó
ella dama valona

Autores Anónimos
146 manuscritos

At the bottom of the image, there are logos for BNE, PROLOPE, READ, and PRHLT, along with the text "© 2018 TEATRO DEL SIGLO DE ORO (BNE) / PROLOPE, READ, PRHLT".

Probabilistic Index basic statistics (as of Nov-2018)

Computed:

#Manuscripts	328
#Page images/ Indexed	41,122 / 36,010
#Spots	42,477,144
Average #Spots / Page	1,180

Estimated from index probabilities:

Running words	5,396,497
Running words / Page	150
Average #Spots / Running word	7.9

TSO Large Scale Demonstrator: Query Examples

PRHLT TSO search interface: <http://prhlt-carabela.prhlt.upv.es/tso>

A small sample of query possibilities:

Austria

ottomano

teniente || alferez || sargento

sol && español

(valor || dolor) && (amor || honor)

Isabel (belleza || hermosura || nobleza)

[Don Juan Austria]

[Lope de Vega]

[Calderon de la Barca]

TSO Large Scale Demonstrator: Query Examples

PRHLT TSO search interface: <http://prhlt-carabela.prhlt.upv.es/tso>

A small sample of query possibilities:

Austria

ottomano

teniente || alferez || sargento

sol && español

(valor || dolor) && (amor || honor)

Isabel (belleza || hermosura || nobleza)

[Don Juan Austria]

[Lope de Vega]

[Calderon de la Barca]

TSO Large Scale Demonstrator: Query Examples

PRHLT TSO search interface: <http://prhlt-carabela.prhlt.upv.es/tso>

A small sample of query possibilities:

Austria

ottomano

teniente || alferez || sargento

sol && español

(valor || dolor) && (amor || honor)

Isabel (belleza || hermosura || nobleza)

[Don Juan Austria]

[Lope de Vega]

[Calderon de la Barca]

TSO Large Scale Demonstrator: Query Examples

PRHLT TSO search interface: <http://prhlt-carabela.prhlt.upv.es/tso>

A small sample of query possibilities:

Austria

ottomano

teniente || alferez || sargento

sol && español

(valor || dolor) && (amor || honor)

Isabel (belleza || hermosura || nobleza)

[Don Juan Austria]

[Lope de Vega]

[Calderon de la Barca]

TSO Large Scale Demonstrator: Query Examples

PRHLT TSO search interface: <http://prhlt-carabela.prhlt.upv.es/tso>

A small sample of query possibilities:

Austria

ottomano

teniente || alferez || sargento

sol && español

(valor || dolor) && (amor || honor)

Isabel (belleza || hermosura || nobleza)

[Don Juan Austria]

[Lope de Vega]

[Calderon de la Barca]

Other KWS Demonstrators

Several others *Handwriting Text Keyword Indexing and Search* PRHLT demonstrators can be tried at:

<http://transcriptorium.eu/demots/KWSdemos>

Handwriting Keyword Indexing and Search PRHLT Demonstrators

Laboratory, small & medium-scale datasets (many with ground truth):

[Various collections indexed together](#)

[Bentham](#)

[Plantas](#)

[Jane Austen](#)

[Wiensanktulrich](#)

[La contienda \(HMMs\)](#)

[Lope and the Spanish Theatre Golden Age \(OLD\)](#)

[Passau miscellaneous collection](#)

[Chancery-Guerin \(95 Acts, with GT\)](#)

[Chancery-Guerin \(444 Acts, without GT\)](#)

[Chancery-Guerin lemmatized \(444 Acts, without GT\)](#)

Handwritten music symbol indexing and symbol-sequence search:

[Vorau-253 Music manuscript\(490 page images, 44 with partial GT\)](#)

Large scale (no ground truth):

[Chancery \(Trésor des Chartes registers -- 199 manuscripts, 82,000 page images\)](#)

[Bentham papers \(193 boxes, 90,000 page images\)](#)

[Teatro del Siglo de Oro \(Spanish Golden Eage Theatre -- 328 manuscripts 41,000 page images\)](#)

Thanks for Your Attention !

Average Precision (AP) versus Mean Average Precision (mAP)

	AP	mAP
Type of Ranking	Global spot list	One spot list for each query
Averaging type	Micro: over all query events	Macro: over the APs of isolated queries
Is it invariant to monotonic score transformations?	no	yes
Does score consistency have an impact?	yes	no
Need all queries be relevant?	no	yes