

- Professional searchers (e.g., lawyers) still like Boolean queries:
 - You know exactly what you're getting.



Dip. di Matematica Pura ed Applicata







C	00K × 1M matrix has half-a-trillion 's and 1's.
	ut it has no more than one billion 1's. Matrix is extremely sparse. Why?
	Vhat's a better representation? We only record the 1 positions.
Dip. di Matem	atica F.Aiolli - Sistemi Informativi 11
Inve	rted index each term <i>t</i> , we must store a list of all ments that contain t.
docu	
docu Do w	e use an array or a list for this?
docu Do n Brutus	e use an array or a list for this? ====>248163264128
docu Do w 3rutus Calpurnia	The use an array or a list for this?
docu Do w Brutus Calpurnia Caesar	The use an array or a list for this?
docu Do w Brutus	e use an array or a list for this?



Indexer st	reps		Term I did enact julius caesa	t ar	c# 1 1 1 1
			l was killed		1 1 1
Sequence of (Document ID)	Modified token) pairs.	,	i' the capito brutus killed me	ol s	1 1 1 1 1 1
Doc 1	Doc 2		so let it be with		2 2 2 2 2 2
I did enact Julius Caesar I was killed i' the Capitol; Brutus killed me.	So let it be w Caesar. The n Brutus hath to Caesar was amb	vith oble Id you pitious	caesa the noble brutus hath told you caesa was	ar s ar	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
Dip. di Matematica					15
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	tivi			
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	tīvī			
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc #		erm [Doc # 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did	1 ar 1 bb	erm [mbitious e	Doc # 2 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact	1 1 1 1 1 1	erm [mbitious e rutus	Doc # 2 2 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar	1 au 1 bu 1 bu 1 cu	erm [mbitious e rutus rutus anitol	Doc # 2 1 2 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I	T 1 au 1 bu 1 bu 1 cu	erm [mbitious e rutus apitol aesar	Doc # 2 2 1 2 1 1 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I was	T 1 au 1 bu 1 bu 1 cu 1 cu 1 cu 1 cu	erm [mbitious e rutus apitol aesar aesar	Doc # 2 2 1 2 1 1 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I was killed i'	T 1 au 1 bu 1 bu 1 cu 1 cu 1 cu 1 cu 1 cu	erm E mbitious e rutus apitol aesar aesar aesar id	Doc # 2 2 1 2 1 1 2 2 1 1 2 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius ⊂ caesar I Was killed i' the □	1 au 1 bu 1 bu 1 cu 1 cu 1 cu 1 cu 1 cu 1 cu 1 cu 1 c	erm E mbitious e rutus apitol aesar aesar aesar id nact	Doc # 2 2 1 2 1 1 2 2 1 1 1 1 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I was killed i' the capitol	1 au 1 bu 1 bu 1 cc 1 cc 1 cc 1 cc 1 cc 1 cc 1 cc 1 c	erm [mbitious e rutus rutus apitol aesar aesar aesar aesar id nact ath	Doc # 2 2 1 1 1 1 2 2 2 1 1 1 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I was killed I' the capitol brutus killed	1 au 1 bu 1 bu 1 ca	erm [mbitious e rutus rutus apitol aesar aesar aesar id nact ath	Doc # 2 2 2 1 1 2 2 1 1 1 1 1 1 1 1 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I was killed i' the capitol brutus killed me _	1 au 1 bu 1 bu 1 ca 1 l 1 l 1 l 1 l	erm [mbitious e rutus rutus apitol aesar aesar id nact ath	Doc # 2 2 1 1 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I Killed i' the capitol brutus killed me so i	1 au 1 bu 1 bu 1 ca 1 l 1 l 1 l 1 l 1 l 1 l 1 l 1 l 1 l	erm [mbitious e rutus apitol aesar aesar id nact ath	Doc # 22 22 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 1 2 2 2 1 1 1 1 1 1 1 1 2 2 2 1 1 2 2 2 1 1 2 2 2 1 1 2 2 2 2 1 1 1 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I vas killed i i' the capitol brutus killed me so let it	1 au 1 bu 1 bu 1 bu 1 ca 1 la	erm [mbitious e rutus apitol aesar aesar aesar id nact ath illus	Doc # 22 22 11 11 12 22 11 11 11 11 11 11 11 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I killed i' the capitol brutus killed me so let it be	1 ar 1 br 1 br 1 br 1 cr 1 r 2 juit 2 juit	erm E mbitious e rutus apitol aesar aesar aesar id nact ath illus illed	Doc # 22 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I killed i' the capitol brutus killed me so let it ibe with is in the second se	T aa 1 bb 1 bb 1 cc 1 cc 1 cc 1 cc 1 cc 1 cc	erm E mbitious e rutus apitol aesar aesar aesar id nact ath illus illus	Doc # 22 22 11 11 12 22 11 11 11 11 11 11 12 11 11
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I killed i' the capitol brutus killed me so let it it be with caesar it	1 aa 1 bb 1 b 1 ca 2 ca 2 ca	erm E mbitious e rutus apitol aesar aesar aesar id nact ath illus illed illed illed illed	Doc # 22 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I killed i' the capitol brutus killed ime so iet it it it iet it it iet it it iet it it it iet it it it iet it	1 aa 1 bb 1 b 1 ca 2 ca aa ca	erm E mbitious e rutus apitol aesar aesar aesar id nact ath illed illed illed illed oble o	Doc # 2 2 1 1 1 1 1 2 2 1 1 1 1 1 1 1 1 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 1 1 1 1 2 2 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I killed i' the capitol brutus killed me so i let it be with caesar the noble brutus i	1 aa 1 bb 1 b 1 ca 2 ca 2 ca a ca ca <td>erm E mbitious e rutus apitol aesar aesar aesar id nact ath illus illed illed illed oble o o e</td> <td>Doc # 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 2 1 1 2 1 1 2 1 1 2 1</td>	erm E mbitious e rutus apitol aesar aesar aesar id nact ath illus illed illed illed oble o o e	Doc # 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 2 1 1 2 1 2 1 1 2 1 1 2 1 1 2 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I Values in the capitol brutus killed in the capitol brutus killed in the capitol brutus killed in the caesar it the caesar the noble brutus hath it caesar it the caesar it	1 T 1 bin 1 bin 1 ci 1 it 1 it 2 it 2 it 2 si	erm E mbitious e rutus apitol aesar aesar aesar id nact ath illius illius illied illed et ne oble o o ne ne e	Doc # 22 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 2 1 2 2 1 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I i so killed i' the capitol brutus killed me so let it be with caesar the caesar lut brutus killed me so let it be with caesar let it be so let it caesar let it be so let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar let caesar caesar let caesar caesar let caesar caesa	T au 1 bb 1 bb 1 bc 1 cc 1 cc	erm E mbitious e rutus apitol aesar aesar aesar id nact ath illius illius illius illius et oble o o e e ie ie ie	Doc # 22 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 1 2 2 1 1 1 1 2 2 1 1 1 1 1 2 2 1 1 1 1 1 1 2 2 2 1
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I i so killed i' the capitol brutus killed me so let it be with caesar the noble brutus hath told you caesar	T au 1 bb 1 bb 1 bb 1 cc 1 cc 2 cc 1 cc	erm E mbitious e rutus apitol aesar aesar aesar aesar id nact ath illius illius illius e oble o o e e e o le e e o le o u asa	Doc # 22 1 1 2 2 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 1 1 2 2 2 1 1 2 2 1 1 2 2 2 1 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I Values in the capitol brutus killed in the caesar the caesar the noble brutus hath caesar the noble brutus hath itold you caesar was in the caesar in the caesar the caesar in the cae	1 T 1 bi 1 bi 1 bi 1 ci 1 it 2	erm E mbitious e rutus apitol aesar aesar aesar aesar id nact ath illius illius illius oble o o bole o o ne ne ne sid o u aesar aes aesar aes aes aes aes aes aes aes aes aes aes	Doc # 22 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 2 1 1 2 2 1 1 2 2 2 1 2 2 2 1 2 2 2 1 2 2 1 2 2 2 2 1 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # 1 did enact julius caesar l vas killed i' the capitol brutus killed me so let it be with caesar the noble brutus hath told you caesar was ambitious	1 T 1 bi 1 bi 1 bi 1 ci 1 it 1 it 2 it it	erm E mbitious e rutus apitol aesar aesar aesar aesar id nact ath illed illed at ne oble o o e e o ld o u aesar aes aes aes aes aes aes aes aes aes aes	Doc # 22 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 1 1 2 2 1 1 2
Pura ed Applicata	F.Aiolli - Sistemi Informa 2007/08	Term Doc # I did enact julius caesar I Values cites the capitol brutus killed me so i so i the capitol brutus killed me so i let it be with caesar the noble brutus hath itold you caesar was ambitious i	1 T 1 bit 1 bit 1 bit 1 cit 2 it 2 www 2 wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww	erm E mbitious e rutus apitol aesar aesar aesar aesar id nact ath illed illed t t e oble o o le e e o le e o le o u a s ar aesar id aesar aesa aesa	Doc # 2 2 1 1 2 2 1 1 1 1 1 1 1 1 1 1 1 1 1



Dip. di Matematica Pura ed Applicata

Aiolli - Sistemi Informa 2007/08





What about an arbitrary Boolean formula? (Brutus OR Caesar) AND NOT			
(Antony OR Cleopatra)			
Can we always merge in "linear" time?			
 Consider (Brutus AND Caesar) AND Calpurnia Can we do better? 			
Dip. di Matematica Pura ed Applicata	F.Aiolli - Sistemi Informativi 2007/08	23	
Beyond to	erm search		
Beyond to What abou	erm search t phrases?		
Beyond to What abou	erm search t phrases? Find Gates NEAR Microsoft.	- 4	



Advantages of the BM

- Possibility to formulate structured queries. E.g. to distinguish 'synonymy' (or t₁ t₂) from noun phrases (and t₁ prox[0,F] t₂)
- In the case of expert users, intuitivity. To those users familiar with Boolean Logic, it is immediately clear why a doc has been or not been retrieved following a given query. This allows query refinement and query reformulation on the part of the user.
- Efficiency obtained through the use of inverted files (Ifs) the data structures in secondary storage in which document representations are physically stored.

Dip. di Mate	ematica
Pura ed App	plicata

F.Aiolli - Sistemi Informativi 2007/08

Disadvantages of the BM

- Intimidating (in general, there is not automatic query acquisition). The proponent of the BM where mathematicians; but "lay users" find the language of the BM unnatural.
- 2. Lack of output magnitude control after a given query, unless the user knows well the distribution of the topics in the collection. This is a consequence of the fact that RSV takes binary values.
- Output obtained as a result of a given query is not ranked with respect to the estimated degree (probability) of relevance.

27

4.	No importanc in the IREPs	e factors can be attached to the of documents and query.	index terms
5.	"Flattened", H discrimination	nence unintuitive, results (lack of n)	output
Bo int ne se	olean system termediaries eds of 'casuc ekers.	ns have thus mainly been use and are widely considered u al' (non professional) informa	d by expert nfit the tion