Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

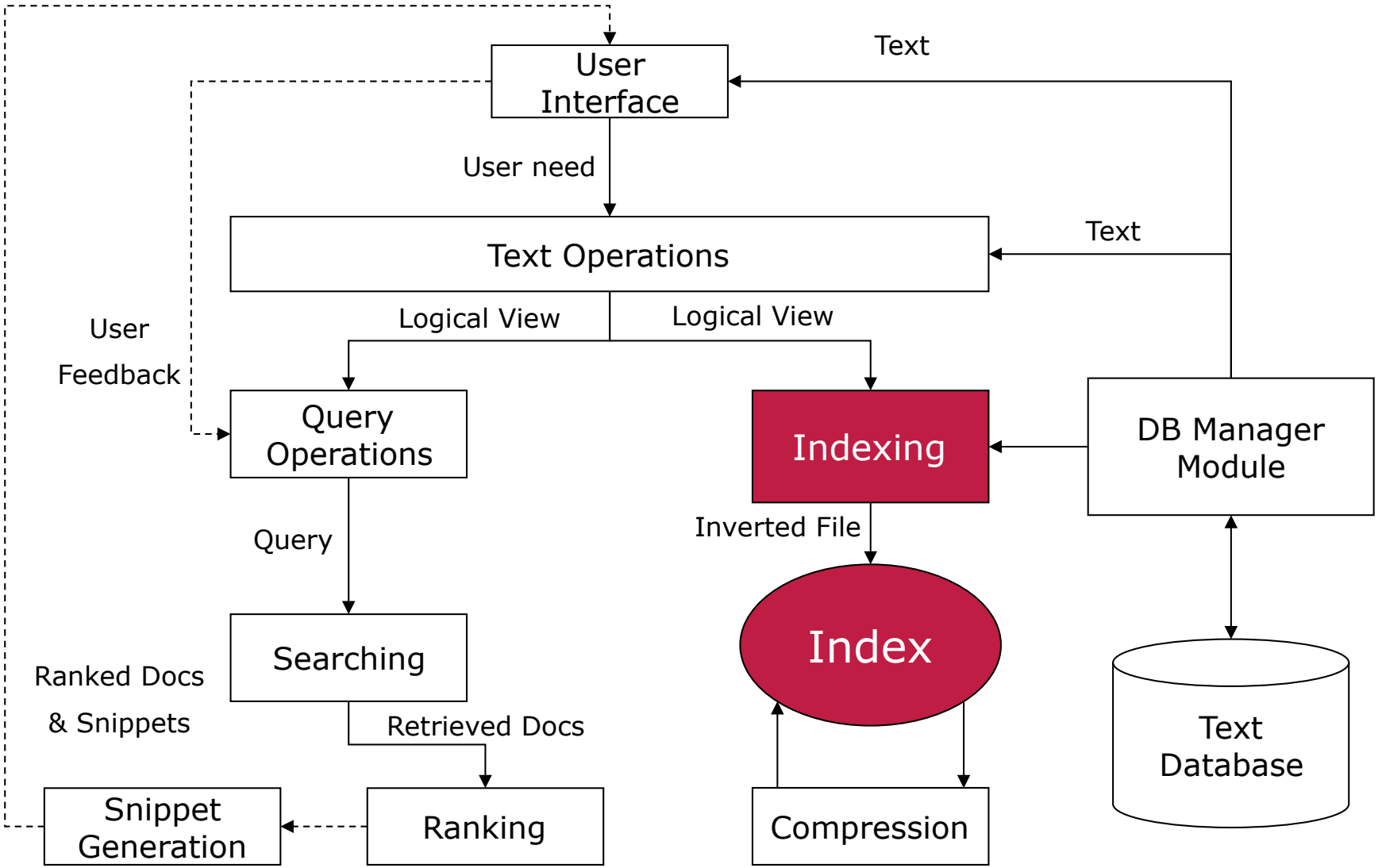
Index Construction

Michael Leben, Martin Lorenz

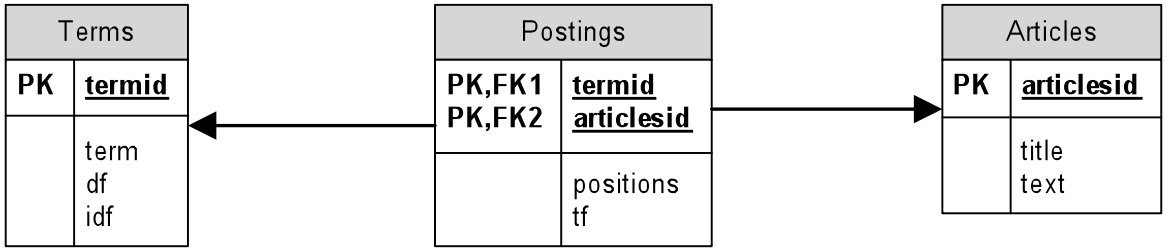
Agenda

- Task Definition
- Database Schema
- Design Decisions
- System Architecture
- Wikiparser
- Stopping / Stemming
- Statistics
- Demo
- Learnings

Task Definition



Database Schema



Terms:

termid	term	documentfrequency	inversedocumentfrequency
105441	jpg	6001	2.6026
-980216425	pratik	1	11.2253
-1102999388	licens	8617	2.27257
50511086	categori	31046	1.22845
374104928	gujarat	30	7.82451
101491	fly	482	5.05377
3169703	gfdl	1717	3.79961

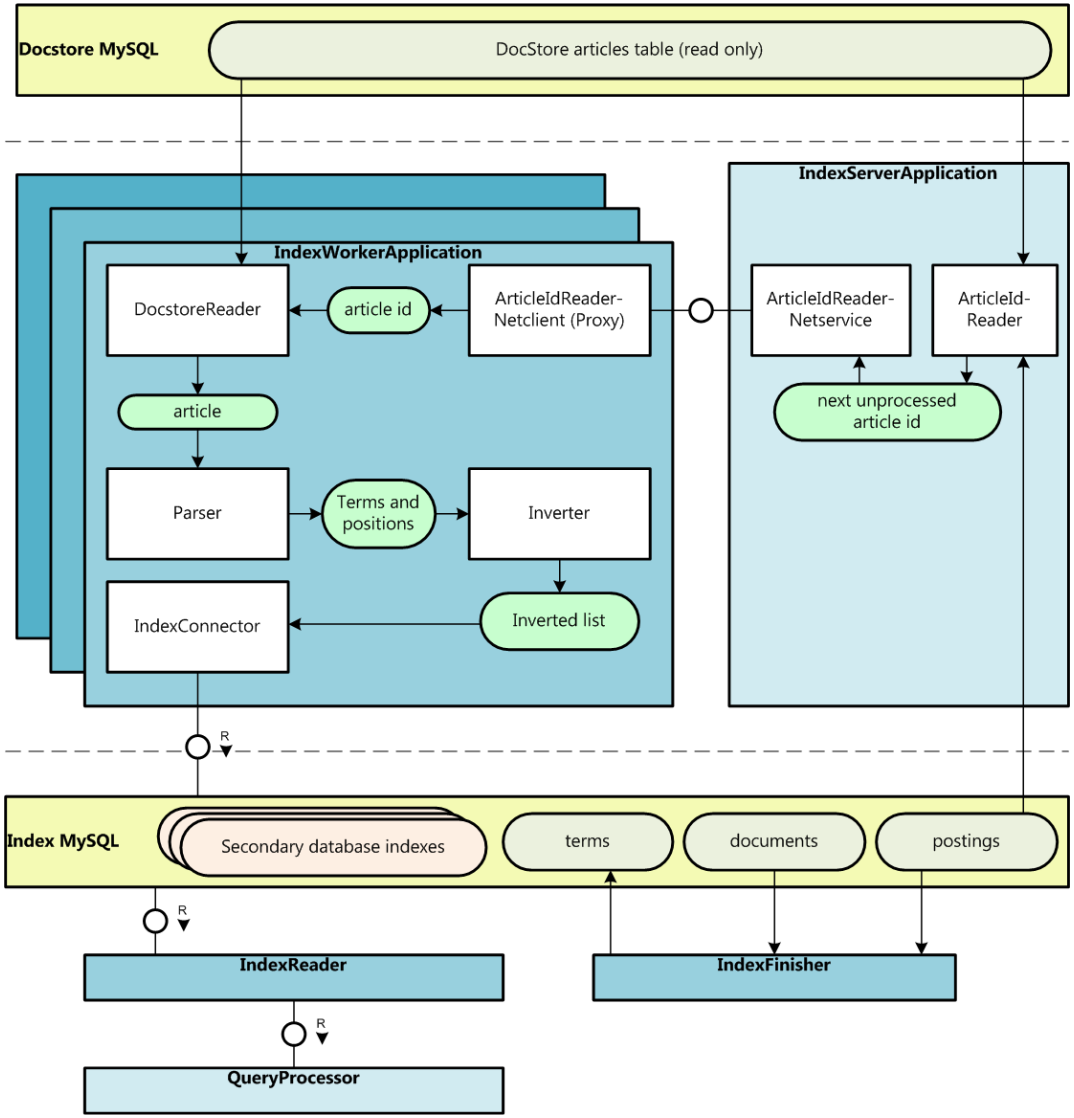
Postings:

termid	articlesid	positions	termfrequency
1347878211	17273230	357	1
93746367	17273230	516	1
99651	17273230	461,464	2
-2111524382	17273230	390	1
108696186	17273230	305,153,149,303	4
-1021901383	17273230	448	1

Design Decisions

- Distributed Indexing Application (Assumption DB is „read-only“)
  - task distributor (IndexServerApplication)
  - indexing (Multiple IndexWorkerApplications)
- Apache Lucene
  - parsing wiki-markup
  - word stemming
- MySQL (InnoDB vs. MyISAM)
  - InnoDB
    - row lock instead of table lock

System Architecture



Wikiparser

Before

'''[[Albania]]''' {{Audio-IPA|en-us-Albania.ogg|/Ã?lË^beÉªniÉ^(TM)/}},
officially the '''Republic of Albania''' ([[Albanian
language|Albanian]]: ''Republika e ShqipÃ«risÃ«''
{{pronounced|É3/4É>Ë^publika É> ÊfcipÉ^(TM)Ë^É3/4iË?s}}, or simply ''ShqipÃ«ria'', [[Gheg Albanian]]: ''Shqipnija''),
is a country in [[Balkans|South Eastern Europe]]. It is bordered by [[Greece]] to the south-east, [[Montenegro]] to the
north, [[Kosovo]]<ref>[[Kosovo]] is recognized as a [[Serbia]]n province by the UN and most UN member states,
however a minority of nations have recognized the ''Republic of Kosovo'' as an independent nation. For more on Kosovo's
status see:
[[Kosovo status process]]</ref> to the northeast, and the [[Republic of Macedonia]] to the east. It has a coast on
the [[Adriatic Sea]] to the west\, and on the [[Ionian Sea]] to the southwest. It is less than 72 km
(45 miles) from [[Italy]], across the [[Strait of Otranto]] which links the [[Adriatic Sea]] to the [[Ionian Sea]].

After

albania ogg ã lë^beé nié officially the republic of albania albanian language albanian republika e shqipã risã pronounced é
é ë^publika é êfcipé ë^é ië s or simply shqipã ria gheg albanian shqipnija is a country in balkans south eastern europe it
is bordered by greece to the south east montenegro to the north kosovo lt ref&gt kosovo is recognized as a serbia n
province by the un and most un member states however a minority of nations have recognized the republic of kosovo as
an independent nation for more on kosovo's status see kosovo status process lt ref&gt to the northeast and the republic of
macedonia to the east it has a coast on the adriatic sea to the west and on the ionian sea to the southwest it is less than
72 amp nbsp km 45 amp nbsp miles from italy across the strait of otranto which links the adriatic sea to the ionian sea

Stopping / Stemming

Our Implementation:

- CSV with static list of stopwords

Improvements:

- add statistical set of stopwords
    - by df

a	allow	anyhow	available	below	certain	course	downwards
able	allows	anyone	away	beside	certainly	c's	during
about	almost	anything	awfully	besides	changes	currently	e
lt	alone	anyway	b	best	clearly	d	each
gt	along	anyways	back	better	c'mon	dare	edu
above	alongside	anywhere	backward	between	co	daren't	eg
abroad	already	apart	backwards	beyond	co.	definitely	eight
according	also	appear	be	both	com	described	eighty
accordingly	although	appreciate	became	brief	come	despite	either
across	always	appropriate	because	but	comes	did	else
actually	am	are	become	by	concerning	didn't	elsewhere
adj	amid	aren't	becomes	c	consequently	different	end
after	amidst	around	becoming	came	consider	directly	ending
afterwards	among	as	been	can	considering	do	enough
again	amongst	a's	before	cannot	contain	does	entirely
against	an	aside	beforehand	cant	containing	doesn't	especially
ago	and	ask	begin	can't	contains	doing	et
ahead	another	asking	behind	caption	corresponding	done	etc
ain't	any	associated	being	cause	could	don't	even
all	anybody	at	believe	causes	couldn't	down	…

Stopping / **Stemming**

- Lucene Stemmer (Porter Stemmer)

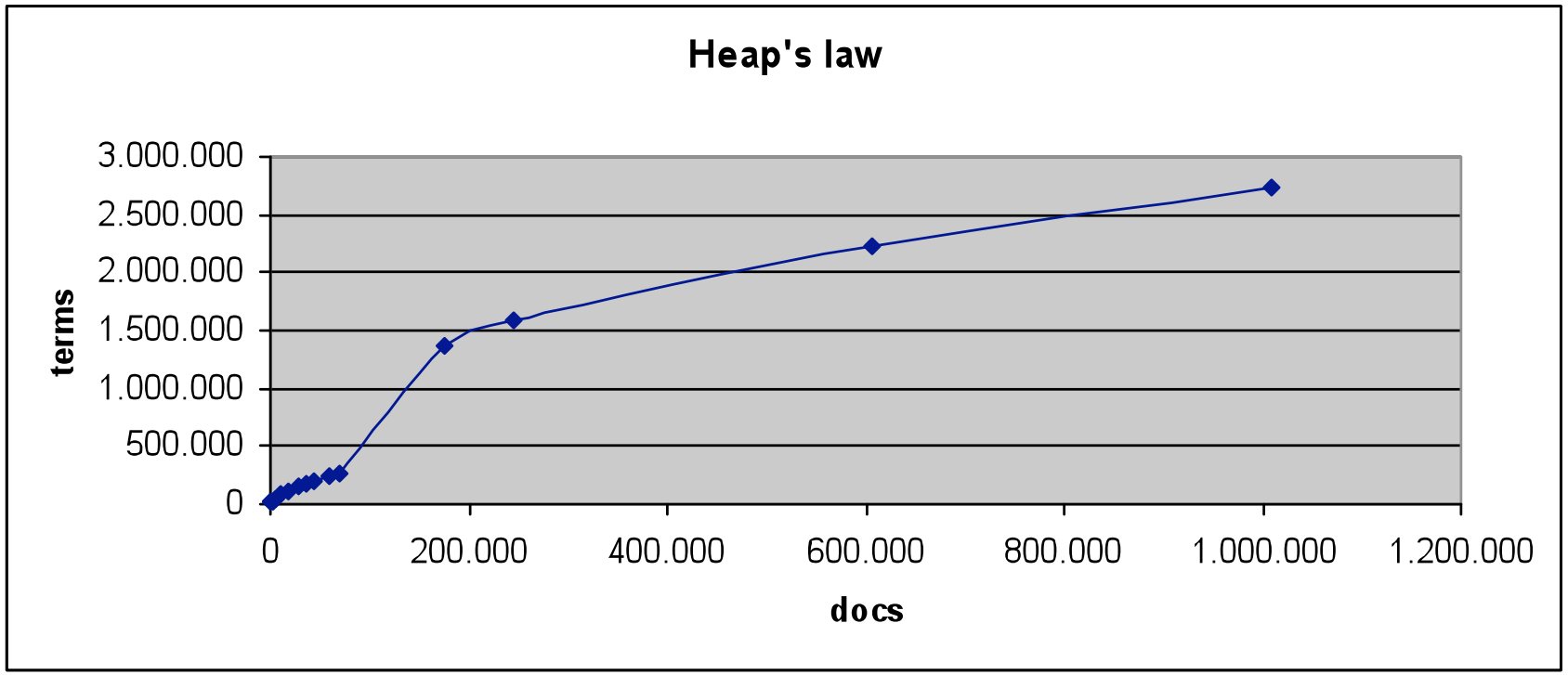
```
String term = "Banking";
Set<String> nt = TextOperator.getNormalizedTerms(word);

System.out.println(nt)          // Print: [bank]
```

- Examples

bank	[bank]		official	[offici]
banking	[bank]		offices	[offic]
banks	[bank]		fish	[fish]
Search-Engines	[search, engin]		fished	[fish]
police	[polic]		fishing	[fish]
policy	[polici]		Heap‘s	[heap]
policies	[polici]		to be or not to be	[]
policed	[polic]			
office	[offic]			

Statistics (I / II)

Heap's Law for Wikipedia



Statistics (II / II)

Where is the time?

Call Tree

Thread name	<Percent Per Thread
main[3656]	100,00 %
main(java.lang.String[]) void	99,73 %
start() void	83,18 %
indexArticle(de.unipotsdam.hpi.wikipediaparsing.WikipediaArticle) int	81,33 %
insertDocumentIntoIndex(de.unipotsdam.hpi.database.util.Document) boolean	79,17 %
addTermToIndex(java.lang.String) void	38,47 %
executeBatch() int[]	33,31 %
setString(int, java.lang.String) void	2,57 %
setInt(int, int) void	0,87 %
addBatch() void	0,43 %
clearBatch() void	0,14 %
clearParameters() void	0,00 %
executeBatch() int[]	29,74 %
executeBatchSerially() int[]	29,74 %
getMutex() java.lang.Object	0,00 %
clearBatch() void	0,00 %
getRewriteBatchedStatements() boolean	0,00 %
isReadOnly() boolean	0,00 %
clearWarnings() void	0,00 %

Demo


```
SELECT * FROM wikipedia.articles
LEFT JOIN (postings, terms)
ON (terms.termid = postings.termid AND postings.articlesid = articles.id)
WHERE terms.term = 'hasso'
```

Learnings

bad infrastructure slows down application

- database inserts a lot slower than document processing
- parallelized document processing doesn't improve indexing speed