

Информационный ПОИСК

Индексирование,
структуры данных,
булев поиск



Петрозаводский государственный
университет



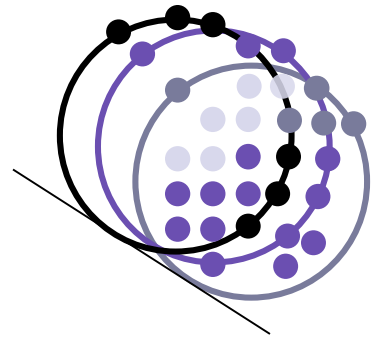
Крижановский Андрей Анатольевич

andrew.krizhanovsky@gmail.com¹





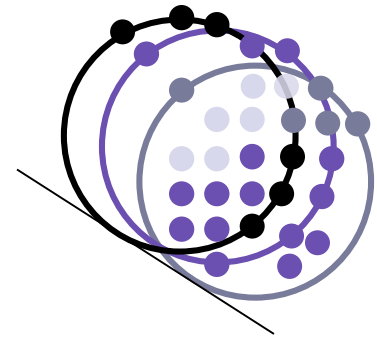
Информационный поиск



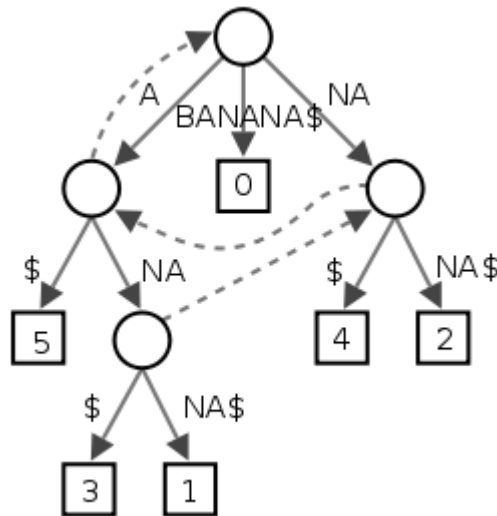
(Information retrieval) — это процесс поиска в большой **коллекции** (хранящейся, как правило, в памяти компьютеров) некоего **неструктурированного** материала ("обычно — документа), удовлетворяющего **информационные потребности**



Основные структуры данных в ИПС

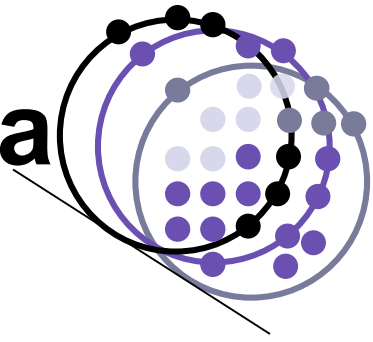


- Словарь
- Инвертированный индекс
- Term-document incidence matrices
- N-gram index
- Suffix tree
 - Suffix array (memory)





Представление документа



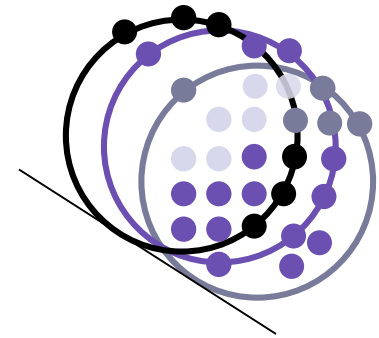
- Весь текст (самое полное представление)

- Список всех слов



- Список нормализованных слов

- *Индекс* (слово -> документ)



Term-document incidence matrices

МАТРИЦЫ ИНЦИДЕНЦИЙ ТЕРМИН-ДОКУМЕНТ

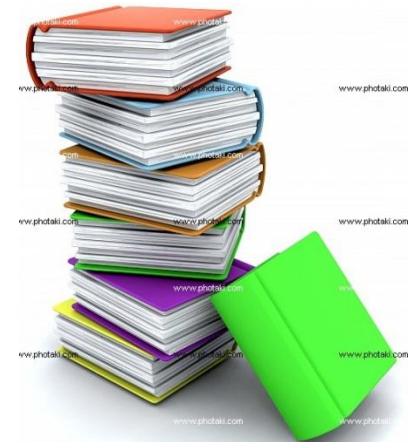
Пример информационного поиска



- Пьеса? (37)

Brutus AND Caesar AND NOT Calpurnia

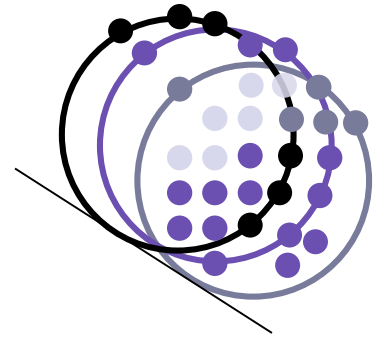
- Linear scanning: **grepping**
(прямой поиск)
 - + джокеры
(wildcard pattern matching)



<1 млн слов



Усложнение задачи



1. Большая коллекция документов
2. Гибкость запроса

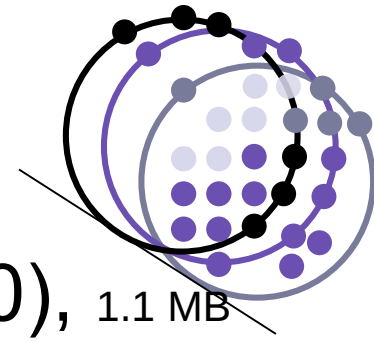
“Romans NEAR countrymen”, $\otimes$ grep

3. Ранжированный поиск



Ускорение поиска

(модель булева поиска)



- $\forall$ документ $\rightarrow$ слово (лексикон 32 000), 1.1 MB
- term-document incidence matrix
 - термин (IR) $\leftrightarrow$ слово
 - вектор термина (по строкам)
 - вектор документа (столбец)

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Brutus AND Caesar AND NOT Calpurnia

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

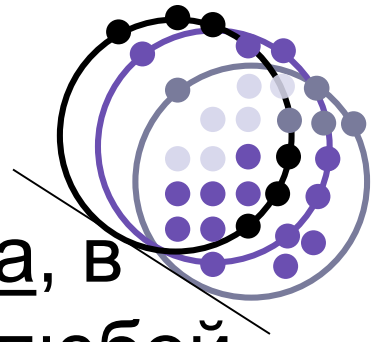
110100 AND
110111 AND
NOT 010000

⇒

110100 AND
110111 AND
101111=
100100



Модель булева поиска

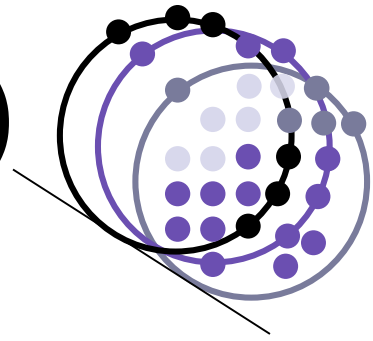


- это модель информационного поиска, в ходе которого можно обрабатывать любой запрос, имеющий вид булева выражения, т.е. выражения, в котором термины используются в сочетании с операциями AND, OR и NOT.
- В этой модели – документ:
 - множество слов





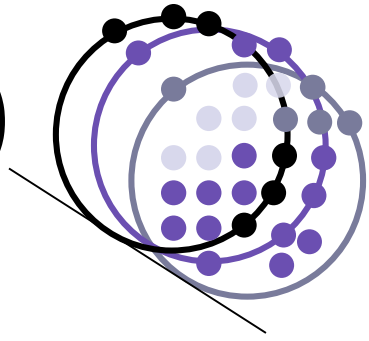
Реалистичный пример (1)



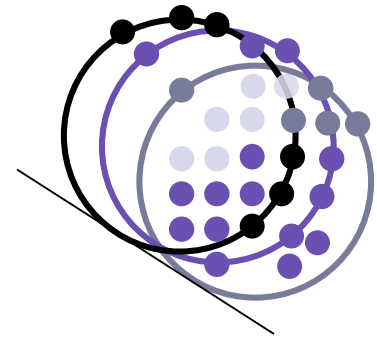
- $N = 1$ миллион документов
 - группа документов: коллекция, корпус, массив текстов (body of texts)
 - $\exists$ документ в 1000 слов (2-3 стр)
 - если слово 6 байт, то коллекция ~ ? Гб
 - м.б. обычно ~ 500 000 разных терминов
- Цель – разработать систему, выполняющую поиск по произвольному запросу (ad hoc retrieval)



Реалистичный пример (2)



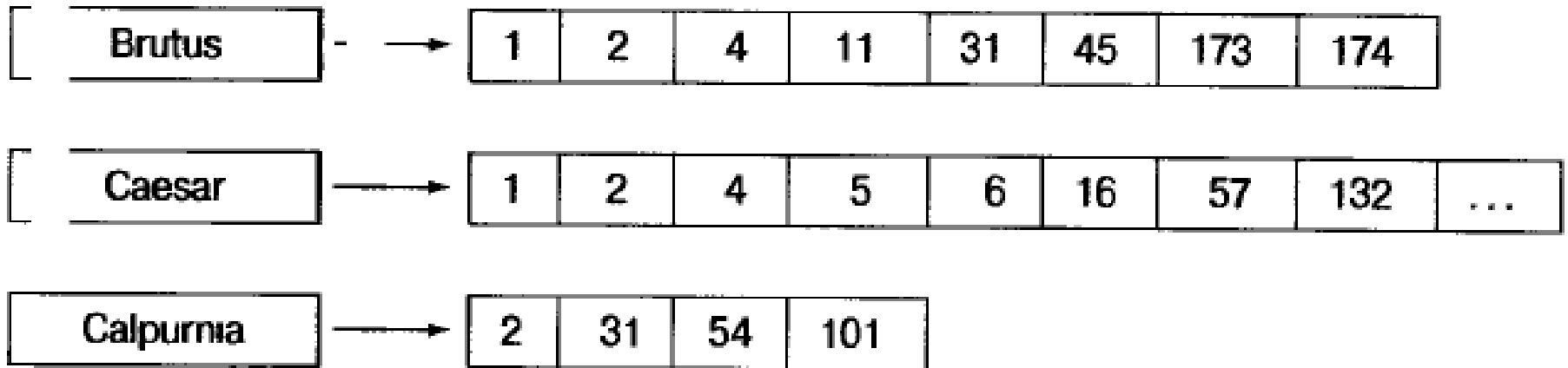
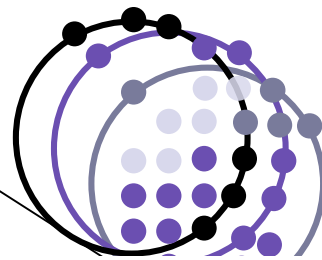
- Матрица $500 \text{ тыс} \cdot 1 \text{ млн} = 500 \text{ млрд} = 500 \text{ Гб}$
 - разреженная, 99.8% нули
- Структура получше?
 - будем хранить в памяти только единицы



Инвертированный индекс

Ключевая структура данных
в современных ИПС

(не)координатный инвертированный индекс



∃ термин
в документе

координата термина в этом
документе (словопозиция, posting)

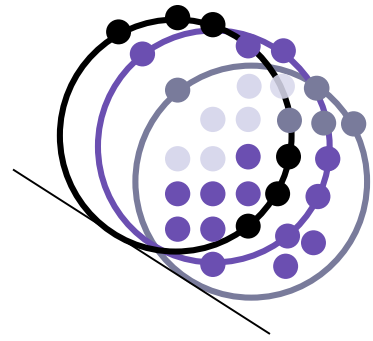
Словарь

Словопозиции

Рис. 1.3. Две части инвертированного индекса. Словарь обычно находится в памяти вместе с указателями на каждый список словопозиций, которые хранятся на диске



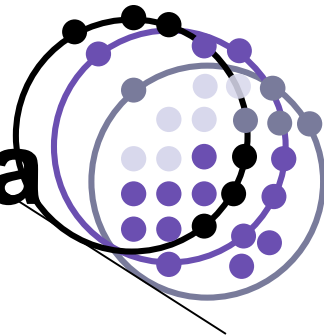
Handout



©1996 by J. Wood



Этапы построения инвертированного индекса



1. Собираем документы

Цыганы шумною толпой	Для вас, души моей царицы,	В ауле, на своих порогах,
----------------------	----------------------------	---------------------------

2. Размечаем текст, превращая $\forall$ документ в список лексем (tokens)

Цыганы	шумною	толпой	...	Для	вас
--------	--------	--------	-----	-----	-----

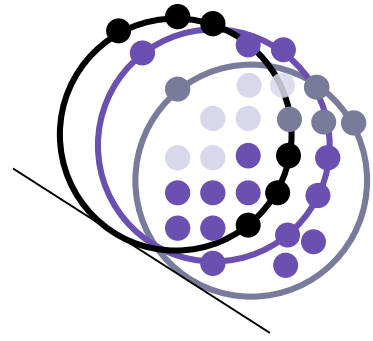
3. Предварительная лингвистическая обработка

цыган	шумный	толпа	...	для	вас
-------	--------	-------	-----	-----	-----

4. Индексируем документы (рис.)



Построение инв. индекса



- docID – идентификатор документа
- $\forall$ документ, на входе:
 - список нормализованных лексем для каждого документа, т.е. список пар «термин-docID»

термин	docID	термин	docID
1	1	ambitious	2
did		be	2
enact		br utus	1
julius		br utus	2
caesar		capitol	1
1		caesar	1
was	1	caesar	2
killed	1	caesar	2
i'		did	1
the		enact	1
capitol		hath	1
br utus		'	
killed	,	.	
me	1	i'	1
so	2	it	2
let	2	julius	1
it	2	killed	
be	2	killed	
with	2	let	2
caesar	2	me	1
the	2	noble	2
noble	2	so	2
br utus	2	the	1
hath	2	the	2
told	2	told	2
you	2	you	2
caesar	2	was	1
was	2	was	2
ambitious	2	with	2

sort



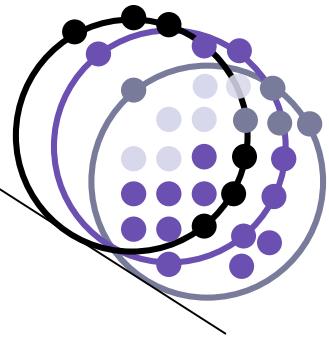
документная частота
(document frequency)

термин	док. частота	→	список словопозиций
ambitious	1	→	2
be	1	→	2
br utus	2	→	1 2
capitol	1	→	1
caesar	2	→	1 2
did	1	→	1
enact	1	→	1
hath	1	→	2
1	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 2
told	1	→	2
you	1	→	2
was	2	→	1 2
with	1	→	2

Группировка по слову, по документам



Инвертированный индекс

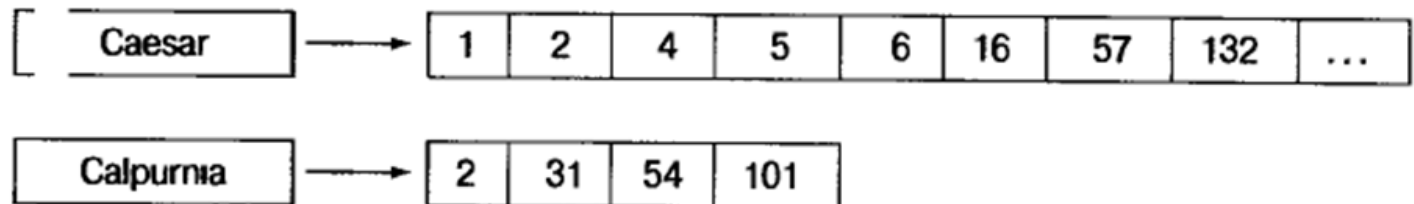


- Основные структуры данных:

- Словарь

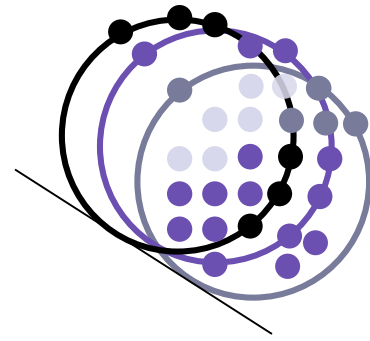
- Списки словопозиций

- Массив фиксированной длины?..



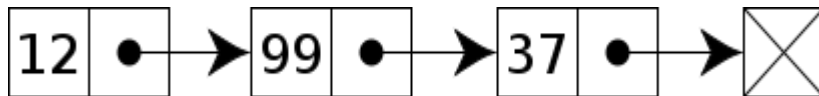


Структура данных для списков словопозиций

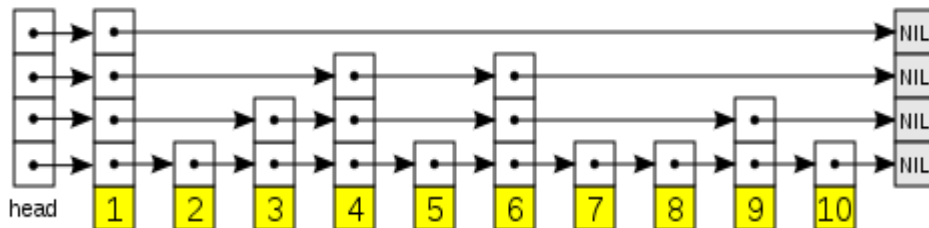


- Односвязные списки (singly linked lists)

- вставка $O(1)$



- списки с пропусками (skip list)



- Массивы переменной длины

- память ↘

(нет указателей)

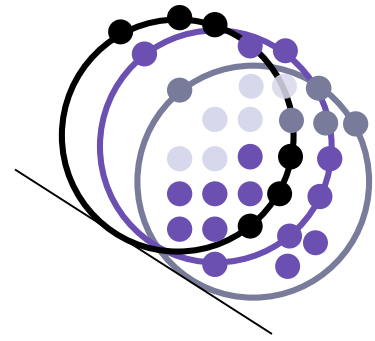
- $t \nearrow$

(процессор с кэшем)

См. анимацию в статье ВП «Список с пропусками»



Литература



- **Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск.** — Вильямс, 2011. — ISBN 978-5-8459-1623-5. <http://rutracker.org/forum/viewtopic.php?t=3887364>

Задачи IR

- Проектирование ИПС
- Классификация / кластеризация док-в
- Классификация веб-запросов
- Извлечение информации (Information extraction)
 - Извлечение терминологии / концептов
 - Выделение именованных сущностей (*Named-entity recognition*)
 - Разрешение кореферентности (Coreference resolution)
 - Извлечение семантических отношений
 - **Анализ терминальных данных**

Спасибо за внимание!



<http://vk.com/club41102811>
“Интернет-математика в ПетрГУ”