

Информационный ПОИСК

Proximity search

Сравнение расширенной булевой
модели и ранжированного поиска



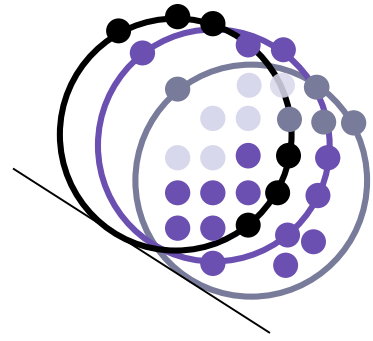
Петрозаводский государственный
университет



Крижановский Андрей Анатольевич

andrew.krizhanovsky@gmail.com¹



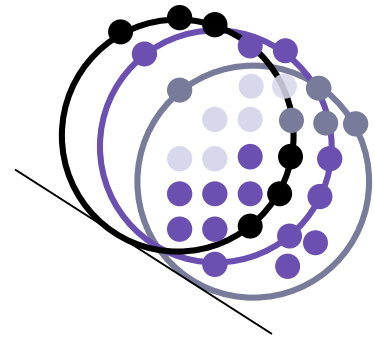


by JESS3

THE STATE OF WIKIPEDIA



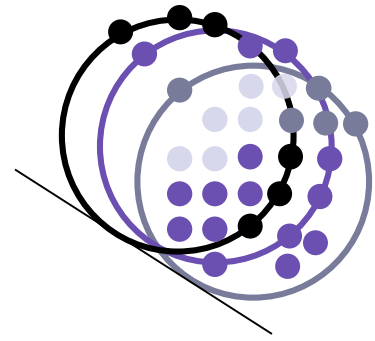
Логические операции



- Конъюнкция
- Дизъюнкция



Логические операции



- Конъюнкция
(лат. *conjunctio*)
логическое "И",
логическое умножение
- Таблица истинности

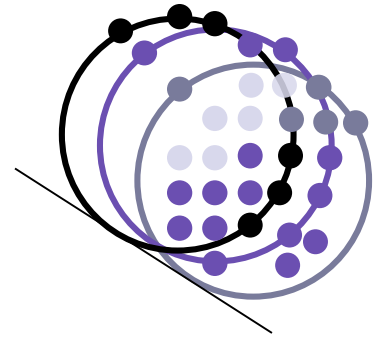
a	b	$a \wedge b$
0	0	0
0	1	0
1	0	0
1	1	1

- Дизъюнкция
(лат. *disjunctio*)
логическое «ИЛИ»,
включающее «ИЛИ»,
логическое сложение

a	b	$a \vee b$
0	0	0
0	1	1
1	0	1
1	1	1



Альтернативные модели (к булеву поиску)

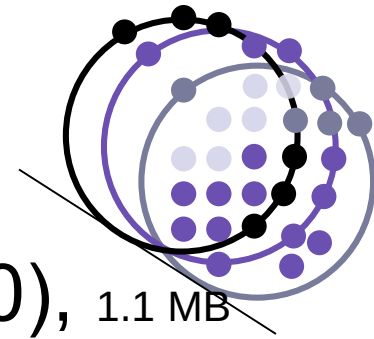


- Модель булева поиска



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

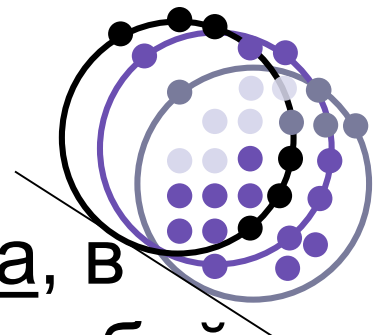
	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

110100 AND
110111 AND
NOT 010000

⇒

110100 AND
110111 AND
101111=
100100

Модель булева поиска



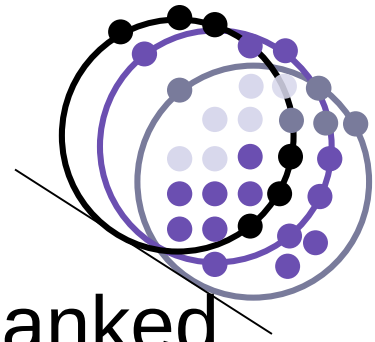
- это модель информационного поиска, в ходе которого можно обрабатывать любой запрос, имеющий вид булева выражения, т.е. выражения, в котором термины используются в сочетании с операциями AND, OR и NOT.

- представление документа
 - список всех слов





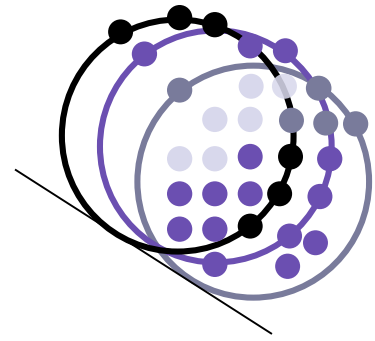
Альтернативные модели (к булевому поиску)



1. Модели поиска с ранжированием (ranked retrieval models)
 1. Модель векторного пространства
: *свободные текстовые запросы* (free text queries)
: *<s> булевы операторы </s>*
2. Расширенные модели булева поиска, доп. операторы:
 - ❖ **оператор близости** терминов (*proximity operator*)
 - ❖ Ex. : слова, предложения, абзацы



Proximity search



Поиск с учётом близости слов

Поиск с учётом соседства слов

**Поиск с учётом расстояния между
словами**

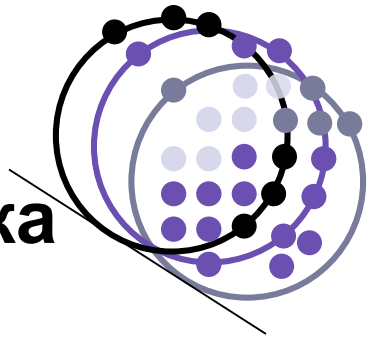
Метод поиска по близости

Поиск с расстоянием



Westlaw (1)

Коммерческая служба булева поиска

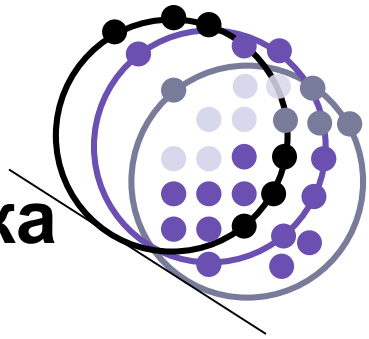


- *Информационные потребности:*
юр. теории, связанные с
предотвращением раскрытия коммерч.
тайны уволенными сотрудниками,
перешедшими на службу в конкурирующие
компании
- *Запрос:*
“trade secret” /s disclos! /s prevent /s employee!



Westlaw (2)

Коммерческая служба булева поиска

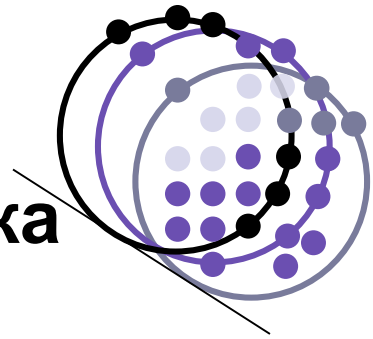


- *Информационные потребности:*
дела, касающиеся ответственности хозяев
за поведение пьяных гостей.
- *Запрос:*
host! /p (reponsib! liab!) /p (intoxicat! drunk!) /p guest



Westlaw (3)

Коммерческая служба булева поиска



- *Запросы:*

“trade secret” /s disclos! /s prevent /s employee!

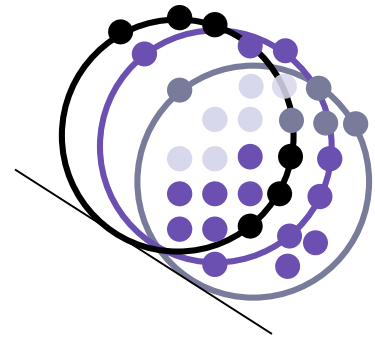
host! /p (reponsib! **liab!**) /p (intoxicat! drunk!) /p guest

- *Язык:*

- пробел = дизъюнкция (^Веб)
- /s /p /к поиск совпадений в одном и том же предложении, абзаце или в окрестности *k* слов
- фразовый поиск (“trade secret”)
- замыкающий джокер (!)
 - liab! = liable || liability || liabilities



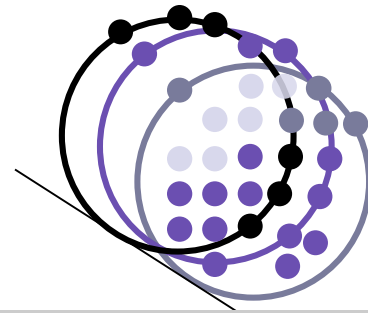
Проблемы булевых запросов



- оператор AND
 - ↗ точность запроса (P, Precision)
 - ↘ полноту поиска (Q, Recall)
- оператор OR
 - ↘ ТОЧНОСТЬ
 - ↗ полноту поиска,
- ☹️ трудно или невозможно найти компромисс

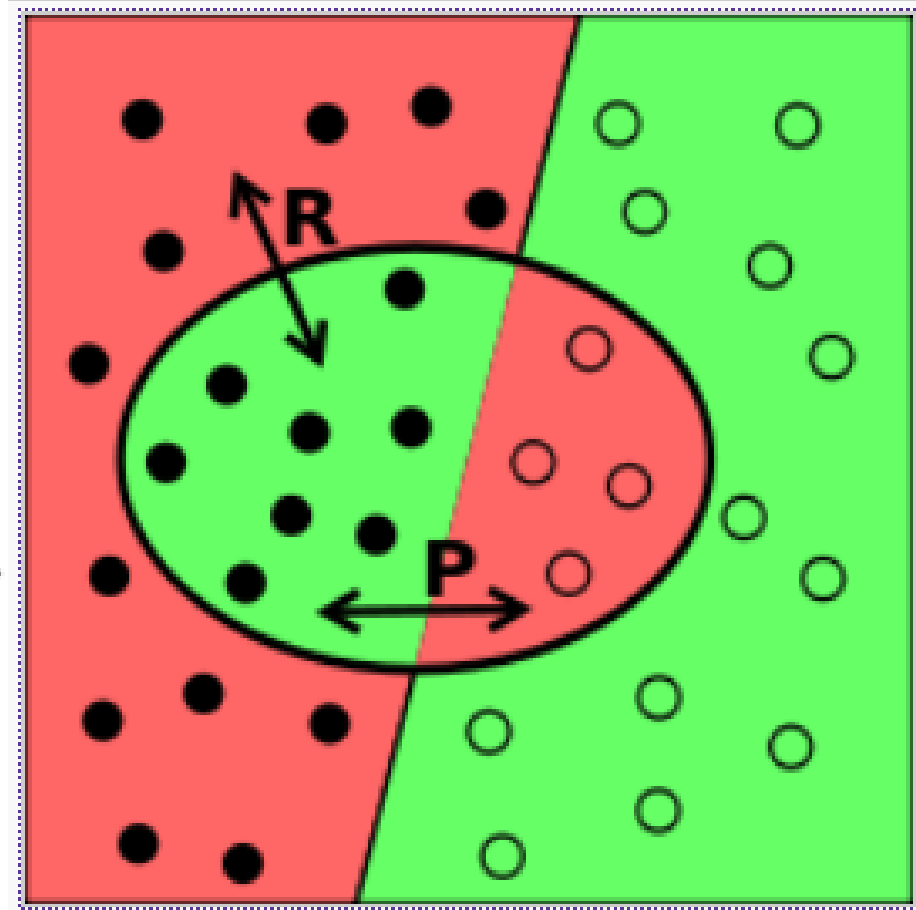


Метрики



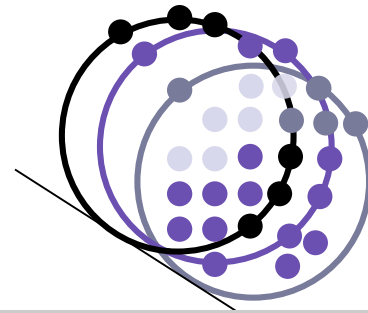
$$\text{Precision} = \frac{|D_{rel} \cap D_{retr}|}{|D_{retr}|}$$

$$\text{Recall} = \frac{|D_{rel} \cap D_{retr}|}{|D_{rel}|}$$

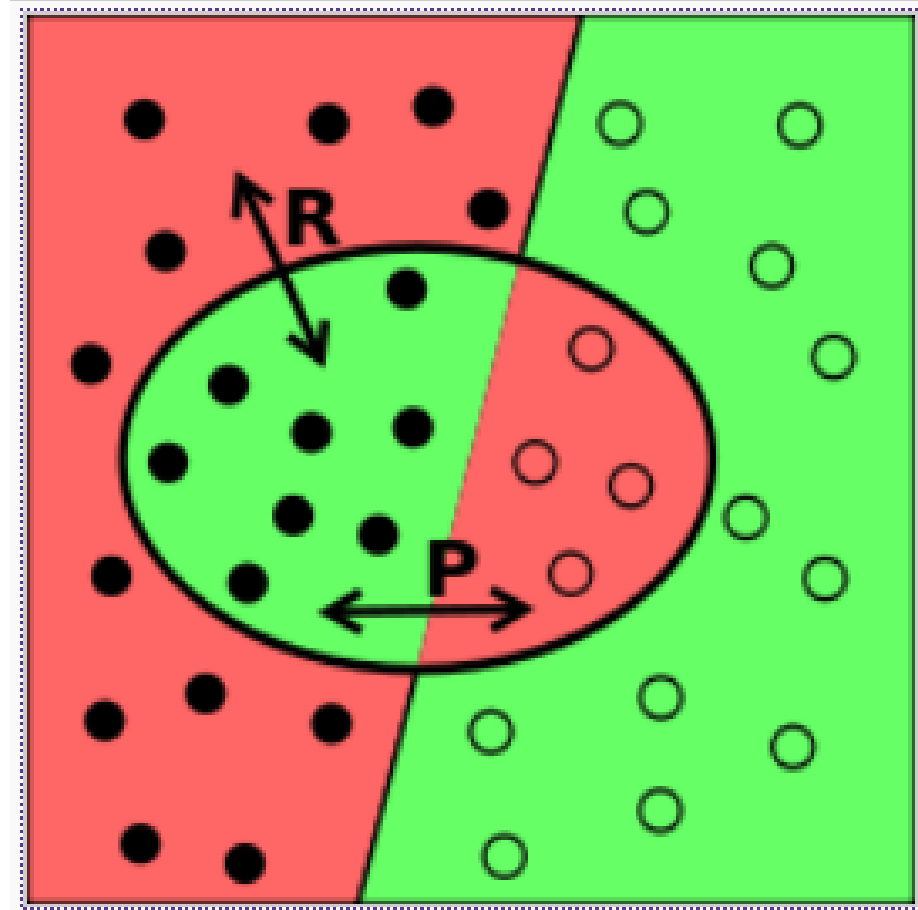




Проблемы булевых запросов



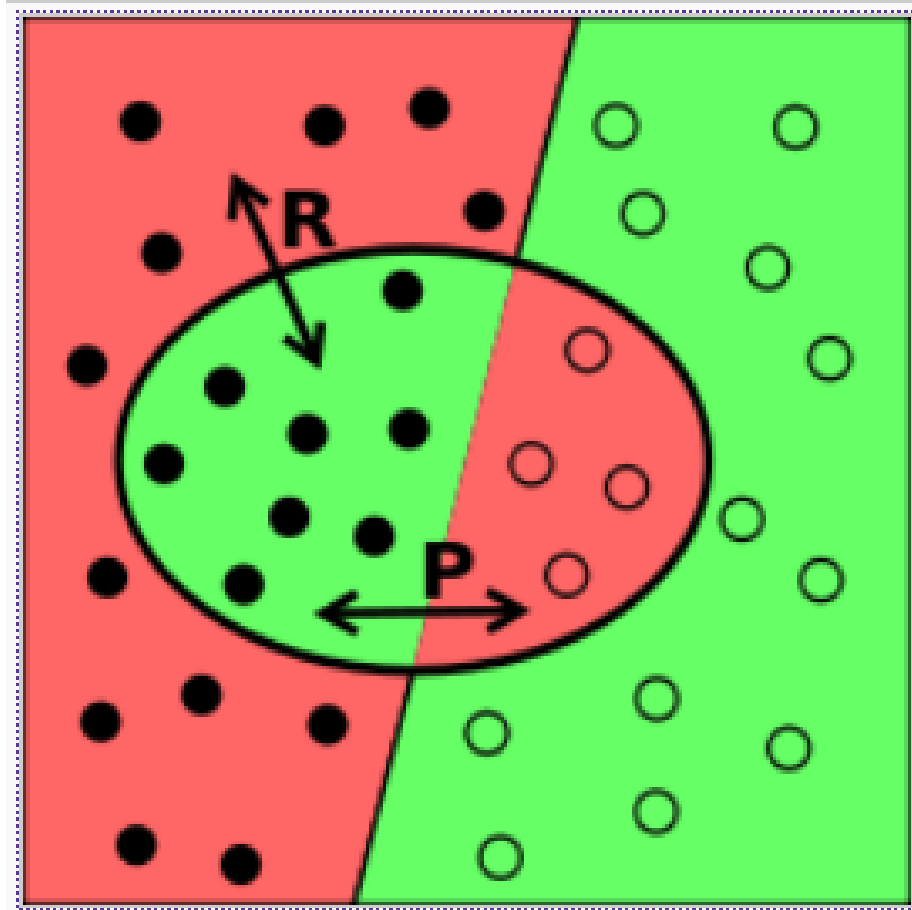
- оператор AND
 - ↗ точность запроса (P, Precision)
 - ↘ полноту поиска (R, Recall)
- оператор OR
 - ↘ ТОЧНОСТЬ
 - ↗ полноту поиска,
- трудно или невозможно найти компромисс





Выпадение (fall-out)

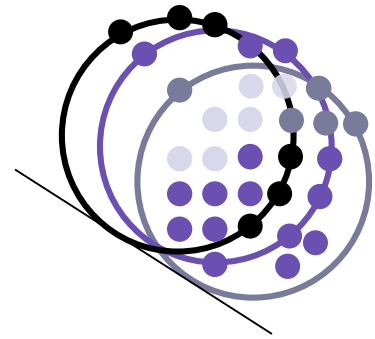
- вероятность нахождения нерелевантного документа



$$\text{Fall-out} = \frac{|D_{nrel} \cap D_{retr}|}{|D_{nrel}|}$$



Альтернативные модели (к булеву поиску)

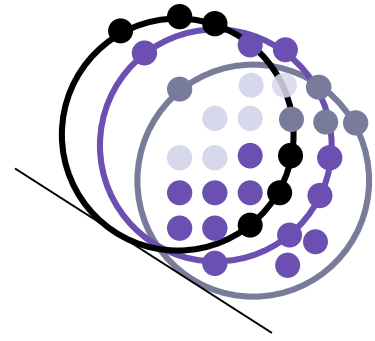


Расширенные модели булева поиска,
дополнительные операторы:

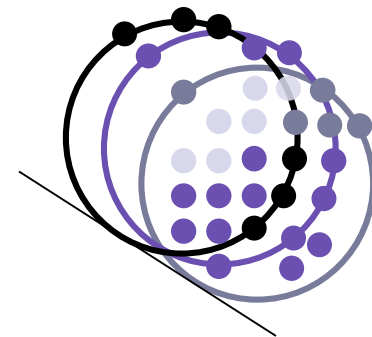
- ❖ **оператор близости** терминов (*proximity operator*)
- ❖ Найти похожие (Query-By-Example, find similar)
- ❖ Поиск в определенных полях (Search by Field)
 - ❖ Title, author... (расширенный поиск)



Cheat sheet

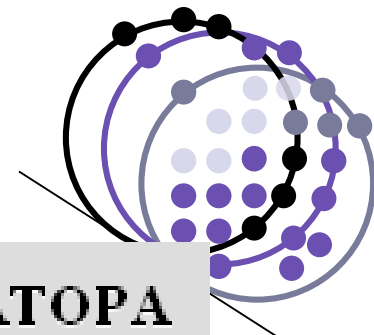


- Spanish:
 - machete (es) m,
 - chuleta (es) f,
 - apunte escondido





Запросы в Google



biking Italy

recycle steel OR iron

"I have a dream"

salsa –dance

Louis "I" France

castle ~glossary

fortune-telling

define:imbroglio

ПРИМЕР ОПЕРАТОРА

отдых Сочи

Сочи **OR** Анапа

"Я помню чудное мгновенье"

вирус –компьютерный

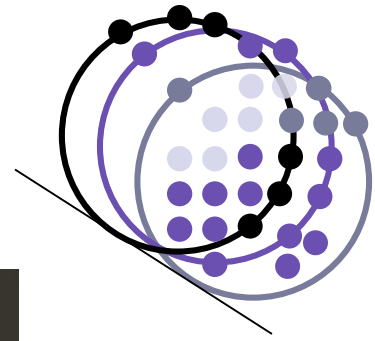
Алексей **+**II

~авто кредит

define:интернет

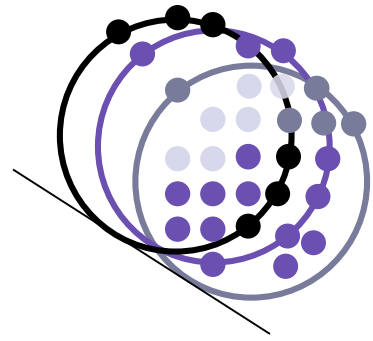
президент * федерации

Мне повезёт!





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>
- Interactive online Google tutorial and references - Google Guide: http://www.googleguide.com/advanced_operators_reference.html
- Растолкованный язык запросов Google <http://adresator.org/cheatsheet.html>

Спасибо за внимание!



http://vk.com/imath_petrSU
“Интернет-математика в ПетрГУ”

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**



Рецензенты без работы?

