«Теория и практика информационно-аналитической работы»

Семинар 5

2018

Добыча данных в массивах неструктурированной информации инструментами лексического поиска

- Добыча данных (data mining)
- Неструктурированной информации
- Лексический поиск

Добыча данных

- **Добыча данных (data mining)** – это нахождение в тексте (фотографии, видеосюжете) элементов информации, о которых мы говорили на первых семинарах:
  - **Фактов (и их взаимоотношений, которые сами по себе отдельный факт)**
  - **Мнений и суждений**
  - **Авторских характеристик**
  - **Обладателей компетенций**
  - **Дискурса**

  С другой стороны, добыча данных – это еще и **отнесение текста** целиком **к какой-то группе** (например, по признаку тональности)

  Это **две разные** задачи, но обе – добыча данных

Добыча данных – основной алгоритм для систем лексического поиска[1]

Основной алгоритм, применяемый при лексическом поиске (и не только в этом наборе инструментов) – мы выделяем **массив «контейнеров»**, в котором требуем наличия **лексем**, связанных между собой **буллевской логикой**. Перед этим строим **рабочую гипотезу**, что именно эти лексемы именно в этих связях делают появление нужных нам данных в массиве более **вероятным**.

Добыча данных – основной алгоритм для систем лексического поиска2

- **массив «контейнеров»**,
- **лексем**, связанных между собой **буллевской логикой**
- **рабочая гипотеза**, что именно эти лексемы именно в этих связях делают появление нужных нам данных в массиве
- более **вероятным**.

Добыча данных – основной алгоритм для систем лексического поиска3

- **массив «контейнеров» + «более вероятным»:**

То есть в результате поиска мы получаем не данные (!), а некоторый массив публикаций, фотографий, документов и телесюжетов, где внутри эти данные содержатся с большей вероятностью, чем если бы мы просто читали тексты случайным образом.

А дальше – возможности поиска заканчиваются, включаются глазки и мозг – смотреть и выбирать нужное.

Добыча данных – основной алгоритм для систем лексического поиска[4]

- **лексем…. булевой логикой**

Алгебра логики (булева алгебра) — это раздел математики, изучающий высказывания, рассматриваемые со стороны их логических значений (истинности или ложности) и логических операций над ними. Алгебра логики позволяет закодировать любые утверждения, а затем манипулировать ими подобно обычным числам в математике.

В нашем случае роль переменных выполняют не числа, а лексемы (слова и

Добыча данных – основной алгоритм для систем лексического поиска[5]

- **булевой логикой**

Булева алгебра названа по имени великого английского математика Джорджа Буля, который в 1854 г. опубликовал ставшую впоследствии знаменитой книгу «Исследование законов мышления». В начале гл. 1 он написал: *«Назначение настоящего трактата — исследовать основные законы тех операций ума, посредством которых производится рассуждение; выразить их на символическом языке некоторого исчисления»*

То есть Буль за полтора столетия до компьютеров решал компьютерную задачу – **применить математический аппарат к процессу рассуждений и умозаключений**.

Добыча данных – основной алгоритм для систем лексического поиска[6]

- **основа булевой логики - логические операторы.**

«СловоА вместе со словомБ и все это вместе на расстоянии семи слов от словаВ, которое, в свою очередь, на расстоянии пяти слов от словаВ рядом со словомГ. Все это вместе – только в том случае, если на расстоянии пяти слов нет словаД, но если рядом со словомД есть словоЕ с любой стороны, то можно».

Разумеется, так никто не пишет, хотя тоже можно – есть **языки** и **операторы**.

Добыча данных – основной алгоритм для систем лексического поиска[7]

- **Пример записи поискового выражения**

(законопроект | (проект /3 закона) && (((внесен | «на рассмотрении») /10 (госдума | «ГД РФ» | (совет /2 депутатов) | закс | заксобрание | «законодательное собрание» | совфед | «совет федерации»)) | (подписал /5 президент)) | (отзыв /5 (минюста | правительства)) | ((первое | второе | третье) /2 чтение))

Это один из простейших профессиональных запросов на изменения в законодательстве – укороченный под поисковую систему, допускающую только 450 знаков (в конкретно – Яндекс)

Добыча данных – основной алгоритм для систем лексического поиска8

- **Пример записи поискового выражения - продолжение**

(Лексика (операторы) не имеют значения, этот язык у каждой поисковой системы свой.

Но:

Самый сложный запрос переводится с языка одной поисковой системы на язык другой, - разумеется, с учетом ограничений конкретной системы.

Запрос остается тот же, меняется только внешний вид и значки.

Запрос – не черта поисковой системы, он универсален, в этом смысл булевой логики.

Добыча данных – основной алгоритм для систем лексического поиска[9]

- **Рабочая гипотеза**

Вы ищете некоторые новые данные, но их искать невозможно – для этого нужен ваш мозг и опыт. Поэтому вы **предполагаете**, что если в тексте есть определенные слова и словосочетания в определенных отношениях, то весь этот текст – про что-то новое в законотворчестве.

Это предположение в деталях – и есть рабочая гипотеза. Ее реализация на практике – **поисковый запрос**.

Результат применения поискового запроса – массив контейнеров.

Добыча данных – основной алгоритм для систем лексического поиска10

- **Рабочая гипотеза - продолжение**

Результат применения поискового запроса – массив контейнеров. Какой?

- ✔ В нем есть **большая часть** документов массива, в которых говорится про законотворчество
- ✔ В нем **по возможности меньше** документов, в которых про законотворчество не говорится.

Любой поисковый запрос (рабочая гипотеза) – **баланс** между **полнотой**, с одной стороны, и захватом **ненужной информации**, с другой.

Абсолютно точных запросов не бывает – мы не в теории, мы инженерная дисциплина.

Лексический профессиональный поиск или «естественный язык»1

Пример:

(законопроект | (проект /3 закона) && (((внесен | «на рассмотрении») /10 (госдума | «ГД РФ» | (совет /2 депутатов) | закс | заксобрание | «законодательное собрание» | совфед | «совет федерации»)) | (подписал /5 президент)) | (отзыв /5 (минюста | правительства)) | ((первое | второе | третье) /2 чтение))

Или:

«Новое в законотворчестве»

И то, и другое работает в одной и той же поисковой системе!

Лексический профессиональный поиск или «естественный язык»2

Естественный язык:

1. Хорошо отрабатывает бытовые потребности: найти товар, человека, узнать ключевые новости
2. Современные системы умеют думать за нас:

   - выделяют темы и сюжеты, отсекают дубли, запоминают что мы искали ранее…
3. Не требует квалификации в написании запроса, не требует оптимизации и шлифовки поиска

Профессионалами не используется – потому что никогда нет возможности **понять, что тебе показали, а что нет, и по какому закону прошел этот отбор**.

На этом построена, в частности, вся скрытая интернет-реклама.

Лексический профессиональный поиск или «естественный язык»3

Лексический профессиональный поиск:

1. Ты всегда интуитивно понимаешь, до чего дотянулся, а от чего отказался
2. Можно настроить размер выдачи под выделенные ресурсы – сто документов или тысячу
3. В процессе отладки рабочей гипотезы (запроса) формируется аналитическая гипотеза

Дилетантами не используется – потому что слишкоммногобуков.

Схема действий подготовки рабочей гипотезы и запроса

1. Самый простой запрос – читаем все подряд, примерно 50-100 документов
2. Выделение лексем (предметной области) – уникальные слова, фразы плюс подходящие слова, фразы; но минус явно лишние слова, фразы.
3. Описание логики поиска (создание языковой модели) – какие сочетания и пересечения слов (фраз) использовать, на каком расстоянии.
4. Перевод запроса на технический язык нужной системы.

Если работаем в одной поисковой системе – можно п.3 и п.4 объединить

5. Проверка запроса в системе – насколько полученные тексты (выборка) соответствуют вашим ресурсам.
6. Проверка уровня информационного шлака
7. Корректировка запроса, если необходимо (назад к пункту 1 и повтор всего цикла).

Тренируемся в подготовке рабочей гипотезы1

Читаем подряд слова на запрос «**Производство автомобилей КАМАЗ»** и отбираем первый набор лексем, которые описывают понятие (тему):

Надо искать слово КАМАЗ рядом со словами:

- **производство**
- **Завод, предприятие, холдинг**
- **Конвейер**
- **Продукция …**
- **ОАО**
- **Компания**
- **Сергей Когогин (владелец)**
- **Директор, гендиректор**
- **Рабочий, забастовка,**
- **Профсоюз, профсоюзный**
- **…**

Тренируемся в описании понятий 2

Что мы не учли?

- Значительный инф. шум от прочих у поминаний КАМАЗа.

Как его уменьшить?

1. Задать **жесткие условия** на употребление ключевых слов, например:

«завод КАМАЗ», «конвейер КАМАЗа», «директор КАМАЗа»…

2. Исключить из получаемых текстов «бытовые» упоминания, а это значит необходимо…

Тренируемся в описании понятий 3

Описать новое понятие:

«Бытовые» упоминания автомобиля КАМАЗ»

Нам не надо искать слово КАМАЗ рядом со словами:

- **ДТП**
- **«Дорожно-транспортное происшествие»**
- **ГИБДД**
- **Сбил, Наезд, наехал**
- **Водитель, шофер**
- **Угон, угонять, угонщик**
- **ПДД, «правила дорожного движения»**
- **…**

Тренируемся в описании понятий 4

Что еще можно сделать?

- **Расширить запрос за счет неявных, но эффективных смысловых ключей.**

КАМАЗ – крупнейшее предприятие

- Визиты **Путина, Медведева**, лоббирование в **Госдуме, Совете Федерации, министерствах**.

- Уменьшить объем выборки за счет выкидывания текстов про **футбол, ралли (но это пиар)**.

- **Можно и далее продолжать совершенствовать запрос в этом ключе, - до уровня, который нас удовлетворит**

Тренируемся в описании логики 1

Вариант 1.

Мы ищем в текстах слово КАМАЗ, находящееся в одном **предложении** с любым из следующих слов:

Завод, производство, предприятие, холдинг, директор, гендиректор, Когогин, ОАО, компания, конвейер, продукция, профсоюз, рабочий…

Вариант 2.

Тоже, что вариант 1, но не в одном предложении, а еще ближе – например, **не далее 2 слов** друг от друга (более строгое условие).

Тренируемся в описании логики 2

Вариант 3

Берем вариант 1 или вариант 2 и добавляем к нему условие:

Нам НЕ НУЖНЫ тексты, где слово КАМАЗ встречается рядом (например, в одном предложении) с любым из слов: ГИБДД, ДТП, сбил, наезд, авария, врезался, наехал, водитель, угонщик, угонять, ПДД…

Составляем запрос 1

Какие бывают операторы?

1. **«И» – пересечение - ВСЕ ключи, соединенные через этот оператор должны быть в текстах.**
2. **«ИЛИ» - объединение – любой из ключей, соединенных через этот оператор может быть в тексте**
3. **«НЕ» – отрицание – любой из ключей после «НЕ» не должен быть в тексте.**
4. **Логические скобки и расстояния между словами, число ключей в предложении и тексте, ключи в одном предложении или абзаце и т.д.**

Переводим запрос на язык системы Яндекс

Вариант 1

КАМАЗ **&** **(**Завод | производство | предприятие | холдинг | директор | гендиректор | Когогин | ОАО | компания | конвейер | продукция | профсоюз | рабочий**)**

Красным выделены все операторы. Скобки – тоже оператор.

Переводим запрос на язык системы Яндекс

Вариант 2

КАМАЗ **/2 (**Завод | производство | предприятие | холдинг | директор | гендиректор | Когогин | ОАО | компания | конвейер | продукция | профсоюз | рабочий**)**

Переводим запрос на язык системы Яндекс

Вариант 3

(КАМАЗ **& (**Завод **|** производство **|** предприятие **|** холдинг **|** директор **|** гендиректор **|** Когогин **|** ОАО **|** компания **|** конвейер **|** продукция **|** профсоюз **|** рабочий**))** **~~ (**КАМАЗ **& (**ГИБДД **|** ДТП **|** сбил **|** наезд **|** наехал **|** авария **|** врезался **|** водитель **|** угонщик **|** угонять **|** ПДД**))**

.

Как корректировать запрос? 1

Переписывать не надо!

1. Работаем по смысловым ключам (убираем-меняем-добавляем слово, словосочетание или группу – смотрим на результат)
2. Изменяем расстояния между словами (50 слов, предложение, 100 слов и более; тонкая настройка: 3 – 10 слов).
3. Используем скобки и строки (как при программировании), чтобы запрос был понятен не только вам, но и тому, кто его увидит впервые.

Как корректировать запрос? 2

Как ответить на вопрос когда остановиться в совершенствовании запроса?

Цель поиска: – максимально точная и полная выборка

Но, чем полнее, тем меньше точность и наоборот.

Часто используют правило 20Х80:

Если на 100 текстов выборки – 80 соответствуют задаче – результат достаточный.

В потоковой (регулярной) работе можно выбирать более строгие критерии (например, лишних – не более 5%).

Поиск в Яндекс – повтор-памятка

Оператор	Описание, особенности
Операции с отдельными словами	
нет оператора	ищет слово с учетом морфологии (день - деть)
_!!	задает поиск слова в его словарных формах (день – ~~деть~~) *«_» здесь и далее обозначает пробел!*
_!	учитывает регистр слова (День или день) и отключает морфологию!
Операторы объединения и пересечения выборок	
_|_	«ИЛИ»
_&&_	в одном документе
Пробел	«нежесткое И» ищет слова сначала близко, потом далеко, потом ооочень далеко (в разных текстах - не использовать в строгих запросах!)
_+	обязательное присутствие слова в запросе имеет смысл в сочетании с пробелом или если ищем «стоп-слово».

Поиск в Яндекс – повтор-памятка

Оператор	Описание, особенности
Поиск с учетом контекста (окружения)	
_&_	в одном предложении
_&&/N_	поиск на расстоянии не более N предложений
_/N_	поиск на расстоянии не более N-1 слов
_/+N_	поиск на расстоянии **СТРОГО** N-1 слов **строго по порядку** слева направо
_/-N_	поиск на расстоянии **СТРОГО** N-1 слов строго **в обратном порядке**
_/(N +M)	**поиск на расстоянии от N до M слов. Причем числа могут быть отрицательными. СОБЛЮДАЕТСЯ ПОРЯДОК СЛОВ!**
«»	поиск **строгих форм** слова (морфология отключается!), в **строгом порядке**
«слово_*_ слово»	замена слов **внутри кавычек**
Операторы отрицания (исключения)	
_~_	одиночная тильда – слов не должно быть **в предложении**
_~~_ **_-**	двойная тильда – слов не должно быть **в тексте** **одиночное тире – аналог (но работает не всегда корректно)**

Поиск в Яндекс – повтор-памятка

Оператор	Описание, особенности
Операторы ограничения региона поиска	
title:	поиск в заголовке
url:	поиск по заданному адресу
inurl:	поиск по фрагменту адреса
site:	поиск по всем поддоменам
domain:	поиск по домену
lang:	поиск по языку
mime:	поиск по типам файлов
date: **date:<**	поиск по дате: YYYYMMDD вместо дней или месяцев можно ставить * поиск до даты, если вместо < поставить знак > - поиск после даты **date:**YYYYMMDD..YYYYMMDD - поиск по диапазону дат – сначала старые

Внимание:

Длина запроса в Яндекс.Новостях – не более 400 символов с пробелами