СПБГУ

Кафедра системного программирования

Отчет по курсовой работе

«Повышение прозрачности сайта госзакупок РФ»

Студент: Коноплев Юрий 445гр.

Научный руководитель: кандидат физ-мат. наук Сергей Сысоев

2012г.

Содержание:

Введение.

В условиях постоянно растущего объема накапливаемой и используемой человечеством информации остро встают проблемы организации информации и информационного поиска.

Сайт госзакупок РФ (http://zakupki.gov.ru) имеет одной из целей своего создания повышение прозрачности процесса проведения аукционов. Но, к сожалению, сам сайт не обладает функциональностью для обработки больших объёмов данных (несколько сотен тысяч договоров).

- Нет удобного способа представления информации.
- Отсутствие удобного поиска по имеющимся данным.

Исходя из этого, появляется задача исправления недостатков.

Предоставление пользователю самостоятельно составлять отчёты по интересующим его показателям.

Основная задача, которая встаёт перед нами – это классификация договоров. Договор – это текст, а значит, мы имеем дело с классификацией текстовых документов.

Большинство методов автоматической классификации текстов, так или иначе, основаны на предположении, что тексты каждой тематической рубрики содержат отличительные признаки (слова или словосочетания) и наличие или отсутствие таких признаков в тексте говорит о принадлежности или непринадлежности исследуемого текста той или иной категории. Задача методов классификации состоит в том, что бы наилучшим образом выбрать такие отличительные признаки и сформулировать правила, на основе которых будет приниматься решение об отнесении текста к категории.

Задача заключается именно в автоматической классификации. Классификация происходит без помощи специально обученных экспертов, которые могут определить, к какому разделу относится данный документ.

1.Синтаксический анализатор XML файлов

Все данные с сайта находятся в свободном доступе на ftp сервере в формате XML файлов.

Из этого вытекает первая задача: парсинг этих файлов.

Существует несколько классификаций парсеров:

- Парсеры с проверкой и без проверки
- Парсеры, поддерживающие один или несколько языков описания схем XML
- Парсеры, поддерживающие объектную модель документа (DOM)
- Парсеры, поддерживающие простой API для XML (SAX)

Парсеры с проверкой (проверяющие парсеры) проверяют XML-документы в процессе их анализа, в то время как парсеры без проверки (непроверяющий парсер) этого не делают. Другими словами, если XML-документ составлен грамотно, парсер без проверки не обращает внимание, соответствует ли он правилам, заданным в DTD или в схеме, или существуют ли для этого документа вообще какие-либо правила.

В нашем случае парсер с проверкой не используется по двум причинам:

1. Большой размер входящих данных
2. XML-документы можно считать корректными

Анализируя XML-документ с помощью парсера DOM, вы получаете *древовидную структуру*, представляющую содержание XML-документа. Весь текст, элементы и символы включены в эту структуру. Кроме того, DOM предоставляет разнообразные функции, которые вы можете использовать для изучения и для работы с содержанием и структурой дерева.

Для анализа полученных с ftp сервера файлов мною был использован DOM-модель парсера, реализованная на языке программирования Java.

2.База данных

Для хранения информации о контрактах, полученной после анализа XML-документов, используется реляционная модель базы данных MySQL. (см рис.1)

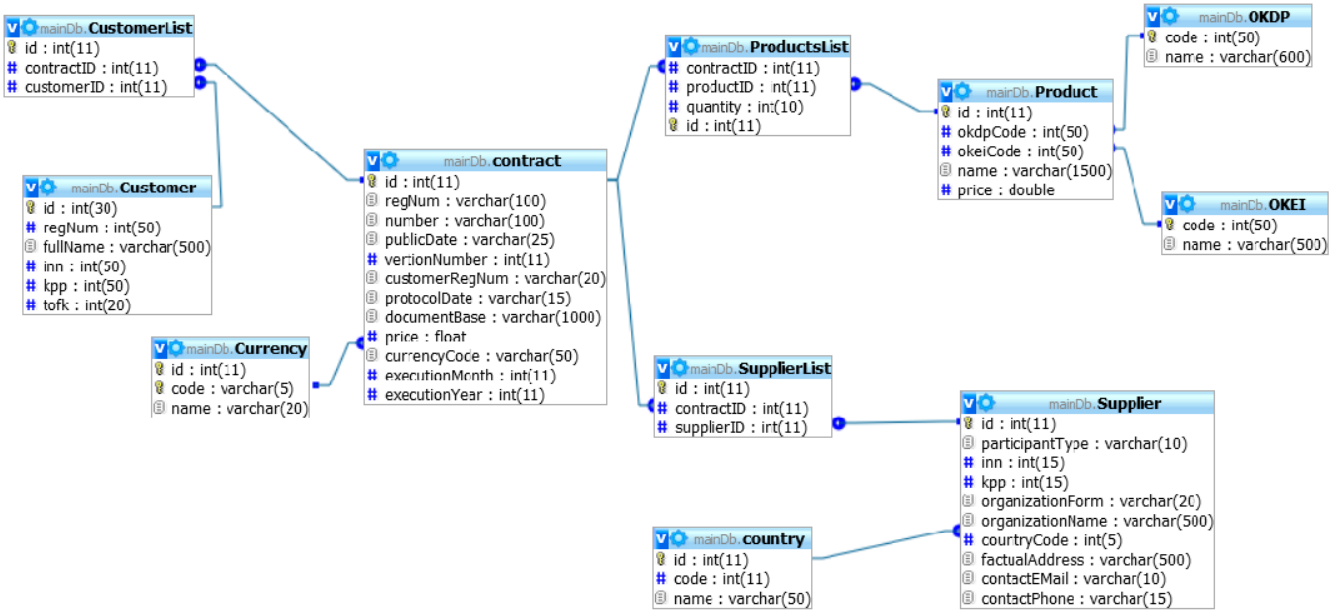

Рис.1 Схема БД

3.Общая постановка задачи классификации.

Имеется множество категорий (классов, меток) $\mathfrak{C} = \{c_1, ..., c_{|\mathfrak{C}|}\}$.

Имеется множество документов $\mathfrak{D} = \{d_1, ..., d_{|\mathfrak{D}|}\}$.

Неизвестная целевая функция $\Phi : \mathfrak{C} \times \mathfrak{D} \to \{0,1\}$.

Необходимо построить классификатор Φ', максимально близкий к Φ.

Имеется некоторая начальная коллекция размеченных документов $\mathfrak{R} \subset \mathfrak{C} \times \mathfrak{D}$, для которых известны значения Φ. Обычно её делят на «обучающую» и «проверочную» части. Первая используется для обучения классификатора, вторая — для независимой проверки качества его работы.

Классификатор может выдавать точный ответ $\Phi' : \mathfrak{C} \times \mathfrak{D} \to \{0,1\}$ или степень подобия $\Phi' : \mathfrak{C} \times \mathfrak{D} \to [0,1]$.

Различают два вида классификаторов: с учителем и без. Второй тип – это кластеризация. В данной работе была использована классификация с учителем.

Методы классификации с учителем полагаются на наличие коллекции Ω= {d1, . . . , d|Ω|} заранее отклассифицированных документов, то есть таких, для которых уже точно известно значение целевой функции. Для того, чтобы после построения классификатора можно было оценить его эффективность, Ω разбивается на две части, не обязательно равного размера:

1. Учебная (training set) коллекция. Классификатор строится на основании характеристик этих документов.

2. Тестовая (test set) коллекция. На ней проверяется качество классификации. Документы их этой коллекции не должны участвовать в процессе построения классификатора.

4.Метод Байеса

Как и все статистические методы классификации, метод Байеса заключается в вычислении вероятностей сопоставления документа каждой из рубрик и выборе рубрики, вероятность для которой будет максимальной. Апостериорная вероятность принадлежности документа *d* рубрике *ci* по теореме Байеса вычисляется так:

$$P(c_i \mid d) = P(c_i \mid x_1 = d_1 \wedge x_2 = d_2 \wedge \ldots \wedge x_n = d_n) =$$
$$= \frac{P(x_1 = d_1 \wedge x_2 = d_2 \wedge \ldots \wedge x_n = d_n \mid c_i) P(c_i)}{\sum_{c' \in C} P(x_1 = d_1 \wedge x_2 = d_2 \wedge \ldots \wedge x_n = d_n \mid c') P(c')} ,$$

где *x j* и *d j* -- признаки и их значения соответственно, в нашем случае признак *x j* соответствует *j*-му слову из словаря и принимает значение 1 если слово присутствует в рассматриваемом тексте, иначе принимает значение 0.

Делая предположение о том, что переменные *x* независимы получаем:

$$P(c_i \mid d) = \frac{P(c_i)\prod_{j=1}^{n} P(x_j = d_j \mid c_i)}{\sum_{c' \in C} P(c')\prod_{j=1}^{n} P(x_j = d_j \mid c')}$$

Далее задача состоит в нахождении оценок априорных вероятностей $P(ci)$ и $P(x\,j = d\,j \mid ci)$.

$$\hat{P}(c_i) = \frac{|c_i|}{|D|}$$

где ci – количество документов из обучающей выборки, которым приписана рубрика ci, D – количество документов в обучающей выборке.

$P(x\,j = d\,j \mid ci)$ можно оценить как отношение документов из рубрики ci, содержащих i-е слово к общему числу документов в рубрике ci:

$$\hat{P}(x_j = d' \mid c_i) = \frac{count(d : d_j = d' \wedge d \in c_i) + 1}{count(d : d \in c_i)},$$

в числителе добавлена единица для избегания нулевых вероятностей, *d'* в нашем случае принимает значения 0 и 1.

Запишем решающее правило для метода Байеса:

$$c^*(d) = \arg\max_{c_i \in C} P(c_i \mid d)$$

На практике существует два подхода к использованию метода Байеса для классификации:

1. для каждой рубрики в отдельности принимать решение относится документ к ней или нет – бинарная классификация. При этом множество рубрик C сокращается до двух – ci и ci, в которую входят все документы не вошедшие в ci.
2. вычислять $P(ci \mid d)$ для всех рубрик и выбирать те, для которых эта вероятность будет максимальной.

5.Итоги работы

На настоящий момент разработан пакет приложений, позволяющий:

1. Загружать данные с сайта госзакупок в реляционную БД под управлением MySQL.
2. Классифицировать данные контрактов по предмету контракта (обучение проводится пользователем).
3. Выводить список контрактов, удовлетворяющих критериям пользователя.

Из-за отсутствия в свободном доступе лингвистических материалов на русском языке, позволяющих производить обучение классификации, был отобран набор текстов контрактов, являющиеся обучающим множеством с классификацией проведённой вручную.

Список используемой литературы

[1] Fabrizio Sebastiani. Machine Learning in Automated Text Categorization http://nmis.isti.cnr.it/sebastiani/Publications/ACMCS02.pdf

[2] Steven Bird, Ewan Klein, Edward Loper Natural Language Processing with Python

[3] Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В."Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа" НПЦ «ИНТЕЛТЕК ПЛЮС»

[4] М.С. Агеев. Методы автоматической рубрикации текстов, основанных на машинном обучении и знаниях экспертов: Диссертация на соискание ученой степени к.ф.-м.н. – М.: МГУ, 2004

[5] Е.В. Дунаев А.А. Шелестов "Автоматическая рубрикация web-страниц в интернет-каталоге с иерархической структурой",Томский государственный университет систем управления и радиоэлектроники, кафедра автоматизированных систем управления

[6] *О.Г. Шевелев, А.В. Петраков «*КЛАССИФИКАЦИЯ ТЕКСТОВ С ПОМОЩЬЮ ДЕРЕВЬЕВ РЕШЕНИЙ И НЕЙРОННЫХ СЕТЕЙ ПРЯМОГО РАСПРОСТРАНЕНИЯ"