Интернет-журнал «Наукovedение» ISSN 2223-5167 <http://naukovedenie.ru/>

Том 7, №1 (2015) <http://naukovedenie.ru/index.php?p=vol7-1>

URL статьи: <http://naukovedenie.ru/PDF/71TVN115.pdf>

DOI: 10.15862/71TVN115 (<http://dx.doi.org/10.15862/71TVN115>)

УДК 004.85

Молчанов Артем Николаевич

ГКОУ ВПО «Академия Федеральной службы охраны Российской Федерации»

Россия, Орел

Сотрудник

E-mail: Tom_Silence@mail.ru

Скурнович Алексей Валентинович

ГКОУ ВПО «Академия Федеральной службы охраны Российской Федерации»

Россия, Орел

Доцент кафедры

Кандидат технических наук

E-mail: alexeymail2010@mail.ru

Алгоритм распознавания веб-страницы, оценивающий когерентность размещенного на ней текста

Аннотация. В настоящей статье приводится описание алгоритма распознавания веб-страницы, оценивающего когерентность размещенного на ней текста.

В статье обосновывается актуальность разработки такого алгоритма, раскрывается содержание основных процедур, приводятся практические результаты распознавания веб-страниц тестовой выборки.

Работа алгоритма строится на представлении объекта распознавания его математической моделью: в виде набора параметров текста веб-страницы, статистический анализ которых позволяет количественно оценить свойство когерентности текста и определить классовую принадлежность рассматриваемой веб-страницы.

Алгоритм предполагает последовательную реализацию двух этапов: этапа обучения классификаторов и этапа распознавания веб-страницы с помощью обученных классификаторов. В целях диверсификации риска ошибочного принятия решения в результате распознавания веб-страницы только одним обученным классификатором, классификация объекта распознавания в предлагаемом алгоритме осуществляется тремя классификаторами разного типа. Итоговое решение о классовой принадлежности объекта выносится на основе частных решений классификаторов по мажоритарному способу.

Основываясь на практических результатах исследования и наборе достоинств, которыми обладает алгоритм, авторы делают вывод о его пригодности для решения задач, связанных с разработкой эффективных систем автоматического поиска сетевой информации, способных устойчиво работать в условиях изменения элементов оформления текста на веб-страницах информационных порталов.

Ключевые слова: веб-страница; распознавание; математическая модель текста на естественном языке; параметры текста; когерентность; мажоритарный способ принятия решения; диверсификация риска.

Ссылка для цитирования этой статьи:

Молчанов А.Н., Скурнович А.В. Алгоритм распознавания веб-страницы, оценивающий когерентность размещенного на ней текста // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 7, №1 (2015)
<http://naukovedenie.ru/PDF/71TVN115.pdf> (доступ свободный). Загл. с экрана. Яз. рус., англ. DOI: 10.15862/71TVN115

Рост числа источников информации, распространяющих свой контент во всемирной сети Интернет, обусловил необходимость создания специальных инструментов для поиска такой информации.

Для решения данной задачи в современных поисковых системах сетевой информации используется, так называемые, программы-посредники, работа которых основана на применении шаблонов разбора веб-страниц информационного портала. Программа-посредник – это программа, которая с определенной периодичностью заходит на заданный пользователем информационный ресурс, и, переходя по гиперссылкам, находящимся на его веб-страницах, извлекает определенные фрагменты информации только с тех страниц портала, для которых в базе данных программ-посредников есть соответствующий шаблон разбора. При этом сами шаблоны разрабатываются на языке регулярных выражений [1].

Очевидными достоинствами такого метода поиска информации является точность и полнота извлекаемой из сети информации, относительная простота реализации процесса извлечения. Вместе с тем, опыт использования современных систем [2, 3], реализующих данный метод, выявил ряд их существенных недостатков. При большом количестве источников информации первичная настройка системы, осуществляемая в ручном режиме, превращается в достаточно трудоемкую по временным ресурсам задачу. Разработчики сайтов пользуются неограниченной свободой в выборе средств и способов верстки веб-страниц. При этом дизайн сайта, то есть оформление его страниц, пусть незначительно, но достаточно часто меняется. Изменение элементов оформления страниц сайта делает недоступным его контент для системы поиска информации. Для обеспечения стабильной работы системы ее администратору требуется вести мониторинг изменения дизайна сайтов и участвовать в постоянной перенастройке шаблонов разбора страниц.

Несовершенство известного метода послужило стимулом для разработки альтернативного подхода к поиску сетевой информации, основанного на решении задачи распознавания веб-страницы портала с помощью обученных классификаторов. В настоящей статье приводится описание алгоритма, который раскрывает содержание и основные этапы решения данной задачи.

Задача распознавания веб-страницы. Алфавит классов. Суть метода извлечения.

Задача распознавания веб-страницы заключается в определении ее принадлежности к одному из заранее заданных классов [4]. Такими классами в данном случае являются класс информационных и класс навигационных веб-страниц. В основе такого деления лежит следующее рассуждение. Любой информационный портал можно представить в виде совокупности взаимосвязанных веб-страниц двух типов: навигационных и информационных. Первый тип страниц предоставляет пользователю возможность быстрой и удобной навигации по структуре сайта, второй тип страниц – непосредственно саму информацию. Текст навигационных веб-страниц, в отличие от информационных, не содержит информацию об описываемых событиях или явлениях в полном виде, что делает его бесполезным для получения информации. Исходя из этого, концептуально суть предлагаемого метода извлечения можно сформулировать следующим образом: доступным программным средством извлечения, например офлайн-браузером, надлежит извлечь все страницы интересующего информационного портала, после чего из общего числа извлеченных удалить все навигационные веб-страницы.

Таким образом, процедура извлечения сводится к процедуре распознавания веб-страницы, эффективность которой зависит от адекватности выбранной математической модели текста и построенного на ее основе алгоритма распознавания.

*Математическая модель текста на естественном языке,
учитывающая свойство когерентности*

Распознавание веб-страницы осуществляется по результатам количественной оценки свойства когерентности [5, 6], размещенного на ней текста. Когерентность означает, что между отдельными фрагментами текста должны быть сохранены смысловые связи. Если для текста информационной веб-страницы это условие выполняется, то в случае с текстом навигационной веб-страницы – нет. Между отдельными фрагментами текста навигационной страницы наблюдается разрыв смысловых связей, что обусловлено особенностями его формирования: текст страницы формируется из начальных фрагментов текста нескольких информационных страниц.

Аналитическая запись модели текста представлена группой выражений 1 – 5.

Текст веб-страницы можно представить в виде множества слов, мощность которого равна их количеству

$$T = \{t_1, t_2, \dots, t_i, \dots, t_w\}; \quad w = |T| \quad (1)$$

где T – множество слов текста на ЕЯ, t_i – i -ое слово текста, w – количество слов в тексте.

В свою очередь, все множество слов текста можно представить в виде объединения k -подмножеств различных частей речи, при этом каждое слово текста может быть отнесено к одному из этих подмножеств

$$T = \bigcup_{j=1}^k C_j; \quad t_i \in C_j; \quad i = \overline{1, w}; \quad j = \overline{1, k}, \quad (2)$$

где C_j – подмножество слов j -ой части речи, k – количество частей речи в языке.

Для биективного отображения множества слов во множество частей речи X вводится функция $F(T)$. Результат работы этой функции есть вектор, i -ым элементом которого является количество i -ой части речи в описываемом тексте

$$F : T \rightarrow X, F(T) = (f_1(T), \dots, f_k(T)) = (x_1, \dots, x_k); \quad i = \overline{1, k}; i \in N, \quad (3)$$

где f_i – функция определения количества частей речи в тексте, x_i – количество i -ой части речи в тексте.

Некогерентный текст можно записать в виде объединения n -разностей двух множеств: множества слов когерентного текста и множества слов текста, удаляемого из когерентного в результате формирования текста навигационной веб-страницы

$$T_{\Sigma} = \bigcup_{i=1}^n (T_i^{\text{КТ}} \setminus T'_i), \quad (4)$$

где T_{Σ} – множество слов некогерентного текста (текста навигационной веб-страницы),

$T^{\text{КТ}}$ – множество слов когерентного текста (текста информационной веб-страницы),

T' – множество слов удаляемых из текста информационной веб-страницы при формировании текста навигационной веб-страницы,

n – количество информационных текстов, используемых для формирования текста навигационной веб-страницы.

Вариант описания модели текста при утрате им свойства когерентности представлен выражениями 5 и 6. Выражение 5 есть результат применения функции определения количества частей речи к выражению 4:

$$F(T_{\Sigma}) = F\left(\bigcup_{i=1}^n (T_i^{\text{КТ}} \setminus T'_i)\right) = \sum_{i=1}^n (F(T_i^{\text{КТ}} \setminus T'_i)) = \sum_{i=1}^n (F(T_i^{\text{КТ}}) \ominus F(T'_i)) = \sum_{i=1}^n ((x_{i1}^{\text{КТ}}, \dots, x_{ik}^{\text{КТ}}) \ominus (\Delta x_{i1}, \dots, \Delta x_{ik})) \quad (5)$$

$$\ominus : \bar{x} \ominus \bar{\Delta x} = (x_1, \dots, x_k) \ominus (\Delta x_1, \dots, \Delta x_k) = (x_1 - \Delta x_1, \dots, x_k - \Delta x_k)$$

где $x_{ik}^{\text{КТ}}$ – количество k -ой части речи в i -ом когерентном тексте, Δx_{ik} – количество k -ой части речи, удаляемое в i -ом когерентном тексте, $\ominus$ – операция поэлементного вычитания векторов.

Выражение 5 позволяет наглядно увидеть, что при утрате текстом свойства когерентности, численные значения его параметров будут меняться. Таким образом, сравнивая значения указанных параметров неизвестного текста с образцами параметров текстов, классовой принадлежности которых известна, можно количественно оценить его когерентность.

Для практического применения модели, задаваемой группой выражений 1–5, требуется конкретизация языка и количества используемых параметров. Проведенные практические исследования текстов публицистического стиля на английском языке, размещенных на веб-страницах различных зарубежных информационных порталов, показали, что для эффективного решения задачи распознавания необходимо и достаточно учитывать 14 параметров текста: (глаголы в форме *PresentSimple*, существительные, местоимения, наречия, союзы, глаголы в 3 лице единственном числе, глаголы, глаголы в форме *PastSimple*, основные определители существительных, отношение глаголов к существительным, отношение уникальных слов к общему количеству слов, *PastParticipleII*, личные местоимения). Предлагаемая математическая модель текста позволила разработать алгоритм распознавания веб-страницы.

Алгоритм распознавания веб-страницы

Алгоритм предполагает последовательную реализацию двух этапов: этапа обучения классификаторов и этапа распознавания веб-страницы с помощью обученных классификаторов.

Важно отметить, что на всех этапах алгоритма параметры текста с помощью компонентного анализа [7] преобразуются в признаки, в качестве которых выступают главные

компоненты ξ_i . Объем обучающей выборки N формируется исходя из заданных параметров точности и надежности интервальных оценок случайной величины. В качестве классификаторов для построения алгоритма используется классификатор Байеса (КБ), линейный дискриминантный анализ (ЛДА) и метод опорных векторов (МОВ) с радиальной базисной функцией Гаусса. Выбор в пользу перечисленных классификаторов был сделан на основе анализа их преимуществ и недостатков, а также с учетом возможного изменения статистических характеристик исходных данных обучающей выборки.

Содержание этапа обучения классификаторов раскрывает блок-схема на рисунке А.1 (приложении А), которая описывает отдельные процедуры этого процесса.

Результатами обучения классификаторов являются значения среднего $\bar{x}_i$ и среднеквадратического отклонения σ_i для параметра текста в каждом классе, значения коэффициентов для расчета главных компонент b_{ij} наборы признаков из состава признаков рабочего словаря, используемые классификаторами $\vec{\xi}_{КБ}, \vec{\xi}_{ЛДА}, \vec{\xi}_{МОВ}$ и коэффициенты значимости классификаторов $k_{\text{знач.}i}$:

$$k_{\text{знач.}i} = \frac{(1 - P_{\text{с.ош.}i})}{\sum_{i=1}^q (1 - P_{\text{с.ош.}i})}, \quad (6)$$

где $P_{\text{с.ош.}i}$ – вероятность средней ошибки классификации, допущенная i -ым классификатором на этапе обучения, q – количество классификаторов.

Важно заметить, что каждому классификатору соответствует свой собственный набор признаков из состава рабочего словаря $\vec{\xi}$, при котором он классифицирует объекты распознавания с минимальной средней ошибкой. Выбор данного набора осуществляется методом полного перебора всех возможных комбинаций признаков рабочего словаря.

Этап распознавания также проводится в несколько последовательных процедур, схематично представленных на рисунке 1. Это процедура вычисления признаков на основе выбранных параметров текста $\vec{S}^m$, процедура классификации и процедура принятия решения.

Рисунок 1. Обобщенная схема работы алгоритма на этапе распознавания (разработано авторами)

В целях диверсификации риска ошибочного принятия решения в результате распознавания веб-страницы только одним обученным классификатором, классификация объекта распознавания в предлагаемом алгоритме осуществляется тремя классификаторами разного типа. Каждый классификатор выносит частное решение о классовой принадлежности ($A_{\text{инф.}}$, $A_{\text{навиг.}}$) веб-страницы $m_{\text{анализ.}}$, присваивая соответствующее значение коэффициенту принадлежности α_i :

$$\alpha_i = \begin{cases} -1, & m_{\text{анализ.}} \in A_{\text{инф.}}, \\ +1 & m_{\text{анализ.}} \in A_{\text{навиг.}} \end{cases}, \quad (7)$$

Итоговое решение об отнесении объекта распознавания к определенному классу выносится исходя из значения мажоритарной суммы R , рассчитываемой с учетом весовых коэффициентов значимости используемых $k_{\text{знач. } i}$:

$$R = \sum_{i=1}^q \alpha_i k_{\text{знач. } i}, \quad (8)$$

где q – количество классификаторов.

Содержание отдельных процедур этапа распознавания раскрывает блок-схема, представленная на рисунке Б.1 (приложение Б).

Результаты распознавания веб-страниц на английском языке

Экспериментальные данные расчета вероятности обнаружения и вероятности ложной тревоги для отдельных классификаторов, используемых в алгоритме, и алгоритма распознавания в целом представлены на рисунках 2 и 3 (параметры точности и надежности интервальных оценок 0,01 и 0,95 соответственно).

Рисунок 2. Экспериментальные данные расчета вероятности обнаружения для отдельных классификаторов, используемых в алгоритме, и алгоритма распознавания в целом (разработано авторами)

Рисунок 3. Экспериментальные данные расчета вероятности ложной тревоги для отдельных классификаторов, используемых в алгоритме, и алгоритма распознавания в целом (разработано авторами)

Рабочие характеристики алгоритма распознавания при классификации веб-страниц отдельными классификаторами, используемыми в алгоритме, и при совместном применении данных классификаторов представлены на рисунке 4.

Рисунок 4. Рабочие характеристики алгоритма распознавания при классификации веб-страниц отдельными классификаторами, используемыми в алгоритме, и при совместном применении данных классификаторов (разработано авторами)

Анализ достоинств разработанного алгоритма

Достоинства алгоритма во многом определяются достоинствами используемой модели объекта распознавания и подходом к организации процедур классификации и принятия решения.

Модель текста, лежащая в основе алгоритма распознавания, применима ко всем языкам индоевропейской языковой группы. При этом адаптация алгоритма для распознавания веб-страниц на другом языке будет сводиться только к конкретизации количества анализируемых параметров текста. В известном смысле, это делает разработанный алгоритм универсальным (массовым), пригодным для решения широкого круга задач.

Используемая модель текста обладает более низкой размерностью в сравнении с размерностью большинства известных моделей представления текста на естественном языке. Это, с одной стороны, позволяет говорить о достаточном уровне ее обобщающей способности [8] в контексте ее использования для описания всей генеральной совокупности

объектов распознавания. С другой стороны – позволяет избежать ряда проблем, связанных с «проклятием размерности» [9, 10] и феноменом Хьюджа [11], негативно влияющих на показатели качества распознавания и затрудняющих обоснованное применение некоторых классификаторов [12]. Кроме того, низкая размерность модели позволяет снизить вычислительную сложность процедур, используемых в алгоритме. В данном случае все процедуры алгоритма вычисляются за полиномиальное время.

Подход к организации процедуры классификации, при котором различные классификаторы компенсируют недостатки друг друга, ожидаемо не позволил добиться максимальных показателей качества распознавания. В частности, вероятность обнаружения у МОВ оказалась несколько выше, чем у предлагаемого алгоритма (рисунок 4). Однако это обстоятельство можно считать соразмерной платой за возможность диверсифицировать средний риск принятия неправильного решения при классификации и повышение устойчивости работы алгоритма в условиях возможного изменения статистических характеристик обучающей выборки. В современных обзорах по методам распознавания [13] и алгоритмическим композициям [14] данный подход к классификации характеризуется как «перспективный», с которым исследователи связывают дальнейшее развитие теории машинного обучения.

Отметим, что разработанный алгоритм не использует объединение однотипных классификаторов, которые отличаются друг от друга только параметрами настройки (например, классификаторы семейства МОВ с различными вариантами функции ядра). В рамках алгоритма объединяются разнотипные классификаторы, отличающиеся друг от друга самим принципом классификации. Такой подход с одной стороны, позволяет значительно снизить зависимость результатов работы алгоритма от статистических характеристик обучающей выборки, но с другой – в ряде случаев вызывает необходимость преобразования исходного пространства параметров текста в пространство, которое по своим характеристикам отвечает требованиям, предъявляемым классификаторами к исходным данным. В частности, между параметрами в исходном пространстве параметров текста наблюдаются сильные корреляционные связи, что делает необоснованным их применение для классификации КБ. В алгоритме для устранения корреляции исходное пространство с помощью компонентного анализа преобразуется в пространство безразмерных независимых признаков – пространство главных компонент, что позволяет устранить указанную проблему.

Итоговое решение о классовой принадлежности объекта выносится на основе частных решений классификаторов по мажоритарному способу. Отметим, что в отличие от классического способа принятия решения по большинству голосов, голоса классификаторов не считаются равными. Для определения веса голоса в алгоритме рассчитываются коэффициенты значимости классификаторов, задаваемые выражением 6. При этом вес голоса классификатора тем больше, чем меньше ошибался классификатор на этапе обучения. Такая процедура позволяет, гарантировать, с одной стороны, что решение ни одного из классификаторов не будет проигнорировано, а, с другой – что итоговое решение будет вынесено в основном за счет более точных классификаторов.

Вывод:

Таким образом, основываясь на практических результатах исследования и наборе достоинств, которыми обладает описанный алгоритм, можно говорить о его пригодности для решения задач, связанных с разработкой эффективных систем автоматического поиска сетевой информации, способных устойчиво работать в условиях изменения элементов оформления текста на веб-страницах информационных порталов.

ЛИТЕРАТУРА

1. Фридл Дж. Регулярные выражения, 3-е издание. – Пер. с англ. – СПб.:Символ-Плюс, 2008. – 608 с., ил.
2. Голубев С.А., Толчеев Ю.К. Шаров Ю.Л. Опыт внедрения и использования информационно-поисковой системы ODB-Text в Совете Федерации Федерального Собрания РФ// Современные технологии в управлении и образовании – новые возможности и перспективы использования. Сборник научных трудов. ФГУП НИИ «Восход», МИРЭА. – М., 2001. – с.58–61.
3. Подсистема сбора сообщений с сайтов новостей сети Интернет/ А.М. Андреев, Д.В. Березкин, В.В. Морозов и др.// Труды №1 молодых ученых, аспирантов и студентов «Информатика и системы управления». – М.: Изд. МГТУ им. Н.Э. Баумана. – 2003. – с.409 – 410.
4. Тихонов В.И. Оптимальный прием сигналов. – М.: Радио и связь, 1983. – 320 с.
5. Робер де Богранд и Вольфганг Дресслер Введение в лингвистику текста. – М., 1981.
6. Дейк ван Т.А., Кинч В. Стратегии понимания связного текста // Новое в зарубежной лингвистике. – Вып. 23:Когнитивные аспекты языка. – М., 1988.
7. Юсупов Р.М. Статистические методы обработки результатов наблюдений. – М.: Министерство обороны СССР, 1984
8. Воронцов, К. В. Комбинаторная теория надёжности обучения по прецедентам: Дис. док.физ.-мат. наук: 05-13-17. – Вычислительный центр РАН, 2010. – 271 с. (<http://www.machinelearning.ru/wiki/images/b/b6/Voron10doct.pdf>)
9. Richard Ernest Bellman; Rand Corporation (1957). Dynamic programming. Princeton University Press. ISBN 978-0-691-07951-6., Republished: Richard Ernest Bellman (2003). Dynamic Programming. Courier Dover Publications. ISBN 978-0-486-42809-3.
10. Richard Ernest Bellman (1961). Adaptive control processes: a guided tour. PrincetonUniversityPress.
11. Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers". IEEE Transactions on Information Theory 14 (1): 55–63. doi:10.1109/TIT.1968.1054102.
12. Померанцев А. Классификация Российское хемометрическое общество, 2011 Электронный ресурс:<http://rcs.chemometrics.ru/Tutorials/classification.htm>
13. Jain A. K., Duin R. P. W., Mao J. Statistical pattern recognition: A review //IEEE Transactions on Pattern Analysis and Machine Intelligence.–2000–. Vol. 22, no. 1. – Pp. 4–37.
14. Tresp V. Committee machines // Handbook for Neural Network Signal Processing / Ed. by Y. H. Hu, J.-N.Hwang. – CRC Press, 2001.

Рецензент: Шишкин Николай Викторович, профессор кафедры, доктор технических наук, Академия ФСО России.

Molchanov Artem Nikolaevich

The Academy of the Federal Guard Service of the Russian Federation
Russia, Orel
E-mail: Tom_Silence@mail.ru

Skurnovich Aleksey Valentinovich

The Academy of the Federal Guard Service of the Russian Federation
Russia, Orel
E-mail: alexeymail2010@mail.ru

Web-page recognition algorithm estimating text coherence

Abstract. This paper describes web-page recognition algorithm estimating text coherence. It grounds actuality of algorithm design, describes main procedures and shows test results of web-page recognition. The algorithm uses mathematical model of the object recognition. The mathematical model of the object recognition is a set of web-page text parameters used for the text coherence estimating and web-page classification.

The algorithm is made up of two consecutive stages. The first stage is the training classifiers. The second one is the web-page recognition by means of trained classifiers. The algorithm uses three different classifiers for risk diversification of wrong make-decision. The final decision is adopted by majority voted system.

With the results of survey the authors draw a conclusion that the described algorithm may be implemented in the up-to-date systems aimed at gathering text information in Internet. It is noted that the described algorithm keeps its workability even through changing of web-page design.

Keywords: web-page; recognition; mathematical model of natural language text; text parameters; coherence; majority voted system; diversification of risk.

REFERENCES

1. Fridl Dzh. Regulyarnye vyrazheniya, 3-e izdanie. – Per. s angl. – SPb.:Simvol-Plyus, 2008. – 608 s., il.
2. Golubev S.A., Tolcheev Yu.K. Sharov Yu.L. Opyt vnedreniya i ispol'zovaniya informatsionno-poiskovoy sistemy ODB-Text v Sovete Federatsii Federal'nogo Sobraniya RF// Sovremennye tekhnologii v upravlenii i obrazovanii – novye vozmozhnosti i perspektivy ispol'zovaniya. Sbornik nauchnykh trudov. FGUP NII «Voskhod», MIREA. – M., 2001. – s.58–61.
3. Podсистема sbora soobshcheniy s saytov novostey seti Internet/ A.M. Andreev, D.V. Berezkin, V.V. Morozov i dr.// Trudy №1 molodykh uchenykh, aspirantov i studentov «Informatika i sistemy upravleniya». – M.: Izd. MGTU im. N.E. Bauman. – 2003. – s.409 – 410.
4. Tikhonov V.I. Optimal'nyy priem signalov. – M.: Radio i svyaz', 1983. – 320 s.
5. Rober de Bogrand i Vol'fgang Dressler Vvedenie v lingvistiku teksta. – M., 1981.
6. Deyk van T.A., Kinch V. Strategii ponimaniya svyaznogo teksta // Novoe v zarubezhnoy lingvistike. – Vyp. 23:Kognitivnye aspekty yazyka. – M., 1988.
7. Yusupov R.M. Statisticheskie metody obrabotki rezul'tatov nablyudeniya. – M.: Ministerstvo oborony SSSR, 1984
8. Vorontsov, K. V. Kombinatornaya teoriya nadezhnosti obucheniya po pretsedentam: Dis. dok.fiz.-mat. nauk: 05-13-17. – Vychislitel'nyy tsentr RAN, 2010. – 271 s. (<http://www.machinelearning.ru/wiki/images/b/b6/Voron10doct.pdf>)
9. Richard Ernest Bellman; Rand Corporation (1957). Dynamic programming. Princeton University Press. ISBN 978-0-691-07951-6., Republished: Richard Ernest Bellman (2003). Dynamic Programming. Courier Dover Publications. ISBN 978-0-486-42809-3.
10. Richard Ernest Bellman (1961). Adaptive control processes: a guided tour. PrincetonUniversityPress.
11. Hughes, G.F. (January 1968). "On the mean accuracy of statistical pattern recognizers". IEEE Transactions on Information Theory 14 (1): 55–63. doi:10.1109/TIT.1968.1054102.
12. Pomerantsev A. Klassifikatsiya Rossiyskoe khemometricheskoe obshchestvo, 2011 Elektronnyy resurs:<http://rsc.chemometrics.ru/Tutorials/classification.htm>
13. Jain A. K., Duin R. P. W., Mao J. Statistical pattern recognition: A review //IEEE Transactions on Pattern Analysis and Machine Intelligence.–2000–. Vol. 22, no. 1. – Pp. 4–37.
14. Tresp V. Committee machines // Handbook for Neural Network Signal Processing / Ed. by Y. H. Hu, J.-N.Hwang. – CRC Press, 2001.

Приложение А

Рисунок А. 1 – Блок-схема работы алгоритма на этапе обучения

Приложение Б

Рисунок Б.1 – Блок-схема работы алгоритма на этапе распознавания