На правах рукописи

ЧЕКИНА Александра Валерьевна

ГЕНЕТИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ТЕХНИЧЕСКОЙ ДОКУМЕНТАЦИИ В ПРОЕКТНЫХ РЕПОЗИТОРИЯХ САПР

05.13.12 – Системы автоматизации проектирования (промышленность)

Автореферат

диссертации на соискание ученой степени
кандидата технических наук

Ульяновск – 2012

Работа выполнена на кафедре «Информационные системы» в Ульяновском государственном техническом университете.

Научный руководитель: доктор технических наук, профессор
Ярушкина Надежда Глебовна

Официальные оппоненты: доктор технических наук, профессор
Соснин Петр Иванович

кандидат технических наук
Черкашин Сергей Витальевич

Ведущая организация: ФГБОУ ВПО ВолгГТУ

Защита диссертации состоится « » _____ 2012 г. в 12 часов на заседании диссертационного совета Д212.277.01 при Ульяновском государственном техническом университете по адресу: 432027, г. Ульяновск, ул. Северный Венец, 32 (ауд. 211, Главный корпус).

С диссертацией можно ознакомиться в библиотеке Ульяновского государственного технического университета.

Автореферат разослан « » _____ 2012 г.

Ученый секретарь
диссертационного совета,
доктор технических наук,
профессор Смирнов В.И.

Актуальность работы

Большинство крупных проектных организаций обладает значительным архивом успешных проектов. Новые проекты должны использовать ранее разработанные решения, так как повторность использования позволяет сократить сроки проектирования. Однако для решения задачи поиска проектного прототипа при хранении больших объемов информации необходима содержательная классификация проектных документов, которая позволит реализовать поиск похожих проектных документов. Следовательно, возникает задача создания проектного репозитория, автоматизирующего процессы классификации имеющихся и вновь поступающих в архив документов. Причем построение системы классов можно выполнить с помощью методов кластеризации.

В настоящее время существующие методы классификации проектных документов в архивах конструкторско-технической документации основаны на ручной процедуре присвоения кода проектному документу на основе справочника-классификатора. Поэтому существует проблема формирования автоматизированного метода кластеризации технической документации на основе лексики документа. Следовательно, для реализации поиска прототипов проектного решения в интеллектуальном проектном репозитории САПР требуется содержательная классификация проектных документов.

Решение научно-технической задачи создания проектного репозитария, автоматизирующего процессы классификации имеющихся и вновь поступающих в архив документов, в полном объеме не достижимо существующими методами и средствами.

Современный проектный репозитарий должен представлять собой интеллектуальное хранилище проектных документов, чтобы обеспечить поиск необходимого проектного решения. Единицей обработки и хранения в репозитарии является проектный документ, понимаемый как информационный ресурс. Информационный ресурс – это файл или совокупность файлов, объединенных общей семантикой и имеющих текстовую аннотацию. Основу индексирования проектных документов традиционно составляет лексический портрет текстового дескриптора информационного ресурса.

Задача кластеризации относится к классу задач оптимизации. Наиболее распространенный алгоритм кластеризации – алгоритм k-средних – является итерационным, достаточно медленным и слабо учитывает особенности пространства поиска.

Одним из методов решения задачи кластеризации могут служить генетические алгоритмы (ГА), так как стандартный генетический алгоритм сходится к глобальному оптимуму (по теореме схем Холланда). Кроме сходимости, ГА позволяет учесть особенности пространства поиска за счет настройки параметров и тем самым улучшить скорость сходимости. Поэтому целесообразно адаптировать ГА к решению задачи кластеризации проектных документов.

Существует много способов реализации идеи биологической эволюции в рамках генетических алгоритмов. Сегодня термином «генетический алгоритм» называют не одну модель, а широкий класс алгоритмов, подчас мало похожих друг на друга. Они различаются в первую очередь типами представления хромосом, операторами скрещивания (кроссинговера), мутации, различными подходами к воспроизводству и отбору. Гибкость структуры генетических алгоритмов, возможность настройки параметров позволяют получить проектные решения, отличающиеся высокой эффективностью.

Так как ГА является стохастическим, то зависимости скорости сходимости от его параметров необходимо исследовать. Результаты исследования должны позволить разработать ГА с управляемой сходимостью.

Основы современной теории кластеризации созданы трудами таких ученых, как С. Макнаотон, Гюстафсон, Кессель, Т. Кохонен, Г. Болл, Д. Холл, Дж. Мак-Кин, Г. Ланс, У. Уильямс, М. Жамбю, Г. Миллиган, М. Брюинош, Р. Дженсен, Х. Фридман, Дж. Рубин, Н.Г. Загоруйко, В.Н. Елкина и других. Основы построения интеллектуальных САПР, в том числе вопросы построения проектного репозитория проектов в САПР, рассмотрены в трудах Хилла П., Дж. Джонса, Норенкова И.П., Борисова А.Н. и др.. Значительный вклад в развитие методов генетической оптимизации в САПР внесли следующие ученые: Голберг Д., Холланд Дж., Курейчик В.М., Растригин Л.А., Курейчик В.В., Редько, Зинченко Л.А., Букатова И.Л. и др.

Цель диссертационной работы

Целью диссертационной работы является разработка новых и эффективных методов и алгоритмов решения задачи кластеризации технической документации проектного репозитория САПР.

Задачи исследования

Для достижения поставленной цели необходимо решить следующие задачи:

1. Выполнить сравнительный анализ существующих методов и систем кластеризации проектных документов;
2. Адаптировать схему генетической оптимизации к прикладной задаче кластеризации проектных документов как информационных ресурсов, для чего построить меру содержательного сходства проектных документов как расстояние между ними;
3. Разработать основные генетические операторы (селекция, кроссовер, мутация, формирование начальной популяции) применительно к задаче кластеризации проектных документов;

4. Разработать адаптивный алгоритм генетической кластеризации проектных документов, обеспечивающий быструю сходимость решения;
5. Предложить методику настройки параметров генетической кластеризации, обеспечивающую быструю сходимость и высокое качество решения на основе вычислительных экспериментов;
6. Разработать и реализовать программную систему генетической кластеризации проектных документов как базовую часть интеллектуального архива проектной документации;
7. Исследовать результативность и сходимость генетической оптимизации кластеризации проектных документов с помощью вычислительных экспериментов и внедрения в практику проектной организации.

Методы исследования

Для решения поставленных задач использовались следующие методы исследования: теория кластеризации, теория генетической оптимизации, методы математической статистики, методы концептуального и лексикографического анализов, метод экспертной оценки специалистов, объектно-ориентированный подход при создании комплекса программ.

Научная новизна

Адаптация схемы генетического алгоритма к прикладной задаче кластеризации проектных документов на основе построенной меры лексического сходства документов.

1. Разработка модифицированных генетических операторов: селекции, мутации и кроссинговера.
2. Разработка адаптивного управляемого генетического алгоритма обеспечивающего быструю сходимость.
3. Разработка методики управления адаптивным параметризованным генетическим алгоритмом.
4. Разработка структурно-функционального решения программной системы генетической кластеризации проектных документов для проектного репозитория САПР.

Все перечисленные положения являются новыми.

Достоверность результатов диссертационной работы

Достоверность научных положений, выводов и рекомендаций подтверждена результатами вычислительных экспериментов, корректным использованием формализованных методов, а также результатами использования материалов диссертации и разработанной системы в проектной организации в соответствии с актами внедрения.

Практическая значимость

На основе разработанных методов и алгоритмов создан программно-алгоритмический комплекс для решения задачи кластеризации информационных ресурсов. При построении программного комплекса использовался объектно-ориентированный язык Java и СУБД MS SQL Server. Программная система генетической кластеризации прошла апробацию в ФНПЦ ОАО «НПО МАРС», МУП «Ульяновская городская электросеть», что подтверждено соответствующими актами внедрения.

Апробация результатов исследования

Основные положения и результаты диссертации докладывались, и обсуждались на: «Interactive Systems and Technologies» (Ульяновск, 2007, 2009), на всероссийской конференции «Проведение научных исследований в области обработки, хранения, передачи и защиты информации ОИ-2009» (Ульяновск, 2009), на одиннадцатой и двенадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2008 (Дубна, 2008), КИИ-2010 (Тверь, 2010), на научных сессиях МИФИ-2007, 2008 (Москва, 2008), на второй всероссийской научной конференции с международным участием «Нечеткие системы и мягкие вычисления» НСМВ-2008 (Ульяновск, 2008), на семинаре с международным участием «Интеллектуальный анализ временных рядов» по результатам НИР, поддержанной ФЦП, проект № 02.740.11.5021 (Ульяновск, 2010), на международных «Конференции по логике, информатике, науковедению» (Ульяновск 2004 - 2007)

Основные положения неоднократно докладывались и обсуждались на научно-технических конференциях УлГТУ «Вузовская наука в современных условиях».

Основания для выполнения работы

Данная научная работа выполнялась в рамках тематического плана научных исследований Федерального агентства по образованию в 2005, 2006, 2007, 2008 г., была поддержана грантами РФФИ № 06-01-02012 и 06-01014087 в 2006 г., № 08-01-97006 в 2008 г.; ряд задач исследования решался в рамках х/д НИР № 100/05 УлГТУ по заказу ФНПЦ ОАО НПО МАРС.

Публикация результатов работы

По теме диссертации опубликовано 23 работы, в том числе 5 тезисов докладов, 18 статей.

Шесть статей опубликованы в изданиях, входящих в перечень ВАК.

Личный вклад

Все результаты, составляющие содержание диссертации, получены автором самостоятельно.

Структура и объем диссертационного исследования

Работа изложена на 215 страницах машинописного текста, содержит 28 рисунков и 43 таблицы, состоит из введения, четырех глав, заключения, списка использованной литературы и 4 приложений.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во **введении** обоснована актуальность темы диссертационного исследования, сформулированы цели, приведены сведения о полученных научных и практических результатах, реализации и внедрении работы, апробации, дано общее описание выполненной работы.

В **первой главе** содержится обзор состояния исследований в области кластеризации.

В параграфе 1.1 приведены определения информационного поиска и информационного ресурса.

Под информационным поиском (ИП) принято называть последовательность операций, выполняемых с целью отыскания и выдачи фактических данных, удовлетворяющих сформулированному запросу.

В настоящее время, в связи с геометрическим ростом информации, решение задач информационного поиска (ИП) электронных информационных ресурсов является актуальной задачей, имеющей существенную научную и практическую ценность. Для решения задач ИП электронных информационных ресурсов (ЭИР) применяют специальный класс автоматизированных систем: информационно-поисковые системы (ИПС).

Единицей обработки и хранения в репозитарии является информационный ресурс. В диссертации принято следующее определение информационного ресурса (ИР): файл или совокупность файлов, объединенных общей семантикой и имеющих текстовую аннотацию. В частном случае, информационный ресурс – это один или несколько текстовых файлов. Текст аннотации (или текст самого ресурса) однозначно отражает смысловое содержание данного ресурса. При кластеризации мы полагаемся на гипотезу о том, что смысловое содержание текста кодируется статистическим распределением слов. То есть, по частотному распределению слов, составляющих текст ресурса (или аннотации), мы можем определить его категорию.

В параграфе 1.2 приводится формализованная задача кластеризации ИР. Поясняется структура задачи. Рассматриваются традиционно используемые в кластерном анализе меры близости между объектами.

Меру близости (сходства) объектов удобно представить как обратную величину от расстояния между объектами. Было выбрано Евклидово

расстояние $d(x,y) = \{\sum_i (x_i - y_i)^2\}^{1/2}$, которое является частным случаем Обобщенного степенного расстояния Минковского для p=1

$$d_{ij} = \left(\sum_{k=1}^{v} \left| x_{ik} - x_{jk} \right|^{p} \right)^{1/p}$$

Параграф 1.2 первой главы включает классификацию и анализ алгоритмов и методов кластеризации. В таблице 1 представлены краткие характеристики каждого вида.

Таблица 1. Обзор алгоритмов кластеризации

Алгоритм кластеризации	**Форма кластеров**	**Входные данные**	**Результаты**
Иерархический	Произвольная	Число кластеров или порог расстояния для усечения иерархии	Бинарное дерево кластеров
k-средних	Гиперсфера	Число кластеров	Центры кластеров
c-средних	Гиперсфера	Число кластеров, степень нечеткости	Центры кластеров, матрица принадлежности
Выделение связных компонент	Произвольная	Порог расстояния R	Древовидная структура кластеров
Минимальное покрывающее дерево	Произвольная	Число кластеров или порог расстояния для удаления ребер	Древовидная структура кластеров
Послойная кластеризация	Произвольная	Последовательность порогов расстояния	Древовидная структура кластеров с разными уровнями иерархии

На данный момент существует множество методов, осуществляющих кластеризацию или классификацию документов. Некоторые методы могут использовать несколько альтернативных алгоритмов. В таблице 2 представлена сводка основных характеристик методов.

В параграфе 1.2.4 сформулированы основные подходы к решению задачи сравнения работы автоматических классификаций, приведены характеристики оценки работы систем автоматической кластеризации ИР. Определены некоторые основные свойства идеального алгоритма автоматической классификации.

Также в параграфе приведены методики оценки качества автоматической кластеризации, статистические формулы функционалов качества.

Во **второй главе** описываются методы, методики и алгоритмы решения задачи классификации ИР на основе эволюционных вычислений.

В параграфе 2.1 рассматривается индекс ИР на основе его лексического

портрета, как основа входных данных для алгоритма генетической кластеризации.

Классическое описание процесса индексация представляется в виде одной или нескольких операций:

1. Отбор индексационных терминов или ключевых слов (используемых для описания содержания документа),
2. Приписывание терминам некоторого веса (отражающего предполагаемую важность термина),
3. Отнесение терминов к определенному типу («к классу действий, свойств или объектов»),
4. Определение отношений между терминами (синонимических, иерархических, ассоциативных).

Взвешивание терминов

Взвешивание терминов подразумевает, что каждому дескриптору x_i в документе D ставится в соответствие некоторый неотрицательный вес w_i.

Вес представляет собой количественную характеристику, которая, как правило, варьируется в интервале [0,1] или [0,100] (интервал выбирается в зависимости от условий анализа и для него могут задаваться другие границы).

Метод сигнал-шум

Для оценки значимости слов используется методы определения частот слов каждого документа и частот, рассчитанных по формуле Шеннона (сигнал-шум):

$$w_i = \frac{S^k}{N^k},$$

где N^k – шум термина,

$$N^k = \sum_{i=1}^{n} \frac{f_i^k}{F^k} \log \frac{F^k}{f_i^k},$$

где f_i^k – частота k – го термина в i– м документе,

F^k – частота k – го термина по всем документам,

S^k – сигнал термина

Таблица 2. Обзор методов кластеризации текстов

Методы	**Вид метода**	**Ограничения**	**Пересекаемость кластеров**	**Инкрементность**	**Используемые числовые характеристики документов**	**Предварительное обучение**	**Скорость работы**
LSI	Числовой, кластеризующий	Количество кластеров	-	+	tfidf	-	N^2 x k, N=terms+docs, k – factors
STC	Нечисловой, кластеризующий	Нет ограничений	+	+	-	-	$O(k^2 N)$, N-число документов, k – число кластеров
Single Link, Complete Link, Group Average	Числовой, bottom-up, кластеризующий	Количество документов в кластере	-	+	Similarity matrix	-	Single Link ~ $O(N^2)$ Complete Link ~ $O(N^3)$ Group Average ~ $O(N^2)$
Scatter/Gather	Числовой, bottom-up, кластеризующий	Количество кластеров	-	-	Similarity matrix	-	Buckshot ~ O(kN) Fractionation ~ O(mN), m=O(k), k – число кластеров
K-means	Числовой, кластеризующий	Количество кластеров, центроиды	-	-	tfidf	-	O(n)
CI – необучаемый вариант	Числовой, top-down, кластеризующий	Количество кластеров	-	В случае кроме рекурсивной бисекции	Similarity matrix	-	O(N*log k), k – число кластеров
CI – обучаемый вариант	Числовой, классифицирующий	Количество кластеров	-	-	Similarity matrix или tfidf	+	
SOM	Числовой, классифицирующий	Количество кластеров	+	+	Similarity matrix или tfidf	+	

Сформулированы особенности кластеризации ИР:

1. ИР представляется его лексическим портретом. В результате процесса индексации получены частоты встречаемости в ИР всех термов из составленного ранее словаря. Словарь может быть очень большим, так как чем он больше, тем наиболее полный лексический портрет ИР будет получен. Это значит, что объект кластеризации (ИР) будет иметь множество атрибутов, в качестве которых выступают частоты терминов.
2. В лексическом портрете ИР могут иметь место нулевые частотности или частотности с очень малыми числовыми значениями встречаемости терминов. Это означает, что данный термин в документе либо не встретился, либо встретился малое (по сравнению с другими терминами из словаря) число раз. Но, несмотря на малые или нулевые значения, они играют большую роль для определения индекса документа. Алгоритм должен работать и с малозначимыми терминами.
3. Проектная организация, как правило, имеет широкий спектр деятельности. Выпускается много продуктов, разработка каждого из которых сопровождается существенным объемом документации. Вся документация, в свою очередь, подразделяется на тематики, рубрики. Как следствие, число кластеров, на которые разбивается документация очень велико.

В параграфе 2.2 приведена адаптация стандартного генетического алгоритма к решению задачи кластеризации.

Адаптация заключается в уточнении параметров стандартного генетического алгоритма для решения прикладной задачи.

В случае кластеризации ИР необходимо определить следующие параметры ГА:

1. Способ кодировки решения (хромосомы)
2. Содержание операторов отбора (селекции), рекомбинации и мутации
3. Функция оптимальности (оценки) каждой хромосомы
4. Условие завершения эволюции
5. Вероятностные параметры управления сходимостью эволюции

Хромосома представляет собой массив пар (документ, кластер).

Документ	1	2	3	4	5	6	7	…	m
Кластер	3	5	n	n-1	n-3	n-7	…	…	n

Каждая хромосома оценивается мерой ее «приспособленности» (fitness–function). Фитнесс – функция для каждой хромосомы определяется суммой евклидовых расстояний от каждого ИР до центра соответствующего кластера, т.е.

$$f = \sum_{j=1}^{m} \sqrt{\sum_{i=1}^{n} (x_i^{u} - x_i^{up})^2}$$

где x^{u} – координата центра i-го кластера,

x^{up} – координата i-го информационного ресурса,

m – количество информационных ресурсов, которое одновременно определяет и длину хромосомы.

n – количество координатных осей, по которым формируется общая координата информационного ресурса.

Наиболее приспособленные особи получают большую возможность участвовать в воспроизводстве потомства.

Пропорциональный отбор назначает каждой i-й хромосоме вероятность P(i), равную отношению ее приспособленности к суммарной приспособленности популяции.

$$P(i) = \frac{fi}{\sum_{i=1}^{n} fi}$$

В алгоритме используется равномерный многоточечный кроссовер, в результате которого все биты хромосом обмениваются с некоторой вероятностью. Значение каждого бита в хромосоме потомка определяется случайным образом из соответствующих битов родителей. Для этого вводится некоторая величина $0<p_0<1$, и если случайное число больше p_0, то на n-ю позицию первого потомка попадает n-й бит первого родителя, а на n-ю позицию второго – n-й бит второго родителя. В противном случае к первому потомку попадает бит второго родителя, а ко второму – первого. Такая операция проводится для всех битов хромосомы. Т.е. при выборе от какого родителя потомок возьмет следующий ген, предпочтение отдается наиболее приспособленному.

Оператор мутации представляет собой обмен двух случайных номеров кластеров.

Документ	1	2	3	4	5	6	7	…	m
Кластер	3	5	n	n-1	n-9	n-1	n-2	…	1

В параграфе 2.3 представлен алгоритм кластеризации ИР на основе схемы генетической адаптации (рис.1).

В параграфе 2.4 описана разработка адаптивного генетического алгоритма.

Адаптивность – это способность/особенность организма приспосабливаться к меняющимся внешним условия.

Адаптивный алгоритм – это алгоритм, который меняет свои параметры на основе подстройки под входные данные.

Сходимость алгоритма - это способность итерационного алгоритма достигать оптимума целевой функции или подходить достаточно близко к

нему за конечное число шагов. Скорость сходимости алгоритмов – один из важнейших показателей качества.

Понятие скорости сходимости

Пусть $\{x_n\}$— последовательность приближений рассматриваемого алгоритма нахождения корня x^* некоторого уравнения, тогда:

Говорят, что метод обладает *линейной сходимостью*, если $\exists \alpha \in [0,1]: \exists N \in \mathrm{N}, \forall n \geq N \left\| x_n - x^* \right\| < \alpha \left\| x_{n-1} - x^* \right\|$

Говорят, что метод обладает *сходимостью степени* β, если.

$\exists \alpha \in [0,1]: \exists N \in \mathrm{N}, \forall n \geq N \left\| x_n - x^* \right\| < \alpha \left\| x_{n-1} - x^* \right\|^{\beta}$

Обычно скорость сходимости методов не превышает квадратичной).

Стагнация генетического алгоритма – это такое состояние алгоритма, при котором на протяжении большого числа поколений не было изменения лучшего значения функции приспособленности у популяции, но текущее решение сильно отличается от глобального минимума. На рис.2 представлен график сходимости генетического алгоритмы. Периоды стагнации выделены красным цветом, они характеризуются горизонтальными участками графика.

Период стагнации характеризуется двумя параметрами:

- ***Длина периода стагнации***

Под длиной периода стагнации понимается количество итераций алгоритма, на протяжении которых не происходило существенных изменений величины функции приспособленности лучшей хромосомы.

- ***Порог стагнации***

Под порогом стагнации понимается величина изменения функции приспособленности лучшей хромосомы (в процентах), которая признается не значимой.

Если на протяжении заданного числа шагов функция приспособленности изменяется не более чем на указанный порог стагнации, возникает необходимость введения режима преодоления стагнации, так как улучшения найденного решения не происходит. Именно на таких участках не меняется величина Fitness–функции, а значит, лучшее решение не находится.

Адаптивный генетический алгоритм исключает периоды стагнации либо сводит их длительность к минимуму за счет увеличения разнообразия популяции.

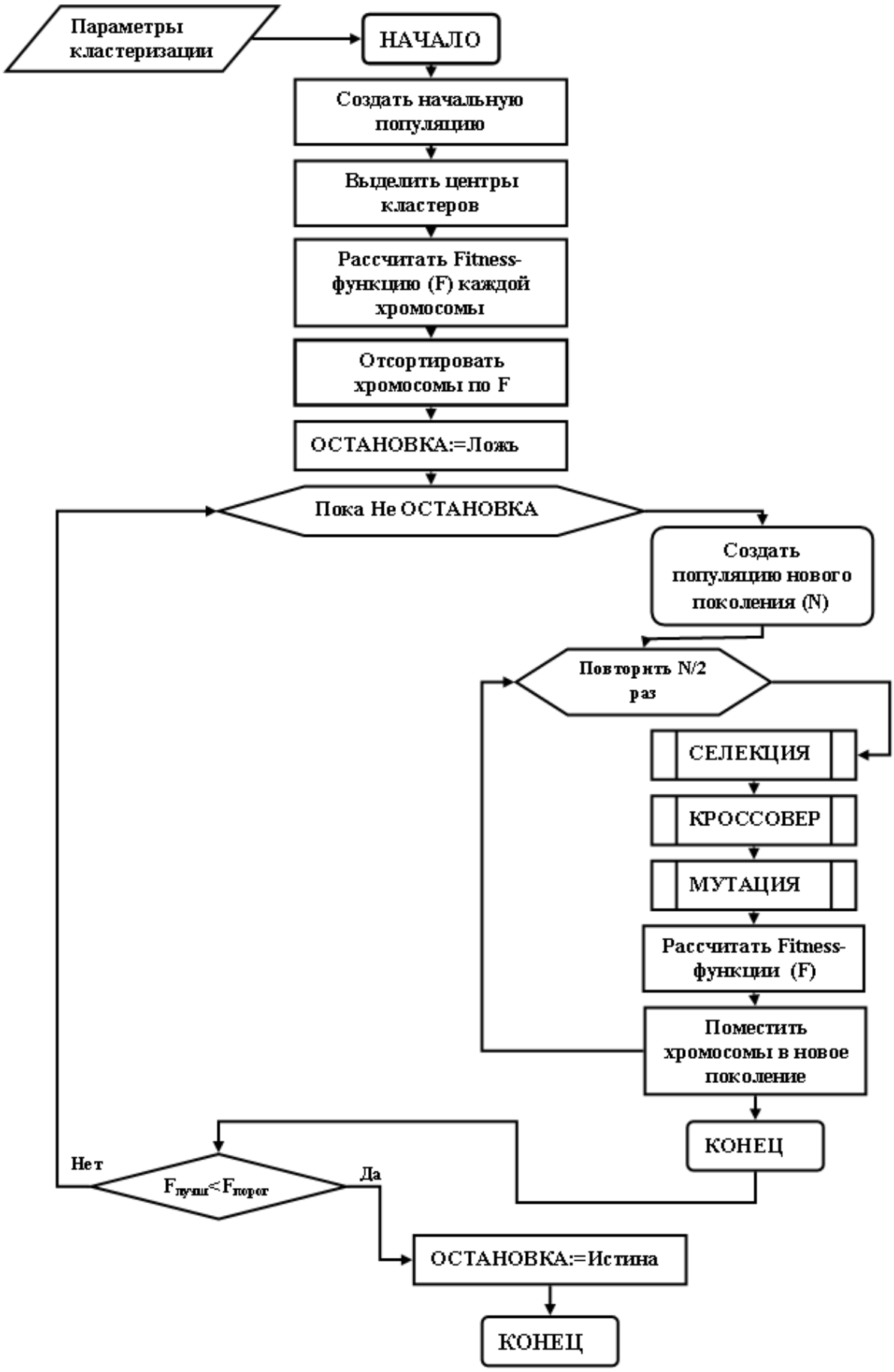

Рис. 1. Алгоритм кластеризации ИР на основе схемы генетической адаптации

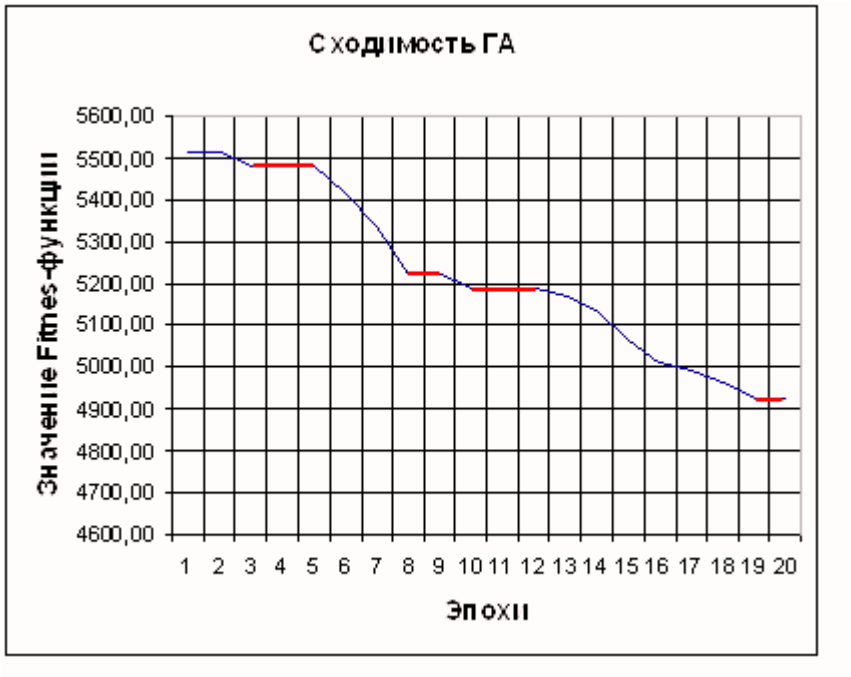

Рис. 2. Сходимость ГА

Адаптивность генетического режима заключается в возможности применения двух режимов преодоления стагнации:

- ***Управление мутацией через степень разнообразия популяции***

Введение такого режима основывается на предположении, что если увеличить процент мутации, что соответственно приведет к большему разнообразию хромосом в популяции, то решение будет найдено быстрее.

- ***Управление элитой через оператор селекции***

Второй возможностью выхода из периода стагнации является режим автоматического изменения количества особей, участвующих в воспроизводстве (элиты). В моменты застоя включается алгоритм автоматической корректировки параметров, который в данном случае добавляет в элиту несколько хромосом идущих по списку ниже черты, отсекающей элиту. В популяции появляются новые гены, позволяющие найти новые лучшие решения.

Адаптивный генетический алгоритм для решения задачи кластеризации представлен на рис.3

В параграфе 2.5 приведена методика настройки параметров генетической кластеризации.

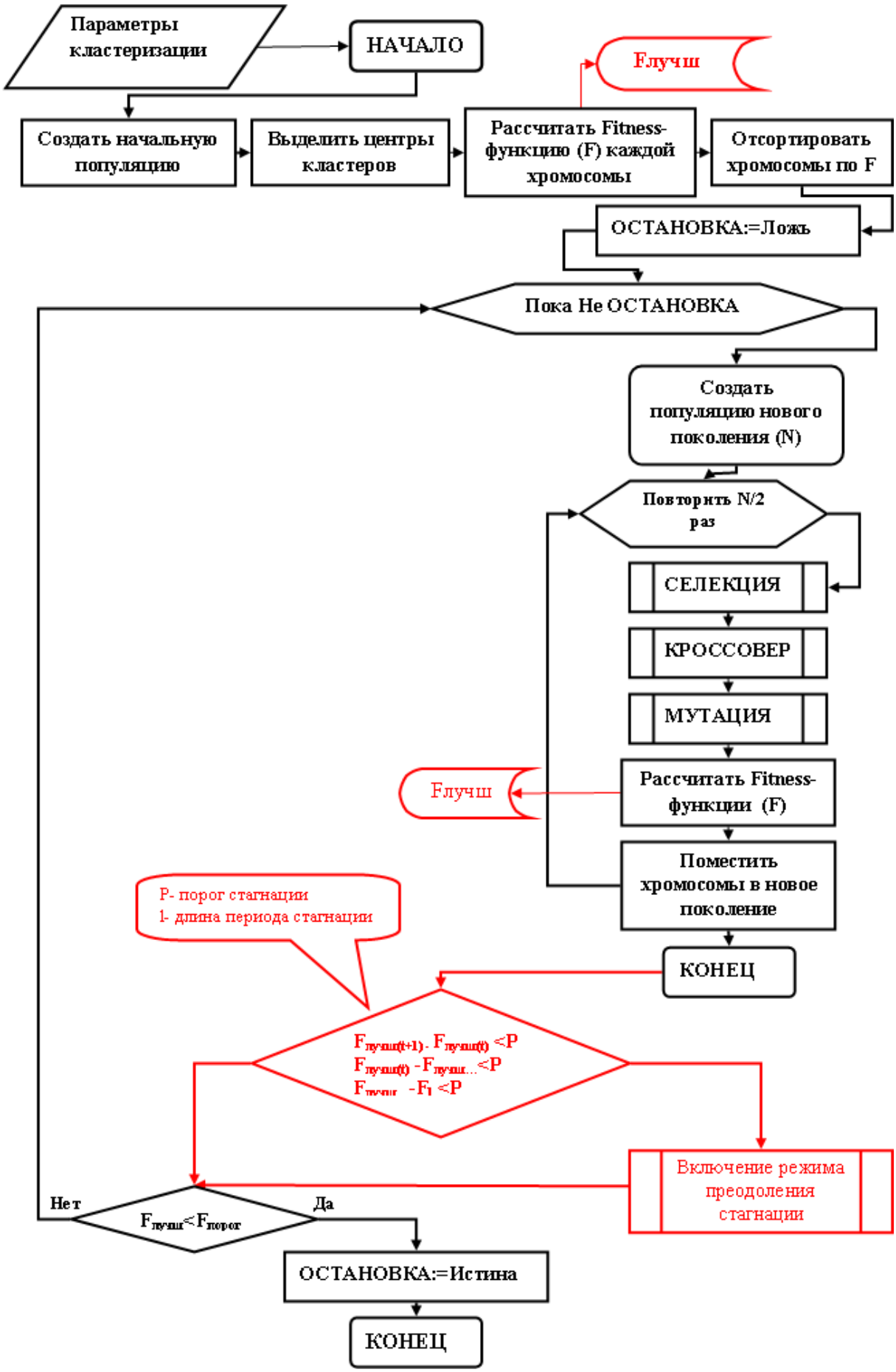

Рис. 3. Адаптивный алгоритм кластеризации ИР на основе режимов управления сходимостью

Методика настройки параметров работы алгоритма предусматривает два случая:

Выбор обоснованных параметров, установленных по умолчанию

Параметры кластеризации:

- Количество кластеров. Задается пользователем в зависимости от стоящей перед ним задачи.
- Количество итераций

При проведении кластеризации небольшого объема документов (50-100 документов) нет необходимости ставить большое число итераций алгоритма. Экспериментально определено, что при 65 документах алгоритм сходится приблизительно за 20-30 итераций. Дальнейшая работа системы существенный изменений на результат не оказывает.

При среднем количестве кластеризуемых документов (200-1000) алгоритм сходится в среднем за 45-50 итераций. Дальнейшая работа системы существенного влияния на результат не оказывает, поэтому в ней нет необходимости.

При работе с большими объемами документов (порядка 5000) следует ставить большие значения кол-ва итераций. Сами значения определяются вычислительной мощностью имеющейся в распоряжении техники.

- Предел популяции=100
- Размер первого поколения=100
- Целевое качество. Определяется стоящей перед пользователем задачей.
- Плодовитость=1.5
- Вероятность мутации=3,2
- Порог стагнации=0,3
- Длина стагнации=3
- Повышение мутации=6%
- Глубина=3

Выбор параметров в случае, если пользователь хочет управлять работой системы

Улучшение значение Fitness-функции

На основе полученных значений коэффициентов корреляции можно говорить о том, что на конечное значение Fitness-функции из рассмотренных параметров большее влияние оказывает плодовитость (R(F,Pl)= -0,42; R(F,Pl)= -0,23). Причем наблюдаемая зависимость является обратной (рис.4), об этом свидетельствует отрицательное значение коэффициента корреляции.

Если для пользователя значимым параметром является Fitness-функция, то для улучшения достигаемого значения следует увеличивать значение плодовитости с Pl=1,5 (устанавливается по умолчанию) до Pl=3, Pl=6.

Мутация хотя и оказалась по результатам проведенных экспериментов менее влияющим параметром, также оказывает некоторое воздействие на

значение достигнутой Fitness-функции (R(F,M)= -0,13; R(F,M)= -0,15). Наблюдаемая зависимость также является обратной (рис.5).

Существенные изменения вносит увеличение процента мутации до 3,2%, дальнейшее увеличение значимого воздействия не оказывает.

Если пользователь хочет достичь лучшего значения Fitness-функции, можно увеличить значение мутации до M=3,2%, M=20%. Следует также учитывать экспериментально определенную зависимость: при малом кол-ве кластеров влияние процента мутации незначительно; при увеличении кол-ва кластеров влияние процента мутации возрастает.

Улучшение скорости сходимости

На скорость сходимости алгоритма наибольшее влияние оказывает плодовитость (рис.4). Об этом свидетельствуют достаточно высокие значения коэффициентов корреляции (R(V, Pl)=-0,67; R(V, Pl)=-0,82).

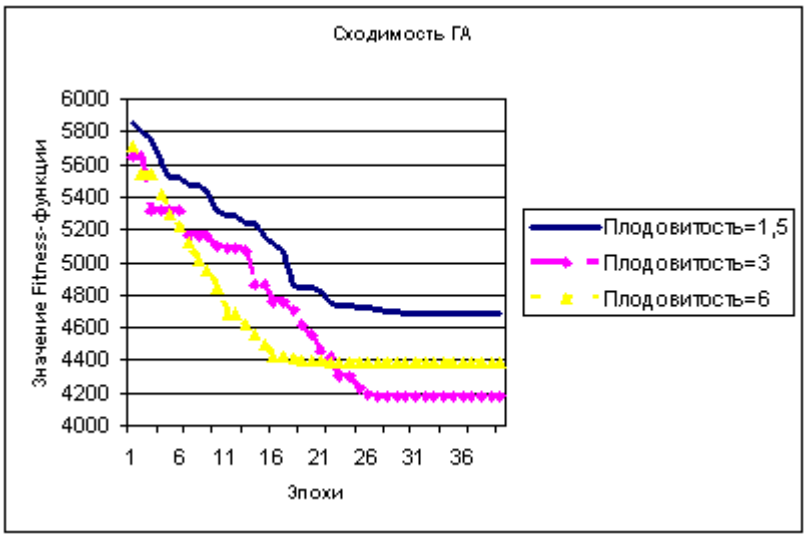

Рис. 4. Влияние плодовитости

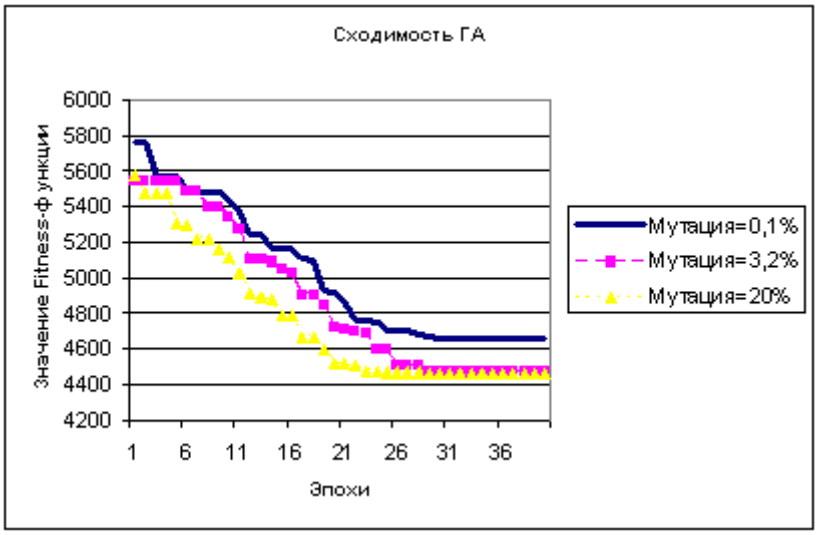

Рис. 5. Влияние мутации

Полученные значения коэффициентов отрицательные, что говорит об обратной зависимости. Причем, судя по значениям коэффициентов корреляции, зависимость более ярко выражена, чем зависимость Fitness-функции от плодовитости.

Если для пользователя наиболее значимым параметром является скорость сходимости алгоритма, необходимо увеличивать плодовитость до Pl=3, Pl=6.

Мутация оказывает меньшее влияние на скорость сходимости (R(V,M)=0,24; R(V,M)=0,15). Но следует отметить, что зависимость здесь прямая. Коэффициенты корреляции положительны.

Если пользователь хочет добиться быстрой сходимости алгоритма, ему следует уменьшать значение процента мутации до M=0,1.

Конечное достигнутое качество (F) и скорость сходимости (V) являются конфликтующими показателями по значению установленного процента мутации. В одном случае наблюдается положительная корреляция, в другом - отрицательная. Лучшее значение Fitness-функции чаще всего достигается за большее число итераций алгоритма. Соответственно меньшей скорости сходимости можно достигнуть за счет ухудшения качества найденного решения. Поэтому необходимо достижение некоторого баланса качества и скорости работы алгоритма. Такими компромиссными значениями параметров являются устанавливаемые по умолчанию значения.

Варьирование параметрами преодоления стагнации

Для включения режима преодоления стагнации пользователь должен определить:

1. Длину стагнации
2. Порог стагнации

Если пользователь хочет включать режим преодоления стагнации только при горизонтальных участках графика сходимости, считая даже небольшие улучшения значения Fitness-функции достаточными, то ему следует устанавливать значение порога стагнации минимальным. Порог стагнации 0,3

Если режим преодоления стагнации должен включаться при незначительном улучшении Fitness-функции, то есть за стагнацию принимаются участки очень плавного изменения значения, то порог стагнации стоит увеличивать. Соответственно, чем более вертикальные участки считаются стагнацией, тем больше значение порога. Порог стагнации 0,5; 1;3;5.

Длина стагнации отвечает за длину горизонтальной полки на графике сходимости. Поэтому, если для пользователя допустимо на протяжении нескольких итераций алгоритма ждать, что система сама выйдет из стагнации без дополнительных вмешательств, то значение длины стагнации можно ставить больше. Длина стагнации 5. При этом существенных изменений в элиту популяции вносится меньше. Но следует понимать, что в этом случае уменьшается скорость сходимости. Если критическим является скорость сходимости, значение длины стагнации следует уменьшать. Длина стагнации 3.

В третьей главе рассматривается программная система генетической кластеризации информационных ресурсов, обосновывается выбор инструмента реализации, описывается информационное обеспечение

комплекса, приведены основные алгоритмы программной реализации.

Программная система, реализующая идеи интеллектуального хранилища, создана в рамках научно-исследовательского проекта по разработке автоматизированной системы «Интеллектуальный сетевой архив информационных электронных ресурсов» (ИСА ЭИР) (х/д НИР №100/05 УлГТУ по заказу ФНПЦ ОАО «НПО «МАРС»). Применяемость и функциональность ИСА ЭИР протестирована в работе архивной службы ФНПЦ ОАО «НПО «МАРС».

Генетический кластеризатор представляет собой отдельный модуль программы «Интерактивный сетевой архив электронных информационных ресурсов» (ИСАЭИР) [Наместников и др., 2008] предназначенный для классификации электронных информационных ресурсов, с целью формирования данных для проведения информационного поиска.

Программа состоит из следующих подсистем: подсистема индексирования электронных информационных ресурсов (ЭИР) (индексатор), подсистема кластеризации ЭИР на основе нейронной сети (нейросетевой кластеризатор), подсистема кластеризации на основе fuzzy-c-means метода (fcm-кластеризатор), классификатор и подсистема кластеризации на основе генетического алгоритма (рис. 6).

На модуль индексации возложены задачи предобработки текстовых документов или аннотаций к ЭИР и построение частотных словарей встречающихся терминов. Для сохранения частотных таблиц используется СУБД MS SQL 2000. Далее, в рамках модулей кластеризации и классификации, на основе значений относительных частот (полученных в результате индексации) создаются предметно-ориентированные кластеры, которые организуются в виде иерархии. В процессе классификации выполняется задача соотнесения вновь заносимого ЭИР с определенным кластером.

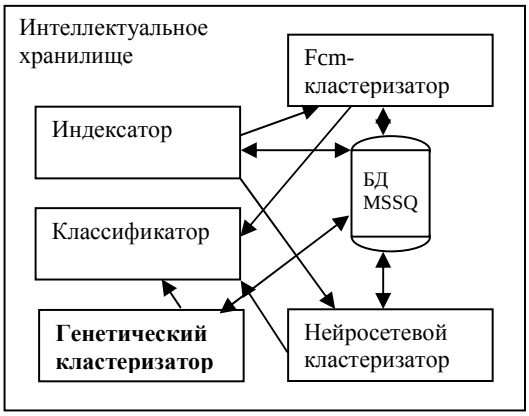

Рис. 6. Структура интеллектуального хранилища

Поскольку система является частью одного программного комплекса, необходимо, чтобы в дальнейшем система была открытой для модификаций. Решением данного требования могут стать использование стандартизированных и открытых решений используемых при инжиниринге программного обеспечения.

В качестве языка программирования приложения был выбран — Java. Платформа Java относится к OpenSource лицензиям и свободна для распространения и использования. Microsoft SQL Server — система управления реляционными базами данных (СУБД), разработанная корпорацией Microsoft. Платформа Java позволяет использовать различные СУБД. Выбор современной реляционной СУБД MS SQL Server обоснован требованием, предъявленным заказчиком.

Диаграмма вариантов использования приложения приведена на рис.7

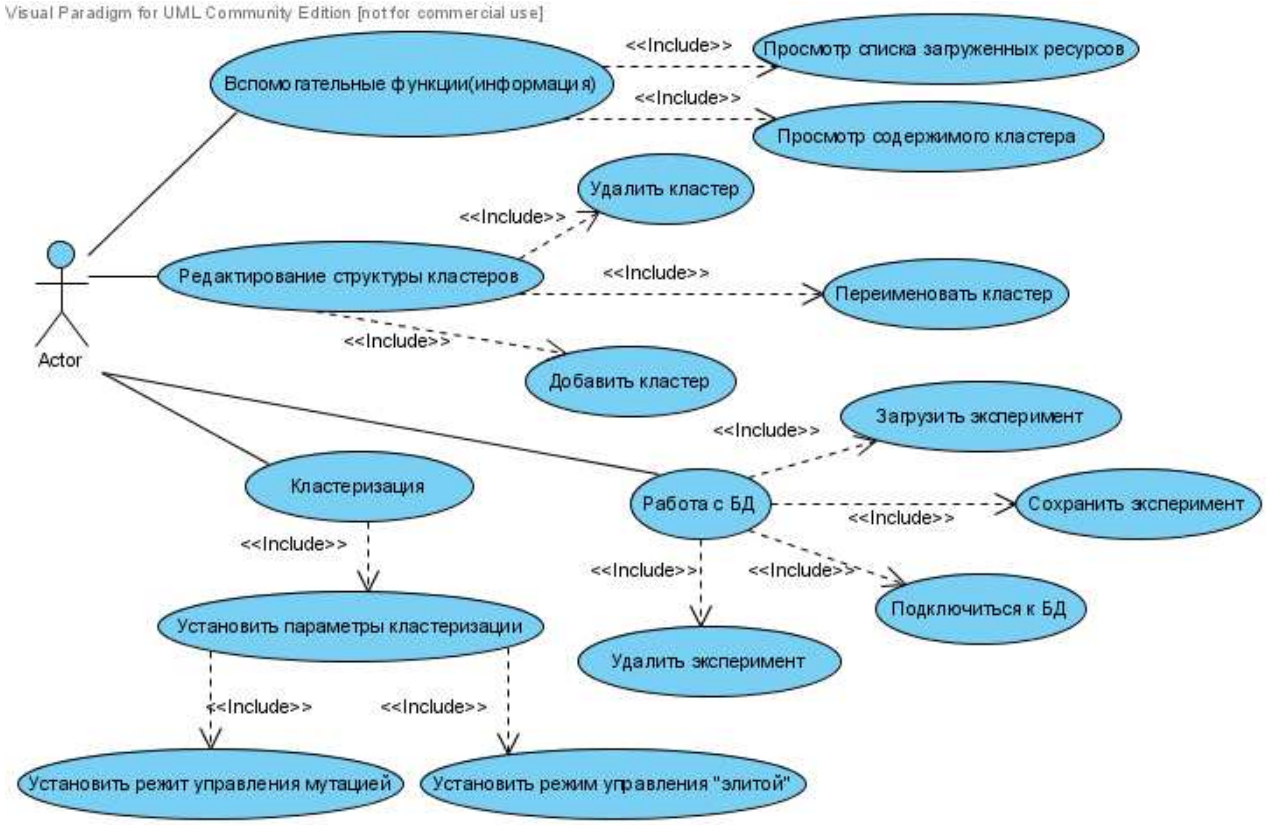

Рис. 7. Диаграмма вариантов использования приложения

Основные функции приложения можно разделить на 4 части:

1. Работа с базой данных
2. Кластеризация
3. Редактирование структуры кластеров
4. Вспомогательные функции

Структура выходных данных генетического кластеризатора и взаимосвязи между ними представлены на ER-диаграмме (рис. 8).

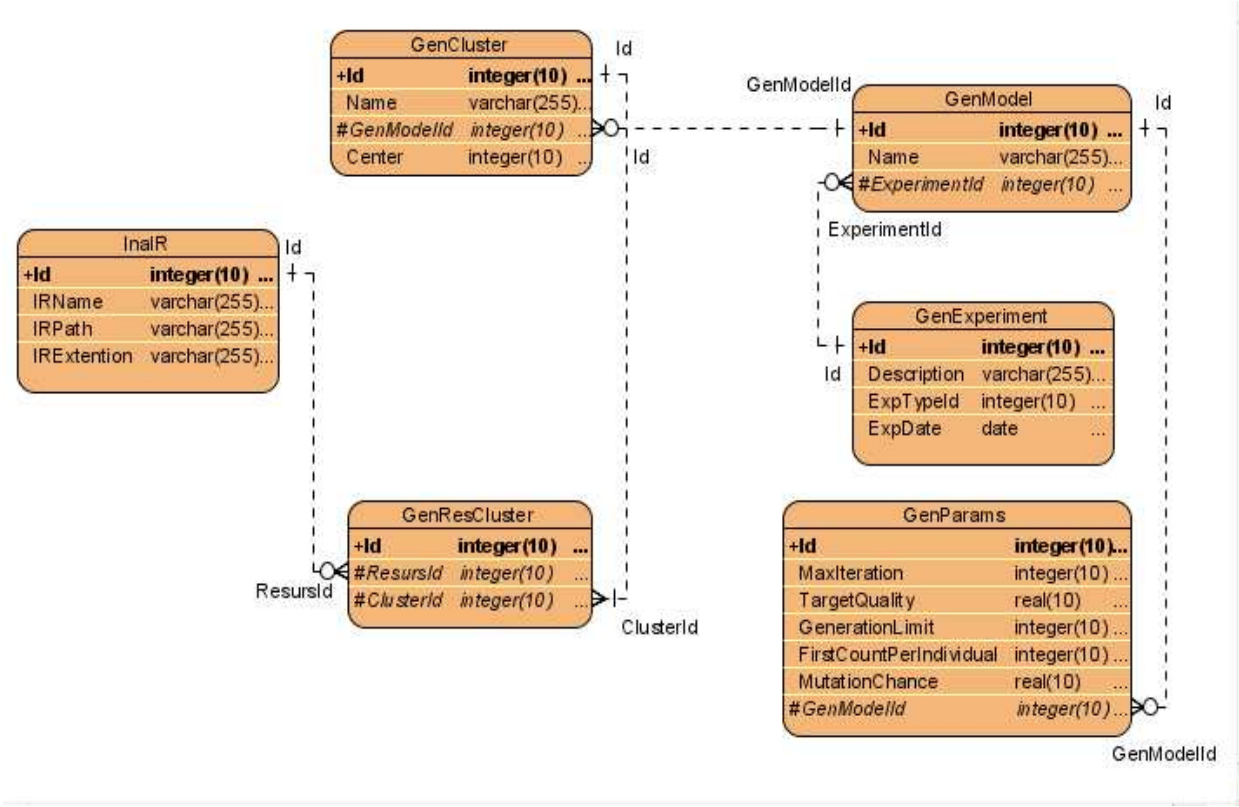

Рис.8. Структура данных генетического кластеризатора

В таблице 3 приведено описание данных сущностей

Таблица. 3. Описание структуры БД (таблицы генетического кластеризатора)

GenCluster(хранит сведения о кластерах)	Id- идентификатор кластера Name-имя GenModelId-идентификатор модели Center-центр
GenModel(хранит параметры модели эксперимента)	Id- идентификатор модели эксперимента ExperimentId-идентификатор самого эксперимента Name-имя
GenExperiment(хранит общие параметры эксперимента)	Id- идентификатор эксперимента Description-текстовое описание ExpTypeId-тип эксперимента(может еще быть fcm и som) ExpDate-дата проведения эксперимента
GenParams(хранит параметры эксперимента, заданные пользователем)	Id- порядковый номер GenModelId-идентификатор модели MaxIteration-максимальное число итераций TargetQuality-значение фитнесс - функции, по достижении которого алгоритм остановится GenerationLimit-лимит величины одного поколения FirstCountPerIndividual-среднее число скрещиваний на особь MutationChance-шанс мутации
GenResCluster(основная таблица, хранящая отнесение документа к кластеру)	Id- порядковый номер ClusterId-идентификатор кластера ResursId-идентификатор ресурса

Диалоговое окно генетического кластеризатора представлено на рис. 9. В левой части перечислены полученные в результате работы модуля кластеры. Щелкнув по выбранному кластеру мышью, справа появляется его содержимое.

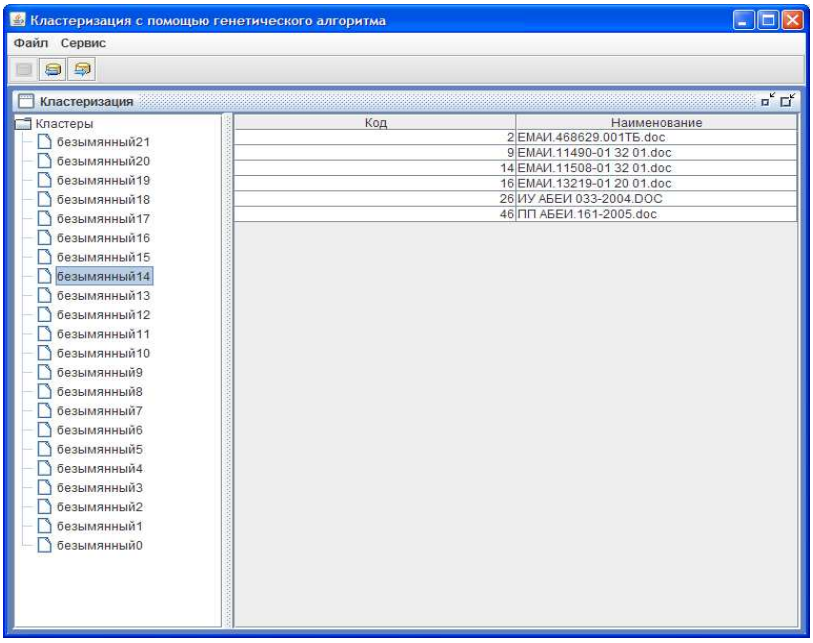

Рис. 9. Диалоговое окно генетического кластеризатора

В четвертой главе приведены результаты вычислительных экспериментов.

В параграфе 4.1 описаны эксперименты по оценке качества кластеризации в сравнении с экспертной классификацией, проведенной архивариусом ФНПЦ ОАО НПО МАРС. Эксперименты проводились на выборках объемом 65, 265 и 5013 документов.

На выборке в 65 документов экспертом получены 4 вида классификаций:

- классы документации(3 класса);
- виды документации(16 классов);
- разделы документации(22 класса);
- тематика работ(21 класс).
- Небольшой пример экспериментальных данных приведен в таблице 4.

Таблица. 4. Эксперименты

№ эксп.	Кол-во кластеров	Плодовитость	Мутация	Кол-во итераций	Fitness-функция
1	3	1,5	0,0001	100	6539,309
2	3	1,5	0,0001	150	6456,679
3	3	1,5	0,032	10	6557
4	3	1,5	0,032	100	6452,17
5	3	1,5	0,1	100	6389,70
6	3	1,5	1	10	6608,7
7	3	1,5	1,5	150	6463,7
8	3	1,8	10	10	6548,68
9	3	1,8	0,032	10	6465,207
10	16	1,5	0,01	100	4799
11	16	1,5	0,02	50	4204,407

12	16	1,5	1	50	4548,896
13	16	1,5	0,0001	50	4767,27
14	16	1,5	0,01	50	4439,699

В соответствии с моделью оценки качества кластеризации были рассчитаны значения целевой функции. Значения целевой функции показывают качество автоматической кластеризации с помощью генетического алгоритма в сравнении с экспертной классификацией той же выборки. Чем меньше величина целевой функции, тем ближе результат автоматической кластеризации и экспертной.

Альфа – коэффициент важности критерия. Чем больше альфа, тем более важна полнота поиска, чем меньше альфа, тем более важна точность поиска. Пример оценки с альфа=0,1 представлен в таб. 5.

Таблица 5. Фрагмент оценки результатов эксперимента

№ экп.	Кол-во кластеров	Виды документации	Изделия	Разделы	Классы
001	3	0,914944444444	0,910121875	0,911645714285	0,709925
002	16	0,662961224489	0,723378846153	0,741216981132	0,286738461538
003	21	0,577830612244	0,646880357142	0,662005357142	0,210348387096

Сходимость алгоритма при кластеризации 65 документов на 22 кластера представлен на рис. 10. Видно, что после определенного этапа значение Fitness-функции остается неизменным (F=6452,172).

Рис. 10. Сходимость ГА

Результат автоматического разбиения наиболее близок к экспертной кластеризации по классам документации.

По аналогичной форме представлены в диссертации результаты экспериментов на 265 и 5013 документах.

В параграфе 4.2 описаны эксперименты по исследованию эффективности предложенного адаптивного генетического алгоритма.

Эксперименты по оценке эффективности адаптивного генетического алгоритма проводились на выборке из 65 документов, содержащей

документы преимущественно организационно-нормативного характера. Ранее комплект документации был классифицирован архивариусом-экспертом. Для обеспечения сопоставимости результатов, 65 документов всегда разбивались на 16 классов, что соответствует экспертной кластеризации по видам документации.

Было проведено 150 экспериментов, из них:

1. С использованием режима «Управление сходимостью через оператор мутации» – 75 экспериментов
   - С длиной стагнации 3 – 25 экспериментов
   - С длиной стагнации 4 – 25 экспериментов
   - С длиной стагнации 5 – 25 экспериментов
2. С использованием режима «Управление сходимостью через размер «элиты»» –75 экспериментов
   - С длиной стагнации 3 – 25 экспериментов
   - С длиной стагнации 4 – 25 экспериментов
   - С длиной стагнации 5 – 25 экспериментов

Результаты экспериментов с адаптивным генетическим алгоритмом приведены в Приложении 1.

Обобщенные результаты экспериментов с адаптивным генетическим алгоритмом приведены в таб.6 и таб.8.

В параграфе 4.3.4. приведена оценка результатов автоматической кластеризации методом ГА.

При интерпретации результатов экспериментов значимыми параметрами являются количество поколений (скорость сходимости V), за которое найдено минимальное значение Fitness-функции F, и сама величина достигнутого значения F. Они представлены парами чисел на пересечении параметров кластеризации.

По результатам проведенных экспериментов, опираясь на модель, представленную в Главе 2 диссертации, для обоснования методики, сформированной во многом эмпирическим путем, была рассчитана корреляция

- между скоростью сходимости алгоритма значением мутации
- между скоростью сходимости алгоритма и значением плодовитости
- между значением Fitness-функции и значением мутации
- между значением Fitness-функции и значением плодовитости

Коэффициенты корреляции:

Алгоритм с мутацией

R(F,M)=-0,13 очень слабая отрицательная корреляция
R(F,Pl)=-0,23 слабая отрицательная корреляция
R(V,M)=0,24 слабая положительная корреляция
R(V,Pl)=-0,67 средняя положительная корреляция

Алгоритм с элитой

R(F,M)=-0,15 очень слабая отрицательная корреляция
R(F,Pl)=-0,42 слабая отрицательная корреляция

R(V,M)=0,15 очень слабая положительная корреляция
R(V,Pl)=-0,82 сильная отрицательная корреляция

Для оценки эффективности применения адаптивного ГА используется некоторая мера качества классификации, которая является функционалом. Наилучшим по выбранному функционалу следует считать такое разбиение, при котором достигается его экстремальное (минимальное или максимальное) значение.

Для оценки эффективности применения адаптивных режимов работы системы были рассчитаны функционалы качества для алгоритма без режима управления сходимостью и с применением режима.

В основе многих характеристик лежит формула:

Vtot=Vint+Vout

$$\sum_{i=1}^{n}\left|x_i-\bar{x}\right|^2=\sum_{l=1}^{k}\sum_{x_i\in S_l}\left|x_i-\bar{x}_l\right|^2+\sum_{l=1}^{k}n_l\left|\bar{x}_l-\bar{x}\right|^2,$$

где Vtot- момент инерции относительно общего центра масс всех объектов;

Vint- сумма моментов инерции Il классов Sl относительно их центров масс;

Vout-межклассовый разброс.

Наиболее распространенными являются следующие функционалы качества:

1. Общая внутриклассовая инерция:

$$F1=\sum_{l=1}^{k}I_l=V_{\text{int}}=\sum_{l=1}^{k}\sum_{x_{ie}\in S_l}\left|x_i-\bar{x}_l\right|^2$$

2. Сумма внутриклассовых дисперсий:

$$\mathrm{F2}=\sum_{\mathrm{l}=1}^{\mathrm{k}}\frac{1}{\mathrm{n}_1}I_l=\sum_{l=1}^{k}\frac{1}{n_l}\sum_{x_i\in S_l}\left|x_i-\bar{x}_l\right|^2$$

3. Сумма попарных внутриклассовых расстояний между элементами:

$$F3=2\sum_{l=1}^{k}n_lI_l=\sum_{l=1}^{k}\sum_{x_i,x_j\in S_l}\left|x_i-x_j\right|^2$$

4. Функционал

$$F4=\left[\frac{1}{n-k}V_{\text{int}}\right]\Big/\left[\frac{1}{k-1}V_{out}\right]$$

Таблица. 6. Результаты работы алгоритма «Управление сходимостью через мутацию»

				Мут=0,1	Мут=3,2	Мут=20	Плод=3	Плод=6
Длина периода стагнации	3	Порог стагнации	0,3	4762,124	4410,108	4436,236	4438,123	5001,816
				30	39	40	31	18
			0,5	4630,988	4678,781	4140,133	4172,599	4387,553
				28	36	40	31	23
			1	4193,236	4193,236	4486,757	4289,903	4240,304
				34	34	40	29	26
			3	4459,48	4654,828	4465,448	4453,749	4625,848
				33	34	38	26	25
			5	4623,646	4481,133	4476,599	4968,698	4324,195
				32	35	39	14	22
	4	Порог стагнации	0,3	4365,924	4776,76	4383,716	4412,286	4382,429
				38	27	39	29	21
			0,5	4551,156	4536,393	4427,673	4297,848	4265,733
				37	32	36	26	19
			1	4413,993	4509,35	4637,946	4873,293	4161,885
				39	29	40	20	25
			3	4517,397	4732,714	4582,119	4214,269	4400,897
				37	32	35	34	37
			5	4569,286	4387,058	4597,736	4395,606	4275,63
				37	39	34	28	20
	5	Порог стагнации	0,3	4507,024	4322,011	4567,274	4358,034	4369,916
				35	40	39	30	22
			0,5	4318,552	5054,674	4410,956	4672,537	4294,345
				37	29	39	22	29
			1	5352,994	4172,327	4494,479	4263,955	4379,558
				11	35	33	31	20
			3	4653,447	4437,943	4588,453	4026,399	4121,9
				37	39	31	32	21
			5	4262,522	4963,854	4583,619	4311,398	4785,838
				39	30	34	30	19

Таблица. 7. Значения функционалов качества

	Алгоритм без адаптации	Алгоритм «Управление сходимостью через оператор мутации»		
		Длина стагнации		
		3	4	5
F1	396396,065	382146	305797	312430
F2	91641,068	75486,8	53829,4	44643,3
F3	5671032	4751348	4710692	5410480
F4	0,2127	0,2093	0,1391	0,14157

Таблица. 8. Результаты работы алгоритма «Управление сходимостью через увеличение «элиты»»

				Мут=0,1	Мут=3,2	Мут=20	Плод=3	Плод=6
Длина периода стагнации	3	Порог стагнации	0,3	4535,12	4795,33	4864,158	4869,356	4481,415
				39	34	36	25	19
			0,5	4635,259	4484,747	4386,715	4508,382	4356,76
				31	34	38	25	26
			1	4258,027	4643,458	4480,642	4720,323	4061,233
				36	33	37	30	25
			3	4918,408	4462,621	4536,114	4337,768	4248,128
				29	33	35	30	19
			5	4554,424	4462,621	4439,799	4568,256	4443,965
				33	33	38	25	24
	4	Порог стагнации	0,3	4730,7	4599,975	4314,228	4371,77	4238,38
				36	36	35	30	23
			0,5	4535,191	4411,554	4458,356	4366,692	4588,89
				39	40	38	26	19
			1	4414,591	4316,376	4464,005	4575,869	4305,279
				33	32	33	31	24
			3	4405,722	4658,286	4568,536	4547,149	4445,515
				38	40	35	24	20
			5	4573,292	4430,016	4451,052	4432,281	4644,245
				40	40	39	28	19
	5	Порог стагнации	0,3	4684,778	4632,305	4589,153	4639,802	4498,193
				33	37	35	26	24
			0,5	4676,576	4772,541	4571,463	4699,391	4441,819
				35	31	33	25	18
			1	4447,128	4399,278	4424,926	4281,897	4219,126
				30	35	38	35	24
			3	4953,032	4373,354	4715,196	4452,483	4559,502
				30	39	35	21	22
			5	4450,471	4698,078	4515,078	4547,239	4198,799
				40	35	35	21	23

Таблица 9. Значения функционалов качества

	Алгоритм без адаптации	Алгоритм «Управление сходимостью через размер элиты»		
		Длина стагнации		
		3	4	5
F1	437028,805	266157	316645	263117
F2	81109,934	43851,2	40911,6	40697,9
F3	6783018	4930380	6026192	5461670
F4	0,2645	0,113	0,14428	0,11088

Проанализировав изменение значений функционалов качества, можно сделать следующие выводы:

Применение алгоритма управления сходимостью через оператор мутации

- ✓ При малой длине стагнации (3) дает улучшение показателя
    - Суммы попарных внутриклассовых расстояний между элементами F1 на 3,6 %
    - Суммы внутриклассовых дисперсий F2 на 17,6 %
    - Общей внутриклассовой инерции F3 на 16,2 %
    - Функционала $F4 = \left[\frac{1}{n-k} V_{\text{int}}\right] \Big/ \left[\frac{1}{k-1} V_{out}\right]$ 1,6 %
- ✓ При средней длине стагнации (4) дает улучшение показателя
    - Суммы попарных внутриклассовых расстояний между элементами F1 на 22,9 %
    - Суммы внутриклассовых дисперсий F2 на 41,3 %
    - Общей внутриклассовой инерции F3 на 16,9 %
    - Функционала $F4 = \left[\frac{1}{n-k} V_{\text{int}}\right] \Big/ \left[\frac{1}{k-1} V_{out}\right]$ 34,6 %
- ✓ При большой длине стагнации (5) дает улучшение показателя
    - Суммы попарных внутриклассовых расстояний между элементами F1 на 21,2 %
    - Суммы внутриклассовых дисперсий F2 на 51,3 %
    - Общей внутриклассовой инерции F3 на 4,6 %
    - Функционала $F4 = \left[\frac{1}{n-k} V_{\text{int}}\right] \Big/ \left[\frac{1}{k-1} V_{out}\right]$ 33,4 %

Применение алгоритма управления сходимостью через размер элиты

- ✓ При малой длине стагнации (3) дает улучшение показателя
    - Суммы попарных внутриклассовых расстояний между элементами F1 на 39,1 %
    - Суммы внутриклассовых дисперсий F2 на 45,9 %
    - Общей внутриклассовой инерции F3 на 27,3 %
    - Функционала $F4 = \left[\frac{1}{n-k} V_{\text{int}}\right] \Big/ \left[\frac{1}{k-1} V_{out}\right]$ 57,3 %
- ✓ При средней длине стагнации (4) дает улучшение показателя
    - Суммы попарных внутриклассовых расстояний между элементами F1 на 27,5 %
    - Суммы внутриклассовых дисперсий F2 на 49,6 %
    - Общей внутриклассовой инерции F3 на 11,2 %

- Функционала $F4 = \left[\frac{1}{n-k} V_{\text{int}}\right] \Big/ \left[\frac{1}{k-1} V_{out}\right]$ 45,6 %

✓ При большой длине стагнации (5) дает улучшение показателя

- Суммы попарных внутриклассовых расстояний между элементами F1 на 39,8 %
- Суммы внутриклассовых дисперсий F2 на 49,8 %
- Общей внутриклассовой инерции F3 на 19,5 %
- Функционала $F4 = \left[\frac{1}{n-k} V_{\text{int}}\right] \Big/ \left[\frac{1}{k-1} V_{out}\right]$ 58,1 %

Для оценки эффективности алгоритмов автоматической кластеризации необходимо сделать вывод о том, насколько близко разбиение массива документации в результате кластеризации к разбиению этого же массива, полученному в результате экспертной классификации. Необходимо объединить в единую структуру разнородные данные различных кластеризаций. Для этого было проведено сравнение результатов генетической кластеризации с результатами кластеризации алгоритмом К-средних.

При анализе результатов автоматической кластеризации, полученных с помощью ГА, и результов автоматической кластеризации, полученных с помощью алгоритма К-средних, были сделаны следующие выводы.

При кластеризации с приоритетом точности поиска (альфа=0,1) при малом количестве кластеров существенных преимуществ одного из алгоритмов не наблюдается.

При увеличении числа кластеров наблюдается существенное преимущество адаптивного генетического алгоритма, что видно из графиков.

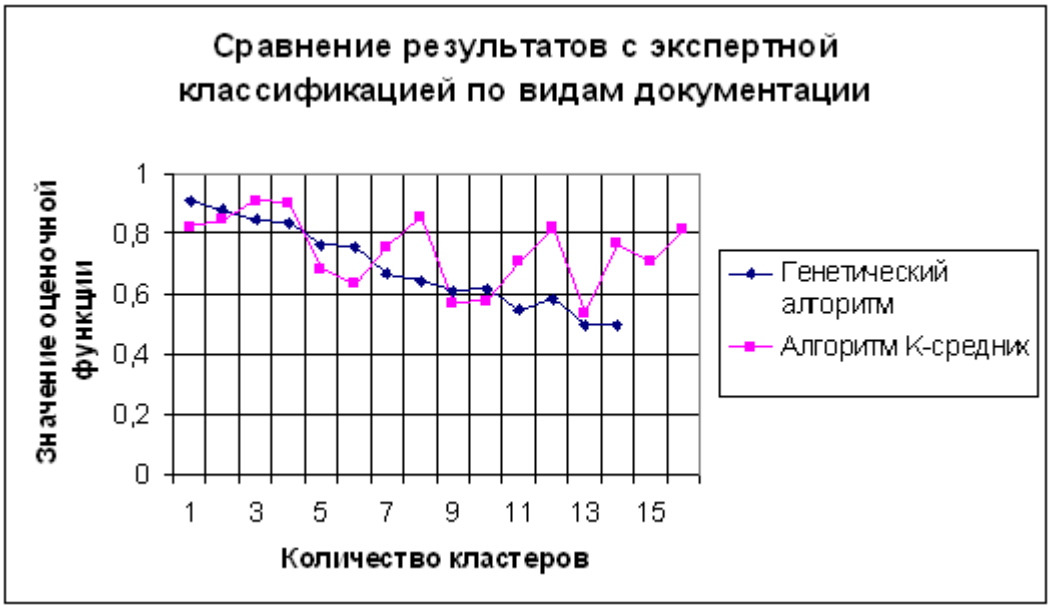

Рис. 11. Сравнение результатов по видам документации

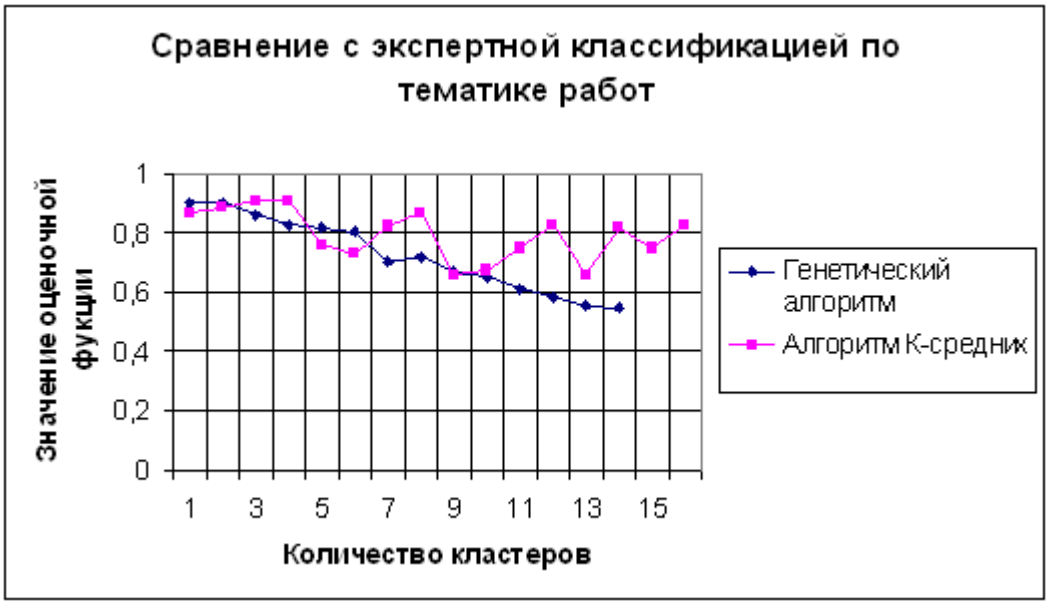

Рис. 12. Сравнение результатов по тематике работ

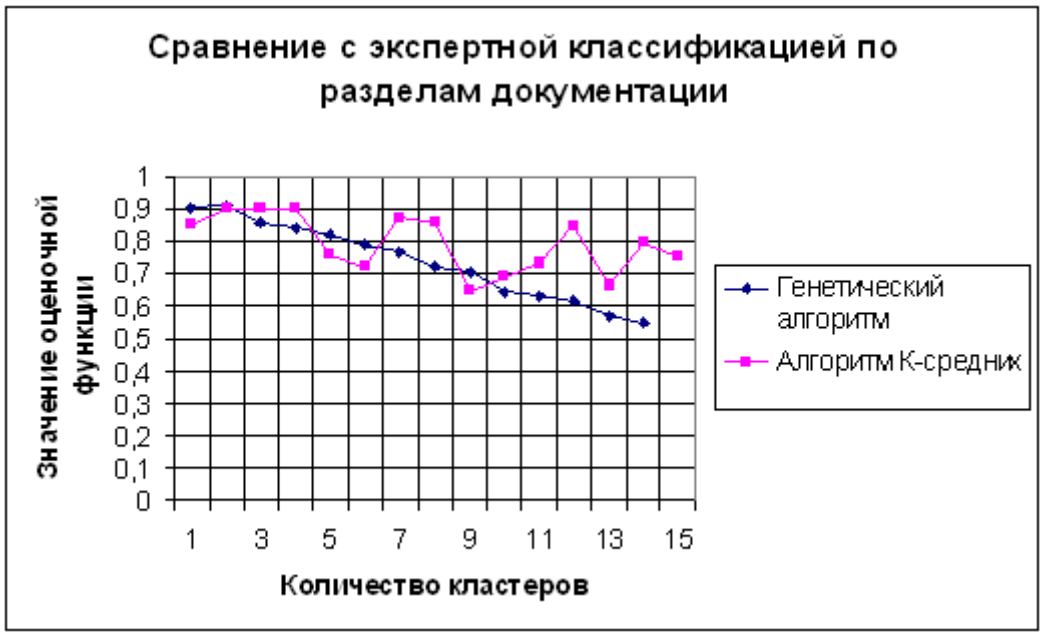

Рис. 13. Сравнение результатов по разделам документации

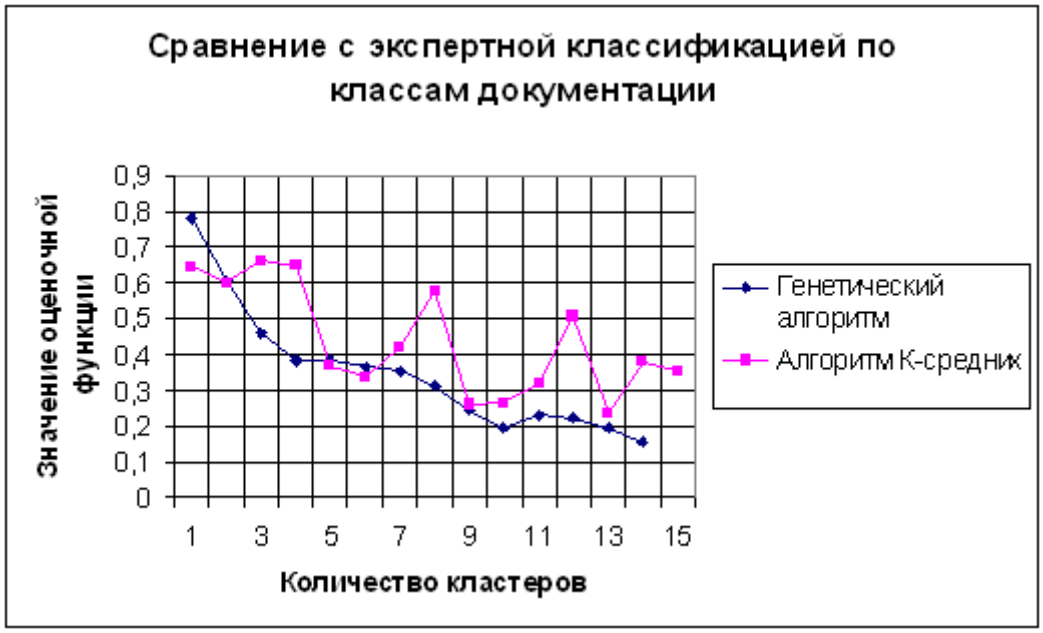

Рис. 14. Сравнение результатов по классам документации

ОСНОВНЫЕ РЕЗУЛЬТАТЫ РАБОТЫ

1. Выполнен научный анализ современных работ, посвященных методам и алгоритмам кластеризации проектных документов. Определены перспективные для исследования методы и алгоритмы эволюционного моделирования, методы генетического поиска.
2. Адаптирована схема генетической оптимизации к прикладной задаче кластеризации проектных документов, как информационных ресурсов, для чего построена мера содержательного сходства проектных документов, как расстояние между ними;
3. Разработаны основные генетические операторы (селекция, кроссинговер, мутация, формирование начальной популяции) применительно к задаче кластеризации проектных документов;
4. Разработан адаптивный алгоритм генетической кластеризации проектных документов, обеспечивающий быструю сходимость решения, среднее сокращение количества проходов генетического алгоритма снизилось в среднем на 30 %;
5. Предложена методика настройки параметров генетической кластеризации, обеспечивающая быструю сходимость и высокое качество решения на основе вычислительных экспериментов;
6. Разработана и реализована программная система генетической кластеризации проектных документов, как базовая часть интеллектуального архива проектной документации
7. Исследована результативность и сходимость генетической оптимизации кластеризации проектных документов с помощью вычислительных экспериментов и внедрения в практику проектной организации.
8. Разработанная программная система внедрена в ФНПЦ ОАО НПО «МАРС» (Ульяновск 2007, 2008), в МУП «УльГЭС» (Ульяновск, 2012).

Список публикаций:

Публикации в изданиях, рекомендованных ВАК

1. Наместников, А.М. Интеллектуальный сетевой архив электронных информационных ресурсов / А.М. Наместников, Н.В. Корунова А.В. Чекина // Программные продукты и системы – 2007. – №4.
2. Ярушкина, Н.Г. Возможности мониторинга динамики развития проекта в интеллектуальном проектном репозотарии / Н.Г. Ярушкина, А.В. Чекина. // Программные продукты и системы – 2008. – №4.
3. Чекина, А.В. Мониторинг динамики развития проекта в интеллектуальном проектном репозитарии / А.В. Чекина // Известия Самарского научного центра Российской академии наук. Специальный выпуск: Четверть века изысканий и экспериментов по созданию уникальных технологий и материалов для авиаракетостроения УНТЦ-ФГУП ВИАМ.– 2008. – Том 4.
4. Наместников, А.М. Интеллектуальный проектный репозитарий / А.М. Наместников, Н.В. Корунова, А.В. Чекина // Известия Самарского научного центра Российской академии наук. Специальный выпуск: Четверть века изысканий и экспериментов по созданию уникальных технологий и материалов для авиаракетостроения УНТЦ-ФГУП ВИАМ. – 2008. – Том 1.
5. Чекина, А.В. Мониторинг динамики развития проекта в интеллектуальном проектном репозитарии /А.В. Чекина // Вопросы современной науки и практики. Университет им. В.И.Вернадского. – 2008. – №4.Том 2.
6. Ярушкина, Н.Г. Кластеризация информационных ресурсов на основе генетического алгоритма / Н.Г. Ярушкина, А.В. Чекина // Автоматизация процессов управления. – 2010. – №4.

В остальных изданиях:

1. Ярушкина, Н.Г. Имитационное моделирование клиент-серверных систем / Н.Г. Ярушкина, А.В. Чекина // Международная конференция «Континуальные алгебраические логики, исчисления и нейроинформатика в науке и технике»: Труды конференции Том 2. Ульяновск: УлГТУ, 2006.
2. Чекина, А.В. Нечеткая временная логика как инструмент мониторинга версий в проектно-конструкторском репозитарии / А.В. Чекина // Международная «Конференции по логике, информатике, науковедению»: Труды конференции. Том 2. Ульяновск: УлГТУ, 2007.
3. Ярушкина, Н.Г. Разработка и реализация интеллектуального репозитария проектной организации / Н.Г. Ярушкина, А.А. Стецко, А.Г. Селяев, Н.В. Корунова, А.В. Чекина // Научная сессия МИФИ-2007: Сборник научных трудов. Том 3. Интеллектуальные системы и технологии. М. – 2007.
4. Chekina, A. INFORMATION WAREHOUSE CONTROL SYSTEMS. ANALYSIS OF PRESENT-DAY REALIZATIONS / A. Chekina,

N. Korunova // Infformation Technologies: Proceeding of Russian-German scientific conference devoted to 10-years cooperation of Ulyanovsk State Technical University and Darmstad University of Applied Science – Ulyanovsk. ULSTU – 2007.
5. Ярушкина, Н.Г. Возможности мониторинга динамики развития проекта в интеллектуальном проектном репозотарии / Н.Г. Ярушкина, А.В. Чекина // Информационные технологии: межвузовский сборник научных трудов // отв. ред. В.В. Шишкин. – Ульяновск: УлГТУ, 2008.
6. Чекина, А.В. Нечеткая темпоральная логика как инструмент мониторинга версий в проектно-конструкторском репозитарии / А.В. Чекина // Научная сессия МИФИ-2008: Сборник научных трудов. Том 10. Интеллектуальные системы и технологии. М. – 2008.
7. Ярушкина, Н.Г. Интеллектуальный проектный репозитарий / Н.Г. Ярушкина, Н.В. Корунова, Ю.А. Родионова, А. Г. Селяев, А.А. Островский, А.В. Чекина // Одиннадцатая национальная конференция по искусственному интеллекту КИИ-2008 с международным участием: Труды конференции. Т. 3. – М.: ЛЕНАНД, 2008.
8. Наместников, А.М. Возможности мониторинга динамики развития проекта в интеллектуальном проектном репозотарии / А.М. Наместников, А.В. Чекина // Одиннадцатая национальная конференция по искусственному интеллекту КИИ-2008 с международным участием: Труды конференции. Т. 3. – М.: ЛЕНАНД, 2008.
9. Нуруллин, А.Ю. Структура и состав Internet интегрированной среды для экспертизы экономического состояния предприятия на основе системы нечеткого вывода / А.Ю. Нуруллин, И.В., Семушин, А.В. Чекина // Одиннадцатая национальная конференция по искусственному интеллекту КИИ-2008 с международным участием: Труды конференции. Т. 3. – М.: ЛЕНАНД, 2008.
10. Ярушкина, Н.Г. Возможности мониторинга динамики развития проекта в интеллектуальном проектном репозитарии / Н.Г. Ярушкина, А.В. Чекина // Нечеткие системы и мягкие вычисления (НСМВ-2008): сборник научных трудов второй всероссийской научной конференции с международным участием. Том 2. – Ульяновск: УлГТУ, 2008.
11. Чекина, А.В. Задачи динамического мониторинга в интеллектуальном проектном репозитарии / А.В. Чекина // Тезисы докладов 43 научно-технической конференции УлГТУ «Вузовская наука в современных условиях». Ульяновск: УлГТУ, 2009.
12. Чекина, А.В. Генетическая кластеризация информационных ресурсов в интеллектуальном проектном репозитарии / А.В. Чекина // Всероссийская конференция с элементами научной школы для молодежи «Проведение научных исследований в области обработки, хранения, передачи и защиты информации» (ОИ-2009): Труды конференции, Ульяновск: УлГТУ, 2009.

13. Чекина, А.В. Алгоритмы кластеризации информационных ресурсов на основе схемы генетической адаптации / А.В. Чекина //Тезисы докладов 44 научно-технической конференции УлГТУ «Вузовская наука в современных условиях». Ульяновск: УлГТУ, 2010.
14. Ярушкина, Н.Г. Кластеризация информационных ресурсов на основе генетического алгоритма / Н.Г. Ярушкина, А.В Чекина. // Интеллектуальный анализ временных рядов: сборник научных трудов семинара с международным участием «Интеллектуальный анализ временных рядов» по результатам НИР, поддержанной ФЦП, проект № 02.740.11.5021. / под ред. Н.Г. Ярушкина. – Ульяновск: УлГТУ, 2010.
15. Ярушкина, Н.Г. Кластеризация информационных ресурсов на основе генетического алгоритма / Н.Г. Ярушкина, А.В. Чекина // Двенадцатая национальная конференция по искусственному интеллекту с международным участием КИИ – 2010: Труды конференции. Том 4.
16. Корунова, Н.В. Интеллектуальный репозиторий проектных документов / Н.В. Корунова, А.М. Наместников, А.А. Островский, Ю.А Родионова, А.В. Чекина, Н.Г. Ярушкина // Двенадцатая национальная конференция по искусственному интеллекту КИИ-2010 с международным участием: Труды конференции. – Т. 2.
17. Чекина, А.В. Генетическая кластеризация информационных ресурсов / А.В. Чекина // Шестая международная научно-практическая конференция «Интегрированные модели и мягкие вычисления в искусственном интеллекте» (Коломна, 16-20 мая 2011г.): Труды конференции – Т.1. – М.: Физматлит, 2011.

Свидетельства:

1. Свидетельство о государственной регистрации программы для ЭВМ №2012612446. Генетическая кластеризация информационных ресурсов / А.В.Чекина. - 6.03.2012 г. – М.: Роспатент, 2012.