УДК 510.63

Н.Г. Ярушкина, А.В. Чекина

КЛАСТЕРИЗАЦИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ НА ОСНОВЕ ГЕНЕТИЧЕСКОГО АЛГОРИТМА

Ярушкина Надежда Глебовна, *доктор технических наук, профессор. Проректор по научной работе, заведующая кафедрой «Информационные системы» Ульяновского государственного технического университета. Имеет статьи и монографии в области интеллектуальных систем нечеткого моделирования. [e-mail: jng@ulstu.ru].*

Чекина Александра Валерьевна, *окончила Ульяновский государственный технический университет, аспирант кафедры «Информационные системы» УлГТУ. Сфера научных интересов – интеллектуальные информационные системы. Имеет статьи по интеллектуальным хранилищам данных. [e-mail: a.sladkova@ulstu.ru].*

Аннотация

В данной статье предложен метод решения задачи кластеризации электронных информационных ресурсов (ЭИР), основанный на генетическом алгоритме (ГА). Все документы проектного репозитария описаны частотными распределениями встречающихся терминов. Исходные данные представлены структурами генетических алгоритмов.

Ключевые слова: информационный ресурс, кластеризация, индексация, генетический алгоритм, кроссовер, функция приспособленности.

Abstract

The article offers a solution method for the task of clustering of electronic information resources on basis of genetic algorithm. All the documents from project repository are described by frequency distributions of met terms. Input data are presented by genetic-algorithm structures.

Keywords: informational resource, clustering, indexing, genetic algorithm, crossover, suitability function.

Введение

Большинство крупных проектных организаций обладают значительным архивом успешных проектов. В новых проектах, как правило, применяются ранее разработанные решения, так как повторность использования позволяет сократить сроки проектирования.

Однако при хранении таких больших объемов информации архивные подразделения сталкиваются с множеством проблем: потеря документов, дублирование, временные затраты на обработку входящей документации. Возникает задача создания проектного репозитария, автоматизирующего процессы классификации имеющихся и вновь поступающих в архив документов, атрибуты которых хранятся в электронной базе данных (БД).

Современный проектный репозитарий должен представлять собой интеллектуальное хранилище информационных ресурсов, чтобы обеспечить поиск необходимого ресурса на основе «гибкого» запроса. Основу индексирования информационных ресурсов традиционно составляет лексический портрет текстового дескриптора ресурса.

Единицей обработки и хранения в репозитарии является информационный ресурс. Информационный ресурс – это 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 [1, 2].

Данная статья посвящена алгоритму автоматической кластеризации электронных информационных ресурсов, основанному на использовании идеи биологической эволюции. Среди эволюционных алгоритмов был выбран традиционный генетический алгоритм для реализации одной из подсистем кластеризации, входящей в состав интеллектуального проектного репозитария.

1 Структурно-функциональная схема интеллектуального хранилища

Программная система, реализующая идеи интеллектуального хранилища, создана в рамках научно-исследовательского проекта по разработке автоматизированной системы «Интеллектуальный сетевой архив электронных информационных ресурсов» (ИСА ЭИР). Применяемость и функциональность ИСА ЭИР протестирована в работе архивной службы ФНПЦ ОАО «НПО «Марс».

Архив крупной проектной организации хранит огромное количество различного рода документов: технические задания, проектно-конструкторскую документацию, инструкции, руководства, стандарты, приказы, распоряжения,

служебные записки и т. д. Так как проект носил исследовательский характер, ставилась задача сравнить эффективность применения различных алгоритмов кластеризации для разных типов документов. Предполагалось, что для разных типов документов могут оказаться наиболее эффективными разные алгоритмы.

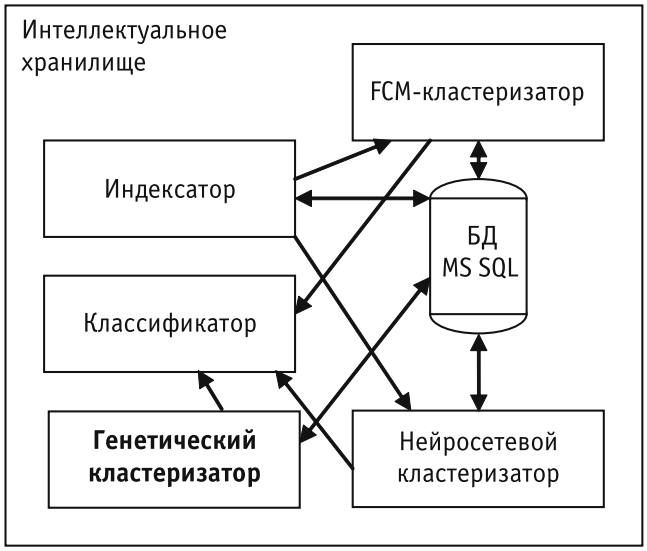

Рис. 1. Структура интеллектуального хранилища

Программа состоит из следующих подсистем (рис. 1):

- подсистема индексирования ЭИР (индексатор),
- подсистема кластеризации ЭИР на основе нейронной сети (нейросетевой кластеризатор),
- подсистема кластеризации на основе метода Fuzzy-C-Means (FCM-кластеризатор),
- классификатор и подсистема кластеризации на основе генетического алгоритма.

На модуль индексации возложены задачи предобработки текстовых документов или аннотаций к ЭИР и построения 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.

1.1 Подсистема индексации

Индексирование документов является важнейшей операцией, обеспечивающей возможности информационного поиска. Сам процесс индексирования документа заключается в определении его центральной темы или предмета на информационно-поисковом языке.

Для оценки значимости слов в индексаторе используются методы определения частоты встречаемости слов в каждом отдельном документе и частоты встречаемости слов, рассчитанных по формуле Шеннона (сигнал-шум):

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

где S^k – сигнал термина, 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.

$$S^k = logF^k - N^k.$$

Данные показатели хранятся вместе со словами в результирующей таблице частоты встречаемости терминов.

Таблица 1

Описание таблиц базы данных, участвующих в процессе индексирования

Таблица	Атрибуты
дескриптор проиндексированного документа	идентификатор ЭИР; путь к ресурсу; расширение файла
частотный словарь терминов в документах	идентификатор ЭИР; термин; частота встречаемости термина в данном ЭИР (тексте) (f_i^k); вес термина в тексте по формуле Шеннона (w_i)
термины, распознанные в процессе индексирования	термин; суммарная частота встречаемости термина по всем ЭИР (F^k)
полный частотный словарь, включая нулевые термы	идентификатор ЭИР; термин; частота встречаемости термина в данном ЭИР (тексте) (f_i^k); вес термина в тексте по формуле Шеннона (w_i)

Значимость подсистемы индексации заключается в том, что результаты ее работы, сохраняемые в вышеуказанных таблицах БД, являются входными данными для работы всех подсистем кластеризации и классификации интеллектуального проектного репозитария. Полный частотный словарь хранит соотношение электронного информационного ресурса (его уникального идентификатора) с частотой встречаемости каждого термина (f_i^k), который в нем используется. Именно эта таблица является ключевой и для генетического кластеризатора.

1.2 Подсистема нейросетевой кластеризации

Для кластеризации применяется нейронная сеть, использующая метод обучения без учителя (unsupervised learning) – самоорганизующие карты Кохонена (Self-Organizing Map – SOM).

Сеть SOM имеет набор входных элементов (частотные портреты текстовых документов, которые необходимо инициализировать из базы данных), и набор выходных элементов (иерархию кластеров), отображающихся в виде дерева результатов. Обучение нейронной сети происходит при анализе каждого документа.

1.3 Подсистема FCM-кластеризации

В качестве входных данных кластеризатор использует набор ресурсов с терминами и весами этих терминов в ЭИР. На выходе должно быть построено дерево кластеров.

Информация об ЭИР и терминах запрашивается из БД MS SQL. После загрузки ресурсов формируется набор уникальных терминов из терминов загруженных ресурсов, который будет характеризовать центры кластеров. Затем пользователь выбирает параметры кластеризации, на основе которых происходит формирование первоначальной матрицы принадлежности ресурсов кластерам.

Метод FCM не является иерархическим. Для иерархической кластеризации требуется участие пользователя.

1.4 Подсистема генетической кластеризации

Для решения задачи кластеризации используется стандартный генетический алгоритм, представляющий собой адаптивный метод поиска, использующийся для функциональной и структурной оптимизации.

Задача кластерного анализа формулируется следующим образом: на основе исходных данных определить такое разбиение $R(A)=\{A_k \mid A_k \subseteq A\}$ множества $A=\{a_1,\ldots,a_n\}$ на заданное число c кластеров, которое доставляет экстремум некоторой целевой функции $f(R(A))$ среди всех разбиений. Все потенциально существующие варианты разбиения множества А на кластеры находятся в пространстве поиска между двумя двумя крайними случаями: все элементы множества А попадают в один кластер *(c = 1)*, каждый элемент множества А отнесен в отдельный кластер *(c = n)*. Критерием пригодности того или иного разбиения ($f(R(A))$) выступает суммарное расстояние от центров кластеров до объектов, входящих в этот кластер.

Таким образом, задачу кластеризации можно рассмотреть как оптимизационную задачу, критерием решения которой служит нахождение экстремума целевой функции.

2 Адаптация стандартного генетического алгоритма к задаче кластеризации ЭИР

Представим задачу кластеризации в терминах эволюционных вычислений. Рассмотрим стандартный генетический алгоритм. Генетические алгоритмы работают с популяцией, каждая из хромосом которой представляет собой возможное решение данной задачи. В нашем случае решение – это разбиение неупорядоченного набора электронных информационных ресурсов на кластеры.

Характерной особенностью использования стандартного ГА для решения практических задач является необходимость адаптации ГА. Как правило, адаптация заключается в уточнении параметров стандартного генетического алгоритма для решения прикладной задачи. В случае кластеризации ЭИР необходимо определить следующие параметры ГА:

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

Хромосома представляет собой массив пар (документ, кластер). Длина хромосомы всегда будет зависеть от того, сколько документов требуется разбить на кластеры. Эта информация представлена в базе данных идентификатором информационного ресурса (документа) и номером кластера. Соответственно, если стоит задача разбить информационные ресурсы на N кластеров, то его значения варьируются от 1 до N.

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-функцией). Наиболее приспособленные особи получают большую возможность участвовать в воспроизводстве потомства.

После того как для каждой хромосомы получено значение fitness-функции, они упорядочиваются в соответствии с его величиной. В самое начало списка попадают те хромосомы, мера приспособленности которых наибольшая, в конец – с наименьшей мерой приспособленности. Далее при выборе потенциальных родителей применяется линейно убывающая функция случайного числа.

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

В качестве оператора рекомбинации используется многоточечный кроссовер. Точка разрыва представляет собой границу между соседними элементами массива (т. е. случайным образом выбирается номер документа). Количество их будет на единицу меньше, чем количество генов в хромосоме или количество кластеризуемых документов. Родительские хромосомы разрываются в этих точках на сегменты. Затем соответствующие сегменты различных родителей склеиваются, и получаются генотипы потомков. При выборе, от какого родителя потомок возьмет следующий ген, предпочтение отдается наиболее приспособленному.

Точка разрыва

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

После стадии кроссовера выполняется операция мутации, которая в данной задаче представляет собой обмен двумя случайными номерами кластеров.

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

Номера документов, для которых значения кластеров меняются местами, выбираются случайным образом.

В результате применения генетических операторов получается хромосома, представляющая собой возможный вариант решения. Для принятия решения об остановке алгоритма необходимо оценить данный вариант решения.

Представим электронный информационный ресурс точкой в n-мерном пространстве терминов.

Индексатор формирует список слов документа по принципу «каждое слово отделяется от другого пробелом». Для каждого термина рассчитывается его вес в данном ЭИР. В базе данных эта информация содержится в соответствующей таблице (частотный словарь терминов в документах), где атрибутами являются идентификатор ЭИР, термин, частота встречаемости термина в данном ЭИР (тексте), вес термина в тексте по формуле Шеннона.

Таким образом, для каждого документа мы можем определить его координату, состоящую из частот встречаемости терминов в ЭИР. Координатными осями в данном случае выступают термины. Число их определяется количеством терминов, по которым проводилось взвешивание, т. е. каждому дескриптору x_i в документе D ставился в соответствие некоторый неотрицательный вес w_i. Документ представлялся точкой в n-мерном пространстве:

$$D = \begin{matrix} x_1 w_1 \\ \dots \\ x_i w_i \\ \dots \\ x_n w_n \end{matrix}$$

В каждый кластер входит определенное количество электронных информационных ресурсов. Для определения центра кластера № 1, предполагаем, что центр – это первый ЭИР. Рассчитываем сумму расстояний от него до всех остальных электронных информационных ресурсов (документов), входящих в кластер № 1. Сохранив полученную величину, предполагаем, что второй ЭИР из кластера № 1 – это центр. Рассчитываем сумму расстояний для него. Сохраняем результат. Проделываем то же самое для каждого ЭИР, представленного в хромосоме и входящего в кластер № 1. Тот документ, для которого сумма расстояний до всех остальных ЭИР выборки будет минимальной, признается центром кластера № 1. Аналогично находятся центры остальных кластеров.

Fitness-функция для каждой хромосомы определяется суммой евклидовых расстояний от каждого ЭИР до центра соответствующего кластера, т. е.

$$f = \sum_{j=1}^{m} \sqrt{\sum_{i=1}^{n} (x_i^{ц} - x_i^{ЭИР})^2},$$

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

$x^{ЭИР}$ – координата i-го ЭИР,

m – количество ЭИР, которое одновременно определяет и длину хромосомы.

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

3 Реализация подсистемы генетической кластеризации

На основе адаптированного генетического алгоритма реализован генетический кластеризатор, представляющий собой отдельный модуль программы «Интерактивный сетевой архив электронных информационных ресурсов» [3], предназначенный для классификации электронных информационных ресурсов, с целью формирования данных для проведения информационного поиска.

В качестве языка программирования приложения был выбран Java. В настоящее время Java является одним из самых распространеных языков для создания кроссплатформенных програмных решений. 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.

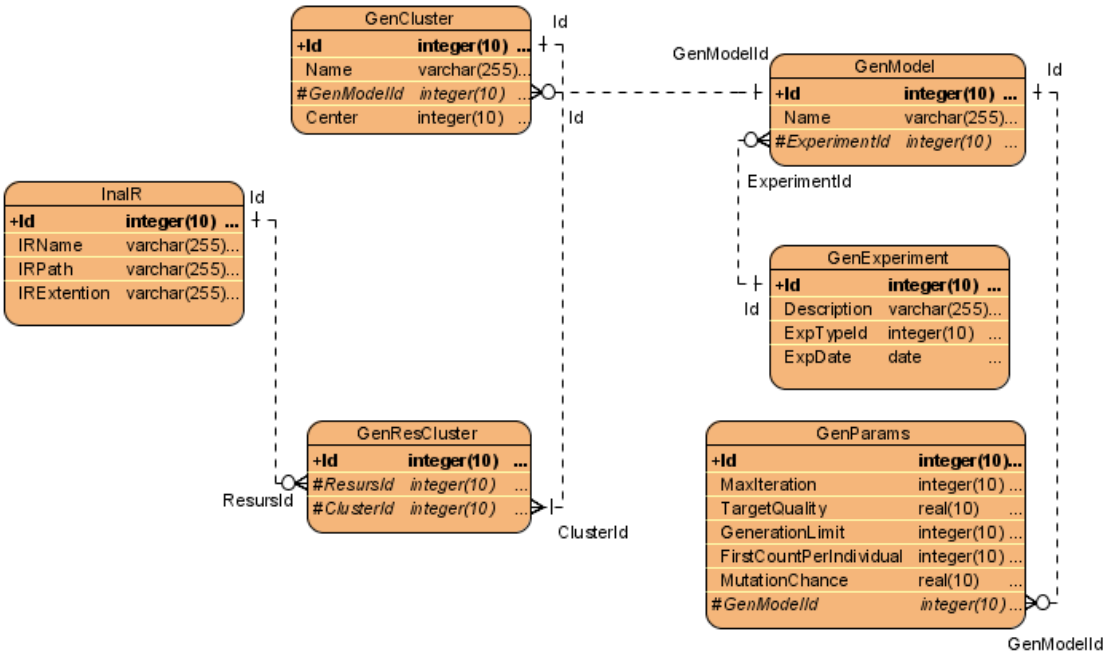

Рис. 2. Модель данных генетического кластеризатора

Таблица 2

Описание таблиц генетического кластеризатора

Таблица	Атрибуты
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 fitness-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

Все данные, полученные в результате кластеризации, хранятся в базе данных и представляются пользователю в диалоговом окне. Структура данных генетического кластеризатора и взаимосвязи между ними представлены на рисунке 2 и таблице 2.

Заключение

Предложенная схема адаптации генетического алгоритма для решения задачи кластеризации электронных информационных ресурсов внедрена в составе интеллектуального проектного репозитария в ФНПЦ ОАО «НПО «Марс» и демонстрирует результативную обработку поступлений новых документов [3].

СПИСОК ЛИТЕРАТУРЫ

1. Ярушкина Н.Г. Основы теории нечетких и гибридных систем. – М.: Финансы и статистика, 2004. – 320 с.

2. Нечеткие гибридные системы. Теория и практика / И.З. Батыршин [и др.] ; под ред. Н. Г. Ярушкиной. – М.: ФИЗМАТЛИТ, 2007. – 208 с.

3. Наместников А.М, Чекина А.В., Корунова Н.В. Интеллектуальный сетевой архив электронных информационных ресурсов // Программные продукты и системы. – 2007. – № 4. – С. 10–13.