УДК 510.63

КЛАСТЕРИЗАЦИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ НА ОСНОВЕ ГЕНЕТИЧЕСКОГО АЛГОРИТМА

Н.Г. Ярушкина (*jng@ulstu.ru*)
Ульяновский государственный технический университет, Ульяновск

А.В. Чекина (*a.sladkova@ulstu.ru*)
Ульяновский государственный технический университет, Ульяновск

В данной статье предложен метод решения задачи кластеризации информационных ресурсов, основанный на генетическом алгоритме. Все документы информационного хранилища описаны частотными распределениями встречающихся терминов.

Введение

Большинство крупных проектных организаций с давней историей обладает значительным архивом успешных проектов. Новые проекты должны использовать ранее разработанные решения, так как повторность использования позволяет сократить сроки проектирования. Современный проектный репозитарий представляет собой интеллектуальное хранилище информационных ресурсов, обеспечивающий поиск необходимого ресурса на основе гибкого запроса [Ярушкина, 2004]. Основу индексирования информационных ресурсов традиционно составляет лексический портрет текстового дескриптора ресурса.

1. Описание объектов кластеризации

Единицей обработки и хранения в репозитарии является электронный информационный ресурс (ЭИР). 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. Следовательно, предполагаем, что по частотному распределению слов,

составляющих текст ресурса (или аннотации), мы можем определить его категорию.

2. Структурно-функциональное решение интеллектуального хранилища

Программная система, реализующая идеи интеллектуального хранилища, состоит из следующих подсистем: подсистема индексирования ЭИР (индексатор), подсистема кластеризации, которая использует три метода: fuzzy-c-means метод (fcm-кластеризатор), нейронную сеть и генетический алгоритм; подсистема поиска ЭИР.

На подсистему индексации возложены задачи предобработки текстовых документов или аннотаций к ЭИР и построение частотных словарей встречающихся терминов. В рамках функций подсистемы кластеризации и классификации, на основе значений относительных частот должны создаваться предметно-ориентированные кластеры, которые организуются в виде иерархии. В процессе классификации выполняется задача соотнесения вновь заносимого ЭИР с определенным кластером.

3. Индексирование документов

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 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 $S^k = \log F^k - N^k$.

4. Подсистема генетической кластеризации

Представим задачу кластеризации ЭИС в терминах эволюционных вычислений. Рассмотрим стандартный генетический алгоритм. Генетические алгоритмы работают с популяцией, каждая их хромосом которой представляет собой возможное решение данной задачи. В нашем

случае решение – это разбиение неупорядоченного набора информационных ресурсов на кластеры

4.1. Адаптация стандартного генетического алгоритма к задаче кластеризации ЭИР.

Для того чтобы применить стандартный генетический алгоритм в качестве метода решения задачи кластеризации, должны быть определены следующие элементы алгоритма [Ярушкина, 2004, Батыршин, 2007]:

- способ кодировки решения (хромосомы);
- функция оптимальности (оценки) каждой хромосомы;
- содержание операторов отбора (селекции), рекомбинации и мутации;
- условие завершения эволюции;
- вероятностные параметры управления сходимостью эволюции.

4.2. Структура хромосомы.

Хромосома представляет собой массив пар (документ, кластер). Длина такого массива всегда будет равна количеству документов, на которое требуется множество ЭИР. Эта информация может быть представлена идентификатором информационного ресурса (документа) и номером кластера. Соответственно, если стоит задача разбить информационные ресурсы на N кластеров, то значения номера кластера 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 …	…

4.3. Селекция.

Каждая хромосома оценивается мерой ее «приспособленности» (*fitness-function*). Наиболее приспособленные особи получают большую возможность участвовать в воспроизводстве потомства. Пропорциональный отбор назначает каждой i-ой хромосоме вероятность $P(i)$, равную отношению ее приспособленности к суммарной приспособленности популяции.

4.4. Кроссовер.

Многоточечный кроссовер в данном случае работает следующим образом. 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. При выборе, от какого же родителя потомок возьмет следующий ген, предпочтение отдается наиболее приспособленному представителю

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

4.5. Оценка приспособленности (*Fitness-function*).

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

Представим 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 информационном ресурсе. 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 числом терминов, по которым проводилось взвешивание, то есть каждому дескриптору x_i в документе *D* ставился в соответствие некоторый неотрицательный вес w_i. Документ представлялся точкой в *n*-мерном пространстве *D*.

$$D = x_1 w_1$$

$$\ldots$$

$$x_i w_i$$

$$\ldots$$

$$x_n w_n$$

В каждый кластер входит какое-то количество информационных ресурсов. Для определения центра кластера №1, предполагаем, что центр - это первый ЭИР. Рассчитываем сумму расстояний от него до всех остальных 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.

Фитнесс - функция для каждой хромосомы определяется суммой евклидовых расстояний от каждого ИР до центра соответствующего кластера. $f = \sum_{j=1}^{m} \sqrt{\sum_{i=1}^{n} (x_i^{u} - x_i^{up})^2}$, где $x^{Ц}$ – координата центра i-го кластера,

$x^{ИР}$ - 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.

4.6. Реализация подсистемы генетической кластеризации

Генетический кластеризатор представляет собой отдельный модуль программы «Интерактивный сетевой архив электронных информационных ресурсов», [Наместников, 2007] предназначенный для классификации электронных информационных ресурсов, с целью формирования данных для проведения информационного поиска. В качестве среды реализации использована Java. - MS SQL Server.

Благодарности. Работа выполнена при финансовой поддержке ФЦП (проект № 02.740.11.5021*)*

Список литературы

[Батыршин, 2007] Батыршин И. З., Недосекин А. О., Стецко А. А., Тарасов В. Б., Язенин А. В.,Ярушкина Н. *Г.* Нечеткие гибридные системы. Теория и практика // Под ред. Н. Г. Ярушкиной. – М.: ФИЗМАТЛИТ, 2007. – 208 с.

[Наместников, 2007] Наместников А.М,, Чекина А.В., Корунова Н.В. Интеллектуальный сетевой архив электронных информационных ресурсов // Программные продукты и системы, №4, 2007, с. 10-13.

[Ярушкина, 2004] Ярушкина Н. Г. Основы теории нечетких и гибридных систем.- М.: Финансы и статистика, 2004. - 320 с.