ОСНОВНЫЕ ШАГИ ГЕНЕТИЧЕСКОГО АЛГОРИТМА ФИЛЬТРАЦИИ РЕЗУЛЬТАТОВ ТЕМАТИЧЕСКОГО ПОИСКА ДОКУМЕНТОВ

Иванов Владимир Константинович
канд. техн. наук, доцент, директор Центра научно-образовательных электронных ресурсов Тверского государственного технического университета, г. Тверь
E-mail: mtivk@mail.ru

BASIC STEPS OF GENETIC ALGORITHM FOR FILTERING OF DOCUMENTS SEARCH RESULTS

Vladimir Ivanov
candidate of Science, Assistant professor, Director of Center of eScience&Learning of Tver State Technical University, Tver

АННОТАЦИЯ

В статье представлены основные особенности предложенного автором подхода к организации поисковых запросов и фильтрации результатов поиска документов, основанного на использовании идей генетических алгоритмов. Описываются основные шаги модифицированного генетического алгоритма, предлагаются решения, учитывающие специфику поисковых процедур. Обсуждаемый подход является частью исследований проекта интеллектуальной системы информационной поддержки инноваций в науке и образовании.

ABSTRACT

The article represents the main features of the proposed approach to organization of search queries and filtering of documents search results. This approach based on the genetic algorithms and describes the main steps of the modified genetic algorithm, proposed solutions, tailored to the search procedures. The considered algorithm is one of the elements of an intelligent system of information support of innovation in science and education.

Ключевые слова: генетический алгоритм, поисковый запрос, релевантность, фильтрация, ранжирование, популяция, скрещивание, мутация, селекция, приспособленность.

Keywords: genetic algorithm, search query, relevancy, filtering, ranking, population, crossover, mutation, selection, fitness.

Введение. Одним из видов поиска в Интернет является поиск достаточной информации о каком-либо объекте или явлении. В отличие от поиска определенных сведений и фактов, касающихся отдельных сторон искомой сущности, решить нетривиальную задачу формулировки поискового запроса для подбора релевантных документов непросто (иногда, может быть, и невозможно). Например, требуется узнать экономические показатели шахты ОАО «Распадская» за первое полугодие 2013 год? Используя эту фразу в качестве поискового запроса, можно получить релевантный ответ в первом десятке результатов поиска Google. Но как подобрать материалы, например, для анализа научно-технических, экономических и социальных факторов влияния на угледобычу в восточных районах России?

Для решения подобных задач пользователи вынуждены применять множество сочетаний ключевых слов и понятий, уточняя их в ходе анализа промежуточных результатов. Неочевидно, что при этом будет строго использована какая-либо обоснованная методика. В ходе поиска неизбежно встает ряд вопросов. Как совместно оценить релевантность документов, найденных разными запросами? Как отфильтровать документы, не относящиеся по сути к искомой тематике? Все ли релевантные документы будут показаны в результатах поиска? Правильно ли определялась релевантность документов? Невозможно однозначно ответить на эти вопросы в рамках тривиальных решений.

В настоящей статье описываются основные особенности предложенного автором подхода к организации поисковых запросов и фильтрации результатов поиска документов, основанного на использовании идей генетических алгоритмов.

Работы проводились при финансовой поддержке РФФИ (договор № НК13-07-00342\13). Они являются частью исследований, касающихся проектных

спецификаций модулей интеллектуальной системы информационной поддержки инноваций в науке и образовании [1].

Постановка задачи. Предположим, что задано множество ключевых слов $K = k$, $K = m$, которые формируют поисковый образ документа (ПОД). Например, $K = \{'генетический\ алгоритм', 'поиск', \dots, 'ранжирование'\}$. Любой поисковый запрос q может быть сформирован, как некоторая совокупность ключевых слов из множества K. Результатом выполнения запроса в какой-либо поисковой системе является некоторое множество адресов найденных документов. Очевидно, что множества результатов различных запросов могут пересекаться.

Задача состоит в том, чтобы из множеств результатов, полученных после выполнения нескольких запросов, выбрать такое множество адресов документов (целевое множество результатов поиска), которое будет наиболее релевантным заданному ПОД.

Можно предположить, что целевое множество результатов поиска с большой вероятностью должно формироваться такими адресами документов, которые (а) находятся в первых позициях ранжированного списка, построенного поисковой системой, и (б) присутствуют в списках результатов, полученных при выполнении всех или большинства запросов.

Опишем подход к решению этой задачи с помощью генетического алгоритма. Отметим, что теория и практика применения генетических алгоритмов — «адаптивных поисковых методов, основанных на селекции лучших элементов в популяции» [2] — в настоящее время является обширным направлением в решении задач оптимизации и моделирования.

Исходная популяция. Предположим, что исходная популяция будет состоять из особей — поисковых запросов. Пусть каждый поисковый запрос образуется парой ключевых слов (генов). В этом случае исходную популяцию из N запросов можно представить множеством $Q = q_i$, $Q = N$, $N < m/2$, $q_i = (k_r, k_s)$, где $(k_r \in K, k_s \in K)$ — случайно образованная пара и $k_r \neq k_s$.

Количество запросов в исходной популяции N является параметром алгоритма и должно быть задано.

Целевая функция. Значения целевой функции должны определять приспособленность особей популяции. В нашем случае приспособленность особи может быть интерпретирована как способность соответствующего поискового запроса сгенерировать такие результаты, которые попадают в следующее поколение популяции.

Пусть f_a — частота появления адреса документа в результатах выполнения N поисковых запросов, вычисляемая следующим образом: $f_a = A_c \ N$, где A_c — количество появлений адреса документа в результатах выполнения N поисковых запросов. Отметим, что $A_c \leq N,\ 0 < f_a \leq 1$, а также $A_c = N$, если адрес документа появился в результатах выполнения всех запросов.

Далее, пусть p_i — позиция в списке первых P результатов выполнения поискового запроса q_i, а $\overline{p}$ — средний номер позиции адреса документов в списке результатов выполненных поисковых запросов, где данный адрес документа присутствует. Тогда $\overline{p} = \frac{1}{n_e} \sum_{i=1}^{n_e} p_i$, где n_e — число запросов, в результатах которых присутствует данный адрес документа.

Для нормировки $\overline{p}$ на диапазон от 0 до 1 можно использовать линейное преобразование $\overline{p}' \ \overline{p} \ = \frac{\overline{p}-\overline{p}_{min}}{\overline{p}_{max}-\overline{p}_{min}}$, где $\overline{p}'$ — нормированное значение $\overline{p}$, $\overline{p}_{min}$ и $\overline{p}_{max}$ — соответственно минимальное и максимальное значения $\overline{p}$ из всех рассчитанных на текущем шаге, $1 \geq \overline{p}' \geq 0$.

Вес каждого результата, полученных после выполнения всех запросов можно вычислить по следующей формуле: $w_i = (\overline{p}' + f_a)\ 2$. Значение w_i определяет позицию p' результата в общем для всех запросов списке результатов.

Значение целевой функции для каждого запроса (приспособленность особи) вычисляется как средний вес результатов запроса $\overline{w} = \frac{1}{P} \sum_{i=1}^{P} w_i$, где P —

количество адресов документов, рассматриваемых как результат поискового запроса. Эта величина является параметром алгоритма и должна быть задана.

Селекция особей. Операция селекции должна обеспечивать участие в формировании следующего поколения только тех особей, у которых значение целевой функции $\overline{w}$ не меньше некоторой пороговой величины. Например, среднего значения $\overline{W}$ по текущей популяции: $\overline{W} = \frac{1}{N} \sum_{i=1}^{N} \overline{w_i}$, где $\overline{w_i}$ — значение целевой функции (приспособляемости особи) для каждого из N запросов.

Выбор родительских пар. Здесь целесообразно использовать метод генотипного аутбридинга. Первая родительская особь выбирается случайно, а второй особью будет являться максимально «далекая» от первой. Расстояние между особями может быть вычислено как $\Delta\overline{w} = \overline{w}_1 - \overline{w}_2$. Такой подход позволит обеспечить максимально полное участие всех текущих запросов в формировании следующего поколения запросов.

Скрещивание. Как указывалось выше, каждая особь (поисковый запрос) популяции состоит из пары ключевых слов, то есть $q_i = (k_r, k_s)$. Отметим, что каждому ключевому слову k_t соответствует множество его синонимов S_t. Для скрещивания будем применять операцию дискретной рекомбинации, которая соответствует обмену генами между особями (в нашем случае обмену ключевыми словами между запросами).

Особенность предлагаемой реализации операции скрещивания состоит в том, что ключевое слово k_t запроса-родителя замещается не ключевым словом другого запроса-родителя, а его синонимом $k_{st} \in S_t$. Это позволяет генерировать существенно больше запросов-потомков при сохранении свойств (семантики) запросов-родителей.

Проиллюстрируем предлагаемую реализацию операции скрещивания на примере. Пусть на предыдущем шаге алгоритма отобраны две родительских пары: $q_1 = (\text{'паровоз'}, \text{'светофор'})$ и $q_2 = (\text{'рельс'}, \text{'сталь'})$.

Пусть для запроса-потомка q_{12} в качестве первого гена отобран второй ген первого запроса-родителя, а в качестве второго гена – первый ген второго

запроса-родителя. Тогда, в результате скрещивания с учетом синонимии можем получить следующий запрос $q_{12} = (\text{'семафор'}, \text{'путь'})$.

Мутация. Суть мутации в рассматриваемом подходе заключается в изменении случайно выбранного гена в особи (ключевого слова запроса). Вероятность мутации p_m может быть фиксированным случайным числом на отрезке [0; 1]. Как правило, $p_m \ll 1$.

Далее, поскольку число ключевых слов в запросе $q_i = (k_r, k_s)$ фиксировано, невозможно применение таких операторов мутации, как: присоединение нового гена, вставка нового гена, удаление гена. Также, обмен местами членов пары (k_r, k_s) в контексте выполнения поисковых запросов лишен смысла.

Поэтому наиболее адекватной операцией мутации можно считать замена ключевого слова в запросе (или гена в особи). В этом случае порядок шагов мутации особи может быть таким:

- Для каждого запроса замена с вероятностью p_{ms} случайного ключевого слова запроса его синонимом.
- Если на предыдущем шаге мутация не произошла, то замена с вероятностью p_{mk} случайного ключевого слова запроса ключевым словом, случайно выбранным из множества K.

Вычисление значений целевой функции. Вычисление значений целевой функции для запросов проиллюстрируем на примере. Пусть исходная популяция включает три запроса: $Q = q_1, q_2, q_3$, а значение $P = 3$.

Результаты выполнения этих запросов могут быть сведены в табл. 1

Таблица 1.

Результаты выполнения этих запросов

p	**Адреса найденных документов**		
	q_1	**q_2**	**q_3**
1	<адрес 1>	<адрес 4>	<адрес 1>
2	<адрес 2>	<адрес 1>	<адрес 4>
3	<адрес 3>	<адрес 2>	<адрес 5>

Результаты расчетов промежуточных параметров целевой функции представлены в табл. 2, а значения целевой функции для запросов – в табл. 3.

Таблица 2.

Промежуточные параметры целевой функции

Адреса документов	A_c	$\overline{p}$	$\overline{p}'$	f_a	$\overline{w}$	p'
<адрес 1>	3	1,33	1	1	1	1
<адрес 2>	2	2,50	0,299	0,667	0,483	3
<адрес 3>	1	3,00	0	0,333	0,167	4
<адрес 4>	2	1,50	0,898	0,667	0,782	2
<адрес 5>	1	3,00	0	0,333	0,167	5

Таблица 3.

Значения целевой функции для запросов

q_1	q_2	q_3
0,550	0,755	0,650

Ясно, что запрос q_1 — наименее приспособленная особь и не будет включена в следующую популяцию.

Предположим далее, что после выполнения операций отбора родителей, скрещивания и мутации был получен запрос q_4, который заменил запрос q_1 в популяции. Результаты выполнения запроса q_4 таковы (позиции p сохранены): <адрес 6>, <адрес 1>, <адрес 2>. Результаты расчетов промежуточных параметров целевой функции на следующем проходе алгоритма представлены в табл. 4, а значения целевой функции для запросов — в табл. 5.

Таблица 4.

Промежуточные параметры целевой функции (2 проход)

Адреса документов	A_c	$\overline{p}$	$\overline{p}'$	f_a	$\overline{w}$	p'
<адрес 1>	3	1,67	0,667	1,000	0,833	1
<адрес 2>	2	3,00	0,000	0,667	0,333	4
<адрес 4>	2	1,50	0,750	0,667	0,708	2
<адрес 5>	1	3,00	0,000	0,333	0,167	5
<адрес 6>	1	1,00	1,000	0,333	0,667	3

Таблица 5.

Значения целевой функции для запросов (2 проход)

q_4	q_2	q_3
0,611	0,625	0,569

Теперь наименее приспособленная особь — запрос q_3, который исключается из числа претендентов на попадание в следующую популяцию.

Формирование новой популяции. Может быть использован элитарный отбор — метод, который не допускает потерю лучших решений. Создается промежуточная популяция, которая включает в себя как родителей, так и их потомков. Из всех членов этой популяции выбираются N с лучшими значениями целевой функции $\overline{w}$. Селектированные запросы войдут в следующее поколение.

Условие остановки алгоритма. В общем случае, условием остановки алгоритма можно считать стабильность популяции. Критерием может быть достижение среднеквадратичным отклонением σ значений целевой функции (пригодности запросов) $\overline{w}_i$ некоторой пороговой величины Δ, задаваемой параметром алгоритма: $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\overline{w}_i - \overline{W})^2}$, $\sigma < \Delta$.

В предельном случае процесс выполнения алгоритма может продолжаться до тех пор, пока значения $\overline{w}_i$ не станут одинаковыми для всех N запросов в популяции.

В частных случаях алгоритм может быть остановлен после получения определенного числа новых популяций, заданного как параметр.

Заключение. Таким образом, в настоящей статье описаны важные, по мнению автора, особенности использования генетического алгоритма для фильтрации результатов тематического поиска документов. Например, при выполнении обзоров источников коммерческой, научно-технической, социальной информации в заданной сфере, подборке материалов для патентных исследований, поиске описаний инновационных решений в отрасли,

определении характеристик новых областей и направлений при бизнес-планировании.

Развивая предложенный подход, можно зафиксировать две ключевых задачи, решение которых позволит применить разработанную технологию в поисковых системах. Во-первых, необходимо произвести настройку (обучение) алгоритма, подобрав эффективные значения его параметров. Во-вторых, выполнить проверку качества алгоритма с использованием современных метрик оценки качества поиска и методик их применения (см., например, [3]).

Список литературы:

1. Иванов В.К. Архитектура интеллектуальной системы информационной поддержки инноваций в науке и образовании / В.К. Иванов, Б.В. Палюх, А.Н. Сотников // Программные продукты и системы. Тверь, 2013. — № 4.
2. Гладков Л.А., Курейчик В.В., Курейчик В.М. Генетические алгоритмы / Под ред. В.М. Курейчика. 2-е изд., испр. и доп. М.: Физматлит, 2006. — 320 с.
3. Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2010 Приложение А. Официальные метрики / М. Агеев, И. Кураленок, И. Некрестьянов. Казань, 2010. — С. 172—187. [Электронный ресурс]. — Режим доступа — URL: http://romip.ru/romip2010/20_appendix_a_metrics.pdf