ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION-COMMUNICATION TECHNOLOGIES

УДК 004.853 **DOI: 10.214151/1561-5405-2017-22-5-471-477**

Применение генетического алгоритма для повышения качества работы поисковых систем

И.В. Беляев, А.Р. Федоров, Л.Г. Гагарина

Национальный исследовательский университет «МИЭТ», г. Москва, Россия

af123@yandex.ru

Проблема поиска информации в неструктурированном массиве данных актуальна, так как в неструктурированной информация содержится уникальный потенциал для извлечения новых знаний. Сложность обработки неструктурированных данных определяется их разнообразием, сильной контекстной зависимостью и динамичностью. Объемы хранимых и передаваемых данных увеличиваются с каждым годом. Количество параметров, характеризующих данные, также неизменно растет. Существующие алгоритмы информационно-поисковых систем не предоставляют гибкого функционала для поиска по различным коллекциям документов или веб-страниц. Сложность тематического поиска в заданном сегменте документов связана с необходимостью предварительной настройки параметров математических моделей поисковых систем. Цель настоящей работы – определение значений параметров, которые позволяют повысить релевантность результата поискового запроса.

Рассмотрены использование генетического алгоритма и его работа, операции мутации и кроссинговера, определены вероятностные значения для каждой из операций. Хромосомы в данном исследовании – числовые значения коэффициентов, представленные в двоичном виде. По результатам работы генетического алгоритма получены значения корректирующих коэффициентов для трех семейств поисковых систем: Apache Lucene, Xapian, Sphinx. На контрольных выборках проведена оценка метрик качества работы каждой из поисковых систем: точность, полнота, аккуратность,
F-мера и ошибки.

В результате применения генетического алгоритма наблюдается увеличение значений метрик от 7 до 15 % и уменьшение ошибки поиска от 15 до 50 %, что подтверждает уместность его использования для повышения корректности работы поисковых систем.

Ключевые слова: генетический алгоритм; документ; мутация; поисковый запрос; популяция; ранжирование; релевантность; тематический поиск; математическая модель.

Для цитирования: Беляев И.В., Федоров А.Р., Гагарина Л.Г. Применение генетического алгоритма для повышения качества работы поисковых систем // Изв. вузов. Электроника. – 2017. – Т. 22. – № 5. – С. 471–477. DOI: 10.214151/1561-5405-2017-22-5-471-477

Use of Genetic Algorithms for Enhancing Efficiency of Search Systems

I.V. Belyaev, A.R. Fedorov, L.G. Gagarina

National Research University of Electronic Technology, Moscow, Russia

af123@yandex.ru

The problem of search for the necessary information in an unstructured data volume is urgent, because the unstructured information contains a unique potential for the extraction of new knowledge. The complexity of the unstructured data processing is determined by their variety, strong context dependence and dynamic character. The volumes of the stored and transmitted data every year increase. The number of the parameters, characterizing the data, is also steadily increasing. The existing algorithms of the information retrieval systems do not provide flexible functionality for searching through various collections of documents or web pages. The complexity of the subject search in the given document segment is connected with the necessity to pre-configure the parameters of the mathematical models of the search systems. The values of the parameters permitting to improve the relevance of the search query result have been determined. The use of the genetic algorithm and its operation, mutation and crossover operations, and the probabilistic values for each of the operations have been considered. In this study the chromosomes are numerical values of the coefficients represented in a binary form.

Based on the results of the work of the genetic algorithm the coefficients for three families of the search systems have been determined: Apache Lucene, Xapian, Sphinx. On the test samples the performance metrics of each of the search systems have been evaluated: accuracy, completeness, exactness, F-measure and errors.

According to the results of the assessment, the metric values increase from 7% to 15% and the search error is reduced from 15% to 50%.

Keywords: genetic algorithm; document mutation; search query; population; ranking; relevance; subject search; symbolic-form model.

For citation: Belyaev I.V., Fedorov A.R., Gagarina L.G. Use of Genetic Algorithms for Enhancing Efficiency of Search Systems // Proc. of Universities. Electronics. – 2017. – Vol. 22. – № 5. – P. 471–477. DOI: 10.214151/1561-5405-2017-22-5-471-477

Введение. Тематический поиск в заданном сегменте документов в Интернете или в специализированных базах данных – одно из актуальных направлений исследований, несмотря на существование мощных поисковых систем. Это ресурсоемкий процесс, ко-

торый недостаточно поддерживается программно и методологически. В современных поисковых системах учитываются многие внутренние и внешние признаки документов. Отметим, что каждый из признаков тем или иным образом входит в результирующий поисковый алгоритм [1].

Применяемые в популярных поисковых системах алгоритмы разнообразны и эффективны, но при этом имеют ряд недостатков: для математических моделей ранжирования используются табличные значения; набор критериев отбора документов жестко задан; эти алгоритмы сложно применять для настройки более «гибкой» поисковой системы; скорость индексации низкая; работать с базами данных затруднительно.

В существующих поисковых системах алгоритмы не подстраиваются под рассматриваемые документы. Данная проблема актуальна, так как в организациях типы документов быстро меняются, а каждый тип документов необходимо ранжировать по-своему. Такая задача относится к классу *NP*-сложных задач, для решения которых используются эвристические алгоритмы. Поэтому для подбора коэффициентов в математические модели информационно-поисковых систем можно использовать генетические алгоритмы.

Цель настоящей работы – определение наиболее значимых значений коэффициентов для алгоритмов ранжирования поисковых систем с целью улучшения результата поискового запроса.

Характеристики поисковых систем. Наиболее популярными поисковыми системами являются семейства Sphinx, Apache Lucene, Xapian. Рассмотрим воможность улучшения данных поисковых систем с помощью генетического алгоритма.

Семейство для полнотекстового поиска Sphinx – один из самых мощных и быстрых из всех открытых поисковых движков. В процессе ранжирования рассчитывается вес документа по внутренним факторам: алгоритм *BM*25; вес фразы.

*BM*25 представляет собой вещественное число в диапазоне от 0 до 1, которое зависит от частот ключевых слов в текущем документе и в общем наборе документов. Текущая реализация *BM*25 рассчитывается исходя из общей частоты слова в документе, а не только частоты фактических совпадений с запросом [1, 2]:

$$BM\,25 = \frac{\left[\sum_{i=1}^{count\_keywords} \frac{TF_i \cdot IDF_i}{TF_i + k_1}\right]}{2 \cdot count\_keywords} + 0{,}5,$$

где $count\_keywords$ – количество терминов в запросе; TF_i – частота ключевого слова в ранжируемом документе; k_1 – коэффициент, который равен 1,2; IDF_i – обратная частота документов во всей коллекции:

$$IDF_i = \frac{\lg\left(\frac{N - n + 1}{n}\right)}{\lg(1 + N)},$$

где n – количество документов, содержащих i термин; N – общее количество документов в коллекции [1].

Вес фразы (query proximity) не учитывает частоты, но учитывает взаимное расположение ключевых слов в запросе и документе (по умолчанию веса полей равны 1) [1]:

$$doc\_phrase\_weight(query) = \sum_{i=1}^{count\_field} user\_weight(field_i) \cdot LCS(query, field_i),$$

где *count_field* – количество полей, по которым проводится поиск; *user_weight* – вес данного поля, заданный пользователем.

Семейство для полнотекстового поиска Apache Lucene – это библиотека, позволяющая организовать полнотекстовый поиск по множеству документов, т.е. поиск с использованием заданных ключевых слов [2].

Оценка веса документа вычисляется по следующей формуле:

$$score(d) = \left(\sum_{i=0}^{count(ter\min)} tf_i^q \cdot \frac{idf_i^q}{norm_i^q} \cdot tf_i^q \cdot \frac{idf_i^d}{norm_i^{d\_t}} \cdot boost_i^t \right) \cdot coord^{q\_d},$$

$$idf^t = \log\left(\frac{numDocs}{docFreq^t + 1} \right) + 1,$$

$$norm^q = \sqrt{\sum_{i=0}^{count(ter\min)} \left(tf^q \cdot idf^2 \right)^2},$$

где $score(d)$ – оценка документа *d*; tf_i^q tf_i^d – корень квадратный из частоты термина в запросе и в документе соответственно; $norm^{d\_t}$ – квадратный корень из числа символов в документе в том же поле при термине; $boost^t$ – указанное пользователем повышение значимости термина; $coord^{q\_d}$ – фактор, который повышает значимость документов (если они содержат, например, все три термина запроса в документе над теми, которые содержат только два термина); *numDocs* – количество документов в индексе; $docFreq^t$ – количество документов, содержащих термин [1].

Семейство для полнотекстового поиска Xapian – это библиотека поискового движка в открытых исходных кодах, распространяемая по лицензии GPL [1].

Xapian использует традиционную схему взвешивания *BM*25 следующего вида:

$$score(d) = \frac{(k_3 + 1)q}{(k_3 + q)} \cdot \frac{(k_1 + 1)f}{(k_1 L + f)} \cdot \lg \frac{(r + 0{,}5)(N - n - R + r + 0{,}5)}{(n - r + 0{,}5)(R - r + 0{,}5)},$$

где k_1, k_3 – константы; *q*, *f* – частота термина в запросе и в документе соответственно; *N* – общее число документов в коллекции; *n* – число документов в коллекции, содержащих *i*-й термин; *R* – общее количество релевантных документов; *r* – количество релевантных документов в коллекции по термину; *L* – нормализация длины документа.

Факторы $(k_3 + 1)$ и $(k_1 + 1)$ помогают измерить веса таким образом, что первый компонент равен 1, когда $q = 1$, и т.д.

По умолчанию, в Xapian используются следующее значения параметров: $k_1 = 1,\ k_2 = 0,\ k_3 = 1$. Эти значения неоптимальные, их лучше подобрать индивидуально, исходя из документов и запросов в конкретной системе.

Работа генетического алгоритма. Генетический алгоритм – это эвристический алгоритм поиска, который используется для решения задач оптимизации путем случайного подбора, комбинирования и вариации искомых параметров с применением меха-

низмов, аналогичных естественному отбору в природе [3]. Задача формализуется таким образом, чтобы каждая особь k характеризовалась своей хромосомой S_k. Формально хромосома – это цепочка символов (генов), N – длина цепочки. Хромосома определяет приспособленность особи $f_k = f(S_k)$. Цель состоит в том, чтобы максимизировать целевую функцию $f(S_k)$. В данном исследовании наборы коэффициентов – хромосомы, каждый коэффициент представляется в двоичном виде, в записи которого биты – гены.

Из полученного множества решений (поколения) с учетом значения «приспособленности» выбираются решения (обычно лучшие особи имеют большую вероятность быть выбранными), к которым применяются генетические операторы, как правило, это скрещивание и мутация. Результатом является получение новых решений [3]. Для операции кроссинговера берутся две лучшие хромосомы, а для операции мутации – две худшие. Остальные хромосомы в равновероятностном отношении либо мутируют, либо скрещиваются. Этот набор действий повторяется итеративно. Так моделируется «эволюционный процесс», продолжающийся несколько жизненных циклов (поколений), пока не будет найдено глобальное решение, в частности, если разность между значением релевантности алгоритма и значением оценки экспертами не превысит 0,01.

Значение вероятности мутации гена, полученное в ходе эксперимента, в котором оценивалась средняя релевантность документов и не использовалась операция кроссинговера, составляет 40 %. Для операции кроссинговера используются два метода:

1) многоточечный кроссинговер – обмен генами (битами) между хромосомами через одного;

2) деление пополам – половина от одного «родителя», половина от другого (рис.1). Данные методы показали себя наилучшим образом в эксперименте по оценке средней релевантности документа.

Рис.1. Процессы мутации и кроссинговера
Fig.1. Mutation and crossover operations

Оценка качества работы поисковых систем проводилась с помощью таких метрик, как полнота, точность, аккуратность, F-мера, ошибка. Результаты исследования для трех систем представлены на рис.2 и в таблице.

Результаты исследования поисковых систем с помощью метрик
Results of a research of search engines by means of metrics

Поисковая система	Метрики				
	Полнота	Точность	Аккуратность	F-мера	Ошибка
Apache Lucene	0,83 (+9 %)	0,89 (+7 %)	0,95 (+4 %)	0,86 (+10 %)	0,05 (–35 %)
Xapian	0,86 (+12 %)	0,88 (+10 %)	0,92 (+2 %)	0,88 (+13 %)	0,06(–15 %)
Sphinx	0,84 (+15 %)	0,90 (+14 %)	0,94 (+5 %)	0,85 (+14 %)	0,05(–50 %)

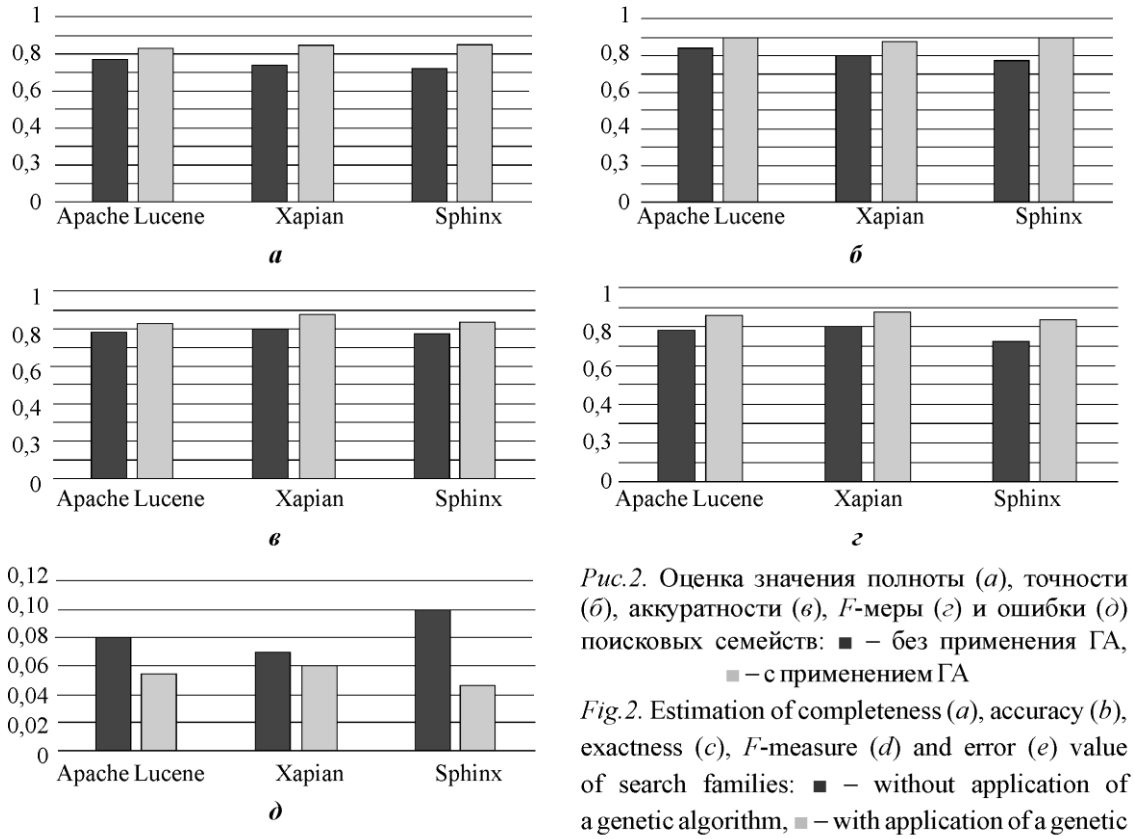

Рис.2. Оценка значения полноты (*а*), точности (*б*), аккуратности (*в*), *F*-меры (*г*) и ошибки (*д*) поисковых семейств: ■ – без применения ГА, ■ – с применением ГА

Fig.2. Estimation of completeness (*a*), accuracy (*b*), exactness (*c*), *F*-measure (*d*) and error (*e*) value of search families: ■ – without application of a genetic algorithm, ■ – with application of a genetic algorithm

Заключение. По результатам оценки полученных алгоритмов по выбранным метрикам наблюдается увеличение значений метрик от 7 до 15 % и уменьшение ошибки поиска от 15 до 50 %. Таким образом, можно сделать вывод о целесообразности дальнейшего использования генетических алгоритмов для улучшения качества поиска.

В дальнейшем планируется исследовать влияние данных алгоритмов на релевантность документов в зависимости от типа поисковых запросов и определить оптимальный алгоритм для целевых решений.

Литература

1. Блог компании «Sphinx Technologies Inc». Как устроено ранжирование. – URL: https://habrahabr.ru/company/sphinx/blog/62287/ (дата обращения: 01.04.2017).

2. WaveAccess. Полнотекстовый поиск с использованием Apache Lucene, 2 сентября 2014. – URL: http://www.waveaccess.ru/blog/2014/september/02/полнотекстовый-поиск-с-использованием-apache-lucene.aspx (дата обращения: 01.04.2017).

3. ***Андреев М.*** Генетический алгоритм. Просто о сложном. – URL: https://habrahabr.ru/post/128704/ (дата обращения: 01.04.2017).

Поступила 04.04.2017 г.; принята к публикации 13.06.2017 г.

Беляев Игорь Валериевич – студент Национального исследовательского университета «МИЭТ» (Россия, 124498, г. Москва, г. Зеленоград, пл. Шокина, д. 1), beligoval@gmail.com

Федоров Алексей Роальдович – кандидат технических наук, доцент кафедры информатики и программного обеспечения вычислительных систем Национального

исследовательского университета «МИЭТ» (Россия, 124498, г. Москва, г. Зеленоград, пл. Шокина, д. 1), af123@yandex.ru

Гагарина Лариса Геннадьевна – доктор технических наук, профессор, заведующая кафедрой информатики и программного обеспечения вычислительных систем Национального исследовательского университета «МИЭТ» (Россия, 124498, г. Москва, г. Зеленоград, пл. Шокина, д. 1), gagar@bk.ru

References

1. *Blog kompanii «Sphinx Technologies Inc». Kak ustroeno ranzhirovanie* [Company Blog «Sphinx Technologies Inc». How is the ranking organized]. Available at: https://habrahabr.ru/company/sphinx/blog/62287/ (accessed: 01.04. 2017). (in Russian).

2. *WaveAccess. Polnotekstovyj poisk s ispol'zovaniem Apache Lucene* [WaveAccess, Full-text search using Apache Lucene]. Available at: http://www.waveaccess.ru/blog/2014/september/02/полнотекстовый-поиск-с-использованием-apache-lucene.aspx (accessed: 01.04. 2017). (in Russian).

3. Andreev M. *Geneticheskij algoritm. Prosto o slozhnom* [Genetic algorithm. Just about the complex]. Available at: https://habrahabr.ru/post/128704/ (accessed: 01.04. 2017). (in Russian).

Submitted 04.04.2017; accepted 13.06.2017.

Belyaev Igor V. – student, National Research University of Electronic Tehnology (Russia, 124498, Moscow, Zelenograd, Shokin sq., 1), beligoval@gmail.com

Fedorov Aleksey R. – phD of technical sciences, associate professor of the Computer Science Department, National Research University of Electronic Tehnology (Russia, 124498, Moscow, Zelenograd, Shokin sq., 1), af123@yandex.ru

Gagarina Larisa G. – Doctor of technical sciences, professor, head of the Computer Science Department, National Research University of Electronic Tehnology (Russia, 124498, Moscow, Zelenograd, Shokin sq., 1), gagar@bk.ru