УДК 004.89;004.048;004.023

Эвристический алгоритм фильтрации и семантического ранжирования результатов поиска документов

Иванов В.К., Виноградова Н.В.

Аннотация

В статье описываются основные идеи предложенного авторами общего алгоритма для фильтрации и ранжирования результатов поиска, выполненного с помощью доступных поисковых систем. При этом используется вычисление показателей близости документов к эталонным множествам, формируемым в ходе выполнения алгоритма.

Рассматриваемый алгоритм является одним из элементов архитектуры интеллектуальной системы информационной поддержки инноваций в науке и образовании. Он будет служить основой модуля поиска этой системы (выполнение поискового запроса для каталога инновационных решений) и модуля уточнения запроса (фильтрация и контроль тематики результатов поиска, уточнение термов поискового запроса). Особенности алгоритма: использует результаты поиска, полученные при штатном применении других поисковых систем; может быть применен с разной степенью автоматизации выполнения различных шагов; инвариантен относительно использования любой поисковой системы; не требует переработки алгоритмов ранжирования поисковой системы, а наоборот, использует их. В статье приведены данные об апробации алгоритма в рамках выполнения патентных исследований - реальной задачи поиска и обработки документов по заданной тематике.

Ключевые слова: алгоритм, векторная модель, качество поиска, мера близости, метрика, поиск решений, ранжирование, релевантность, семантика документа, фильтрация.

UDC 004.89;004.048;004.023

Heuristic algorithm for filtering and semantic ranking of the document search results

Ivanov V.K., Vinogradova N.V.

Abstract

The paper describes the main ideas proposed by the authors of the general algorithm for filtering and ranking search results that are made using the available search engines. It uses the calculation of document similarity indicators to the set of epy standard documents to be formed by the algorithm.

The considered algorithm is one of the elements of an intelligent system of information support of innovation in science and education. It will the basis for the search module of the system architecture (search for innovative solutions directory), and refine query module (filtering and subjects control in search results, refining the search query terms). Features of the algorithm: it uses the search results that obtained in the regular use of other search engines, can be applied with varying degrees of various steps automation, invariant with respect to the use of any search engine, does not require redevelopment search engine ranking algorithms, but rather uses them. The article presents info on the validation of the algorithm in the patent research that is real task for documents searching and processing on a particular topic.

Keywords: algorithm, vector model, search quality, measure of similarity, metric, search for solutions, ranking, relevancy, document semantics, filtering.

Введение

Выполняя поиск в Интернет, мы часто хотим найти определенные сведения, факты, информацию о каком-либо объекте или явлении. Соответственно поисковые запросы, состоящие в подавляющем большинстве случаев из нескольких ключевых слов, включают более или менее точную идентификацию искомого: название, марку, тип, размер и т.п. Но нередко требуется найти информацию, которую невозможно быстро и точно определить несколькими ключевыми словами. Например, выполнить тематические обзоры источников коммерческой, научно-технической, социальной информации, подобрать материалы для патентных исследований, найти описания инновационных решений в отрасли. Здесь в общем случае необходимо определить оценочные критерии, найти (и отобрать из найденного) подходящие альтернативы, выявить потенциальные решения, которые и есть значимый результат информационного поиска. То есть, проще говоря, требуются ответы на вопросы, а не ссылки на источники информации.

Каждый, кто пытался выполнить подобный вид поиска в традиционных или специализированных поисковых системах (см., например, портал для пользователей [6]), неизбежно задавался очевидными вопросами. Все ли релевантные документы показаны в результатах поиска? Все ли потенциально релевантные документы были оценены поисковой системой? Правильно ли определялась релевантность документов? Однозначного ответа на эти вопросы нет.

В настоящей статье описываются основные идеи предложенного авторами общего алгоритма для фильтрации и ранжирования результатов поиска, выполненного с помощью доступных поисковых систем. При этом используется вычисление показателей близости документов к эталонным множествам, формируемым в ходе выполнения алгоритма.

Работы проводились при финансовой поддержке РФФИ (договор № НК13-07-00342\13).

Особенности алгоритма

Рассматриваемый алгоритм является одним из элементов архитектуры интеллектуальной системы информационной поддержки инноваций в науке и образовании [3]. Он будет служить основой модуля поиска (выполнение поискового запроса для каталога инновационных решений) и модуля уточнения запроса (фильтрация и контроль тематики результатов поиска, уточнение термов поискового запроса). Особенности алгоритма следующие:

- Использует результаты, полученные при штатном применении других поисковых систем, работающих как в Интернет, так и со специальными базами данных.
- Может быть применен с разной степенью автоматизации выполнения различных шагов. Например, шаги, использующие модель векторного пространства документов, должны быть полностью автоматизированы, а

основная фильтрация документов может быть частично выполнена экспертом.

- Инвариантен относительно использования любой поисковой системы. Не требует переработки алгоритмов ранжирования поисковой системы, наоборот, использует их.

Описание алгоритма

1. Описание общего запроса Q_0. Выполняется по методике [9].
2. Формирование множества Q поисковых запросов $q \in Q$, $|Q| = N$. Запросы должны определять основные понятия искомого множества документов с помощью ключевых слов и их синонимов и понятий.
3. Создание нижнетреугольной матрицы запросов A_{NxN}, каждый ненулевой элемент которой $a_{ij|i\neq j}$, за исключением элементов главной диагонали, поставлен в соответствие запросу $q_{ij|i\neq j} \in P$, $|P| = N_p$, где $N_p = N(N-1)/2$, $q_{ij} = q_i \cdot q_j$, $1 \le i \le N$, $1 \le j \le N$, $q_i \in Q$, $q_j \in Q$.
4. Формирование множества информационно-поисковых систем (ИПС) S поисковых запросов $s \in S$, $|S| = N_s$.
5. Формирование $k-$паттернов – эталонных текстов, служащих для вычисления мер близости документов. Здесь могут быть сформированы следующие множества $k-$паттернов: P_{kb} (тексты словарных статей из авторитетных источников для каждого из термов запроса) и P_{kd} (описание общего запроса).
6. Выполнение каждого из N_p запроса q_{ij} в каждой поисковой системе s_k из N_s используемых в процедуре поиска документов.
7. Сохранение исходной группы результатов поиска R_{qs} по каждому запросу q_{ij} в каждой поисковой системе s_k. Исходная группа содержит документы, в R первых позициях списка результатов.
8. Предварительная фильтрация результатов поиска в каждой R_{qs}: исключение из результатов ресурсов, зараженных вирусами, а также несуществующих ресурсов.
9. Основная фильтрация документов в каждой R_{qs}. Исключаются документы, тематика которых формально является релевантной, но по некоторым причинам не должна быть предметом поиска. Может быть выполнена как вручную, так и с помощью классификатора, обучающее множество для которого актуализируется в ходе анализа найденных текстов. Примеры фильтруемых документов: определения словарей, глоссариев и справочников; учебные пособия; контрольные работы и рефераты студентов; учебные программы, лекции, тесты, билеты, шпаргалки для студентов и школьников; материалы для раскрутки сайтов; сайты компаний, сайты для закупок, социальные сайты; блоги; рекламные объявления и др.
10. Формирование исходной группы результатов запроса $R_q = \bigcup_{s=1}^{N_s} R_{qs}$ и далее основной группы результатов $R = \bigcup_{q=1}^{N_p} \bigcup_{s=1}^{N_s} R_{qs}$, $|R| = N_r$. Одновременно выполняется удаление результатов-дублей.

11. Формирование второй группы $k-$паттернов. Здесь могут быть сформированы следующие множества k $-$паттернов: P_{ka} (объединение текстов из первых R_p ранжированных ИПС результатов поиска) и P_{kc} (первый из наиболее пертинентных результатов поиска в ИПС).
12. Вычисление значение общих k $-$паттернов P_a, P_b и P_c для всех запросов. Очевидный вариант - средневзвешенное значение нормированных значений $P_{abc} = \sum_{q=1}^{N_p} w_{abc(q)} P_{abc(q)}$, где $P_{abc(q)}$ и $w_{abc(q)}$ – соответственно $k-$паттерн и вес для запроса q_{ij}. Веса $w_{abc(q)}$ могут быть назначены экспертом при формировании множества Q.
13. Вычисление векторов $\overline{v}(d)$ для документов из множества R и $k-$паттернов P_a, P_b, P_c и P_d. Здесь в качестве универсального подхода предлагается использование модели векторного пространства (VSM) документов [2], которая является фундаментальной для многих задач информационного поиска, включая ранжирование документов, фильтрацию данных, классификацию и кластеризацию документов [5]. Каждый документ d интерпретируется как вектор $\overline{v}(d) = (w_{1,d}, w_{2,d}, \dots, w_{1,N_r})$, где $w_{t,d} = tf_{t,d} * idf_{t,d}$. Здесь: $tf_{t,d}$ - частота использования терма в документе, а $idf_{t,d}$ - величина, обратная числу документов, содержащих данный терм, которая вычисляется следующим образом: $idf_{t,d} = \log D_c/D_t$, где D_c - общее число документов в корпусе, а D_t - число документов, содержащих данный терм.
14. Формирование $M_{N_r \text{x} 4}$ - матрицы семантического сходства документов из множества R с общими k $-$паттернами P_a, P_b, P_c и P_d. Сходство двух документов d_1 и d_2 здесь интерпретируется как косинусная мера близости $Sim(d_1 d_2) = (\overline{v}(d_1) \cdot \overline{v}(d_2))/(\|\overline{v}(d_1)\| \cdot \|\overline{v}(d_2)\|)$, где в числителе скалярное произведение векторов документов $\overline{v}(d_1)$ и $\overline{v}(d_2)$, а в знаменателе – произведение евклидовых норм этих векторов.
15. Ранжирование документов из R в соответствие с их мерой близости $Sim(d_1 d_2)$ к общим $k-$паттернам. Выполняется как упорядочение элементов R по значению $Sim(d_1 d_2)$.

Схема алгоритма графически представлена на рис. 1.

Апробация алгоритма

Ниже приведены предварительные результаты апробации алгоритма, проведенной на реальных данных в ходе выполнения патентного исследования по тематике проекта, в рамках которого и был разработан обсуждаемый алгоритм. Цель этого патентного исследования - найти аналоги проектируемой системы и установить степени патентной чистоты и научной новизны.

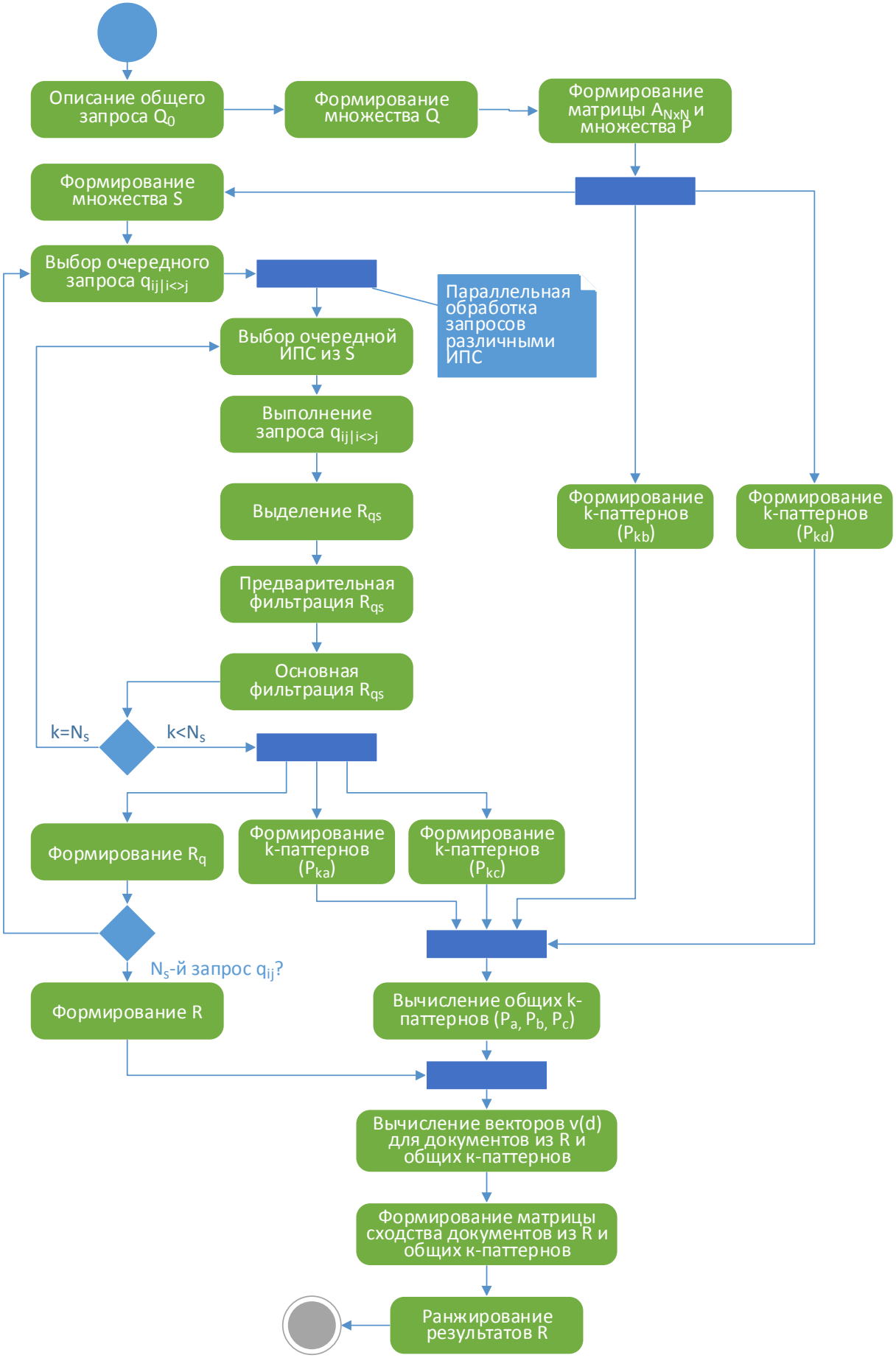

Рис. 1. Схема алгоритма

Алгоритм исследовался на примере общего запроса Q_0, сформулированного следующим образом.

Тема - фильтрация, классификация и упорядочение результатов поиска в Интернет и патентных базах данных. Цель - сбор информации по теме, включая ссылки на ресурсы. Нет единого идеального ответа. Релевантны(+) - статьи о подходах, методах и реализациях задач уточнения запросов, фильтрации и классификации результатов поиска; описания соответствующих программных продуктов, результаты экспериментальных исследований алгоритмов. Релевантны(–) - обзоры методов интеллектуального анализа текстовых данных, исследования семантики связей между объектами. Не релевантны – учебные курсы по теме, рекламные материалы по продвижению сайтов, исследования баз данных.

Множество Q: «уточнение поискового запроса», «классификация результатов поиска», «интеллектуальный анализ результатов поиска», «фильтрация результатов поиска», «семантика связей между объектами», «определение подмножества семантически связанных данных», «отбор результатов поиска», «контроль тематики результатов поиска». Пример запроса $q_{ij|i\neq j}$(элемента матрицы запросов A_{NxN}) - «классификация результатов поиска семантика связей между объектами.

В качестве ИПС были использованы Google и ИПС Роспатента [10]. После выполнения всех запросов было отобрано 1800 документов. Результат предварительной и основной фильтрации - 623 оставшихся документа. После удаления дублей и вторичного анализа было зафиксировано финальное количество документов в основной группе N_r = 217.

Вычисление значение мер близости $Sim(d_1 d_2)$ документов основной группы R и $k-$паттернов P_a, P_b, P_c и P_d было выполнено с использованием реализации VSM, описанной в [4]. Результаты вычислений представлены на рис. 2. Отметим, что очевидная нелинейность изменения $Sim(d_1 d_2)$ для P_d объясняется коротким текстом паттерна.

Оценка качества ранжирования

Для оценки качества ранжирования, выполненного в соответствие с алгоритмом, использовалась метрика DCG (discounted cumulative gain) [1] в модификации, описанной в [8]. Для документов, упорядоченных в соответствие с ранее рассчитанными значениями $Sim(d_1 d_2)$ с каждым $k-$паттерном, вычислялись значения:

$DCG = \sum_{p=1}^{n} 2^{grade(p)} - 1/\log_2(2+p)$, где

$grade(p)$ - средняя экспертная оценка релевантности, выставленная документу, расположенному на позиции p в списке результатов, $grade \in [0,3]$, причем 3 означает «релевантный», 0 – «нерелевантный», 1 и 2 – «частично релевантный» («релевантный(+)» или «релевантный(-)»); $1/\log_2(2+p)$ - дисконт за позицию документа (документы в начале списка имеют больший вес).

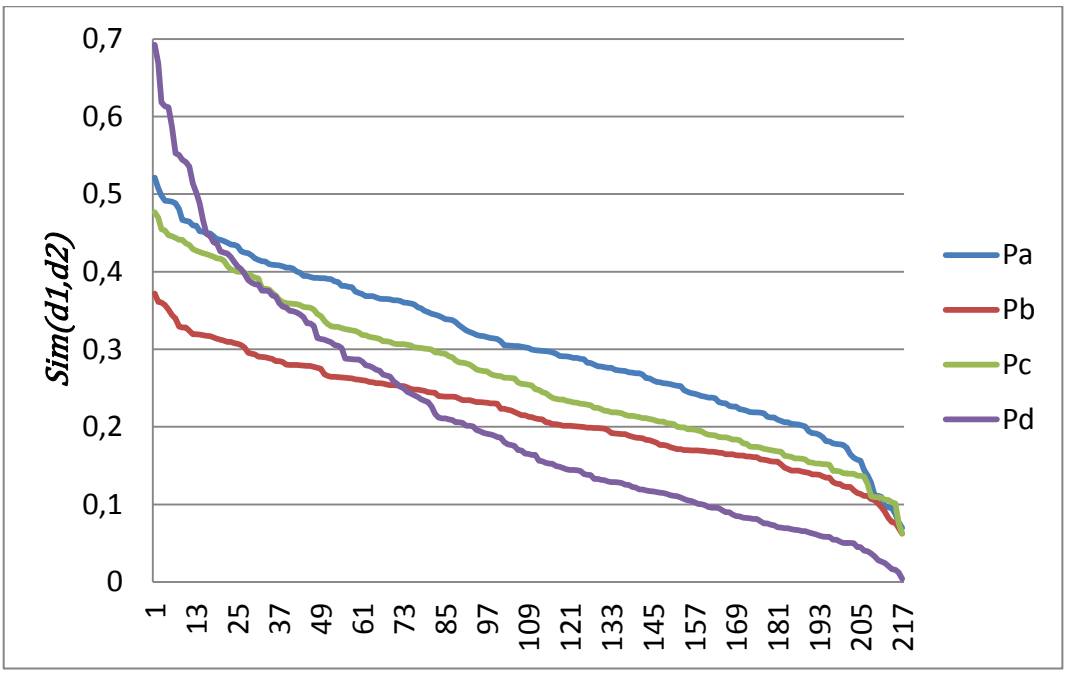

Рис. 2. Результаты вычисления мер близости документов основной группы и k −паттернов

На рис. 3 изображено соотношение метрик DCG идеального ранжирования и различных k −паттернов. Видно хорошее совпадение значений метрик для различных паттернов. В то же время есть резервы для более точного отнесения документов к группам релевантности $grade$. Рассматриваемый алгоритм до 10-15% документов относит к группе со значением $grade$, отличным от идеального ранжирования (см. пики в точках излома графика).

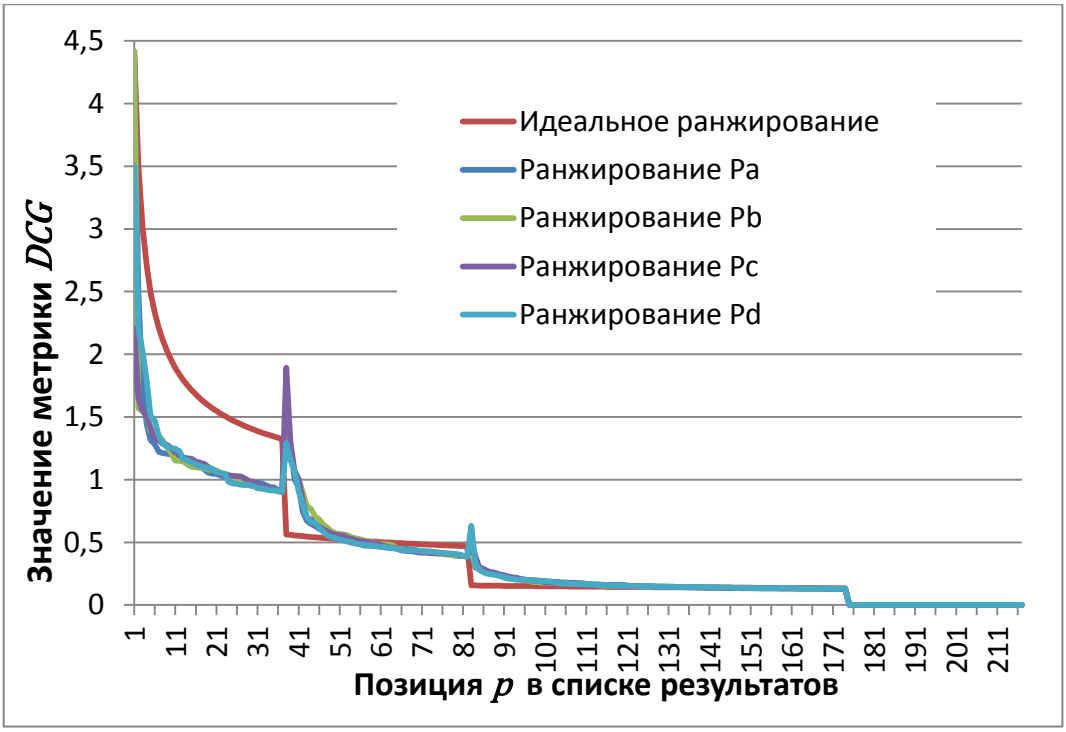

Рис. 3. Соотношение метрик идеального ранжирования и различных k −паттернов

В заключение для каждого k −паттерна были вычислены нормализованные значения $NDCG = DCG/Z$, где Z - фактор нормализации, равный максимально возможному значению DCG, то есть DCG для случая идеального ранжирования в соответствие с оценками эксперта. Показатель $NDCG$ принимает значения от 0 до 1. Соотношение значений $NDCG$ для общего запроса Q_0 и различных k −паттернов представлено на диаграмме (рис. 4). Видно, что наилучшие результаты алгоритм показал при использовании k −паттерна P_a.

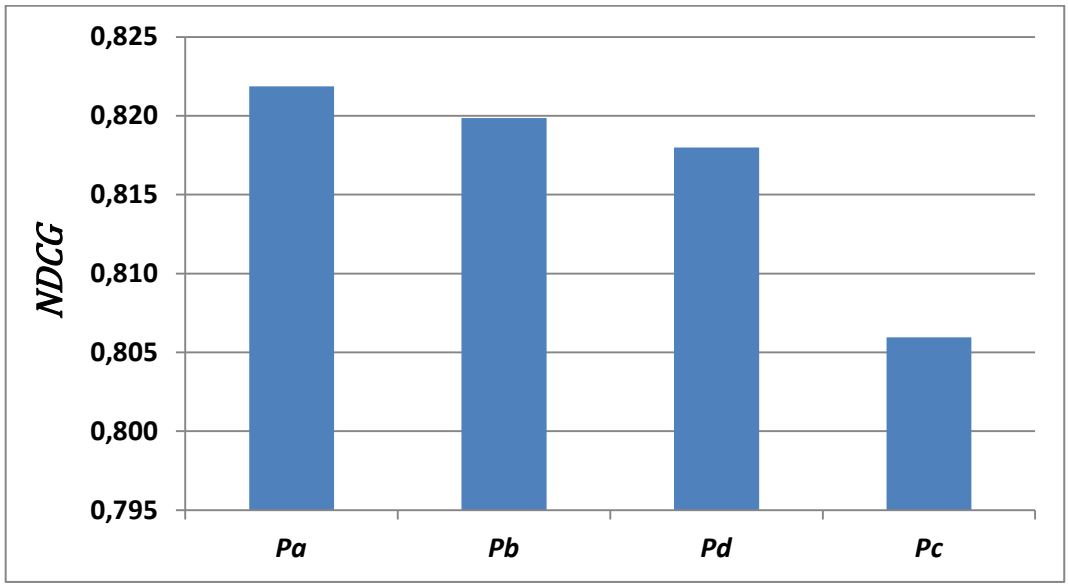

Рис. 4. Соотношение значений $NDCG$ для различных k −паттернов

Выводы и перспективы

Итак, выше определены основные идеи, исходные данные и шаги выполнения эвристического алгоритма фильтрации и семантического ранжирования результатов поиска документов. Приведены данные об апробации алгоритма в рамках выполнения реальной задачи поиска и обработки документов по заданной тематике.

Отметим, что некоторые фрагменты алгоритма естественным образом требуют дальнейшего развития.

Например, формирование множества Q поисковых запросов из описания общего запроса Q_0 может быть автоматизировано генетическим алгоритмом, оптимизирующим суммарную релевантность (или вес) результирующей выборки документов при заданных ограничениях на количество выполняемых операций (глубину эволюционного процесса). Следует также обратить внимание на процедуру формирования k −паттернов. По-видимому, здесь требуются дополнительные экспериментальные исследования, включающие проверку и сравнение различных способов получения эталонных текстов. Также требует уточнения технология основной фильтрации документов – должны быть отработаны процедуры автоматизированного выбора и обучения классификационных алгоритмов. Здесь возможно использование поискового агента, обеспечивающего функции «информированного» поиска в оперативном

режиме, включая обратную связь с пользователем алгоритма (фундаментальные идеи таких агентов см. в [7]).

И, наконец, алгоритм должен быть встроен в проектные спецификации модулей поиска и уточнения запроса интеллектуальной системы информационной поддержки инноваций в науке и образовании [3].

Библиографический список

1. Järvelin, K. Cumulated gain-based evaluation of IR techniques [Electronic resource] / K. Järvelin, J. Kekäläinen // ACM Transactions on Information Systems (TOIS) TOIS. - 2002. - Vol. 20, issue 4, October 2002. - P. 422-446. - Режим доступа: http://doi.acm.org/10.1145/582415.582418.

2. Salton, G. A Vector Space Model for Automatic Indexing [Electronic resource] / G. Salton, A. Wong, C.S. Yang // Communications of the ACM. - 1975. - Vol. 18, nr. 11. - P. 613–620. - Режим доступа: http://dl.acm.org/citation.cfm?id=361220.

3. 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.

4. Иванов, В.К. Критерии интегральной оценки электронных документов в системах подготовки принятия решений // Вестник Тверского государственного технического университета : науч. рецензир. журнал. - Вып.22. - Тверь, 2012. - С. 20-26.

5. Маннинг, К.Д. Введение в информационный поиск : пер. с англ. / К.Д. Маннинг, П. Рагхаван, Х. Шютце. - М. [и др.] : Вильямс, 2011. - 528 с.

6. About.com. Web search. Портал для пользователей поисковых систем [Электронный ресурс]. – Режим доступа: http://websearch.about.com/od/enginesanddirectories/a/searchengine.htm.

7. Рассел, С. Искусственный интеллект: современный подход : пер с англ. - 2-е изд. / С. Рассел, П. Норвиг. – М. [и др.] : Вильямс, 2006. – 1408 с.

8. Российский семинар по Оценке Методов Информационного Поиска (2010; Казань). Труды РОМИП 2010 [Электронный ресурс]. Приложение А. Официальные метрики / М. Агеев, И. Кураленок, И. Некрестьянов. - Казань, 2010. - С. 172-187. - Режим доступа: http://romip.ru/romip2010/20_appendix_a_metrics.pdf

9. Российский семинар по Оценке Методов Информационного Поиска (2010; Казань). Труды РОМИП 2010 [Электронный ресурс]. Приложение В. Инструкция для асессора для дорожки поиска по Веб-коллекции. - Казань, 2010. – С. 188-192. - Режим доступа: http://romip.ru/romip2010/21_appendix_B_WA.pdf.

10. Сайт ФИПС (Федеральное государственное бюджетное учреждение «Федеральный институт промышленной собственности»). Информационно-поисковая система [Электронный ресурс]. – Режим доступа: http://www1.fips.ru/wps/wcm/connect/content_ru/ru/inform_resources/inform_retrieval_system.