УДК 004; 025.4
DOI:10.18413/2411-3808-2018-45-1-176-183

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ПОДСИСТЕМЫ ПОИСКА И РАНЖИРОВАНИЯ ДОКУМЕНТОВ В ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМАХ

MATHEMATICAL MODEL OF SUBSYSTEM OF DOCUMENT RETRIEVAL AND RANKING IN INFORMATION-RETRIEVAL SYSTEMS

Н.Н. Кучукова[1], Н.А. Вершков[2]
N.N. Kuchukova[1], N.A. Vershkov[2]

[1)] ФГАОУ ВО «Северо-Кавказский федеральный университет»
Россия, 355009, г. Ставрополь, ул. Пушкина, 1
[2)] ФЭС-Агро, Россия, 355003, г. Ставрополь, ул. Дзержинского, 162

[1)] FSAEI HE «North-Caucasus Federal University», 1 Pushkin St, Stavropol, 355009, Russia
[2)] FES-Agro, 162 Dzerzhinsky St, Stavropol, 355003, Russia

E-mail: knn.storage@yandex.ru, vernick61@yandex.ru

Аннотация
В статье проведен анализ существующих моделей поиска и ранжирования документов в информационно-поисковых системах (ИПС), указаны основные параметры эффективности ИПС, поставлена задача построения математической модели поиска и сортировки документов с учетом взаимозависимости термов в текстах, рассмотрена возможность применения аппарата корреляции для определения меры схожести поискового запроса и документа и уменьшения вычислительных затрат в ИПС, а также использования показателя корреляции как единого обобщенного показателя ИПС для установления степени схожести поискового запроса и документов, проведения сортировки по релевантности, исключения документов-дубликатов из результатов поиска.

Abstract
The article proposes the problem of fast and accurate information retrieval on the user's request in information retrieval systems. There is analysis of the existing models of searching and ranking documents and specification of the main performance parameters of information retrieval systems. It is revealed that modern search engines use a large number of parameters, according to which the completeness of the search and the relevance of the data obtained are estimated. This leads to an increase in computing and hardware costs. We offered the mathematical models of search and ranking documents in information retrieval systems, given the interdependence of terms in texts, the possibility of application of the apparatus of correlation to measure the similarity between a search query and document, and reduce computational costs. Also it is offered to use metric correlation as a single generalized indicator of information retrieval systems to establish the degree of similarity between the search query and documents, sort by relevance, with the exception of documents-duplicates from the search results.

Ключевые слова: информационно-поисковая система, модель поиска, сортировка, ранжирование, корреляция, коэффициент корреляции.
Keywords: information retrieval system retrieval model, sorting, ranking, correlation, correlation coefficient.

Введение

Бурное развитие корпоративных информационных систем (КИС), связанное со значительным ростом систем документооборота, широкое использование сети Интернет в повседневной деятельности предприятий значительно стимулирует развитие информационно-поисковых систем (ИПС). И если КИС еще предполагает определенное

структурирование данных и предварительное обучение пользователей применению корпоративных поисковых систем, то бурный рост Интернета и невысокий уровень подготовки пользователей требует от производителей ИПС значительных усилий для построения алгоритмов поиска и ранжирования. А учитывая жесткую конкуренцию между отдельными коммерческими ИПС, от разработчиков требуется неимоверная изобретательность, чтобы многомиллионная армия пользователей Интернет отдала предпочтение именно этому «поисковику», а не конкуренту. При этом остается удивительным факт, что 3 из 5 ИПС функционируют без использования математических моделей [Сегалович, 2002]. Хотя этот факт не относится лидерам рынка ИПС, тем не менее, алгоритмы, применяемые для построения ИПС в глобальной сети, также не отличаются значительным разнообразием: прямой поиск, использование инвертированных файлов, суффиксные деревья, сигнатуры, латентно-семантическое индексирование, применение лингвистических алгоритмов, нейронные сети и т.п. [Salton, 1979; Van Rijsbergen, 1979] Жесткая конкуренция между коммерческими ИПС требует постоянно повышать эффективность поиска, а именно: производительность, точность и полноту. И если производительность можно повысить за счет аппаратного обеспечения (увеличение количества процессоров, объема оперативной памяти, пропускной способности сети и т.п.), то точность и полнота поиска достигается за счет применения новых, модифицированных алгоритмов поиска и ранжирования [Сэлтон, 1973]. Кроме того, важнейшим показателем является количество дублированных документов. Количество дублированных документов в Интернете огромно: они могут отличаться кодировкой, рекламными вставками, исправлениями. Для решения этой задачи применяют механизм шинглов. Однако данный механизм требует дальнейшего роста затрат аппаратной части для хранения контрольных сумм. Таким образом, ИПС постоянно нуждаются в совершенствовании алгоритмов, поиске новых унифицированных показателей релевантности документов, распараллеливании работы подсистем.

1. Анализ существующих алгоритмов поиска и ранжирования

Как понятно из введения, современное состояние ИПС определяют, в первую очередь, программисты. Математический аппарат используют обычно на уровне операций над множествами, векторной алгебры или теории вероятностей. Этим и объясняется невысокий интерес математиков к моделям ИПС, а предлагаемые усовершенствования вышеперечисленных алгоритмов обычно сводятся к лингвистическим дополнениям. На практике обычно различают булеву, векторно-пространственную и вероятностную модели поиска [Тявкин и др., 2008; Хорошко, 2014; Шарапов, 2007; Baeza-Yates, Ribeiro-Neto, 1999; Salton et al., 1983]. Для построения модели поиска обычно используют следующий понятийный аппарат:

1) множество документов, на котором осуществляется поиск $\{d_1, d_2, ..., d_n\}$;

2) множество термов $\{T(d_i)\} = U_{i=1}^{n} T(d_i)$, где $T(d_i)$ – набор слов (терминов), входящих в документ d_i по которым осуществляется поиск.

Булева модель поиска подразумевает наличие в базе данных индекса, организуемого в виде инвертированного массива, в котором для каждого терма из словаря базы данных находится соответствие со списком документов, в котором эти термы встречаются. Булева модель подразумевает, как правило, поиск термов из базы данных, входящих в состав запроса и поиск координат всех вхождений термов путем обращения к инверсной таблице. По массивам документов осуществляется выявление релевантных документов путем выполнения теоретико-множественных операций. Булева модель поиска очень широко применяется в КИС, а также поддерживается практически всеми коммерческими ИПС в Интернет.

Векторно-пространственная модель является классической алгебраической моделью [Croft, Harper, 1979; Greiff, 1998]. Данная модель представляет запрос и документы, в которых осуществляется поиск, в виде n-мерного вектора в евклидовом

пространстве термов. При этом модулем вектора является некоторая величина, определяющая его вес (значимость) [Salton, Buckley, 1988]. В качестве веса w_{ij} терма $T_i(d_j)$ часто используют нормированную частоту его использования $freq_{ij}$:

$$w_{ij} = freq_{ij} \Big/ \max(freq_{ij}) \quad (1)$$

Показатель (1) может варьироваться в зависимости от условий его использования. Например, если существует статистика использования термов во всем массиве документов, то вес может рассчитываться исходя из числа документов, в которых используется терм, и логарифма общего количества документов в базе. Значения весов w_{ij} могут нормироваться делением на квадратный корень из суммы квадратов весов всех термов для получения документа как ортонормированного вектора. Данная модель облегчает решение вопроса тематической идентичности документов. Ее можно определить как скалярное произведение, которое пропорционально косинусу угла между векторами. Чем больше эта величина, тем ближе эти документы в тематическом плане.

Вероятностная модель основывается на байесовской теореме об условных вероятностях событий. В этом случае каждому документу d_l ставят в соответствие вектор $l = (t_1, t_2, ..., t_k)$, у которого $t_i = 0$, если i-тый терм не входит в состав документа, и $t_i = 1$, если входит. Если обозначить через W_1 событие, обозначающее, что документ d_l релевантен запросу, а W_2 – событие, обозначающее, что документ d_l не релевантен запросу, то $P(W_i \mid l)$ – это вероятность того, что для документа d_l наступит событие W_i. На основании теоремы Байеса можно перейти к вероятностям, значения которых удобнее оценить:

$$P(W_i \mid l) = P(l \mid W_i) P(W_i) / P(l).$$

В канонической вероятностной модели используется упрощение, заключающееся в предположении независимости вхождения в документ любой пары термов. В этом случае $P(l \mid W_i) = P(l_1 \mid W_i) \times ... \times P(l_n \mid W_i)$. Если использовать следующие обозначения $p_i = P(l_i = 1 \mid W_1)$, $q_i = P(l_i = 1 \mid W_2)$, то $P(l \mid W_1) = \prod_{i=1,...,n} p_i^{l_i}$ и $P(l \mid W_2) = \prod_{i=1,...,n} q_i^{l_i}$. Неравенство, определяющее релевантность документа запросу, можно записать следующим образом [Маннинг и др., 2011]:

$$\log(P(l \mid W_1) P(W_1) / P(l \mid W_2) P(W_2)) > 0.$$

Анализ вышеперечисленных моделей приводит к тому, что все они обладают значительными недостатками. Так, булева модель обладает жестким набором операторов, требующих знания хотя бы основ логической алгебры, а также имеет значительные сложности с ранжированием найденных документов. Векторно-пространственная модель оперирует значительными объемами данных и массивами высокой размерности, что вызывает сложности с обработкой запросов большого объема. Вероятностная модель характеризуется низкой масштабируемостью, требует разработки алгоритмов обучения [Захаров, 2005].

На практике применяются гибридные подходы: например, поиск осуществляется с использованием булевой модели, а последующее ранжирование осуществляется на основе векторно-пространственной модели. Кроме того, часто применяют так называемые расширенные модели, которые характеризуются применением дополнительных алгоритмов анализа запроса: выявление связей между терминами запроса, лингвистический анализ, использование латентно-семантического индексирования и т.п. Нетрудно заметить, что все вышеперечисленные модели объединяет применение значительного количества параметров, по которым оценивается полнота поиска и релевантность полученной информации [Захаров, 2005]. Такой подход накладывает значительную нагрузку на аппаратную часть ИПС, что влечет за собой значительное удорожание системы. Следовательно, при построении математической модели

необходимо уделить внимание выбору таких параметров поиска и оценки релевантности, которые позволят при минимальном их количестве выполнить оценку, как минимум, не хуже, чем у существующих моделей.

Постановка задачи построения математической модели поиска и сортировки

При построении математической модели поиска и сортировки необходимо рассматривать ее как средство повышения эффективности ИПС. В рамках конференции по оценке систем текстового поиска [TREC], созданной Американским Институтом Стандартов (NIST), различают следующие основные параметры эффективности ИПС:

1) коэффициент полноты $k_П$ – часть выданных релевантных документов среди всех найденных релевантных;

2) коэффициент точности $k_Т$ – часть выданных релевантных документов среди всех выданных;

3) коэффициент шума $k_Ш$ – часть выданных нерелевантных документов среди всех выданных;

4) коэффициент осадков k_0 – часть выданных нерелевантных документов среди всех нерелевантных;

5) коэффициент специфичности $k_С$ – часть не выданных нерелевантных документов среди всех нерелевантных.

Важнейшими из них признаны коэффициент полноты и коэффициент точности. Коэффициент полноты определяется как $k_П = R/C$, где C – общее число документов в системе, релевантных запросу s_i, а R – число документов, выданных пользователю и релевантных запросу. Коэффициент точности, определяемый как $k_Т = R/L$, где L – общее количество документов, выданных пользователю вне зависимости от их степени релевантности [Ландэ, 2006]. Иными словами, коэффициент точности характеризует процесс поиска как возможность ИПС отыскать максимально возможное количество релевантных документов, а коэффициент полноты характеризует процесс сортировки документов как способность системы выдавать в первую очередь наиболее релевантные документы. В соответствии с этим можно выделить два основных этапа работы ИПС:

1) этап поиска, когда система в соответствии с полученным запросом осуществляет поиск релевантных документов, хранящихся в базе;

2) этап сортировки, на котором система выстраивает последовательность найденных документов по определенному критерию, определяющему релевантность документа.

К сожалению, среди пяти показателей эффективности, предлагаемых NIST, отсутствуют те, которые указывали бы на эффективность борьбы ИПС с дубликатами. При этом следовало бы рассмотреть третий этап работы ИПС – удаление дубликатов, который можно было бы охарактеризовать коэффициентом дублирования $k_д = D/L$, где D – количество дублей, найденных в результатах поиска.

Подход, описанный при анализе существующих моделей поиска и сортировки, опирается на запрос, как на набор взаимно независимых термов (множество, вектор и т.п.). На практике термы в запросе взаимозависимы, применение одного из них часто связано с применением другого или некоторой группы термов. Поэтому при разработке модели поиска и сортировки необходимо использовать комплексный подход, включающий прежде всего:

1) учет при определении параметров поиска взаимную связь термов в запросе;

2) унификацию параметров, используемых для поиска, сортировки и определения дублей;

3) обеспечение совместимости разрабатываемых моделей со структурами данных, используемыми в существующих ИПС.

Таким образом, ключевым положением является использование гибких стратегий поиска, сравнения и сортировки документов при минимальном количестве параметров.

Все существующие модели объединяет общий подход: поиск и сортировка документов осуществляются в n-мерном пространстве термов, а удаление дублей – в m-мерном пространстве шинглов. Несмотря на то, что при векторном подходе, например, в качестве показателя релевантности используется единый обобщенный показатель (вектор), тем не менее, он рассчитывается как тригонометрическая сумма в пространстве термов и не снижает аппаратных затрат для его вычисления. Кроме того, построение n-мерного евклидова пространства термов предполагает их ортогональность, т.е. их несвязанность, что в большинстве случаев неверно. Сошлемся на краткую и авторитетную формулировку одного из ведущих лингвистов В.Б. Касевича: «будучи целостной единицей, текст обнаруживает по отношению к своим структурным компонентам (сверхфразовым единствам/абзацам, высказываниям, тем более – словам) свойство неаддитивности: характеристики текста не выводимы полностью из признаков его составляющих; в первую очередь, передаваемое текстом значение несводимо к сумме значений компонентов» [Большакова и др., 2011]. Отсюда следует, что найденная сумма термов $\sum_i t_i$ не всегда является искомой по смыслу фразой. Поэтому применение существующих моделей ведет к снижению количества релевантных документов и, как следствие, к снижению коэффициента точности.

Рассмотрим задачу поиска релевантных документов в массиве документов $\{d_1, d_2, ..., d_n\}$ по запросу s в виде картежа:

$$I = \langle d_i, s_j, \delta \rangle$$

где δ – отношение идентичности вида $x\ \delta\ y \Leftrightarrow x \sim y$, I – показатель релевантности [Тявкин и др., 2008]. Часто отношение идентичности рассматривают в смысле $x = y$ для поиска термов в базах данных. Делается это обычно для применения стандартных методов поиска и сортировки, сформулированных в трудах [Кнут, 2002]. Понятно, что требовать отношения идентичности в смысле $x = y$ для поиска релевантных документов не имеет смысла, т.к. сочетания могут использоваться в различных родах и падежах. Поэтому в существующих моделях используют морфологический анализ текста и его предварительную обработку перед помещением в базу данных. В случае применения многопараметрического или векторного показателя релевантности обычно используют количество вхождений (абсолютное или относительное) термов из запроса в состав документа, т.е. $F(s, d_i) \xrightarrow[T(d_i)]{} \max$. Тогда относительной единицей меры релевантности будет одно вхождение одного терма в документ или $\rho_i = 1/N_i$, где N_i – число слов (термов) в документе d_i, а показатель релевантности документа d_i будет выглядеть как сумма числа относительных вхождений термов из запроса, т.е.

$$F = \sum_i \sum_j \rho_{ij} = \sum_i \frac{1}{N_i} \sum_j \{T(d_j)\} \qquad (2)$$

Здесь ρ_{ij} – относительная частота вхождения j-того терма в i-тый документ. Учитывая ранее сформулированный тезис о том, что $\sum_j \{T(d_j)\}$ не всегда является смысловой единицей искомого текста, придем к пониманию, что существующие модели не ведут к повышению коэффициента точности и, как следствие, к повышению эффективности ИПС. Применение гибких стратегий поиска и сортировки сдерживается, с одной стороны, отсутствием единого показателя поиска и сортировки, имеющего понятный физический смысл и имеющего быстрый и эффективный алгоритм его вычисления и, с другой стороны, быстрым удешевлением вычислительных мощностей для реализации существующих моделей. В тоже время лавинообразный рост объемов и разнообразия информации, представленной в электронном виде, требует быстрого роста

эффективности поиска и предоставления информации для пользователей сетей. Высокая доля мультимедийного контента также способствует снижению эффективности существующих моделей [Большакова и др., 2011]. Для разрешения этого противоречия необходимо использование принципиально нового методического аппарата, который позволит значительно повысить эффективность процессов в ИПС.

Математическая модель поиска и сортировки

Одним из путей повышения эффективности ИПС может быть переход от ортогональной модели поиска множеств или векторов независимых термов к поиску, сортировке и поиску дублей, основанных на представлении запросов и документов в виде сигналов, исследованию их степени схожести на принципах корреляционного анализа. Это позволит использовать унифицированный механизм ИПС не только в области текстового, но и в области мультимедийного поиска. Математический аппарат корреляции [Ifeachor, Jervis, 2002] нашел широкое применение в цифровой обработке сигналов, распознавании образов, радиолокации и радиопеленгации. Рассмотрим, как этот аппарат может быть использован для определения меры схожести запроса и документа.

Для определения меры схожести сигналов в радиотехнике используют коэффициент корреляции, представленный в виде

$$r = \frac{1}{2T\sqrt{P_s P_d}} \int_0^T S_s(t) S_d(t) dt,\ t \in [0, T]. \quad (3)$$

Смысл корреляции заключается в том, что если запрос $s(k)$ и i-тый документ $d_i(k)$ являются совершенно различными текстами, не содержащими общих понятий (термов), то взаимная корреляция их представлений будет невелика и обусловлена исключительно избыточностью естественного языка. Если же текст документа отвечает потребностям пользователя и документ изобилует понятиями, содержащимися в запросе, то взаимная корреляция будет значительно больше по значению. Несмотря на внешнюю схожесть выражений (2) и (3) они обладают совершенно различным физическим и лингвистическим смыслом. Если выражение (2) определяет показатель релевантности на основании среднего количества вхождений термов запроса в документы, то в выражении (3) показатель релевантности определяется как сумма произведений представления запроса и документа. Лингвистическая ценность показателя (3) состоит в том, что сравнивается не просто вхождение в документ тех или иных термов из запроса, но учитывается и их «связанность», т.е. использование соответствующего рода и падежа, дополнительных слов-связок, конструирующих не простую сумму термов, но связное текстовое значение [Баклицкий и др., 1986].

Рассматривая поиск информации в одной предметной области, можно заметить, что, несмотря на различие формулировок запросов, набор термов будет примерно одинаков. При этом выражения (2) и (3) будут отражать степень наличия одинаковых термов в запросе и документе, но показатель (3) будет более точным, т.к. учитывает сходство текста в целом, включая и дополнительные слова, не входящие в пространство термов. На основании этого можно утверждать, что результат поиска при использовании корреляционного показателя будет функцией в пространстве термов вида $g = w(t_i) = w_1(t_i) + w_2(t_j)$, где $w_1(t_i)$ – квазидетерминированная составляющая, возникающая за счет совпадения набора термов для процессов в предметной области, а $w_2(t_j)$ – случайная составляющая, возникающая за счет особенностей естественного языка, а также различных языковых конструкций, используемых разными людьми [Баклицкий и др., 1986]. Здесь первая составляющая может быть аппроксимирована полиномом k-той степени вида $w_1(t_i) = \sum_{l=0}^{k-1} a_l t_i^l$, здесь a_i – коэффициенты полинома, а t_i^l – случайная величина, характеризующая отличия в запросах. При этом случайная

составляющая может быть описана случайным законом $w_2(t_j) = \sum_j \xi(t_j)$. Здесь t_j – набор дополнительных слов, не входящих в состав пространства термов $T\{t_i\}$. Таким образом, математическое ожидание описывается полиномом *k*-той степени, а случайная составляющая – некоторым законом распределения, чаще всего нормальным. Величина $g(t)$ будет случайной с математическим ожиданием $w_1(t_i)$ и случайным процессом $w_2(t_j)$ с нулевым математическим ожиданием.

Все пространство документов $D = \{d_i\}$ может быть разделено на два подпространства – множество релевантных документов D_0 и нерелевантных документов D_1. Каждому документу d_i соответствует точка в *n*-мерном пространстве термов в течении времени *t*: $D = F(T, t)$. Существующие методики разделения подпространств опираются, как правило, на раздельный расчет показателя релевантности *I* в многомерном пространстве термов $\{t_j\}$, что требует значительных вычислительных затрат. Использование единого показателя $g(t)$ позволит осуществлять расчет показателя релевантности с меньшими затратами вычислительных ресурсов и, по предварительным оценкам, применение непозиционных систем счисления для расчета показателя даст выигрыш не менее, чем на порядок. Этот же показатель может быть использован для сортировки документов по степени релевантности, а также для поиска дублей среди документов, хранящихся в базе поиска.

Заключение

Коэффициент корреляции запроса и документа можно рассматривать как некий обобщенный показатель, который позволяет решить следующие задачи:

1) установить степень похожести запроса с документами, в тексте которых осуществляется поиск;

2) по значению коэффициента корреляции можно осуществлять сортировку по релевантности;

3) расчет коэффициента корреляции документов между собой позволит отыскать дубли и с высокой степенью вероятности удалить их из результатов поиска.

Таким образом, использование коэффициента корреляции позволит создавать ИПС с высоким быстродействием и менее критичные к аппаратной части.

Список литературы
References

1. Баклицкий В.К., Бочкарев А.М., Мусьяков М.П. 1986. Методы фильтрации сигналов в корреляционно-экстремальных системах навигации. М., Радио и связь, 216.

Baklitskiy V.K., Bochkarev A.M., Mus'yakov M.P. 1986. Metody fil'tratsii signalov v korrelyatsionno-ekstremal'nykh sistemakh navigatsii [Methods of signal filtering in correlation-extreme navigation systems]. M., Radio i svyaz', 216. (In Russian)

2. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. 2011. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. М., МИЭМ, 272.

Bol'shakova E.I., Klyshinskiy E.S., Lande D.V., Noskov A.A., Peskova O.V., Yagunova E.V. 2011. Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika [Automatic processing of texts in natural language and computer linguistics]. M., MIEM, 272. (In Russian)

3. Захаров, В.П. 2005. Информационно-поисковые системы.СПб, 320.

Zakharov V.P. 2005. Informatsionno-poiskovye sistemy [Information retrieval systems]. SPb, 320. (In Russian)

4. Кнут Д.Э. 2002. Искусство программирования, том 2: получисленные алгоритмы. Пер. с англ. М., Издательский дом Вильямс, 832. (Knuth D.E. 1997. The art of computer programming, volume 2: seminumerical algorithms. MA: Addison-Wesley, 762.)

Knut D.E. 2002. Iskusstvo programmirovaniya, tom 2: poluchislennye algoritmy. Per. s angl. M., Izdatel'skiy dom Vil'yams, 832. (Knuth D.E. 1997. The art of computer programming, volume 2: seminumerical algorithms. MA: Addison-Wesley, 762.)

5. Ландэ Д.В. 2006. Основы интеграции информационных потоков. Киев, Инжиниринг, 240.

Lande D.V. 2006. Osnovy integratsii informatsionnykh potokov [Basics of Information Stream Integration]. Kiev, Inzhiniring, 240. (In Russian)

6. Маннинг К.Д., Рагхаван П., Шютце Х. 2011. Введение в информационный поиск. М., Издательский дом «Вильямс», 528.

Manning K.D., Ragkhavan P., Shyuttse Kh. 2011. Vvedenie v informatsionnyy poisk [Introduction to information retrieval]. M., Izdatel'skiy dom «Vil'yams», 528.

7. Сегалович И.В. 2002. Как работают поисковые системы. Мир Internet, 10: 24-32.

Segalovich I.V. 2002. Kak rabotayut poiskovye sistemy [How search engines work]. Mir Internet, 10: 24-32. (In Russian)

8. Тявкин И.В., Тютюнник В.М. 2008. Математическая модель информационного поиска и оценка эффективности поисковой системы. Вестник Тамбовского государственного технического университета, 14(3): 478 – 480.

Tyavkin I.V., Tyutyunnik V.M. 2008. Mathematical model of information retrieval and evaluation of search engine efficiency. Vestnik Tambovskogo gosudarstvennogo tekhnicheskogo universiteta, 14(3): 478 – 480. (In Russian)

9. Хорошко М.Б. 2014. Разработка и модификация моделей и алгоритмов поиска данных в Internet/Intranet среде для улучшения качества поиска. Диссертация на соискание ученой степени кандидата технических наук. Южно–Российский государственный политехнический университет (НПИ) имени М. И. Платова, Новочеркасск, 225.

Khoroshko M.B. 2014. Razrabotka i modifikatsiya modeley i algoritmov poiska dannykh v Internet/Intranet srede dlya uluchsheniya kachestva poiska. [Development and modification of models and algorithms for searching data in the Internet/Intranet environment to improve the search quality]. Dissertatsiya na soiskanie uchenoy stepeni kandidata tekhnicheskikh nauk. Yuzhno – Rossiyskiy gosudarstvennyy politekhnicheskiy universitet (NPI) imeni M. I. Platova, Novocherkassk, 225. (In Russian)

10. Шарапов Р.В., Шарапова Е.В., Саратовцева Е.А. 2007. Модели информационного поиска. URL: http://vuz.exponenta.ru/PDF/FOTO/kaz/Articles/sharapov1.pdf. (дата обращения: 09.08.2017)

Sharapov R.V., Sharapova E.V., Saratovtseva E.A. 2007. Models of information retrieval. Available at: http://vuz.exponenta.ru/PDF/FOTO/kaz/Articles/sharapov1.pdf. (accessed: 9 August 2017). (In Russian)

11. Baeza-Yates R., Ribeiro-Neto B. 1999. Modern Information Retrieval: The Concepts and Technology behind Search. ACM Press Books, 944.

12. Croft W.B., Harper D.J. 1979. Using probabilistic models of document retrieval without relevance information. MCB UP Ltd, Journal of documentation, 35(4): 285-295.

13. Greiff W. R. 1998. A theory of term weighting based on exploratory data analysis. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. ACM: 11-19.

14. Ifeachor E.C., Jervis B.W. 2002. Digital signal processing: a practical approach. Prentice Hall, 960.

15. Salton G. 1968. Automatic information organization and retrieval. McGraw-Hill, 527.

16. Salton G. 1979. Mathematics and information retrieval. Journal of Documentation, 35(1): 1-29.

17. Salton G., Buckley C. 1988. Term-weighting approaches in automatic text retrieval. Information processing & management, 24(5): 513-523.

18. Salton G., Fox E.A., Wu H. 1983. Extended Boolean information retrieval. Communications of the ACM, 26(11): 1022-1036.

19. Text REtrieval Conference (TREC). Available at: http://trec.nist.gov/pubs.html (accessed 13 September 2017).

20. Van Rijsbergen C.J. 1979. Information retrieval. MA, Butterworth-Heinemann Newton, 208.