УДК 004.912+004.051

ПОДХОД К РАНЖИРОВАНИЮ РЕЗУЛЬТАТОВ ДЛЯ ТЕРМИНОЛОГИЧЕСКОГО ПОИСКА НА ОСНОВЕ МЕРЫ БЛИЗОСТИ СТРОК

Д.А. Заикин

Аннотация

В статье проведена модификации полнотекстового подхода к информационному поиску на основе использования семантической информации о терминологических единицах в тексте. Исследован вопрос о выборе ранжирующей метрики для такого вида поиска. Рассмотрены несколько существующих метрик, а также предложена новая, использующая особенности терминологического поиска. Проведены экспериментальные сравнения предложенных метрик ранжирования и оценки эффективности самой поисковой системы.

Ключевые слова: информационный поиск, ранжирование, терминология.

Введение

Взрывообразный рост разнообразных публикаций в сети Интернет приводит к тому, что постоянно повышаются требования к информационно-поисковым системам [1]. Актуальность исследований в области информационного поиска также обусловлена тем, что при поиске информации в сети Интернет число документов, возвращаемых на запрос пользователя, как правило, получается очень большим за счет огромного числа нерелевантных документов, попавших в отклик. Например, в работе Чуна отмечается, что Google, фокусируясь на релевантности результатов, мало заботится о числе ответов [2].

В последние годы появились многочисленные поисковые сервисы, стремящиеся усовершенствовать поисковые технологии, выходя за рамки стандартного поиска по ключевым словам [3]. Разработчики поисковых систем стали использовать более сложные модели представления документов для наиболее эффективного использования имеющихся в нём данных. Текст стал рассматриваться как объект со сложной семантической структурой связей [4]. По этой причине приобретают высокую актуальность исследования, принимающие во внимание семантику текста.

Вышеизложенное позволяет сделать заключение о необходимости проведения исследований по вопросам повышения качества информационного поиска, основываясь на использовании семантической информации в документе. Одним из направлений подобных исследований является использование специальной лексики предметных областей для улучшения качества поиска [5].

Как оказалось, общие подходы, разработанные для индексации и обработки запросов к страницам всемирной паутины, не всегда подходят для решения задач поиска для конкретных предметных областей [6].

Значительным отличием предметных публикаций является наличие в них специальной лексики. С одной стороны, она несёт в себе больше информации, чем окружающая её специальная лексика, а с другой – создаёт дополнительные трудности [7].

1. Терминологический поиск

Главная идея заключается в осуществлении свободного поиска (произвольный порядок слов, любая морфологическая форма) в пределах одного термина-словосочетания. Если все слова запроса принадлежат одному термину-словосочетанию, то вне зависимости от их порядка, морфологической формы и расстояния между друг другом документ, содержащий этот многословный термин, будет признан релевантным и выдан пользователю. Если все слова, входящие в запрос пользователя, встречаются в одном документе и при этом принадлежат разным терминам-словосочетаниям, то данный документ рассматривается как нерелевантный.

Рассмотрим формальную постановку задачи поиска с использованием семантической информации о специальной лексике в тексте.

Определим множества

$D = \{d_1, d_2, \ldots, d_n\}$ – множество документов;

$TT = T_{d_1}, \ldots, T_{d_n} : T_{d_i} = \{t_i^1, \ldots, t_i^{m_i}\}$ – множества терминологических словосочетаний для документов из D;

$t_i^j = \{l_{i,1}^j, \ldots, l_{i,p_{i,j}}^j\}$ – множество лексем терминологического словосочетания t_i^j.

Пусть $Q = \{q_1, q_2, \ldots, q_k\}$, где q_i – ключевые слова запроса, тогда результат поискового запроса в общем случае формулируется как

$$S(Q, D) = \{r_1, \ldots, r_z\} \subset D.$$

Результат терминологического поиска определяется как

$$S_t(Q, D, TT) = \{r_1, \ldots, r_z\}, \quad r_i \in D : \exists\, t \in T_{r_i}\ \forall\, q \in Q\ \exists\, l \in t : q = l.$$

Сравнение качества результатов поискового запроса обычно производится по метрикам точности и полноты [8].

Точность ($Precision$) – это отношение числа релевантных документов, найденных поисковой системой, к общему числу найденных документов:

$$Precision = \frac{|D_{\text{rel}} \bigcap D_{\text{retr}}|}{|D_{\text{retr}}|}, \tag{1}$$

где D_{rel} – это множество релевантных документов в базе, а D_{retr} – множество документов, найденных системой.

Полнота ($Recall$) – это отношение числа найденных релевантных документов к общему числу релевантных документов в базе:

$$Recall = \frac{|D_{\text{rel}} \bigcap D_{\text{retr}}|}{|D_{\text{rel}}|}. \tag{2}$$

Иногда бывает полезно объединить точность и полноту в одной усреднённой величине. В таком случае, как правило, используется F-мера – взвешенное гармоническое среднее точности P и полноты R:

$$F = \frac{1}{\alpha/P + (1-\alpha)/R}, \quad \alpha \in [0, 1]. \tag{3}$$

Сбалансированная F-мера, или F_1-мера, придает одинаковый вес точности и полноте

$$F_1 = \frac{2PR}{P+R}. \tag{4}$$

Исследование ставит целью проверить гипотезу, что использование дополнительной семантической информации о специальной лексике в тексте позволяет улучшить значение F-меры:

$$F_1(S_t(Q, D, TT)) > F_1(S(Q, D)).$$

Для построения индекса используется Apache Solr – кроссплатформенная система для корпоративного поиска со свободным исходным кодом на базе проекта Apache Lucene [9]. Его основные функции включают в себя мощный полнотекстовый поиск, подсветку результатов, фасеточный поиск, динамическую кластеризацию, интеграцию с базами данных и поддержку обработки большого числа форматов документов (например, Word, PDF). Система Solr является самой популярной корпоративной поисковой системой [10].

2. Ранжирование результатов

Терминологический поиск отличается от полнотекстового, поэтому у первого могут быть свои особенности работы ранжирующих метрик. Для выяснения этих особенностей было предложено несколько базовых методов ранжирования результатов и проведено сравнение между ними. Ранжирование однозначно определяется ранжирующей функцией. В дальнейшем будем сравнивать указанные функции, то есть будем проводить сравнение результатов работы поисковых систем, использующих данные ранжирующие функции.

Поисковая платформа Solr по умолчанию ранжирует результаты в зависимости от косинусной меры близости, вычисляемой из значения $TF{\cdot}IDF$. Была составлена ранжирующая функция, использующая данную метрику:

$$Score_1 = (TF{\cdot}IDF)(t, d, D) = tf(t, d) \cdot idf(t, D) = \frac{n_t^d}{\sum\limits_k n_k^d} \cdot \log \frac{|D|}{|(d_i \supset t_i)|},$$

где n_t^d – число вхождений слова t в документ d; $\sum\limits_k n_k^d$ – общее число слов в документе d; $|D|$ – количество документов в корпусе; $|(d_i \supset t)|$ – количество документов, в которых встречается t (когда $n_t^{d_i} \neq 0$).

Одним из популярных в современных поисковых системах подходов к ранжированию является ссылочный. В качестве ссылочной ранжирующей функции было выбрано значение, возвращаемое алгоритмом $PageRank$ [11]:

$$Score_2 = PageRank.$$

Основной единицей, с которой работает терминологический поиск, является терминологическое словосочетание. Поэтому логичным является введение метода ранжирования, основанного на «схожести» пользовательского запроса с терминологическим словосочетанием в тексте. Выбрана функция ранжирования

$$Score_3 = NGramDist(QT, IT, 2) \cdot \sqrt{|IT|},$$

где $NGramDist(\cdot)$ – метрика $N\text{-}GramDistance$ [12]; QT – термин запроса; IT – термин в индексе; $|IT|$ – длина термина в индексе.

С помощью метрики $N\text{-}GramDistance$ [12] оценивается расстояние между строками, которое аналогично известным расстояниям Левенштейна или Яро – Винклера. Особенностью данной метрики является то, что она учитывает не только длину наибольшей общей подстроки, но и контекст окружения. Например, пары

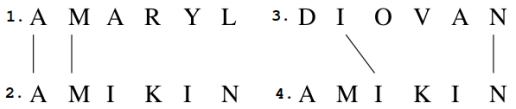

Рис. 1. Пример сравнения строк

строк “AMARYL” и “AMIKIN” будет иметь то же расстояние Левенштейна, что и строки “DIOVAN” и “AMIKIN”, так как в обоих случаях одинаково число совпадающих символов (см. рис. 1). Метрика $N\text{-}GramDistance$ поощряет подряд идущие сходные символы, что приводит к тому, что строки один и два (рис. 1) считаются ближе друг к другу, чем строки три и четыре.

У $N\text{-}GramDistance$ присутствует параметр N, определяющий число символов контекста, принимаемых во внимание при расчёте значения метрики. Используется значение параметра, равное двум, как показавшее лучшие результаты при сравнительной оценке точности метода [12].

Использование именно этой метрики позволяет присваивать больший вес словосочетанию «абелева конечная группа» по сравнению с конструкцией «конечная абелева группа» относительно строки «конечная группа», что является преимуществом.

Значение метрики равно единице при полном совпадении строк. Увеличение длины строки, а также перестановка слов ведут к уменьшению величины $N\text{-}GramDistance$. Таким образом, при ранжировании принимается во внимание порядок слов в запросе и найденном терминологическом словосочетании.

Следует отметить, что метрика в одинаковой степени оштрафует и перестановку слов и удлинение термина. Для компенсации уменьшения величины метрики производится умножение значения на корень длины термина. В результате отдаётся предпочтение увеличению длины терминологического словосочетания по сравнению с перестановкой слов.

Был рассмотрен также смешанный вариант функции ранжирования, основанный на линейной комбинации метрик расстояния между строками и $PageRank$.

Результаты упорядочиваются по величине

$$Score_4 = \frac{PR}{MPR} + NGramDist(QT, IT, 2) \cdot \sqrt{|IT|},$$

где PR – значение $PageRank$; MPR – максимальное значение $PageRank$.

В ранжирующую функцию включено значение метрики ранжирования, основанной на библиографическом цитировании, которое нормируется максимальным значением в индексе и включается в итоговую функцию сортировки. За счёт нормирования вклад этого значения оказывается меньшим, чем функции схожести терминологических словосочетаний.

Таким образом:

• метрика схожести терминов имеет больший вес (так как, вообще говоря, может принимать значения больше 1), чем ранжирующий параметр, основанный на библиографии;

• вхождение запросных слов в длинный термин ранжируется ниже, чем точное совпадение;

• перемена мест слов термина из запроса приведёт к тому, что этот термин будет отранжирован ниже, чем термин с правильным порядком слов, но несколько большей длины;

Табл. 1

Результаты сравнения точности ранжированного поиска

	1	2	3	4	5	6	7	8	9	10
$Score_1$	0.96	0.97	0.967	0.965	0.964	0.96	0.96	0.963	0.962	0.962
$Score_2$	0.94	0.91	0.913	0.915	0.912	0.91	0.917	0.92	0.922	0.918
$Score_3$	1	0.97	0.967	0.96	0.964	0.963	0.966	0.965	0.964	0.96
$Score_4$	1	0.96	0.96	0.955	0.96	0.963	0.966	0.965	0.964	0.962

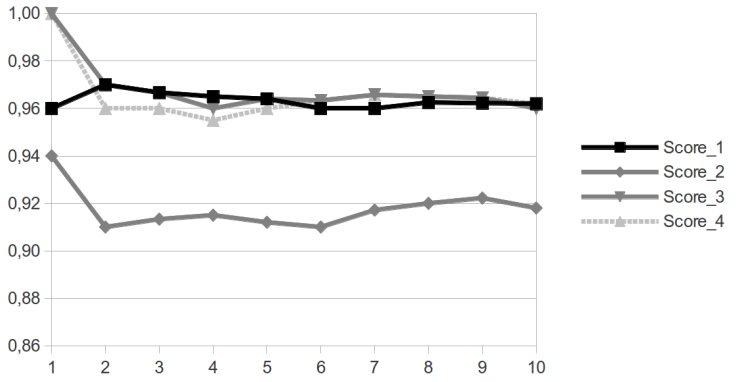

Рис. 2. Результаты сравнения точности ранжированного поиска

• термины с примерно равными значениями метрики схожести терминологических словосочетаний будут проранжированы в соответствие со значением своего библиографического веса.

3. Выбор метрики ранжирования на основании сравнения

Для определения качества работы ранжирования поисковой системы используется показатель «Точность на уровне n документов» ($P@n$) (см. (1)), который определяется как количество релевантных документов среди первых n документов, деленное на n.

Оценки проводились по точности до уровня 10 документов. Всего оценивались результаты для 50 запросов без ограничения на число результатов. Полученные значения проиллюстрированы в табл. 1 и на рис. 2.

Можно видеть, что использование $Score_2$ приводит к худшим значениям точности результатов. По всей видимости, это говорит о низкой эффективности ссылочного ранжирования для научных публикаций при терминологическом поиске.

Остальные ранжирующие метрики показывают примерно сходные результаты, а разброс значений не позволяет однозначно выделить лучшую среди них. Для решения этой проблемы была введена целевая функция

$$H(S) = \sum_{n=1}^{10} c_n \, P@n(S),$$

где S – ранжирующая метрика; $P@n(S)$ – точность на уровне n метрики S; $c_1, \cdot, c_{10}$ – настроечные коэффициенты.

Коэффициенты c_n были выбраны обратно-пропорциональными n: $c_n = 1/n$, таким образом, точность на уровне первых результатов в списке получает больший вес, чем точность на уровне более поздних.

В итоге задач а была сведена к нахождению максимального значения

$$H(S) = \sum_{n=1}^{10} \frac{P@n(S)}{n}$$

среди всех ранжирующих метрик S.

Согласно этой формуле были получены результаты: $H(Score_1) = 2.82184$; $H(Score_2) = 2.70255$; $H(Score_3) = 2.86232$; $H(Score_4) = 2.85325$.

Использование метрики $Score_2$, основанной на ранжировании, вновь приводит к результатам хуже, чем у предложенных аналогов. Лучший результат достигается при использовании метрики $Score_3$, основанной на близости запроса и терминологического словосочетания. Дополнительное использование ссылочного ранжирования, вдобавок к близости строк, лишь ухудшает значение для ранжирования результатов с помощью $Score_4$. Применение $Score_1$, базирующейся на $TF \cdot IDF$, даёт результат немного хуже.

На основе полученных результатов в качестве рабочей была выбрана метрика $Score_3$.

4. Оценки качества информационного поиска

Существует много способов оценить, насколько хорошо документы, найденные поисковой системой, соответствуют запросу. В основе чаще всего лежит понятие релевантности – семантического соответствия поискового запроса и поискового образа документа [13].

К сожалению, понятие степени соответствия запроса, или, другими словами, релевантности, является субъективным понятием, а степень соответствия зависит от конкретного человека, оценивающего результаты выполнения запроса.

В качестве основания для сравнения (*comparison baseline*) часто используются свободно распространяемые поисковые системы: Apache Lucene, Sphinx, Apache Solr, Isearch [14].

Сравнение производилось с системой Apache Solr 4.2.1 [15]. Данные индексировались в единое поле стандартного типа $text\_ru$. Использовался стандартный обработчик запросов со связыванием ключевых слов запроса конъюнкцией и максимальным удалением слов, равным десяти.

Для рассматриваемой предметной области не существует так называемого «золотого стандарта ранжирования» (корпуса документов, размеченного в соответствии с релевантностью эталонному списку запросов). Поэтому сравнение проводилось эмпирически.

Все оценки проводились на случайно составленных запросах-терминах, имеющих более одного результата в выдаче (упорядоченном множестве документов, возвращаемых поисковой системой в ответ на запрос пользователя) по крайней мере в одной из сравниваемых систем. Сравнение качества поиска разработанной системы по запросам, не являющимся терминами, будет неэффективной тратой труда экспертов предметной области: программа не разрешает такие запросы и возвращает пустую выдачу. По этой причине при поиске слов, не являющихся терминами, полнотекстовая поисковая система предпочтительнее описанной в настоящей работе.

При проведении вышеописанного сравнения было выявлено, что результаты на однословные запросы являются идентичными. Этот факт можно объяснить тем, что обе сравниваемые системы используют одну и ту же базовую поисковую платформу. По этой причине в дальнейшем подобные запросы не рассматриваются.

Точность. Оценивалась на 35 запросах с количеством результатов не более 50. Оценка точности работы системы на запросах с более чем 50 статьями в выдаче

слишком трудозатратна. Кроме того, пользователи почти никогда не просматривают такое большое число результатов [16].

Точность результатов терминологического поиска равняется 0.919, в то время как полнотекстовый поиск Solr – 0.749. Разница существенная, однако следует принимать во внимание и особенности методики сравнения. Во-первых, в перечне запросов отсутствует общая лексика, так как терминологический поиск такие запросы не разрешает вообще. Во-вторых, из исследования были исключены однословные запросы, точность результатов на которые, как правило, выше [17].

Довольно часто причиной нерелевантного результата полнотекстовой поисковой системы являлись слова на небольшом расстоянии друг от друга, но из разных контекстов. Например, фрагмент текста «Предлагаемый метод решения спектральной задачи позволяет строить приближенные решения без последующей интерполяции и выбора пробных функций» был возвращен на запрос «методы интерполяции» и релевантным к последнему не является. Терминологический поиск при разборе этого фрагмента относит слова «метод» и «интерполяция» к различным терминологическим словосочетаниям «предлагаемый метод решения спектральной задачи» и «интерполяции и выбора пробных функций» соответственно.

Общий подход не всегда применим к частным предметным областям. Например, в запросе «пространство над алгеброй» предлог «над» расценивался как стоп-слово (или шумовое слово) и игнорировался при обработке. При этом в данном контексте предлог имеет существенное значение и фрагмент «$End(B)$, когда B – банахово пространство, – банахова алгебра $Hom(B,B)$; I_B – единичный оператор из алгебры $End(B)$;» релевантным результатом не является.

Полнота. Для оценки полноты результатов с хорошей надёжностью и достоверностью требуется разметка всей коллекции документов на соответствие запросов. Ввиду большого размера коллекции (более 1400 документов) подобная разметка требует вклада большого ручного труда экспертов, что весьма ресурсозатратно.

О полноте можно судить косвенным образом: считать, что обе системы вместе обеспечивают 100%-ную полноту результатов. Таким образом, относительная полнота есть отношение числа найденных релевантных документов к числу релевантных документов найденных обеими системами.

Подход на основе косвенных данных не позволяет узнать точного значения метрики полноты (истинное значение меньше вычисленного таким образом), но позволяет сравнивать полноту результатов двух поисковых систем друг относительно друга.

По результатам сравнения на 35 запросах относительная полнота результатов поиска равняется 0.892 у терминологического поиска и 0.941 у системы Solr. Этот результат был ожидаем в силу строгого отбора результатов по параметру принадлежности одному терминологическому словосочетанию.

В подавляющем большинстве случаев потери релевантных документов при терминологическом поиске объясняются тем, что терминологические словосочетания разбиваются общей лексикой. Например, фрагмент «квадратичная форма $2Q$ является положительно определенной формой» разбивается общей лексикой «является» на два различных термина и ошибочно считается системой нерелевантным запросу «положительно определённая квадратичная форма».

Следует отметить интересный факт, что число документов, возвращаемых системой Solr, по результатам оценки 50 запросов на 20% выше числа документов в выдаче терминологической поисковой системы.

F-мера. Значение сбалансированной F-меры для терминологического поиска равняется $F_1(S_t(Q,D,TT)) = 0.905$. Полнотекстовый поиск Solr по данным эксперимента получил $F_1(S(Q,D)) = 0.834$.

Табл. 2

Результаты оценки точности ранжированного поиска

Уровень	1	2	3	4	5	6	7	8	9
Поиск по терминам	1	0.97	0.967	0.96	0.964	0963	0.966	0965	0.964
Solr	0.86	0.87	0.86	0.87	0.876	0.87	0.863	0.858	0.858

Можно видеть, что

$$F_1(S_t(Q, D, TT)) > F_1(S(Q, D)).$$

Таким образом, подтвердилась гипотеза о положительном влиянии использования дополнительной семантической информации о специальной лексике предметной области в тексте на результаты поиска в корпусе текстов этой предметной области.

Точность на уровне n документов. Качество ранжированного поиска оценивалось по точности до уровня 10 документов. Всего оценивались результаты для 50 запросов без ограничения на число результатов. Полученные значения приведены в табл. 2.

Для ранжирования результатов Solr использует метрику на основе $TF \cdot IDF$ (частота слова – обратная частота документа) [18]. Как можно видеть из табл. 2, эта метрика уступает метрике, основанной на близости терминологических словосочетаний.

Однако следует отметить, что, как и в случае с общей точностью, на качество результатов Solr существенное влияние оказывает исключение однословных запросов из процедуры оценки.

По результатам оценок можно заключить, что ранжирование на основе терминологической близости оказывается довольно эффективным. Например, на запрос «разностное неравенство» поиск Solr среди первых результатов выдаёт нерелевантный документ, в котором часто упоминаются «разностные схемы и операторные неравенства». Из-за большой встречаемости данного выражения статья получает высокий счёт $TF \cdot IDF$ и попадает в начало списка выдачи. В случае терминологического поиска статья с данным фрагментом также ошибочно попадает в список выдачи, однако по причине низкой оценки метрики близости строк оказывается в конце списка результатов.

Заключение

Предложена и построена система, осуществляющая индексацию в границах терминологических словосочетаний. Разработан способ организации подобной системы на базе поисковой платформы Solr. Для этих целей разработана и реализована в прототипе схема поисковой платформы, а также методы обращения к ней (создание и обновление индекса, обработка запроса).

Исследована проблема ранжирования результатов поиска на основе метрики схожести терминологических словосочетаний. Разработан подход к ранжированию, опирающийся на нововведенный атрибут – меру близости строк запроса и найденных терминов. В качестве метрики близости строк используется $N\text{-}GramDistance$. На примере объединения с ранжирующими метриками на основе пристатейных ссылок продемонстрированы возможность и способ сочетания разработанной терминологической метрики ранжирования с другими.

Проведено экспериментальное сравнение разработанных метрик с известными популярными аналогами. На основании полученных данных выбрана наилучшая метрика и реализована в прототипе.

Произведено сравнение качества результатов терминологического поиска с поисковой платформой Apache Solr. Оценивались метрики точности и точности на уровне n документов. Оба параметра показали некоторое преимущество терминологического поиска над полнотекстовым при условии поиска математических терминов в коллекции математических текстов.

По полученным косвенным данным о полноте можно ожидать, что значение этой метрики будет выше у полнотекстовой поисковой системы. Такой результат был предсказуем по причине существенного ограничения круга поиска границами терминологических словосочетаний.

Полученные экспериментальные данные позволяют утверждать, что введение дополнительного, специфического для терминологического поиска, атрибута при ранжировании позволяет улучшить качество результатов информационного поиска.

Суммируя вышесказанное, можно утверждать, что подход к поиску на основе выделения терминологии позволяет извлекать пользу из специализированности лексики рассматриваемой коллекции статей. Можно ожидать, что такие же схожие результаты будут и в других предметных областях при условии существенного вклада специальной лексики в тексты, относящиеся к этой предметной области. В качестве примеров таких областей можно привести естественные науки (физика, химия и др.), медицину, юриспруденцию.

Summary

D.A. Zaikin. A String Proximity-Based Approach to Result Ranking for Terminological Search.

In this article, a modification of the full-text approach to information retrieval is carried out by using semantic information about technical terms in text. The question of choosing a ranking metric for this type of search is raised. Several existing ranking metrics are discussed, and a new metric is proposed, which uses the features of terminological information retrieval. The proposed ranking metrics are experimentally compared, and the efficiency of the search system itself is estimated.

Keywords: information retrieval, ranking, technical terms.

Литература

1. *Roberts L.G.* Beyond Moore's Law: Internet Growth Trends // Computer. – 2000. – V. 33, No 1. – P. 117–119.
2. *Ding C.H., Buyya R.* Guided Google: A Meta Search Engine and its Implementation Using the Google Distributed Web Services // Int. J. Comput. Appl. – 2004. – V. 26, No 3. – P. 181–187.
3. *Koster C.H.A., Seibert O., Seutter M.* The PHASAR search engine // Proc. 11th Int. Conf. on Applications of Natural Language to Information Systems. – Berlin: Springer-Verlag, 2006. – P. 141–152.
4. *Egozi O., Markovitch S., Gabrilovich E.* Concept-Based Information Retrieval Using Explicit Semantic Analysis // ACM Trans. Inf. Syst. – 2011. – V. 29, No 2. – P. 8:1–8:34.
5. *Johannsson D.V.* Biomedical information retrieval based on document-level term boosting: Ph.D. Thesis. – Trondheim, Norway: Norwegian University of Science and Technology, 2009. – 69 p.

6. *Ramampiaro H.* Retrieving BioMedical Information with BioTracer: Challenges and Possibilities // Proc. Norsk Informatikk Konferanse (NIK 2009). – Trondheim, Norway: Tapir, 2009. – P. 49–60.

7. *Ramampiaro H., Li C.* Supporting BioMedical Information Retrieval: The BioTracer Approach // Transactions on Large-Scale Data- and Knowledge-Centered Systems IV (Lecture Notes in Computer Science. V. 6990). – 2011. – P. 73–94.

8. *Manning C.D., Raghavan P., Schutze H.* Introduction to Information Retrieval. – Cambridge: Cambridge Univ. Press, 2008. – 482 p.

9. *Smiley D., Pugh D.E.* Apache Solr 3 Enterprise Search Server. From technologies to solutions. – Packt Publishing, 2011. – 418 p.

10. DB-Engines Ranking of Search Engines. – URL: http://db-engines.com/en/ranking/search+engine.

11. *Page L., Brin S., Motwani R., Winograd T.* The PageRank citation ranking: Bringing order to the web. – Stanford: Stanford InfoLab, 1998. – 17 p.

12. *Kondrak G.* N-gram similarity and distance // String Processing and Information Retrieval (Lecture Notes in Computer Science. V. 3772). – Berlin: Springer, 2005. – P. 115–126.

13. *Михалевич В.С.* Словарь по кибернетике. – Киев: Гл. ред. Укр. сов. энцикл. им. М.П. Бажана, 1989. – 751 с.

14. *Tang J., Arni T., Sanderson M., Clough P.* Building a diversity featured search system by fusing existing tools // Proc. 9th Cross-language evaluation forum conference on Evaluating systems for multilingual and multimodal information access. – Berlin, Heidelberg: Springer-Verlag, 2009. – P. 560–567.

15. Apache Solr 4.2.1. – URL: http://lucene.apache.org/solr/4_2_1/.

16. *Goldberg J.H., Stimson M.J., Lewenstein M., Scott N., Wichansky A.M.* Eye tracking in web search tasks: design implications // Proc. 2002 symposium on Eye tracking research & applications. – N. Y.: ACM, 2002. – P. 51–58.

17. *Anick P.* Using terminological feedback for web search refinement: a log-based study // Proc. 26th annual int. ACM SIGIR conference on Research and development in informaion retrieval. – N. Y.: ACM, 2003. – P. 88–95.

18. *Aizawa A.* An information-theoretic perspective of tf-idf measures // Information Processing and Management. – 2003. – V. 39, No 1. – P. 45–65.

Поступила в редакцию
24.12.13

Заикин Данила Александрович – аспирант кафедры прикладной информатики, Казанский (Приволжский) федеральный университет, г. Казань, Россия.

E-mail: *ksugltronteal@gmail.com*