РАЗВИТИЕ МОДЕЛИ, ОСНОВАННОЙ НА ЗНАНИИ ОБ АВТОРАХ, ДЛЯ ПОИСКОВЫХ ПРИМЕНЕНИЙ

Молоканов В. О. (vmolokanov@it.ru),
Романов Д. А. (dromanov@it.ru),
Цибульский В. В. (vtsibulsky@it.ru)

НИУ Высшая школа экономики, Москва, Россия

Предлагается новая технология для широких поисковых применений к текстам естественного языка. Данная технология принимает во внимание информацию об авторстве документов и основана на анализе сети коммуникаций между авторами. Подробно рассматривается ее частное применение для задачи поиска экспертов. В качестве коллекций данных для проведения экспериментов используются корпуса TREC Enterprise track. Точность поиска экспертов, даваемая реализованной моделью, сравнима с наиболее эффективными современными информационно-поисковыми системами и движками. Обсуждается возможность применения описанного движка к другим поисково-аналитическим сценариям, таким как поиск плагиата, поиск информационных разрывов и др.

Ключевые слова: поиск экспертов, корпоративные коллекции большого объема, сетевые коммуникации, алгоритмы ранжирования

ADVANCES OF THE AUTHORSHIP-BASED MODEL FOR SEARCH APPLICATIONS

Molokanov V. O. (vmolokanov@it.ru),
Romanov D. A. (dromanov@it.ru),
Tsibulsky V. V. (vtsibulsky@it.ru)

National Research University “Higher School of Economics”, Moscow, Russia

A new technology is proposed for wide search applications to natural language texts. Its particular application to an expert search task is considered in details on the example of TREC Enterprise track. The vocabulary is treated statistically, but, as opposed to a standard TFIDF metric, two special metrics are used. They involve into calculations information about

lexicon usage by authors and communications between them. Calculating connection cardinality between an author and lexicon enables to reveal definite terms which are characteristic for an author so this author can be found with the help of such terms. Lexicon weighing allows to extract from the whole collection a small portion of vocabulary which we name significant. The significant lexicon enables to effectively search in thematically specialized knowledge field. Thus, our search engine minimizes the lexicon necessary for answering a query by extracting the most important part from it. The ranking function takes into account term usage statistics among authors to raise role of significant terms in comparison with others, more noisy ones. We demonstrate the possibility of effective expertise retrieval owing to several rationally built heuristic rating indicators. First, we receive an expert search efficiency that is comparable with the most effective modern information retrieval engines. Second, the chosen indicators allow to distinguish between "good" and "bad" queries. This is essentially important for further optimization of our engine. We discuss the possibility of applying our engine to other search and analytic scenarios such as plagiarism search, information gap retrieval and others.

Keywords: expert search, large-scale enterprise collections, network communications, ranking algorithms

1. Введение

Поиск экспертов является часто выполняемой работой в различных сценариях деятельности организации. В простейшей ситуации задача поиска экспертов становится востребованной при необходимости спросить что-либо у кого-либо по конкретной области профессиональной деятельности. В крупных организациях поиск экспертов неизбежно вовлекается в более сложные и даже глобальные сценарии, например, подбор персонала под новый или расширяемый проект для реализации корпоративных возможностей на рынке. В подобных сценариях задачу поиска экспертов предпочтительно решать автоматизированным способом с целью обеспечения максимально быстрого получения представления о том, кто из сотрудников организации наиболее осведомлен в интересующем вопросе.

Мы изучаем задачу поиска экспертов с помощью автоматизированных движков. Входом в такой задаче являются пользовательский запрос и поисковая коллекция исходных документов, а выходом — ранжированный список потенциальных кандидатов, иными словами, список людей, упорядоченный в порядке убывания вычисленной вероятности их экспертизы по теме запроса.

Задача поиска экспертов является предметом многих исследований. Помимо большого количества публикаций, поиску экспертов была посвящена в 2005–2008 гг. корпоративная дорожка (Enterprise track) серии конференций Text Retrieval Conference (TREC). На TREC участниками были продемонстрированы десятки различных поисковых моделей. Это не случайно, так как на сегодняшний день единой общепринятой модели поиска экспертов не существует.

В имеющихся работах, в основном, описываются два принципиальных подхода к поиску экспертов. Первый подход называется документо-ориентированным и предполагает выполнение поиска экспертов в две стадии: первичный поиск документов в соответствии с пользовательским запросом и последующий поиск людей в найденных документах [12]. Второй подход к поиску экспертов — человеко-ориентированный — подразумевает построение описания (так называемого профиля) для каждого человека, а поиск людей фактически производится в этих описаниях [2]. Несколько достаточно распространенных моделей, идентифицирующих связи между запросом и людьми, результативно применяются в рамках обоих подходов. Например, в широко используемой оконной модели [7], [9], [18], также известной как модель близости, связь считается тем сильнее, чем короче текстовое окно, содержащее упоминание о человеке и термины запроса. Несколько моделей, ориентированных на структуру или разметку документа [8], [13], [16], также могут давать хороший результат [1]. Здесь релевантность человека можно определять, в частности, на основании частоты упоминаний о нем в текстовых блоках, заголовки которых содержат термины запроса [5].

Было показано, что упомянутые два подхода не имеют преимуществ в итоговой эффективности друг перед другом [4]; в то же время экспериментально подтверждено, что дополнительного улучшения точности результата можно достичь путем сбора дополнительных сведений, в частности, вычислением веса документов, рассмотрением графа ссылок между документами, длины URL [17], деталей структуры документов [3], а также привлечением дополнительной информации извне коллекции [14].

Важная идея предлагаемой в настоящей статье модели поиска экспертов заключается в том, что процесс поиска экспертов можно организовать без предварительного нахождения документов по запрашиваемой теме. По существу, наша модель является человеко-ориентированной. Важно при этом отметить, что в ней нет необходимости собирать большие объемы дополнительной информации, учитывать какую-либо структуру текста или привлекать внешние данные. Однако, хотя в этом смысле наша модель проще, чем другие показанные на TREC человеко-ориентированные модели, она не уступает им в эффективности. Наоборот, высокая эффективность поиска экспертов в нашей модели достигается благодаря нескольким рационально построенным эвристическим метрикам, удачно моделирующим связь запрос-эксперт. Этим наша модель уникальна и резко отличается от моделей поиска экспертов, представленных на TREC.

Для проведения экспериментов мы используем две известные коллекции исходных текстовых данных: World Wide Web Consortium (W3C) и Commonwealth Scientific and Industrial Research Organization (CSIRO). Эти корпоративные коллекции были введены на дорожке TREC Enterprise track и предоставлялись каждому участнику дорожки для выполнения задания поиска экспертов. Мы применяем свой поисковый движок также к этим коллекциям. В настоящей статье мы фактически воспроизводим задания TREC 2005–2007, т. е. выполняем поиск экспертов на коллекциях W3C и CSIRO по соответствующим наборам запросов TREC. Это позволяет нам оценить степень эффективности нашего поискового ядра на фоне известных современных движков.

2. Модель, основанная на информации об авторстве документов

2.1. Основные вычисляемые параметры

Основой нашего поискового движка является статистическая модель, предназначенная для ранжирования людей относительно запроса, т. е. их выстраивания по вычисляемому уровню знаний заданной запросом темы. Полное математическое описание модели приведено в [11]. Помимо самих текстов документов, в вычислениях также используется информация об их авторстве, причем мы называем авторами людей, как отправляющих, так и получающих документы. Данная информация позволяет эффективно смоделировать сеть коммуникаций между присутствующими в коллекции людьми.

Используемые алгоритмы ранжирования экспертов опираются на две ключевые метрики, вычисляемые в нашей модели: это значимость лексических элементов (терминов, биграмм, сущностей) и сила их связи с авторами.

Методика вычисления значимости лексики достаточно нестандартна. Для определения значимости лексики мы вводим модификацию к методу TFIDF. Дело в том, что оригинальная мера TFIDF не принимает во внимание никаких сведений сверх информации об употреблении лексики в документах коллекции. Это означает, что она не позволяет делать различия между общеупотребительными, но редко встречающимися словами и словами, часто употребляемыми в небольших группах авторов. Однако в случае задачи поиска экспертов мы сочли важным реализовать такую возможность, т. е. отделить более часто употребляемые слова среди меньшего количества авторов. Ясно, что такие слова составляют профессиональную лексику, конкретизируют область знаний и соответствуют определенному кругу экспертов в ней.

Имея в виду упомянутую цель, при взвешивании лексики мы учитываем дополнительные сведения, присутствующие в коллекции, а именно, информацию об употреблении терминов авторами. Конкретно говоря, для каждого термина мы строим распределение частоты его употребления по авторам и определяем его значимость пропорционально дисперсии такого распределения, т. е. дисперсия частоты употребления термина — важный критерий его значимости. При этом коэффициент пропорциональности есть отношение средней частоты употребления термина к количеству авторов, употребивших термин; таким образом, приоритет явно отдается более часто употребляемым словам среди меньшего количества авторов.

Еще одну важную особенность модели представляет способ моделирования связи между лексикой и человеком. Такая связь выстраивается не только на основе частоты употребления термина автором, но также и в зависимости от топологических особенностей автора в подсети термина. Подсеть термина — это граф, в котором узлы представляют собой людей, отправивших или получивших термин, а ребра моделируют содержащие этот термин сообщения от отправителя к адресату. Ребра имеют вес, определяемый количеством сообщений с этим термином между двумя соответствующими авторами. В зависимости

от количества входящих и исходящих ребер, каждый автор получает свою характеристику в подсети термина, добавляющую вклад в силу его связи с термином. Таким образом, сила связи между термином и автором сравнивает автора с другими авторами и оказывает непосредственное влияние на вероятность его экспертизы по данной теме. Вычисление силы связи автора с лексикой, очевидно, также позволяет выделить наиболее сильно связанные с ним термины из всех терминов, с которыми у него есть связь. Мы полагаем, что такие термины характеризуют данного автора, а использование их в качестве запроса дает высокий шанс «вытянуть» его к первым позициям списка экспертов.

2.2. Функция ранжирования

Функция ранжирования определяет рейтинг человека относительно запроса. По существу, движок вычисляет ее для каждого найденного автора, сортирует авторов в порядке убывания их рейтинга и выполняет некоторую постобработку.

В самом общем виде функция ранжирования в нашей модели задается выражением

$$W(p) = \sum_i \sum_{x_i \in X_i} C_i \, S(x_i) \, L(x_i, \, p) \tag{1}$$

Здесь p — рассматриваемый человек, x_i — различные лексические элементы, X_i — множества лексических элементов различных типов в поисковом запросе, S — значимость лексического элемента, L — сила связи между лексическим элементом и автором, а C_i — числа, используемые в качестве свободных коэффициентов, задаваемых пользователем системы. В настоящей статье мы рассматриваем четыре типа лексических элементов: термины запроса (*tp*), расширяющие термины (*tt*), биграммы (*bp*), сущности (*ep*) (фактически выражение (1) обобщает формулу (6) из [11] на произвольное количество алгоритмов ранжирования авторов по разным типам лексики). Отметим также, что процедура выстраивания множества терминов расширения основана на применении модели близости к каждому случаю употребления термина запроса в коллекции, но в это множество включаются не все термины в текстовом окне, а только сильно связанные с термином запроса [11]. Путем введения дополнительного настроечного коэффициента принципиально возможно установить порог уровня ассоциации терминов расширения с запросом [10], но здесь вместо выполнения такой процедуры мы ограничиваем общее количество расширяющих терминов десятью терминами, рассматривая, таким образом, 10 наиболее сильно связанных расширяющих терминов с каждым запросом.

Итак, мы определяем функцию ранжирования как линейную комбинацию четырех параметров — выражений вида $W(p) = \sum_i \sum_{x_i \in X_i} C_i \, S(x_i) \, L(x_i, \, p)$,

имеющих смысл степени прямой связи человека с данным типом лексики x_i (термины запроса, расширяющие термины, биграммы, сущности). Эти вычисляемые в движке параметры, очевидно, можно рассматривать в качестве отдельных рейтинговых параметров человека, выдаваемых независимыми алгоритмами ранжирования, которые идентифицируют связи людей с соответствующим типом лексики, что и было продемонстрировано в [11].

3. Результаты

Мы провели ряд пусков нашего движка при различных сочетаниях настроечных коэффициентов системы. Производительность движка оценивалась показателем макроусредненной средней точности (MAP). В таблице 1 для пусков по коллекциям W3C (2005, 2006) и CSIRO (2007) приведены наборы таких коэффициентов C_{tp}, C_{tt}, C_{bp}, C_{ep}, которые дают максимально возможные значения MAP, достижимые на каждой из коллекций. Отметим, что коэффициенты C_{tp}, C_{tt}, C_{bp}, C_{ep} изменялись в широком диапазоне — от 0 до 100 с охватом существенно малых значений (до 10–17), а методология поиска максимального значения MAP на всем множестве данных по различным пускам являлась абсолютно эмпирической. Таблица 1 содержит также иные показатели точности получившихся оптимизированных пусков, такие как средняя точность на 5-й и 20-й позициях.

Таблица 1. Оптимальные значения настроечных параметров и соответствующие показатели точности

Пуск	C_{tp}	C_{tt}	C_{bp}	C_{ep}	MAP	P@5	P@20
hse2005q	0,4	0,001	0,57	0	0,1597	0,296	0,203
hse2006qMod	0,4	0,170	0,51	0	0,5954	0,620	0,513
hse2007Ent	0,5	$2{,}8 \cdot 10^{-17}$	2,50	0	0,3930	0,212	0,084

Сравнение с имеющимися результатами TREC 2005 [6], 2006 [15] и 2007 [1] позволяет сделать вывод, что наше поисковое ядро не уступает по точности результатов большинству участников TREC, т. е. сравнимо с наиболее эффективными информационно-поисковыми движками.

Определив наилучшие с точки зрения макроусредненной средней точности настройки нашего движка на коллекции CSIRO и зафиксировав списки первых пяти найденных авторов по каждому запросу (далее — исходные списки), мы выполнили следующий эксперимент по оценке устойчивости этих списков к изменению настроек. Мы провели четыре проверочных пуска, в каждом из которых один из четырех настроечных коэффициентов системы был установлен в 1, остальные — в 0; таким образом, мы задействовали каждый алгоритм ранжирования экспертов по отдельности. В каждом проверочном пуске на каждом запросе мы сравнили получившийся список первых пяти авторов с исходным списком, полученным на том же запросе, при этом в качестве естественной

меры совпадения мы выбрали количество записей, содержащихся в проверочном и исходном списках. В результате каждому запросу мы приписали число, равное сумме количеств совпадающих записей между исходным и каждым из проверочных списков. Это число имеет смысл характеристики устойчивости списка результатов по отношению к настройкам. Вычисление такой характеристики дает нам возможность оценить степень понимаемости запроса в движке. Действительно, если мы сгруппируем все запросы TREC по количеству совпадающих результатов в полученных на нашем движке списках, то ответ системы в среднем тем точнее, чем больше количество указанных совпадений между списками (таблица 2).

Это, в свою очередь, означает следующее. Если мы можем несколькими способами (т. е. применяя разные алгоритмы) получить одинаковые или близкие результаты в первых позициях списка, то вполне естественно считать релевантными авторов, содержащихся в каждом из результатов; применение *четырех* независимых способов поиска экспертов при этом гарантирует высокую надежность релевантности таких авторов. Если же несколько поисковых алгоритмов выдают непересекающиеся списки, то в этом случае запрос следует считать «плохим» для движка. Высокая чувствительность результата относительно настроек, вообще говоря, не дает возможности уверенно прогнозировать релевантность того или иного автора. Конечно, для повышения точности ответа на «плохие» вопросы можно пользоваться внешними по отношению к системе факторами, такими как перезаданием запроса вручную или его пояснением, предоставляемым коллекцией [10]. Однако интересным объектом для специального исследования представляется автоматическое уточнение запроса. Мы полагаем, что наш движок мог бы осуществлять такую функцию с использованием расширяющих терминов, т. е., терминов, имеющих сильную ассоциативную связь с запросом.

Таблица 2. Доля ответов с заданным уровнем средней точности (AP) для заданного числа совпадающих записей в первых пяти результатах списков

число совпадений	**0–4**	**5–8**	**9–14**
AP > 0,3	0,33	0,58	0,62
AP > 0,5	0,22	0,38	0,50
AP > 0,7	0,17	0,17	0,50

4. Сфера применения

Некоторые поисково-аналитические задачи, стоящие перед организациями и сообществами и вовлекающие до нескольких тысяч пользователей, могут быть благополучно решены с помощью описанного ядра. Такой характерный размер достаточно велик, чтобы отдельные пользователи уже не знали о сферах деятельности других пользователей, объединенных с ними в одно образование,

и одновременно недостаточен для предъявления чрезмерных требований к обычным для таких организаций и сообществ информационно-вычислительным ресурсам. В частности, для получения архива корпоративной переписки или форума интернет-конференции не требуется значительно бóльших ресурсов по сравнению с теми, которые обеспечивают их работу. Наше ядро эффективно сжимает получаемую информацию с тем, чтобы выделить и сохранить из информационного потока только значимую для решения задач поиска информацию, отбрасывая ненужные, избыточные сведения, подобно тому, как обычная поисковая система отбрасывает из потока слов частые короткие слова, т. н. стоп-слова. Последние обычно вводятся в поисковик словарным списком и, как правило, характерны для языка в целом. В нашем ядре значимые для поиска слова вычисляются в потоке информации, который далее используется как поисковый ресурс. Значимая лексика играет решающую роль в функции ранжирования, в которой влияние общеупотребимых слов на результат сильно снижено. Обычные для поисковых систем стоп-слова, наряду с другими, постоянно и более-менее равномерно употребляемыми каждым автором, не удаляются принудительно из словаря системы, а сильно понижаются в значимости на основе анализа потока.

Таким образом, движок выделяет слова, употребляемые относительно часто в небольших группах авторов на фоне всего сообщества. Подобное употребление лексики характерно для предметных обсуждений, когда, в отличие от всей совокупности авторов, некоторые из них постоянно обращаются к общей теме, например, приводя доводы и контрдоводы относительно какого-либо общего для них предмета. Вполне возможно представить себе ситуацию, когда в ходе обсуждения перечисляются атрибуты обсуждаемого объекта, часто упоминаются связанные с темой персоны, оценки, места, а также временные, финансовые и иные сущности.

Задачи, в которых в соответствии с запросом такого рода требуется определить связанные с объектом характеристики, могут быть решены нашим ядром лучше, чем при стандартном поисковом подходе, за счет уменьшения шума в ответе.

Указание автора документа в системе, по сути, представляет собой дополнительную разметку, нанесенную на коллекцию документов. Каждому документу приписывается признак, отличающий его по авторству, однако документы также могут быть размечены и по иным признакам, например, по структурным подразделениям авторов, по тематическим веткам форумов, по времени написания — по любым признакам, которые в общем случае могут отсутствовать не только в текстах документов, но и во всей коллекции (в частности, такая информация, как способ пересылки документа от автора к читателям, может быть известна лишь извне коллекции). В этом ключе естественную разметку документов авторами можно воспринимать как *классификацию* коллекции по авторам. Но, поскольку элементы классификатора могут быть иными, следует, что ядро в принципе способно решать задачи классификации документов по заданному элементу. К таким задачам относятся плоская классификация, иерархическая классификация и многомерная классификация коллекции документов.

Информация об авторе документа может быть дополнена информацией о его получателе. Для корпоративной переписки это пары отправитель-адресат, а для форума в сообществе — автор поста-посетитель. Наличие дополнительной (классифицирующей) информации у документа позволяет выявить группы авторов со сходной лексикой, которые не участвуют в прямых коммуникациях друг с другом. При этом реализуется алгоритм поиска информационных разрывов — ситуаций, когда похожие или одинаковые темы обсуждаются отдельными группами участников проекта независимо, изолированно от остальных групп. Такие обсуждения могут впоследствии материализоваться в разных конечных решениях или продуктах, приводя, с одной стороны, к разнообразию спектра деятельности или продукции организации, а, с другой стороны, к значительным непроизводительным потерям ресурсов на унификацию постфактум или устранение противоречий в запущенных стадиях.

Информация об авторах дает возможность реализовать поиск заимствований одного автора у других (плагиата). Для реализации такого алгоритма производится фрагментация документов и анализируется значимая лексика во фрагментах. Применение энтропийного анализа к фрагментам позволяет выделить сходные текстовые отрезки. Построенная на нашем ядре система поиска плагиата проявляет чувствительность к фрагментам, полученным из исходных текстов механическими преобразованиями текста: перестановками, заменой синонимами, перефразированием и добавлением незначимых слов.

Представленное ядро может быть также применено в задачах кластеризации коллекции документов. В одном из применений на коллекции рефератов диссертаций система показывает осознаваемую пользователем тематическую кластеризацию текстов.

5. Заключение

Мы представили поисковое ядро, базирующееся на принципах учета авторства документов. В качестве авторов рассматриваются как человек, отправивший документ, так и человек, получивший документ, т. е. учитываются и источник, и приемник информации. По статистическому принципу производится рассмотрение лексики, но здесь, в отличие от стандартной метрики TFIDF, используются две особые метрики, которые включают в вычисления сведения об употреблении терминов автором и о коммуникациях автора с другими авторами. Вычисление силы связи автора с лексикой позволяет выявить определенные термины, которые являются характеристическими для автора и с помощью которых можно его находить. Взвешивание лексики дает возможность выделить из всей коллекции небольшую порцию лексики, которую мы называем значимой. Значимая лексика позволяет эффективно проводить поиск по тематически специализированным запросам в конкретизированной области знаний. Таким образом, поисковый движок производит минимизацию лексики, необходимой для ответа на запрос, выделяя ее самую важную часть. Функция ранжирования учитывает статистику употребления слов среди авторов, повышая роль

значимых терминов на фоне остальной, более шумовой лексики. Извлеченные слова могут определять конкретные сущности, отношения, и являются подходящими для выявления не только авторов, но и их предметных сфер.

Для того, чтобы выяснить качество работы нашего ядра на фоне других систем, мы сравнили его с движками, представленными на TREC. Результат в целом оказался весьма успешным — даже лучше, чем у большинства передовых поисковых движков. Итак, основная поисковая функция нашего движка выполнена на высоком уровне.

Помимо поиска экспертов, существует принципиальная возможность эффективного применения нашего ядра к иным поисково-аналитическим задачам. Например, на нашем движке построены система поиска информационных разрывов и система поиска плагиата. Кроме того, предварительные результаты моделирования задачи тематической кластеризации авторефератов диссертаций указывают на реализуемость и улучшаемость функциональности кластеризации документов. Наконец, функциональность поиска экспертов в нашем движке можно обобщить на разнообразные задачи классификации информации по произвольным классификационным элементам.

Благодарности

Данная работа выполнена в рамках проекта по созданию сервисов «Поиск экспертов», «Выявление заимствований в текстовых документах» и разработке тиражного продукта «Логика.ECM.Правовая экспертиза» при сотрудничестве с компанией «Логика бизнеса 2.0». Коллектив авторов благодарит руководство компании за оказанное содействие и предоставление технических средств.

Литература

1. *Bailey P., Craswell N., de Vries A. P., Soboroff I.* Overview of the TREC 2007 enterprise track. Proceedings of the 2007 Text REtrieval Conference (TREC 2007), Gaithersburg, MD, 2007, pp. 30–36.
2. *Balog K., Azzopardi L., de Rijke M.* Formal models for expert finding in enterprise corpora. Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, 2006, pp. 43–50.
3. *Balog K., de Rijke M.* Associating people and documents. Proceedings of the European Conference on IR Research (ECIR-08), Berlin, 2008, pp. 296–308.
4. *Balog K., Soboroff I., Thomas P., Bailey P., Craswell N., de Vries A. P.* Overview of the TREC 2008 enterprise track. Proceedings of the 2008 Text REtrieval Conference (TREC 2008), Gaithersburg, MD, 2008, pp. 14–25.
5. *Cao Y., Liu J., Bao S., Li H., Craswell N.* (2008). A two-stage model for expert search. Technical Report MSR-TR-2008-143, Microsoft Research.

6. *Craswell N., de Vries A. P., Soboroff I.* Overview of the TREC-2005 enterprise track. Proceedings of Fourteenth Text REtrieval Conference (TREC 2005), Gaithersburg, MD, 2005, pp. 16–22.
7. *Duan H., Zhou Q., Lu Z., Jin O., Bao S., Cao Y., Yu Y.* Research on enterprise track of TREC 2007 at SJTU APEX lab. Proceedings of the 2007 Text REtrieval Conference (TREC 2007), Gaithersburg, MD, 2007, pp. 489–498.
8. *Fu Y., Yu W., Li Y., Liu Y., Zhang M., Ma S.* THUIR at TREC 2005: enterprise track. Proceedings of Fourteenth Text REtrieval Conference (TREC 2005), Gaithersburg, MD, 2005, pp. 772–779.
9. *He B., Macdonald C., Ounis I., Peng J., Santos R. L. T.* University of Glasgow at TREC 2008: experiments in blog, enterprise, and relevance feedback tracks with Terrier. Proceedings of the 2008 Text REtrieval Conference (TREC 2008), Gaithersburg, MD, 2008, pp. 368–380.
10. *Molokanov V., Romanov D., Tsibulsky V.* Optimization of algorithms and parameter settings for an enterprise expert search system. Proceedings of the International Conference on Computer Science, Information System and Communication Technologies (ICCSISCT 2012), Bangkok, 2012, pp. 79–88.
11. *Molokanov V., Romanov D., Tsibulsky V.* (2013). A new model for enterprise expert retrieval. International Journal of Computer and Communication Engineering, Vol. 2, pp. 201–205.
12. *Petkova D., Croft W. B.* Hierarchical language models for expert finding in enterprise corpora. Proceedings of the IEEE International Conference on Tools with Artificial Intelligence 2006, Washington, 2006, pp. 599–608.
13. *Ru Z., Li Q., Xu W., Guo J.* BUPT at TREC 2006: enterprise track. Proceedings of Fifteenth Text REtrieval Conference (TREC 2006), Gaithersburg, MD, 2006, pp. 151–156.
14. *Serdyukov P., Hiemstra D.* Being omnipresent to be almighty: The importance of the global web evidence for organizational expert finding. Proceedings of the SIGIR Workshop on Future Challenges in Expertise Retrieval, Singapore, 2008, pp. 17–24.
15. *Soboroff I., de Vries A. P., Craswell N.* Overview of the TREC 2006 enterprise track. Proceedings of Fifteenth Text REtrieval Conference (TREC 2006), Gaithersburg, MD, 2006, pp. 32–51.
16. *You G., Lu Y., Li G., Yin Y.* Ricoh research at TREC 2006 enterprise track. Proceedings of Fifteenth Text REtrieval Conference (TREC 2006), Gaithersburg, MD, 2006, pp. 570–582.
17. *Zhu J., Huang X., Song D., Rüger S.* (2010). Integrating multiple document features in language models for expert finding. Knowledge and Information Systems, Vol. 23, pp. 29–54.
18. *Zhu J., Song D., Rüger S.* The Open University at TREC 2007 enterprise track. Proceedings of the 2007 Text REtrieval Conference (TREC 2007), Gaithersburg, MD, 2007, pp. 431–434.