тсюда можно сделать вывод, что у разработанного
ритма минимизирована работа оператора по руч-
тработке результатов, т.е. хотя скорость обработ-
есколько меньше, но алгоритм позволяет суще-
но разгрузить операторов за счёт интеллектуаль-
истемы принятия решений, чего не может предло-
алгоритм прямого сравнения.

ри сравнении экономических характеристик раз-
танного программного обеспечения на основе
ываемого алгоритма с процедурой прямого
ения, для годового объёма идентификации в
000 физических лиц были получены следующие
ые: трудовые затраты на обработку информации
етоду нечёткого сравнения по сравнению с мето-
прямого сравнения уменьшены в 6,7 раза, абсо-
ое снижение трудовых затрат составило 1446
, годовые затраты при использовании метода
кого сравнения уменьшились в 3 раза по сравне-
аналогичным периодом применения метода пря-
сравнения, а годовой экономический эффект пре-
580000 руб. Для наглядности некоторые стои-
ые показатели, формирующиеся при использо-
разработанного и применявшегося до настояще-
мени программного обеспечения, отображены на
амме, приведенной на рис. 2. Величины затрат
ены по оси ординат в рублях.

очение

бучающиеся системы позволяют освободить
ческие ресурсы для выполнения творческих
В этой области технология Data Mining предо-
ет полный набор теоретических и практических
в для выбора, разработки или использования
лектуальных компьютерных систем.

ссмотренную в статье процедуру идентификации
рассматривать как часть системы поддержки
тия решений (СППР). Процедура не требует
тельства оператора, накапливает опыт и само-
тся в процессе работы, позволяя, тем самым,
стью освободить специалистов от низкопро-
ой, неэффективной, ручной работы напрямую с
ми реквизитов физических лиц, хранящимися в
анных. Данная процедура реализована на языке
L СУБД Oracle 11g и успешно функционирует с
2007 г. в муниципальном учреждении
ттинский городской информационный центр».
еская структура разработанного алгоритма (рис.
воляет реализовать его на любом популярном
программирования.

ерспективе данный алгоритм обладает возмож-
успешного внедрения в системы глобального
нения хранилищ государственных или коммер-
организаций для ведения единой базы данных
ния любой страны мира. Масштабируемость алгоритма позволяет применять программные процедуры на его основе как в малых организациях, так и в крупных корпорациях, везде, где ведётся и актуализируется реестр данных физических лиц. Возможные примеры использования: портал госуслуг, медицинские электронные системы, кадровые и бухгалтерские системы учёта служащих, банковские системы хранения данных о клиентах и т.п.

ЛИТЕРАТУРА

1. Чубукова И.А. «Data Mining»: учебный курс. // Издательство Интернет-университета информационных технологий (http://www.intuit.ru/), 2006.
2. Международный фонд автоматической идентификации. Технологии автоматической идентификации. (http://www.fond-ai.ru/art1/art223.html).
3. Подборка материалов о международном опыте законодательного регулирования использования систем идентификации личности (http://www.kongord.ru/Index/Prison/SViP.htm)
4. Отчёт о выполнении научно-исследовательской, опытно-конструкторской работы «Разработка механизмов однозначной идентификации данных о физических лицах и объектах недвижимости, хранящихся в различных информационных системах органов государственной власти и местного самоуправления (http://www.nisse.ru/business/article/article_464.html)
5. Положение о персональном идентификационном номере граждан Российской Федерации, проживающих или пребывающих на территории Санкт-Петербурга. (http://iac.spb.ru/shablon.asp?subpage=171&id=40&dir=0).
6. Проект «Социальная карта москвича» (http://www.soccard.ru).
7. Сборник тезисов городской научно-практической конференции студентов, аспирантов, преподавателей вузов и специалистов муниципальных учреждений г.Тольятти «Информатизация в социальной сфере» (http://it-exclusive.ru/idperson/docs/stat.doc).
8. Урман С. «ORACLE 9i – Программирование на языке PL/SQL»: учебное пособие // Oracle Press – издательство «Лори», 2004.

Наталья Игоревна Лиманова,
д-р техн. наук, профессор.
Тольяттинский гос. университет

Максим Николаевич Седов,
инженер-программист I-ой категории.
Мэрия городского округа Тольятти.

Ю.А. Ставенко, Д.А. Романов, А.И. Громов

ОПРЕДЕЛЕНИЕ ЗНАЧИМОСТИ ЛЕКСИКИ НА ОСНОВЕ СТАТИСТИЧЕСКОГО КОНТЕНТ-АНАЛИЗА

В статье приведен алгоритм контент-анализа корпоративных документов с целью выявления уровня экспертизы сотрудников в той или иной предметной области. В процессе контент-анализа были выявлены наиболее значимые слова, описывающие смысл документов, сгенерированных сотрудником за определенный период времени, а также был предложен способ отсеивания общеупотребительной лексики.

Ключевые слова: контент-анализ, значимость лексики, поиск экспертизы, корпоративный поиск, TF, IDF

Yu.A. Stavenko, D.A. Romanov, A.I. Gromov

ESTIMATION OF TERM INFORMATIVENESS BASED ON A STATISTICAL CONTENT-ANALISIS

The article is dedicated to the content analysis algorithm of corporate documents in order to extract the level of expertise of each expert in a particular domain area. During the process of content analysis the most significant words were identified that describe the meaning of documents generated by the employee for a period of time, and a method of screening vernacular vocabulary was proposed.

Keywords: content analysis, significant words, expertise search, TF, IDF

Введение

Информация, представленная в виде текстов на естественном языке, является одним из самых распространенных видов информации. Соответственно, существует множество способов контент-анализа текстов (например, аннотирование, конспектирование, реферирование), позволяющих извлекать и оценивать значимую информацию (знания) из текстов. Философский смысл контент-анализа как исследовательского метода, состоит в восхождении от многообразия текстового материала к абстрактной модели содержания текста.

Прародителем контент-анализа можно считать проведенный анализ сборника церковных гимнов в Швеции в 18 веке. Эти гимны сначала прошли государственную цензуру и приобрели большую популярность, после чего были обвинены в несоответствии религиозным догматам. Для того чтобы доказать некорректность текстов с точки зрения официального учения церкви, в них был осуществлен подсчет религиозных символов, количество которых было затем сравнено с другими религиозными текстами, в том числе с теми, которые считались еретическими. Частота использования определённых, заранее собранных слов и тем позволяла судить о том, насколько корректен текст [1].

Что касается дальнейшего развития контент-анализа, то огромную популярность этот метод приобрел после появления различного рода политических партий в начале 20 века, например, он использовался Максом Вебером в 1910 году для анализа освещаемости прессой политических акций в Германии. Позднее, в 1937 году метод контент-анализа был использован в США рационных речей американских президентов, а затем в военных целях и в целях исследования направлений пропагандистской деятельности.

Огромный вклад в контент-анализ текстов был внесен французским журналистом Ж. Кайзером, который разработал методику статистического анализа публикаций СМИ, а затем Э. Морэн, которая ввела в научный оборот термин *«единица информации»* - блок, содержание которого определяет смысловую нагрузку текста.

Затем метод контент-анализа получил большое распространение в различных сферах политологии и социологии, и успешно используется в них до сих пор за счет своей эффективности. Эта эффективность достигается с помощью упрощения (редукции) текста к стандартному набору смысловых единиц (терминов), которые могут быть обработаны методами статистики, в результате чего на выходе будет получен значимый результат для целей исследования. Методами обработки наиболее часто являются частота упоминаний терминов в массиве текстов, на основании которой делается вывод об информативности термина и характеристиках текста.

В данной статье будет сделана попытка применения метода контент-анализа к корпоративным текстам, созданным сотрудниками организации. Выбор данного массива текстов не случаен. В организациях любого уровня каждый день встают вопросы поиска сотрудника, имеющего достаточный уровень компетенции для решения вопросов конкретной предметной области. Поиск необходимой компетенции - одна из самых приоритетных задач управления знаниями, которая, к сожалению, во многих организа-

ручного профилирования. Основная проблема заключается в том, что неявные знания, присутствующие у сотрудников, слабо структурированы и должны быть извлечены из корпоративной документации. Соответственно, задача поиска неявных знаний сильно взаимосвязана с задачей контент-анализа текстов документов.

Цель и актуальность исследования

Одной из задач извлечения информации из текста является выделение ключевых терминов, с определённой степенью достоверности отражающих тематическую направленность документа. Автоматическое извлечение ключевых терминов можно определить как автоматическое выделение важных тематических терминов в документе. Оно является одной из подзадач более общей задачи – автоматической генерации ключевых терминов, для которой выделенные ключевые термины не обязательно должны присутствовать в данном документе [2].

В последние годы было создано множество подходов, позволяющих проводить анализ наборов документов различного размера и извлекать ключевые термины, состоящие из одного, двух и более слов. Самыми распространёнными схемами для расчёта весов терминов являются *TF-IDF и различные его модификации*. В общем-то, индекс *TF-IDF* уже можно использовать для ранжирования результатов поиска – чем выше индекс, тем больше релевантность данного термина в данном документе. Применительно к задаче поиска экспертов метрика IDF может быть переопределена как логарифм отношения числа авторов, хотя бы раз употребивших термин, к общему числу авторов. Но поскольку число авторов в корпоративной среде определяется числом сотрудников компании и фактически равно константе (или меняется достаточно медленно), метрика IDF зависит только от числа авторов, употребивших данный термин. Тем не менее, для задачи поиска компетенции этого всё ещё недостаточно, так как при сравнении вектора *TF-IDF* поискового запроса и вектора *TF-IDF множества документов автора* незначимые, часто встречаемые термины вносят слишком большую вариацию в результат. Поэтому представляется важным различать общеупотребительность и частоту встречаемости термина. Поскольку для решения задачи поиска экспертов применяется, и будет применяться ограниченный по объему корпус текстов, накопленный за конечное (и, часто, весьма небольшое) время, некоторые вполне общеупотребительные, но редко встречающиеся слова могут просто «не успеть» проявиться в лексике всех авторов. В этом случае метрика IDF для таких терминов окажется завышенной, хотя сами по себе термины не будут являться значимыми.

Целью исследования является разработка алгоритма для программного сервиса внутрикорпоративного поиска экспертизы на основании контент-анализа. С математической точки зрения необходимо разработать алгоритм нахождения наиболее релевантных экспертов с помощью представления запроса пользователя и контента, генерируемого сотрудниками организации в виде набора значимых терминов $\{t_k\}$ и $\{t_i\}$ соответственно. Каждому термину t_i должна соответствовать мера $Weigth(t_i; p_j)$, определяющая значимость данного термина t_i для эксперта p_j.

Методологическая база. Выделение значимых терминов

Извлечение значимых терминов является базисным этапом для выделения экспертов. Значимыми терминами являются ключевые термины, которые могут дать интерпретацию смысла документов, созданных сотрудниками. Это задача обработки естественного языка, которая решается универсальными методами с помощью подсчета частотных характеристик употребления терминов, например, в поисковых системах.

Методологической базой для алгоритмов, основанных на выявлении статистических закономерностей распределения частоты появления различных слов в текстах, принято считать закон Ципфа (Zipf's Law) [3]. Для его интерпретации необходимо посчитать частоту появления слов в тексте TF (term frequency). Если расположить слова текста в порядке убывания частоты их появления TF, начиная с наиболее часто встречающихся, то произведение частоты слова $(TF)_i$ на порядковый номер частоты будет постоянным для любого данного слова t_i:

$$TF_i \times r_i = C. \qquad (1)$$

где C – некоторая константа, r_i – порядковый номер (ранг) частоты слова.

На рис. 1 представлен график закономерности «ранг-частота» для текста научной статьи [3].

Ципф выявил, что для закономерностей ранг-частота всегда существуют три явно различаемые зоны ранговых распределений: зона ядра рангового распределения (наиболее часто употребляемые слова языка: «the», «a», «that»), центральная зона и зона усечения. В центральной зоне находятся слова, максимально характеризующие данный текст и выражающие его специфичность и тематику, то есть ключевые термины: «zipf, «low». В зоне усечения находятся слова, не несущие основной смысловой нагрузки. От ширины центральной зоны зависит

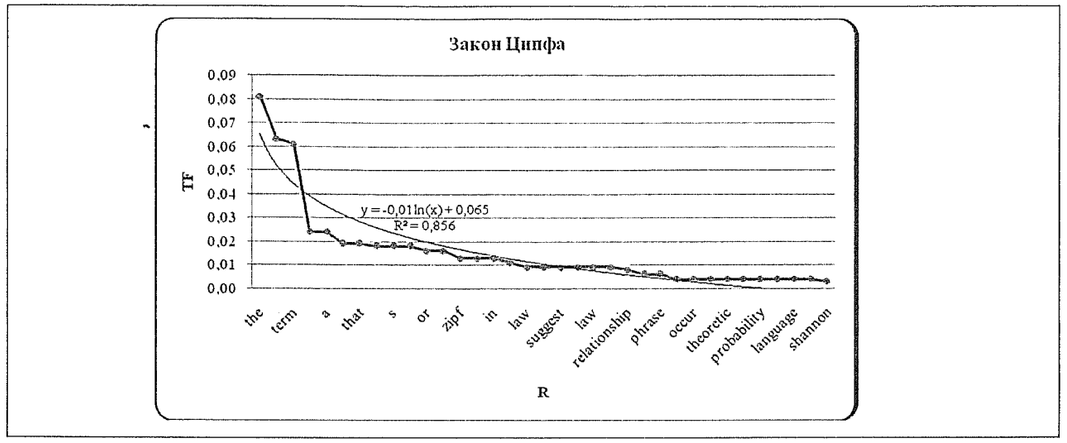

Рис. 1. Закон Ципфа

качество выделяемых ключевых терминов. На этапе определения ширины центральной зоны первостепенной задачей является отсечение информационного шума – нейтральных стоп-слов. Традиционный способ решение данной задачи заключается в применении специализированных словарей, содержащих наиболее часто встречающиеся языковые единицы (местоимения, предлоги, союзы и т.п.).

В настоящем исследовании также стоит вопрос выделения значимых терминов в текстах, написанных сотрудниками, для автоматического профилирования сотрудников организации и выделения их профессиональной экспертизы. Воспользовавшись следствием из закона Ципфа, на первом этапе предварительной обработки из массива выделенных слов из текстов сотрудников были удалены все стоп-слова.

Следующим этапом необходимо выделить ключевые термины, характеризующие профиль сотрудника (его предметную область). При этом подобная задача должна решаться в соответствие со спецификой каждой организации (социальной среды).

Одним из традиционных методов выделения значимых терминов является расчет показателя IDF (inverse document frequency) – инверсия частоты, с которой некоторый термин встречается в документах коллекции, то есть значение IDF тем меньше, чем чаще встречается слово. По сути IDF оценивает количество информации, свойственной слову. Применительно к задаче составления профиля сотрудника показатель IDF будет рассчитываться следующим образом для термина t_i:

$$IDF(t_i) = \frac{\lg N}{nt}, \qquad (2)$$

где N – общее число экспертов; nt – число экспертов, употребивших термин t_i хотя бы один раз.

Ключевыми в данном случае расчета значимости будут являться термины, набравшие наибольший вес IDF. Таким образом, для слов, которые употребляются большим числом экспертов, IDF будет близок к нулю.

Смысл применения подобного алгоритма для поиска экспертов состоит в доказательстве гипотезы, что значимость термина пропорциональна количеству встречаемости термина в документах автора и обратно пропорциональна количеству встречаемости термина во всей коллекции документов. К сожалению, эта гипотеза может быть опровергнута, так как метрика IDF не может учитывать зависимость появления терминов во времени, что критично для анализа динамически изменяющегося информационного поля организации. Поскольку для решения задачи поиска экспертов применяется и будет применяться ограниченный по объему корпус текстов, накопленный за конечное (и, часто, весьма небольшое) время, некоторые вполне общеупотребительные, но редко встречающиеся слова могут просто «не успеть» проявиться в лексике всех авторов. В этом случае метрика IDF и, соответственно, значимость для таких терминов окажется завышенной, хотя сами по себе они не будут являться значимыми. Зависимость от организационно-социальной среды и зависимость от времени обусловливают необходимость динамического определения значимости термина.

Гипотеза исследования

Для разрешения вопроса отделения значимых терминов от общеупотребительных, описанного выше, можно сформулировать гипотезу, что характеристиками рангового распределения обладают не только зависимость ранга от частоты употребления слова в

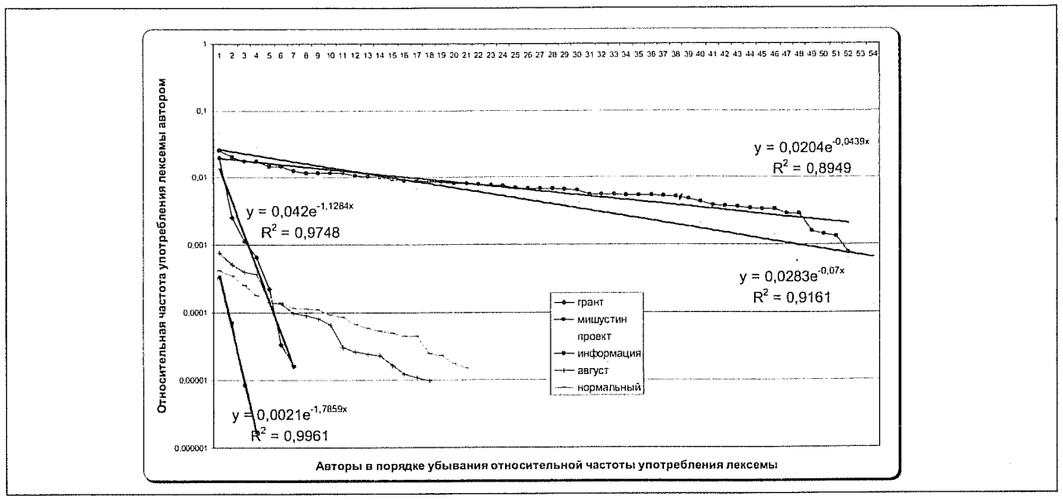

Рис. 2. Зависимость относительной частоты употребления термина автором от ранга в би-логарифмической шкале

тексте (закон Ципфа), но и зависимость ранга от относительной частоты употребления термина автором. Для этого необходимо подсчитать статистику *относительной частоты употребления термина* t_i для всех текстов, написанных конкретным сотрудником p_j:

$$TF(t_i, p_j) = \frac{m(t_i, p_j)}{\Sigma_k m_k}, \quad (3)$$

где $m(t_i, p_j)$ – число употреблений термина t_i персоной p_j, а в знаменателе – общее число употреблений всех терминов персоной p_j.

Можно предположить, что значимые термины должны иметь сильно неравномерное распределение относительной частоты употребления среди сотрудников, а общеупотребительные – примерно одинаковую относительную частоту употребления. Построим подобные зависимости в двойной логарифмической шкале для различных терминов (см. рис. 2), выбранных из текстов сотрудников ИТ-компании[1]:

Интуитивно понятно, что такие термины как «грант» и «мишустин» значимы, так как «грант» - специфический термин предметной области, «мишустин» – имя собственное. Термины «проект», «информация» являются общеупотребительными в выбранной предметной области. Анализируя график видно, что для значимых слов и общеупотребительных отличается характер распределения, а именно дисперсия в наблюдаемом распределении относительных частот употребления терминов. В результате эксперимента можно сделать вывод, что именно более высокие значения дисперсии относительной частоты употребления термина являются тем самым критерием, который позволит отличить значимые термины от общеупотребительных.

Для подтверждения гипотезы в проведенном исследовании было проведено сравнение двух алгоритмов поиска экспертов на основании определения значимости терминов:

- алгоритма, основанного на подсчете IDF и используемого в векторной модели TF*IDF;
- алгоритма, основанного на вычислении дисперсии распределения относительной частоты встречаемости терминов.

Лингвистический анализ текста. Подготовка к работе алгоритмов

В проведенном исследовании поиск экспертов осуществлялся на основе обработки текстового контента, образующегося в корпоративной сети (включая архив электронной почты и массивы других видов текстовых сообщений). Лингвистический анализ текстов реализуется в виде лингвистического процессора, задачей которого является преобразование текста на естественном языке в некоторый набор структур, являющихся формальным представлением контента исходного текста.

Для выделения терминов из текстов применяется парсинг документов, созданных сотрудниками организации, цель которого – провести токенизацию (графематический анализ, лексический анализ), на основании которой происходит выделение токенов – лексических конструкций, состоящих из последовательности символов одного типа (слов, цифровых записей).

[1] Эмпирическую базу исследования составили документы электронной почты компании «АйТи».

После этого происходит морфологический анализ с помощью морфологического анализатора, на вход которого поступают токены. С помощью него происходит приведение выделенных токенов к словарной форме, а также установление грамматических характеристик словоформ (род, число, падеж), которые используются для выделения словосочетаний. Результаты морфологического анализа используются на этапе синтаксического анализа, в частности. Модуль синтаксического анализа позволяет частично разрешить грамматическую омонимию как для идентифицированных, так и для отсутствующих в словаре слов. На следующем этапе происходит синтаксический анализ – выделение не только слов, но и именных групп – это выражение, центральным элементом которого является существительное.

В итоге на выходе получается сформированный список терминов, привязанных к автору – сотруднику организации.

Алгоритм поиска экспертов

Для каждого термина t_i и сотрудника p_j считается статистическая информация по количеству употребления термина сотрудником (см. формулу 3) – относительная частота употребления $TF(t_i, p_j)$. В итоге для каждого термина можно составить выборку относительной частоты употребления этого термина авторами $TF(t_i, p_j)$:

$$\begin{matrix} TF(t_1, p_1) & \dots & TF(t_1, p_N) \\ \dots & \dots & \dots \\ TF(t_n, p_1) & \dots & TF(t_n, p_N) \end{matrix}, \quad (4)$$

где n – количество терминов в информационном поле организации, N – количество авторов.

Можно предложить несколько вариантов метрик для определения значимости термина, основанных на идее нелинейного распределения вероятности. Простейшие метрики для степени изменения значений выборки определяются путем вычисления их дисперсии $D(t_i)$, то есть дисперсии относительной частоты употребления термина t_i и стандартного отклонения σ:

$$\begin{aligned} D(t_i) = & \left((TF(t_i, p_1) - M)^2 + \right. \\ & + (TF(t_i, p_2) - M))^2 + \\ & \left. + (TF(t_i, p_N) - M))^2 \right) / (N - 1) = \\ & 1/(N-1) \sum_{k=1}^{N} (TF(t_i, p_k))^2 \end{aligned} \quad (5)$$

где M – оценка математического ожидания (выборочное среднее) относительной частоты употребления термина, рассчитывающаяся по формуле:

$$M(t_i) = \frac{TF(t_i, p_1) + TF(t_i, p_2) + \dots + TF(t_i, p_N)}{N} = \frac{1}{N} \sum_{k=1}^{N} TF(t_i, p_k) \quad (6)$$

Среднеквадратичное отклонение относительной частоты употребления термина – корень из дисперсии:

$$\sigma(t_i) = \sqrt{D(t_i)} \quad (7)$$

Тогда значимость термина для конкретного автора будет являться превышением над математическим ожиданием $M(t_i)$, измеренным в среднеквадратических отклонениях $\sigma(t_i)$:

$$Impact(t_i : p_j) = \frac{(TF(t_i, p_j) - M(t_i))}{\sigma(t_i)} \quad (8)$$

По формуле (8) получатся, что отрицательные значения $Impact(t_i; p_j)$ будут у авторов, которые редко употребляли термин, то есть не экспертов по этому термину.

Для нормировки веса $Impact(t_i; p_j)$ возьмем арктангенс, тогда чем выше значение арктангенса, тем слово значимее для автора:

$$FinalImpact(t_i; p_j) = \operatorname{arctg} Impact(t_i; p_j) \quad (9)$$

Мерой близости компетенции эксперта к поисковому запросу (тексту) будет являться сумма весов термов, употребленных в тексте для автора:

$$Weight(p_j) = \sum_{i}^{l} FinalImpact(t_i; p_j) \quad (10)$$

где t_i – термин в запросе пользователя; l – количество терминов в запросе пользователя.

Произведя ранжирование сотрудников по мере $Weight(p_j)$ получим релевантный список экспертов.

Алгоритм поиска экспертов на основании векторной модели

Для определения эффективности предложенного алгоритма будем сравнивать его со стандартным алгоритмом TF×IDF, который используется при информационном поиске. Главной идеей этого алгоритма является рассмотрение документов и запросов как векторов в пространстве слов, а релевантность документа и запроса – как расстояние между ними. В качестве меры подобия используется "мера косинуса", которая считается как косинус угла между векторами. Она хороша тем, что принимает значения от 0 до 1, и равна 1 только при полном подобии векторов. В исследовании была использована аналогичная модель для определения релевантности эксперта запросу пользователя.

На первом шаге алгоритма происходит построение матрицы относительной частоты употребления тер-

мина t_i авторами (см. формулу 4), которая умножается на показатель $IDF(t_i)$, для термина t_i (см. формулу 2). Таким образом, каждый эксперт p_j характеризуется вектором $TF(t_i, p_j)\cdot IDF(t_i)$, рассчитанным по всему текстовому информационному полю организации (по всем терминам).

Для того чтобы сопоставить запрос пользователя с компетенцией эксперта, для запроса вычисляется вектор $TF(t_i, d)\cdot IDF(t_i)$, где

$$TF(t_i, d) = \frac{n_i}{\sum_k n_k} \quad (11)$$

где n_i – число вхождений термина t_i в запрос d, а в знаменателе общее число вхождений всех терминов в запрос d.

Мерой близости компетенции эксперта к поисковому запросу (тексту) будет являться мера косинуса между векторами:

$$\cos\theta = \frac{A \cdot B}{|A||B|} = \frac{\sum_i^n \left[TF(t_i, d) * IDF(t_i)\right] * \left[TF(t_i, p_j) * IDF(t_i)\right]}{\sqrt{\sum_i^n \left[TF(t_i, d) * IDF(t_i)\right]^2} * \sqrt{\sum_i^n \left[TF(t_i, p_j) * IDF(t_i)\right]^2}} \quad (12)$$

Произведя ранжирование сотрудников по мере *cosu* получим релевантный список экспертов.

Оценка эффективности алгоритмов

Проверка эффективности алгоритмов была произведена на основании подсчета показателей $Weight(p_j)$ для определенных значимых и общеупотребительных терминов.

Например, для термина **«совершенствование»** в массиве документов сотрудников организации были получены следующие результаты:

$$\sigma(t_i) = 2{,}789E-05$$
$$M(t_i) = 0{,}000589985$$

t_i	p_j	$TF(t_i, p_j)$	$Impact(t_i;p_j)$	$Impact(t_i; p_j)$
24	67	0,004098	125,7934	1,562846951
24	70	0,004016	122,8426	1,562656011
24	74	0,001093	18,03195	1,51539596
24	71	0,000381	-7,50623	-1,438353556
24	12	0,000267	-11,5969	-1,484778878
24	20	0,00019	-14,3279	-1,501115551
24	39	0,00017	-15,05	-1,504448591
24	60	0,000163	-15,3246	-1,505633949
24	72	0,000138	-16,2126	-1,509193863
24	10	0,000119	-16,8936	-1,511671259
24	22	9,87E-05	-17,614	-1,514084188
24	23	9,21E-05	-17,852	-1,514838648
24	68	8,21E-05	-18,211	-1,515939491
24	56	7,59E-05	-18,4327	-1,516597979
24	32	7,55E-05	-18,4484	-1,51664416
24	7	6,06E-05	-18,9794	-1,518156366
24	40	3,86E-05	-19,7702	-1,520258204
24	50	3,48E-05	-19,9052	-1,520600471
24	4	1,7E-05	-20,5434	-1,522157177

Для термина **«деятельность»** в массиве документов сотрудников организации были получены следующие результаты:

$$\sigma(t_i) = 0{,}000202344$$
$$M(t_i) = 0{,}00211234$$

t_i	p_j	$TF(t_i, p_j)$	$Impact(t_i;p_j)$	$Impact(t_i; p_j)$
4375	2	0,000305	8,933093	1,459317134
4375	4	0,004726	-12,9184	-1,493541558
4375	5	0,001293	4,047163	1,328561503
4375	6	0,000362	8,648108	1,455675364
4375	7	0,000882	6,079572	1,407770845
4375	8	0,005848	-18,4617	-1,516683102
4375	9	0,001589	2,587306	1,20197972
4375	10	0,000693	7,01382	1,429175132
4375	11	0,002988	-4,32916	-1,343786453
4375	12	0,001698	2,046931	1,116361764
4375	15	0,000116	9,868227	1,469805741
4375	20	0,001267	4,176676	1,335795223
4375	22	0,001456	3,242271	1,271627523
4375	23	0,000138	9,756665	1,468658944
4375	24	0,001457	3,239629	1,271397806
4375	26	0,002302	-0,93919	-0,754050771
4375	32	0,002656	-2,68762	-1,21459141
4375	34	0,001504	3,006481	1,249692586
4375	36	0,00534	-15,9536	-1,508196542
4375	37	0,000291	9,002924	1,460174747
4375	38	0,001346	3,785081	1,312502807
4375	39	0,002043	0,343173	0,33057958
4375	40	0,00193	0,90234	0,734106685
4375	42	0,00655	-21,9324	-1,525233207
4375	43	0,004022	-9,43831	-1,465238963
4375	45	0,001828	1,404473	0,952054744
4375	47	0,00122	4,408485	1,347735733
4375	48	0,002543	-2,1302	-1,13189379
4375	50	0,001157	4,71942	1,361994247
4375	52	0,00021	9,399844	1,464810221
4375	53	0,001668	2,196851	1,143629045
4375	54	0,000667	7,142445	1,431692463
4375	55	0,011435	-46,071	-1,549094102
4375	56	0,003045	-4,61139	-1,357248441
4375	58	0,000799	6,488858	1,417888956
4375	60	0,001202	4,499738	1,352115031
4375	61	0,0003	8,955146	1,459589403
4375	62	0,000809	6,440915	1,416768639
4375	63	0,000833	6,322334	1,413926445
4375	64	0,001323	3,902215	1,319930215
4375	68	0,001724	1,920625	1,090754721

4375	71	0,001657	2,250875	1,152716238
4375	72	0,002687	-2,84215	-1,232477862
4375	73	0,003392	-6,32294	-1,413941279
4375	74	0,005464	-16,5666	-1,510506976
4375	78	0,000398	8,473009	1,453317949

Для термина «Башкортостан» в массиве документов сотрудников организации были получены следующие результаты:

$$\sigma(t_i) = 1{,}03584E-06$$
$$M(t_i) = 0{,}000186121$$

t_i	p_j	$TF(t_i, p_j)$	$Impact(t_i;p_j)$	$Impact(t_i; p_j)$
7399	43	0,001006	791,0617	1,569532204
7399	73	0,000565	366,0526	1,568064486
7399	9	0,000318	127,0877	1,56292791
7399	48	0,000187	1,240819	0,89245633
7399	47	8,93E-05	-93,4801	-1,560099273
7399	7	5,79E-05	-123,792	-1,56271844
7399	24	5,4E-05	-127,592	-1,562959029
7399	20	4,76E-05	-133,734	-1,563318919
7399	12	4,76E-05	-133,774	-1,563321163
7399	53	1,95E-05	-160,85	-1,564579445
7399	40	1,29E-05	-167,262	-1,564817758
7399	60	9,56E-06	-170,449	-1,564929543
7399	50	5,36E-06	-174,509	-1,565066031

Отрицательные значения $Impact(t_i; p_j)$ в данном случае можно не учитывать.

Тогда по полученным результатам мы можем подсчитать релевантных экспертов по запросу «Совершенствование деятельности в Башкортостане», как сумму весов термов $Impact(t_i; p_j)$, употребленных в тексте для автора.

Заключение

Количество усилий, которые придется приложить, чтобы найти эксперта в организации, зависит от целого ряда факторов, таких как размер организации, уровень автоматизации, сила социальных связей внутри организации и т.д. В случае отсутствия автоматизированной системы поиска экспертов, поиск нужного человека может потребовать приложения значительных усилий. Разработанный алгоритм поиска экспертов на основании анализа контента в организации может применяться в качестве инструмента (сервиса) для решения задач поиска экспертов на корпоративных порталах, в ECM –системе и в любой другой ИС, накапливающей сведения о публикациях сотрудников. На данный момент подобные алгоритмы применяются в разрабатываемом сервисе в рамках договора № 13.G25.31.0096 с Министерством образования и науки РФ о «Создании высокотехнологичного производства кросс-платформенных систем обработки неструктурированной информации на основе свободного программного обеспечения для повышения эффективности управления инновационной деятельностью предприятия в современной России». Применение сервиса поиска экспертов в коммерческом секторе позволит сделать такой поиск прозрачным и эффективным.

ЛИТЕРАТУРА

1. http://lrgp.lt/index.php?option=com_content&task=view&id=275&Itemid=49
2. Turney P. Learning to extract keyphrases from text.: Technical report, National Research Council, Institute for Informational Technology, 1999.
3. Robert M. Losee: Term dependence: A basis for Luhn and Zipf models. JASIST 52(12): 1019-1025 (2001). 2000.
4. Balog K. and de Rijke M. Finding similar experts. / In SIGIR '07, PP. 821–822. ACM Press, 2007.
5. Kun-Woo Yang, Soon-Young Huh. Automatic expert identification using a text categorization technique in knowledge management systems // Expert Systems with Applications 34 (2008). PP. 1445–1455.
6. Hofmann K., Balog K., Bogers T., de Rijke M. Integrating Contextual Factors into Topic-centric Retrieval Models for Finding Similar Experts.

Ставенко Юлия Андреевна,
научный сотрудник НОЦ Технологий управления информацией при НИУ Высшая Школа Экономики.
ystavenko@hse.ru.
Сайт: http://www.hse.ru/org/persons/14296686

Романов Дмитрий Алексеевич,
канд. техн. наук, зам. директора НОЦ Технологий управления информацией при НИУ Высшая Школа Экономики.

Громов Александр Игоревич,
канд. хим. наук, директор НОЦ Технологий управления информацией при НИУ Высшая Школа Экономики.
Сайт:http://www.hse.ru/org/persons/67957