УДК 519.21:681.142

П.А. Помилуйко, Л.А. Тимашова

МЕТОДЫ И МОДЕЛИ ДЛЯ АНАЛИЗА КОРПУСА ДОКУМЕНТОВ НОРМАТИВНЫХ АКТОВ В ЗАДАЧАХ ЭЛЕКТРОННОГО УПРАВЛЕНИЯ

Описаны подходы, используемые для классификации корпуса текстовых документов. Предлагаются новые модели и методы позволяющие осуществлять тематическую классификацию и вычислять степень тематической принадлежности текста к образцу.

Введение

Развитие информационных технологий привело к новой форме общения с государством – электронному управлению, под которым следует понимать способ организации государственной власти с помощью систем локальных информационных сетей и сегментов глобальной информационной сети. В результате обеспечивается функционирование определенных служб в режиме реального времени и делается максимально простым и доступным ежедневное общение гражданина с официальными учреждениями. Считается, что такая форма общения с государством приведет не только к более эффективному и менее затратному администрированию, но и к кардинальным изменениям взаимоотношений между гражданами и органами государственной власти и местного самоуправления.

История создания электронного управления идет параллельно с развитием информационных технологий. Введение информационно-коммуникационных технологий (ИКТ) в государственное управление позволит ускорить развитие экономики, снизить затраты на бюрократические процедуры, повысить эффективность работы и производительность труда государственных ведомств, расширить возможности населения в формировании гражданского общества за счет улучшения доступа к различному роду информации. В настоящее время донесение информации до граждан носит, как правило, пассивный и нерегулярный характер, нет технологии быстрого и качественного ознакомления с документами, когда это действительно требуется. Особые трудности вызваны проблемой понимания как законов, так и их изменений и связаны с недоступностью информации, а также незаинтересованностью служащих госучреждений в предоставлении юридических, производственных и финансовых услуг. Это сложная задача, так как политические, социально-экономические и правовые реформы, осуществляемые на Украине, требуют существенного обновления законодательства.

Опыт последних лет показывает, что значительное увеличение количества законов зачастую сказывается на их качестве, внутренней согласованности и, как следствие, ведет к малой эффективности на практике, что подрывает принцип верховенства законов, являющийся неотъемлемым условием существования правового государства. Среди основных причин, негативно отражающихся на качестве законов, необходимо также, выделить недостаточную исследованность правил и приемов в юридической практике,

а также недооценку ее влияния на качество и эффективность различных видов правовой деятельности. Данное обстоятельство подтверждает необходимость привлечения новых методов обработки и анализа правовой информации, а также потребность создания правил и приемов юридической техники, позволяющих обеспечивать качество нормативно-правовых актов, новых методов решения задачи поиска документов. Чтобы решить подобные проблемы, перечень и описание всех услуг госучреждения, нормативно-правовая информация должны быть доступны всем в любое время. Главная роль в решении этих проблем отводится информационным технологиям, реализуемым с помощью систем локальных информационных сетей и сегментов глобальной информационной сети.

Целью данной статьи является разработка методов повышения эффективности аналитической обработки информации, представленной в виде корпуса распределенных текстовых документов, в частности документов, расположенных на различных веб-ресурсах. В статье предлагается новый метод решения задачи поиска документов, для чего разработаны: модель структурного представления текстовой информации, метод и алгоритмы ее тематического анализа, позволяющие осуществлять тематическую классификацию и вычислить степень тематической принадлежности текста к образцу. Предложенная модель, метод и алгоритмы могут использоваться как для решения конкретных задач поиска документов по образцу, так и для решения общих задач тематического анализа и обработки речевых высказываний.

Под текстовым корпусом в современной лингвистике понимается ограниченный в размере набор текстов, пригодный для машинной обработки и отобранный таким образом, чтобы наилучшим образом представлять языковое множество. Такого рода представленный документально юридический массив текстов можно считать текстовым корпусом. Возникает проблема поиска и анализа правовой информации в корпусе неструктурированных текстов.

Существует ряд средств для автоматизации семантического анализа текстовой информации: Oracle Text, Intelligent Miner for Text, Text Miner, Text Analyst и др. Продукты такого типа осуществляют интеллектуальный анализ текстовых данных или так называемый text mining — «нетривиальный процесс обнаружения действительно новых, потенциально полезных и понятных шаблонов в неструктурированных текстовых данных» [1]. Text mining является полезным инструментом обработки данных, однако возрастающий объем информации вызывает необходимость в более глубоких методах анализа текста.

Для решения этой задачи предлагаются: модифицированный алгоритм выделения из текста доминантных термов, семантическая модель корпуса документов и алгоритм поиска документов по образцу в глобальных сетях.

Постановка задачи

В данной работе текстовый документ рассматривается в виде $D = < T, W >$, где $T = \{t_i \mid i = 1, ..., n\}$ — множество доминантных термов

документа, $W = \{w_j \mid j = 1, ..., n\}$ — множество весов термов, показывающих важность терма t_i для документа *D*.

Неотъемлемой частью любой информационно-поисковой системы является модуль индексации, который в автоматическом режиме создает на базе текста индекс, т.е. переводит текст в записи таблицы базы данных. В данной работе используется следующая структура индекса (рис. 1).

Для каждого терма хранится его идентификационный код, строковое представление, начальная форма, идентификационный код документа, в котором встретился данный терм.

Для каждого вхождения терма в документ хранится его идентификационный код, идентификационный код терма, значение TF · IDF.

Для каждого документа хранится его идентификационный код, путь к документу, дата индексации, количество слов в документе, количество термов в документе, название документа и текст после обработки.

Рис 1. Структура индекса

Индексация текстового документа происходит по следующему алгоритму: из файла или html-документа извлекается текст. Для этого используется модуль импорта, который работает со всеми известными текстовыми форматами (HTML,Word, Excel, Pdf).

1. *Предобработка текстов.* Так как тексты в цифровом виде, как правило, оформлены для чтения, в них содержится много лишней для анализа информации (комментарии, ссылки, введения и т.п.). Кроме того, тексты, взятые из разных источников, могут быть оформлены по-разному, поэтому предварительная чистка и унификация анализируемых текстов крайне желательна, хотя иногда и необязательна. Также необходимо выполнить процедуры декапитализации и деакцентизации. Декапитализация — преобразование всех символов корпуса в нижний регистр. Деакцентизация — прием используемый при обработке текстов на французском языке, в которых существуют буквы с акцентами, к примеру *é*, *è*, *à* и т.п. Из-за особенностей грамматики, одно и то же слово в разных контекстах может появляться с акцентами и без, поэтому все символы с акцентами заменяются на аналоги без акцентов.

2. *Морфологический и синтаксический анализ текста.* Большой корпус текстов может исчисляться в десятках, сотнях тысяч, а иногда и миллионах документов. Количество слов в таких корпусах текстов нередко достигает десятков миллионов. При этом текст, написанный на естественном языке, слабо структурирован и может содержать большое количество

ошибок.

Множество всех уникальных слов и словосочетаний, содержащихся в корпусе *U*, формируется с помощью процедуры токенизации. Алгоритм определяет границы слов с помощью множества стоп-знаков слова — множества знаков, которые позволяют отделять в тексте слова друг от друга, а также некоторых правил, после чего составляется список всех уникальных словосочетаний в корпусе. Размер множества *U*, как правило, не превышает 50000 слов.

Требуется извлечь из текста не только отдельные слова, но и словосочетания, т.к. основные понятия предметной области очень часто представлены сочетаниями слов (например, прилагательное и существительное) [2], к примеру, словосочетание «коммерческая недвижимость» в большей степени характеризует документ, нежели слова «коммерческая» и «недвижимость» по отдельности.

Наиболее важным на этом этапе является классификация всех слов по категориальным (частеречным) признакам: существительное, глагол, прилагательное и т.д. Для этого необходимо привести все слова к нормальной форме, т.е. выполнить процесс лемматизации. Лемматизация слов производится с помощью специального модуля автоматизированной обработки текста, который присваивает каждому слову соответствующую часть речи.

Данный подход основывается на гипотезе о том, что на определение класса в большей степени влияют следующие части речи [3]:

- Существительное
- Прилагательное и существительное

3. *Удаление стоп-слов.* Стоп-лист — список вспомогательных слов, несущих мало информации о содержании документа. Из текста удаляются все слова из стоп-листа.

4. *Определение доминантных термов.* Семантический образ отражает содержание документа и включает в себя множество термов, находящихся в некоторой зависимости друг от друга. Один из вариантов выражения такой зависимости — весовые коэффициенты, отражающие значимость (вес) того или иного слова в конкретной тематике.

Для термов-кандидатов, частота вхождения которых в данный документ превышает заданный порог (в данной работе значение порога было равно единице), вычисляется значение меры $TF \cdot IDF$. Данная мера позволяет осуществить так называемый контрастный тест, понизив вес слов, которые часто встречаются не только в данном документе, но и в других документах корпуса. По этой причине для вычисления данной меры в общем случае необходимо предварительно проиндексировать весь корпус.

Однако при большом объеме корпуса такой расчет потребовал бы значительного времени, поэтому для вычисления уже на этапе индексации одного документа вместо данных о частоте употребления слова в корпусе используются данные, взятые из частотного словаря. При этом используется формула

$$W(t) = \frac{freq(t)}{|d|} \log \frac{N_D}{N_d}, \quad (1)$$

где $freq(t)$ — частота употребления терма t в документе; $|d|$ — количество слов в документе; N_D — количество документов в анализируемом корпусе; N_d — количество документов из корпуса, содержащих терм t.

5. Доминантными для документа терминами считаются те, значения меры $TF \cdot IDF$ для которых превышает заданный порог. После этого все данные записываем в базу.

Алгоритм поиска документа по образцу

Одной из задач информационного поиска является задача поиска по документу-образцу. Именно эта задача представляет наибольший интерес в рамках работы с нормативно-правовыми документами.

Документ-образец выступает в качестве одной из форм представления информационных потребностей пользователя. Целью поиска является обнаружение тематически близких документов. При этом, как правило, речь идет не о поиске идентичных или синтаксически близких документов, а о поиске документов, близких по содержанию, близких по смыслу.

Самым простым подходом к решению задачи поиска документов по образцу является использование всех слов документа-образца в качестве запроса. Однако длина такого запроса может оказаться очень большой, что отрицательно скажется на качестве поиска, т.к. результатом поиска будут все документы, в которых присутствовали данные слова, и таких документов может быть очень много. Это повлияет как на саму поисковую систему — вычислительные ресурсы и трафик не безграничны, и система может оказаться перегруженной, так и на человека — просмотр и анализ найденных документов может занять значительное время.

Приемлемым вариантом в данном случае является выделение тематики документа. Под тематикой понимается множество доминантных термов, описывающих, с некоторой степенью адекватности, содержание документа.

Поиск новых документов

В случае отсутствия заданного корпуса текстов, система использует гипертекстовую архитектуру сети Интернет. Для этого производится анализ структуры графа Интернета (вершинами которого выступают страницы, а ребрами — ссылки). В качестве документа образца выступает некоторая страница (наиболее релевантная для заданного класса), ссылки на данную страницу и ссылки с нее используются в различных алгоритмах локального анализа структуры графа Интернета. В работе [2] рассматривается один из таких алгоритмов — HITS (Hyperlink Induced Topic Search). В рамках этого алгоритма определяется два класса документов:

- «первоисточник» — документ, на который часто ссылаются в контексте некоторой тематики (чем чаще ссылаются, тем лучше «первоисточник»);

- «посредник» — документ, который ссылается на много «первоисточников» (чем больше ссылок на первоисточники, тем лучше «посредник»).

Алгоритм HITS состоит из двух шагов:

- выбор подмножества из Интернета на основе запроса;
- определение лучших «первоисточников» и «посредников» по результатам анализа этого подмножества.

Подмножество строится методом расширения множества найденных по запросу пользователя страниц за счет добавления всех страниц, связанных с ними путем, состоящим из одной (иногда двух) ссылок. Далее для каждого документа рекурсивно вычисляется его значимость как «первоисточника» и как «посредника».

Суть алгоритма HITS состоит в том, чтобы выделить множество, соответствующее тематике запроса, и на основе анализа этого множества определить наиболее авторитетные страницы.

Для проведения информационного поиска новых нормативных документов предлагается использовать агентный подход. Агент — программный модуль, который осуществляет обход веб-ресурсов и загружает с них файлы определенного типа. Поиск, осуществляемый агентом, можно разделить:

- на поиск документов, информация о которых уже содержится в базе данных агента (государственные организации);
- поиск новых URL-документов через базы данных Google.

Для поиска страниц «первоисточников» и «посредников» используются спецкоманды к базе данных Google и Яндекс. Агент производит поиск во всех известных текстовых форматах (HTML,Word, Excel, Pdf) и извлекает содержимое документов в базу данных системы. На этом этапе главное — отобрать релевантные тексты для заданного класса достаточного объема (более 50000 слов). В дальнейшем по заданному пользователем расписанию агент вновь посещает все веб-ресурсы и при наличии на них новых файлов загружает их. Файлы, которые обрабатываются агентом, сохраняются в указанной пользователем директории на сервере.

Латентно-семантический анализ

В данной работе использован латентный семантический анализ (LSA — latent semantic analysis) как метод определения сходства значений слов и документов путем статистических вычислений над большим текстовым корпусом. Он использован, поскольку для вычислений с его помощью не требуется никакой дополнительной информации, такой как построенные вручную словари, семантические сети или базы знаний. В основе метода LSA лежит гипотеза о том, что между словами и тем контекстом, в котором они употребляются, существуют неявные (латентные) взаимосвязи. Предполагается, что семантическое значение документа может быть представлено как сумма значений входящих в него слов.

Метод позволяет вычислить корреляции между парой термов, между парой документов и между термом и документом.

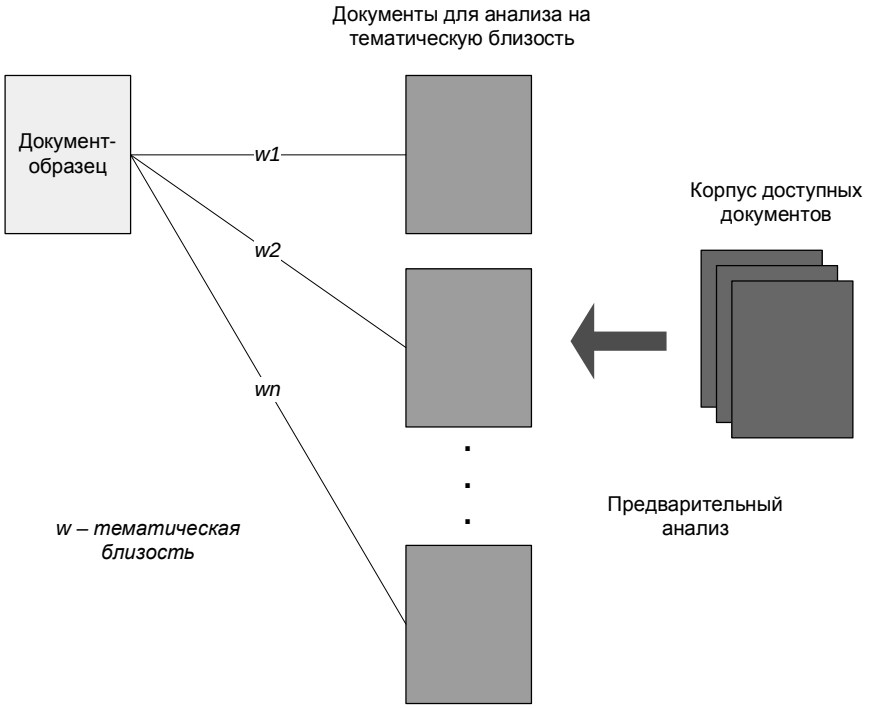

Рис. 2. Поиск документа по образцу

Каждая строка исходной матрицы C — вектор, соответствующий терму и показывающий его связь с каждым из документов корпуса:

$$t_i^T = \{w_{1i}, w_{2i}, \ldots w_{ni}\}. \quad (2)$$

Каждый столбец исходной матрицы C — вектор, соответствующий документу и показывающий его связь с каждым из термов корпуса:

$$d_i = \begin{Bmatrix} w_{1i} \\ w_{2i} \\ \ldots \\ w_{im} \end{Bmatrix}. \quad (3)$$

Скалярное произведение $t_i^T t_p$ двух векторов показывает корреляцию между соответствующими термами. Произведение матриц CC^T содержит все такие скалярные произведения. Элемент (i, p) (равный элементу (p, i)) содержит скалярное произведение $t_i^T t_p = t_p^T t_i$. Аналогично, матрица $C^T C$ содержит скалярные произведения между векторами всех документов, показывающие корреляцию между ними: $d_j^T d_q = d_q^T d_j$.

Метод LSA заключается в сингулярном разложении матрицы C (*SVD* — singular value decomposition) и аппроксимации ее матрицей C_k меньшего ранга k. Тогда матрица C_k, содержащая только k первых линейно

независимых компонентов C, отражает структуру ассоциативных связей, присутствующих в исходной матрице, и в то же время не содержит «шума». Помимо этого, уменьшение размерности матрицы ведет к уменьшению количества k вычислений.

По теореме о сингулярном разложении существует разложение используются $C = UZV^T$, такое, что U и V — прямоугольные матрицы, а Z — диагональная матрица.

Сходство между двумя документами может быть получено на основании следующего выражения:

$$C^T C = (UZV^T)^T UZV^T = (V^{T^T} Z^T U^T)(UZV^T) = VZ^T U^T UZV^T = \\ = VZ^T ZV^T = (VZ)(VZ)^T. \quad (4)$$

Поскольку произведения матриц ZZ^T и $Z^T Z$ являются диагональными матрицами, то матрица U должна содержать собственные векторы CC^T, а матрица V — собственные векторы $C^T C$. Оба произведения должны иметь одинаковые не равные нулю собственные значения при не равных нулю элементах ZZ^T или, что то же самое, при не равных нулю элементах $Z^T Z$.

Разложение матрицы C выглядит следующим образом:

$$\begin{bmatrix} w_{11} & \dots & w_{1n} \\ \dots & \dots & \dots \\ w_{m1} & \dots & w_{mn} \end{bmatrix} = \begin{bmatrix} [u_1] & \dots & [u_l] \end{bmatrix} \begin{bmatrix} z_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & z_l \end{bmatrix} \begin{bmatrix} [v_1] \\ \dots \\ [v_l] \end{bmatrix}, \quad (5)$$

где $z_1, ..., z_l$ — сингулярные числа, $u_1, ..., u_l$ и $v_1, ..., v_l$ — левый и правый сингулярные векторы. Известно, что выбор k наибольших сингулярных значений и соответствующих им сингулярных векторов из матриц U и V дает наилучшую аппроксимацию матрицы C матрицей ранга k:

$$C_k = U_k Z_k V_k^T. \quad (6)$$

Теперь, применяя к уменьшенным матрицам полученное ранее соотношение (4), можно вычислить сходство каждой пары документов как скалярное произведение соответствующих векторов, умноженных на сингулярные значения:

$$C_k^T C_k = (V_k Z_k)(V_k Z_k)^T. \quad (7)$$

Выводы

В работе рассмотрено новое решение задачи поиска документов по образцу. Разработаны модель структурного представления текстовой информации, метод и алгоритмы ее тематического анализа, позволяющие осуществить тематическую классификацию и вычислить степень тематической принадлежности текста к образцу. Предложенная модель, метод и алгоритмы могут использоваться как для решения конкретных задач поиска документов по образцу, так и для решения общих задач

тематического анализа и обработки речевых высказываний. Это направление является весьма перспективным и будет приоритетным для развития систем электронного управления в дальнейшем. Развитие информационных технологий приведет к новой форме общения с государством — электронному управлению, под которым следует понимать способ организации государственной власти с помощью систем локальных информационных сетей и сегментов глобальной информационной сети. В результате функционирование определенных служб в режиме реального времени сделает максимально простым и доступным ежедневное общение гражданина с официальными учреждениями. Считается, что такая форма общения с государством приведет не только к более эффективному и менее затратному администрированию, но и к кардинальным изменениям взаимоотношений между гражданином и органами государственной власти и местного самоуправления.

1. *Gibson D., Kleinberg J.M., Raghavan P.* Inferring web communities from link topology // Proc. of the UK Conf. on Hypertext. — 1998. — P. 225–234.
2. *Salton G., Buckley C.* Term-weighting approaches in automatic text retrieval // Information Processing & Management. — 1988. — № **5** (24). — P. 513–523.
3. *Некрестьянов И.С.* Тематико-ориентированные методы информационного поиска: Дис…. работа канд. техн. наук. — СПб., 2000. — 80 с.

Международный научно-учебный центр
информационных технологий и систем
НАН Украины и Министерства образования
и науки, молодежи и спорта Украины, Киев

Получено 27.11.2012