Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

**Келдыш Наталья Всеволодовна**
Keldysh Natalia Vsevolodovna
к.т.н., доцент НОУ ВПО ИГУПИТ
Ph.D. in Technics, the associate professor IGUPIT

# Анализ существующих методов решения информационных задач, используемых при разработке систем электронного документооборота

# Analysis of existingmethods for solving information problemsthat are usedin the development ofelectronic document management systems

**Аннотация:** В статье рассматриваются методы решения функциональных задач систем электронного документооборота, а так же анализируются существующие методы решения информационных задач. Рассматриваетсязависимость методов от технологии и конкретики особенностей области применения.

**Ключевые слова:** информационная система, система электронного документооборота, методы поиска, логическая поисковая модель, типология задач классификации.

**The Abstract:** This paper discusses methods for solving functionalproblems ofelectronic document management systems, as well as analyzes the existingmethods for solvinginformation problems.The dependenceon technology andmethods ofconcretefeatures ofthe application.

**Keywords:** information system, process of training, a dialogue mode, an educational portal, criteria of an assessment of efficiency of the interface, technology of electronic training.

***

В перечень дисциплин, используемых при создании научно-методического аппарата (НМА), входят математическая и формальная логика, теория вероятностей, математическая статистика, прикладная лингвистика, вариационное исчисление, кластерный и семантический анализ, нечеткие множества и другие. Разнообразие методов и подходов, реализованных при этом, настолько велико, что их анализ далеко выходит за рамки данный статьи и должен служить предметом отдельного исследования.

Основу методического аппарата, создаваемого при разработке любой автоматизированной информационной системы, составляют методы поиска, классификации и фильтрации данных. Первые из них были разработаны в середине прошлого века, с тех пор число их непрерывно увеличивается по мере расширения области применения автоматизированной обработки данных. В терминах математической лингвистики типовой документ как объект обработки представляет собой конечное множество слов (терминов), объединенных лексическими, грамматическими, смысловыми, частотными отношениями и образующих информативное сообщение. В терминах нечетких множеств документ как объект обработки представляет собой слабоструктурированный массив данных, изложенных в цифровой или в естественной языковой форме, при этом принадлежность его к любой конкретной предметной области описывается непрерывной функцией (функцией принадлежности).

К числу особенностей обработки подобных объектов относится, прежде всего, сложность

Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

привязки исходных текстов к типовым синтаксическим структурам, что связано с неоднозначностью реферируемости, нарушением порядка слов и другими средствами выражения коммуникативной организации [2].

Специфически деловой, официальный стиль изложения, присущий большинству документов кадровых органов, также утяжеляет их автоматизированную обработку. В большинстве случаев текстовое описание представляет собой группу длинных предложений, характеризующихся развитой структурой синтаксического дерева, специфической лексикой и конструкцией фигур речи. Необходимо отметить и высокую вероятность употребления продуктивного словообразования при формулировке описания, что также затрудняет герменевтику текста.

Кроме того, к числу особенностей следует отнести и высокую смысловую насыщенность лексических конструкций, вследствие чего значительная часть содержательной информации не выражена эксплицитно. Хотя некоторая часть сведений передается имплицитно, т.е. восстанавливается человеком по догадке, машинный анализ таких информационно насыщенных предложений сопряжен со значительными сложностями.

Проведем типологию задач информационного поиска и основные методы решения.

Поиск данных представляет собой процесс, в ходе которого происходит отбор - соотнесение отыскиваемого (эталонного) образа с каждым объектом, содержащимся в базе данных. При этом сравниваются не сами объекты, а их описания - так называемые поисковые образы. Итерационный алгоритм поиска включает, по крайней мере, следующие операции:

– формирование для заданных условий эталонного поискового образа (эталона);

– выборку объекта из базы данных;

– сравнения выбранного объекта с эталоном;

– определение значения показателя соответствия;

– расчет заданного критерия для формирования решения о соответствии объекта эталону;

– переход к выборке следующего объекта с образцом или завершение процесса поиска.

Эти операции хорошо формализуются и могут выполняться в автоматическом или автоматизированном режиме. Однако для запуска алгоритма необходимо специфицировать поисковый образ эталона, выбрать показатели и критерии соответствия объекта и образца, что и составляет основную специфику и трудность выбора метода решения.

По степени семантической неопределенности или по характеру и степени соответствия в предмете поиска известного и неизвестного выделяют предметный, тематический или проблемный поиск [3].

К задачам предметного или атрибутивного поиска относится поиск по атрибутам заданного объекта, то есть поиск по логическому выражению над именами понятий, задаваемых терминами или их комбинациями (значениями определенного характеристического признака).

Второй тип задач (тематический поиск) состоит в подборе информации по некоторой теме, например, для разрешения возникшей проблемы, обоснования или поиска решения практической задачи. Тематический поиск состоит в нахождении в базе данных информационной системы описаний реально существующих объектов, свойства которых могут быть полностью определены на уже известном множестве атрибутов. Неопределенность отображения объекта на предметную область порождается возможной множественностью системной

Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

основы на уровне среды информационной систем, связям, частично задаваемым комбинацией характеристических признаков. И здесь модель поиска это поиск по характеристикам задаваемых комбинаций, а также по связям или по части определенного понятия.

Тематический поиск реализуется как последовательность атрибутивных поисков, каждый из которых соответствует определенному (априорно заданному) системному основанию представления объекта поиска.

Третий тип задач поиска представляет собой проблемный поиск, который является, по сути, основной составляющей творческого процесса- определения путей решения конкретной задачи пользователя. Проблемный поиск состоит в поиске в базе данных объектов или их составляющих, потенциально существующих в предметной области, и в совокупности, возможно, образующих целое, свойства которого больше суммы свойств частей. То есть этим свойствам в явной форме не сопоставлены «собственные» атрибуты, а новое свойство, например, может быть задано комбинацией уже известных атрибутов. В этом случае к неопределенности отображения объекта на предметную область, свойственной тематическому поиску, добавляется неопределенность представления объекта на уровне «субъект-объект основной деятельности». Например, представление, которое субъект имеет об объекте, может не соответствовать реальности. Логическая поисковая модель для этого случая - поиск «похожих» документов, содержание которых некоторым образом ассоциируется с задачей пользователя.

Данная типология поисковых задач с точки зрения структурной полноты представления объекта поиска становится очевидной в контексте общей теории систем: объект поиска представляется как система $S_i=\{M_i, A_i, R_i, Z_i\}$, определяемая в виде гипотетической комбинации множества первичных элементов ($M_i$) через задание системообразующих признаков ($A_i$) , системообразующих отношений ($R_i$)исистемообразующего закона композиции ($Z_i$).

В этом контексте предметный (атрибутивный) поиск- это нахождение объекта – системы $S_i$по заданному (полностью определенному) системному его основанию $<A_i, R_i, Z_i>$ . Тематический поиск – это нахождение подмножества систем $\{S_i, i= 1, n\}$, для которого задано $Z_i$ и одно из оснований $A_i$или$R_i$. Проблемный поиск – это разновидность тематического поиска с неединственным законом композиции $Z_i$.

Необходимо отметить, что в ходе функционирования системы электронного документооборота ввиду непрогнозируемого характера и многообразия возникающих практических задач, может потребоваться проведение каждого из перечисленных типов поиска.

В целом, методы, применяемые для решения задач поиска, могут быть распределены по трем категориям в соответствии со способом представления предмета поиска:

в виде отдельной единицы хранения, то есть целостного (неделимого) объекта, имеющего заданное наименование и содержание;

в виде композиции «атомарных» единиц информации, каждая из которых является отдельным термином;

в виде информационного объекта, содержание которого характеризуется полным набором входящих терминов.

Наиболее распространенные способы решения задач поиска построены на использовании методов обнаружения контентного вхождения (при поиске целостных объектов хранения и объектов, содержащих уникальный термин) и семантического анализа (для поиска информационных объектов, содержащих заданный смысл) [1].

Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

Наиболее популярными инструментами выявления контентного вхождения являются алгоритмы Кнута-Морриса-Пратта, Бойера-Мура и Рабина. В числе недостатков их применения в интересах обработки базы данных документооборота следует отметить недостаточную временную эффективность, что обусловлено обобщенной универсальностью и отсутствием адаптации к предметной области.

В качестве характеристики качества решения задачи поиска традиционно используются показатели полноты и точности, значения которых определяются характером запроса, методом решения, алгоритмом поиска, а также широтой и составом предметной области исходного массива документов. Согласно экспертным оценкам усредненные значения данных показателей не превышают 40-50% и 60-70%, соответственно. Необходимо отметить, что повышение качества поиска далеко не всегда обеспечивается с помощью интуитивно обоснованных подходов. Так, например, использование в виде информационного запроса контактной пары терминов вместо одного термина значительно ухудшает эффективности поиска, при этом положительный эффект может быть получен только при введении существенного расстояния между терминами [5]. Несмотря на обилие публикаций, в литературе отсутствуют конструктивные подходы, обеспечивающие оптимизацию выбора методов решения поисковой задачи. Таким образом, в ходе разработки методического аппарата решения задач электронного документооборотав организации целесообразно использовать методы решения, показатели и критерии, адаптированные к характеристикам документов и структуре документооборота.

Вторым типом основных задач обработки данных являются задачи классификации, решение которых формально заключается в присвоении булева значения каждой паре $(d_j, c_i) \in DxC$, где $D = \{d_1,..., d_n\}$ - множество данных, $C = \{c_1,..., c_m\}$ - множество тематических категорий (в задаче классификации). Целевая функция $\Phi: D \times C \rightarrow \{T, F\}$ — это классификатор, где $T$и $F$- «истина» и «ложь», соответственно.

В известном смысле классификация является частным случаем поисковой информационной задачи, в ходе которой сравниваются образы классифицируемых данных и заданных категорий классификатора. Итерационный алгоритм классификации включает следующие операции:

- выборка очередной тематической категории из классификатора;
- определение значения показателя соответствия между данными и категорией;
- проверка выполнения заданного критерия соответствия;
- переход к выборке следующей категории или завершение процесса классификации.

Типология задач классификации включает две разновидности.

К задачам первого типа относится классификация документов с обучением или категоризация документов. При этом документы классифицируются по предопределенному классификатору на основании знаний о том, каким критериям должны отвечать документы, принадлежащие к той или иной категории.

Задачи второго типа состоят в классификации документов без обучения или кластеризации документов. При этом документы необходимо классифицировать в условиях отсутствия предопределенной классификационной схемы, проведя их распределение на основе тематического сходства или различия.

Необходимо отметить, что при организации электронного документооборота могут быть востребованы алгоритмы, как категоризации, так и кластеризации документов.

Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

Математическая суть классификации состоит в выборе показателей и в оценке критерия соответствия для задач категоризации документов или в оценке тематического сходства (различия) для задач кластеризации.

Известны логико-аналитические, статистические и лингвистические методы, применяемые для классификации текстовых документов.

Разработки в этой области проводятся с 50-х годов прошлого века. Исторически первой сложилась группа логико-аналитических (экспертных) методов, в настоящее время наиболее используемая в библиотечных и справочных системах (т.е. в «статичных системах»). Основу методов составляет экспертная привязка классифицируемого объекта к тематическим категориям. В задаче поиска выполняется аналогичная процедура, состоящая в экспертном установлении соответствия между критериями поиска и тематическими категориями, а результат решения определяется путем последующего экспертного отбора объектов учета из группы тематических категорий, наиболее адекватной заданным условиям поиска [5]. Преимуществом логико-аналитических методов является простота организации, к недостаткам следует отнести большие временные затраты и недостаточно высокую вероятность правильного решения, в частности, при отсутствии жесткого соответствия «объект-категория» [2]. В целом методы этой группы не могут считаться перспективными для разрабатываемых информационных систем органов государственного управления, в первую очередь, из-за низкой временной эффективности и ограниченной применимости в условиях поточной или групповой обработки данных.

Статистические методы, объединенные во вторую группу, основаны на формировании признаков тематических категорий путем обучения, на известном множестве заранее классифицированных данных, оценке статистического веса данных признаков для анализируемого документа и оценки показателя попарной близости на основе вероятностных, матричных и других подходов. Безусловным достоинством этих методов является их быстродействие. Так, эксперименты для целой группы стандартных алгоритмов (таких как Роккио, SVM , NaïveBayes, DecisionTrees и BayesNets) показали, что время классификации документа не превышает 2 мс [1].

Большое количество статистических методов создает разработчикам НМА нетривиальную задачу априорного выбора, сложность которой состоит в необходимости тестирования сопоставляемых алгоритмов в одинаковых условиях, а, следовательно, на одних и тех же стандартных коллекциях данных и тематик, для которых заранее известен корректный результат классификации, и при наличии сопоставимых показателей эффективности.

Тем не менее, разнообразие статистических методов, по крайней мере, не исключает возможность успешного выбора для конкретных условий. Помимо этого, для повышения эффективности решения классификационной задачи любой из методов данной группы может, во-первых, входить в так называемый комитет классификаторов, а во-вторых, служить основой модифицированного алгоритма, адаптированного к реальным условиям решаемой задачи. Причём в результате модификации эффективность традиционно «более слабых» методов может даже превысить эффективность традиционно «более сильного» метода.

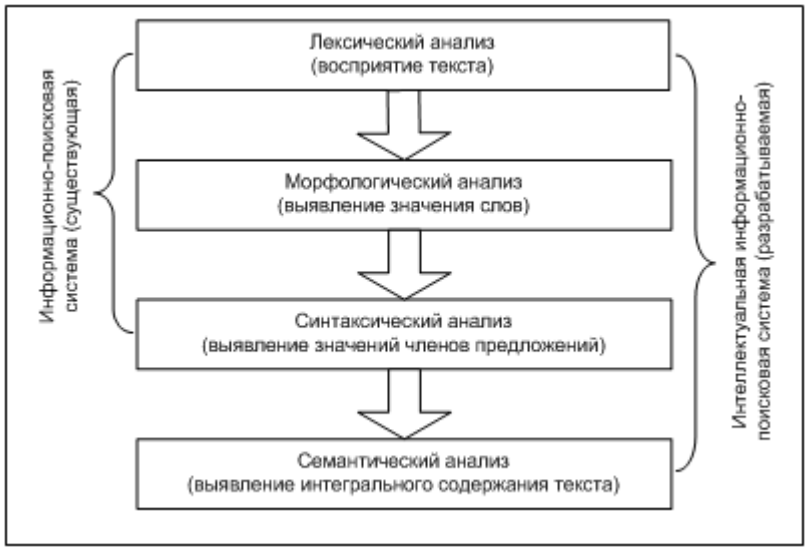
Особенностями статистических методов являются высокая скорость выполнения классификации и поиска, трудоемкость обнаружения ошибок и отсутствие возможности корректировки полученного результата в условиях поточной обработки данных. В целом данные методы являются вполне приемлемыми для применения в разрабатываемых информационных системах органов государственного управления, однако требуют предварительного выбора и оценки эффективности, а также адаптации к конкретной предметной области.

Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

Одним из средств повышения вероятности правильного решения, получаемого на основе статистических методов, состоит в использовании последующей его экспертной проверки, в рамках которой эксперт принимает полученные результаты или организует повторное автоматическое решение. Однако применение данного способа возможно только при отсутствии жестких временных ограничений.

В следующую группу объединены интеллектуальные методы, основанные на лингвистическом анализе.

Структура и назначение его отдельных блоков представлено на рисунке.




*Блок-схема лингвистического анализа текста.*

Лингвистический анализ состоит из четырех этапов.

Основной задачей лексического анализа является выделение из непрерывной последовательности одиночных символов последовательности слов. При этом происходит разбор текста на отдельные предложения, абзацы. Определяется тип изложения информации.

Морфологический анализ сводится к автоматическому распознаванию частей речи каждого слова текста (каждому слову ставится в соответствие лексико-грамматический класс). Данная задача может быть выполнена для русского языка практически со стопроцентной точностью благодаря его развитой морфологии.

Синтаксический анализ заключается в автоматическом выделении семантических элементов предложения – именных групп, терминологических целых, предикативных основ. Это позволяет повысить интеллектуальность процесса обработки тестовой информации на основе обеспечения работы с более обобщенными семантическими элементами.

Семантический анализ заключается в определении информативности текстовой информации и выделении информационно-логической основы текста. Проведение автоматизированного семантического анализа текста включает выявление и оценку смыслового содержания текста.

Лингвистические методы отличаются от остальных возможностью анализа содержания текста, что расширяет число информативных признаков для решения задач классификации и поиска. В этой связи их применение наиболее эффективно в условиях нечетких множеств и многокритериального поиска, что и вызывает к ним особый интерес при разработке интеллек-

Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

туальных систем. Использование лингвистических методов является весьма перспективным в связи с возможностью организации на их основе более глубокой и многоцелевой обработки данных, включающей:

- оценку корректности использования терминологии;
- контроль правильности используемых аббревиатур;
- контроль синтаксиса;
- поиск информации по смысловому значению;
- проверку дублирования информации;
- контроль соответствия содержания документов действующим стандартам;
- формирование гибких классификаторов на основе семантических включений.

Несмотря на активную разработку, в настоящее время не существует достаточно эффективных методов глубокого (более одного предложения) лингвистического анализа текста. Причиной этого является громоздкость расчетных методов лингвистики, связанных с построением *n*-мерных векторов и расчетами показателя их близости. В результате прикладная лингвистика не в состоянии обеспечить выявление смысла основных единиц текста и семантические связи между ними. Все реализованные подходы отличаются невысокой надежностью, весьма громоздки и ресурсоемки [2].

Несмотря на это, лингвистические методы являются наиболее перспективными для решения большинства прикладных задач автоматического анализа текстовой информации (автоматическое аннотирование, тематическая категоризация, классификация и т.д.). Однако их реализация требует разработки методических приемов, снижающих громоздкость вычислений для заданной предметной области.

Третьим типом основных задач обработки данных является фильтрация данных, содержание которой составляет поиск данных в соответствии с определенным критерием. По своей формализации задача фильтрация очень близка к задачам поиска и классификации, отличия состоят в наличии шума и задаваемого набора показателей и критериев. Однако данные отличия не являются критичными и не препятствуют использованию методов решения, изложенных выше.

Общим выводом выполненного анализа является констатация того, что существующие методы решения информационных задач могут быть использованы в ходе разработки методик автоматизированного решения функциональных задач электронного документооборота организации, однако при этом необходимо детальное обоснование системы тематических категорий, набора используемых критериев и показателей. Эти положения в последующем и будут определять научную новизну разрабатываемого методического аппарата.

Множество методов и методик оценки качества НМА формально не входит в состав методического аппарата, однако является неотъемлемым элементом его разработки и в той или иной степени должно быть использовано и на последующих этапах в интересах оптимизации алгоритма, программного обеспечения и информационной системы в целом, разрабатываемых на основе аппарата. Само понятие «качество научно-методического аппарата» не является общепринятым и нуждается в конкретизации. Это вызвано тем, что НМА не может служить конечным продуктом широкого применения, таковым скорее являются алгоритмы и их программные реализации, качество которых в случае информационных систем определяет-

Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

ся временной эффективностью. Основным показателем временной эффективности считается временная сложность или асимптотическая временная сложность алгоритмов [1].

Для оценки временной эффективности алгоритмов применяются вероятностный и статистический подходы. Вероятностный подход строится на анализе вероятностных характеристик входных данных алгоритма (без его реализации). Статистический подход подразумевает статистическую обработку результатов многократного выполнения реализованного алгоритма.

Это означает, что известные методы оценки временной эффективности алгоритмов и программ не могут быть применены к НМА до завершения формирования алгоритма, так как требуют проведения анализа входных данных или статистики выполнения расчетов. Для построения массива методик оценки эффективности решения функциональных задач в каждой конкретной разработке необходимо доопределить понятие качества аппарата и разработать методику его оценки.

Для информационных систем традиционно используются такие характеристики ее качества, как экономические показатели, сложность обучения и использования, качество информации (достоверность, актуальность, целостность и т.п.),релевантность и т.д. Однако все перечисленные параметры определяются техническими и программными решениями, реализованными в системе, и не могут быть оценены с достаточной достоверностью на этапе формирования НМА [3]. Кроме того, предметом оценивания при этом является не система в целом, а эффективность выполнения конкретного вида обработки. В целом современное состояние разработки системы критериев и показателей качества НМА характеризуется недостаточностью методологической основы.

Выполненный анализ существующих методов решения информационных задач показывает, что условием их эффективного применения в интересах автоматизированного решения функциональных задач системы электронного документооборота является адаптация к особенностям предметной области. Это связано со специфическими структурными и лингвистическими особенностями документов и характером задач документооборота, а также с недостаточным развитием существующего методического аппарата.

Таким образом, результаты проведенного анализа свидетельствуют, что эффективность существующих методов решения информационных задач, в число которых входят логико-аналитические, статистические и лингвистические (семантические) методы, существенно зависит от технологии применения и конкретики особенностей области применения. В случае использования для обработки текстовых массивов, соответствующих документам в организации, обеспечение требуемой эффективности может быть обеспечено путем соответствующей адаптации набора классификаторов, показателей и критериев, а также разработки способа структурирования данных при вводе и последующей обработке.

Главный редактор - д.э.н., профессор К.А. Кирсанов
тел. для справок: +7 (925) 853-04-57 (с $11^{00}$ – до $18^{00}$)
Опубликовать статью в журнале - http://publ.naukovedenie.ru

## ЛИТЕРАТУРА

1. Баканова Н. Б. Проектирование подсистем сбора и анализ информации для территориально распределенных информационных систем // Научно-техническая информация. Серия 1. Организация и методика информационной работы / Нижегородский госуниверситет. – Нижний - Новгород, 2006.– С. 26-38.

2. Келдыш Н. В. Концептуальный подход к автоматизации электронного документооборота на примере математической модели интегрального временного показателя. // Научно-метод. сборник № 48. / ВА МО. – М., 2007. – С.110-117 Инв. № 58592.

3. Келдыш Н. В. Методические основы автоматизированного решения задач ведомственного электронного документооборота. // Научно-метод. сборник № 56. / ВА МО. – М., 2009. – С.110-117 Инв. № 58592.

4. Моргунов Е.Б. Перспективы развития и уровни пользовательского интерфейса.— http://psyhological.ucoz.ua/publ/56-1-0-127

5. Серебряков В.А., Галочкин М.П. Основы конструирования компиляторов.— http://citforum.ru/programming/theory/serebryakov/