**Ю. В. Малинина**

# ИС ТРАНСФОРМ: АВТОМАТИЗАЦИЯ НАПОЛНЕНИЯ СИСТЕМЫ*

## ВВЕДЕНИЕ

За последние годы число статей, опубликованных в Интернете, значительно увеличилось, но для исследователей по-прежнему актуальна проблема нахождения публикаций, имеющих отношение к конкретной проблеме или области исследования.

Проект по созданию информационной системы по преобразованиям программ продолжается уже нескольких лет. Его целью являлось создание средств для обеспечения исследователей необходимой информацией. В 1999 году была завершена первая версия системы, которая в течение последних лет находилась в опытной эксплуатации. За прошедшее время процесс индексирования в документных системах прошел развитие от ручного заполнения списка ключевых слов до автоматического полнотекстового индексирования внедряемого сегодня и подразумевающего сохранение всех слов текста [10]. Внедрение автоматического индексирования открывает новые возможности перед информационной системой, однако следует учитывать ограничения этого подхода, т. к. согласно последним исследованиям число получаемых при поиске нерелевантных документов подчас достигает 90%, а размер индекса составляет в среднем не менее 40–60% объема документа. С учетом быстрого роста количества электронных документов актуальнось этих проблем усиливается.

Данная статья рассматривает существующие методы индексации и возможные подходы к решению проблемы адекватного автоматического индексирования документов и извлечения из них сопутствующей информации.

* Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант № 01-01-794) и Министерства образования РФ.

## ОБЗОР ТЕКУЩЕГО СОСТОЯНИЯ СИСТЕМЫ

ИС TRANSFORM предназначена для накопления, систематизации и использования знаний в области преобразования программ, поддержки научных исследований и обеспечения органичной работы с уже сложившимися информационными потоками. Система создана на основе публикаций в области исследования формальных методов оптимизации и преобразования программ. По существу, ТРАНСФОРМ — это соединение хранилища фактов и полнотекстовых источников информации, а также процессов, связанных с обработкой уже имеющихся данных. Если рассматривать указанные части системы: поддержка фактографической информации, возможность работы с полнотекстовыми документами, поддержка процессов обработки данных, то согласно классификации, приведенной в [6], они определяют трехмерное пространство свойств информационной системы. Первая ось (F) характеризует уровень организации хранения фактографической информации, вторая ось (D) — полнотекстовые документы, третье измерение — развитие процессов обработки информации. В рамках этой модели текущее состояние системы находится на начальном уровне и предоставляет следующие возможности.

Первая реализация системы ориентирована на общие описания преобразований программ, однако в дальнейшем собираемые данные предполагается расширить до описания конкретных реализаций преобразований и других аспектов преобразований программ, как то: архитектуры ЭВМ (семейства и модели компьютеров) и языков программирования (оптимизирующих компиляторов). Документальная информация — это библиографические описания публикаций по преобразованиям программ, т.е. авторы, название, источник, издательство, год издания публикации и т.д. Часть публикаций имеет дополнительную информацию: аннотацию, сведения о месте нахождения доступной online-версии, тему публикации и перечень упомянутых в публикации преобразований программ.

БД системы содержит следующие публикации.

- Отечественные публикации с библиографией на русском языке. Часть описаний снабжена аннотацией на русском и/или английском языке. Количество записей — около 572, что составляет более 67,37% от общего объема.

- Зарубежные публикации. Библиографическое описание приводится на языке оригинала и предусматривает перевод заглавия и/или ключевых слов на русский язык. Часть описаний снабжена аннота-

цией на языке оригинала. Количество записей — около 277, что составляет более 32,63% от общего объема.

На сегодня индексировано более 850 публикаций и около 150 преобразований.

К процессам оперирования информацией относятся поиск и ввод данных. Поиск документов осуществляется на основе булевского запроса с логическими выражениями (AND, OR, NOT и т.д.) и возможностью использовать оператор "Like this". Быстрый поиск представляет собой поиск по ключевым словам. Расширенный поиск дополнительно предоставляет возможность настроить более точно параметры поиска по отдельным полям. Поиск с предварительными запросами предоставляет пользователю возможность улучшить формируемый запрос. Все виды поиска разделены также по типу получаемого результата (публикация или преобразование).

Просмотр преобразований можно осуществлять не только при помощи поиска, но и используя дополнительную навигацию в виде рубрикаторов, на которые условно разбивается множество преобразований и публикаций [7].

Текущая версия поддерживает вставку новых библиографических и фактографических данных [1]. На сегодня в системе каждая публикация или преобразование могут быть введены в двух режимах: в пакетном режиме из файла специального формата или через web-интерфейс. Стандарта на обязательные для ввода метаданные на текущий момент не существует, но обычно они включают по крайней мере год публикации, название, автора, краткое содержание — аннотацию и ключевые слова. Стоит отметить, что последние поля (аннотация и ключевые слова) на сегодняшний день заполняются вручную. При этом практически всегда в реальных документах они отсутствуют, поэтому обычно они игнорируются по причине крайне дорогого и медленного их заполнения оператором, вводящим документы в систему. С ведением фактографической информации проблем стало еще больше, т.к. это требует больших усилий по извлечению и предварительной подготовке информации по преобразованиям программ, которые основаны на активном участии специалистов по преобразованиям программ, обладающих достаточной квалификацией.

Следует также отметить, что в текущей версии ИС в базе данных хранится только ссылка на полный текст публикаций, который размещен на некотором сайте. В системе также существует модуль, осуществляющий через определенные промежутки времени доступность этих ссылок и в качестве результата работы при обнаружении недоступной ссылки возвращающий причины, по которым ссылка недоступна. В процессе использования этого подхода была обнаружена следующая проблема: основные свежие

версии научных статей еще слабо организованы и в основном доступны через архивные сайты, сайты научных учреждений, журналов и домашние страницы исследователей, а информация, предоставляемая этими сайтами, постоянно перемещается или уничтожается. Проверка доступности ссылок показала, что 60% ссылок через год после внесения были недоступны и не могли быть впоследствии актуализованы.

## ОБЗОР МЕТОДОВ ИНДЕКСИРОВАНИЯ

Прежде чем статьи будут доступны для поиска, их необходимо индексировать. Индекс — это набор слов документа или о документе, по которым этот поиск производится. Основными критериями качества индексирующе-поисковых подсистем являются качество поиска (процент релевантных документов в списке найденных), размер индекса по отношению к размеру документа и скорость поиска по нему.

Рассмотрим в общих чертах проблемы, которые необходимо решить при реализации процесса автоматического индексирования.

Индексирование документа обычно организуется через автоматическую обработку его текста и заполнение метаданных. Автоматическая обработка с полнотекстовым индексированием включает в себя: предварительную конвертацию документа в текстовый формат для публикаций в специальном формате (например, Postscript, Acrobat PDF, TEX или Latex) и преобразование текста документа в набор слов. Причем обычно для слов сохраняется их позиция в документе для обеспечения возможности поиска по словосочетаниям. Существуют два принципиально различных метода такого индексирования с учетом применяемых в дальнейшем методов поиска:

- ***бинарное индексирование*** — не зависит от языка документа по причине бинарной или словарной индексации;

- ***морфологическое индексирование*** — производится с учетом морфологии и семантики языка.

При бинарном индексировании (контекстно-независимом по классификации [8]) поиск ведется на основе алгоритмов “нечеткого поиска”, т.е. поиска с ошибками. В этом случае допускается неполное (с заданным количеством ошибок в начале, середине и конце слова) совпадение слов с шаблоном. Объем индексной информации, полученной из текстовой, может быть в два раза больше, чем исходный текст, и предполагает использование обширного словаря. Поэтому достаточно неэффективно сохранение в базе

данных всего словарного множества документа, предлагаемого в работе [4], при этом, несомненно, нужно отдать должное простоте реализации и быстродействию алгоритма индексирования.

При втором методе индексации (контекстно-зависимом по классификации [8]) слова преобразуются в словоформы с отсечением суффиксов и окончаний, что позволяет искать склонения и спряжения шаблонов. Заметим, что эта же проблема должна решаться и при поиске. Морфологический анализ можно реализовать, оценивая окончания слов в документе [3]. Но если принять, что морфология слов русского языка определяется по окончанию и суффиксу, остается достаточно много слов, выпадающих из этого правила, т.е. есть слова, которые имеют окончание, подходящее для некоторой формы слова, но являются совершенно другой формой. Например, окончание "-ать" указывает на то, что слово является глаголом (прыгать, бежать). Но слово "кровать", оканчивающееся на "-ать", является существительным. Значит, из правил морфологического разбора могут быть исключения. Также есть слова, которые не изменяют свою форму, например: предлоги, наречия, и т.д. Значит, есть дополнения к правилу морфологического разбора. Для неоднозначно трактуемых слов можно использовать специальную таблицу в БД с атрибутами *слово* и *часть речи*, и при анализе просматривать сначала ее, а затем (если слова там нет) выполнять оценку по окончанию слова. В этой же таблице будет находиться и незначительное число слов, принадлежащих неизменяемым частям речи, таким как междометие, наречие и т.п. Аналогичный подход применим и к англоязычным текстам [11].

Заметим, что несмотря на несомненные плюсы, полнотекстовое индексирование в любом своем виде имеет и ряд существенных минусов, т.к. дело не только в том, что в базу данных попадут одинаковые слова, имеющие разные падежные окончания и т.п., но и в более глубокой проблеме, включающей следующие моменты:

- ***много излишней информации в индексе***, т.е. слов, никак не характеризующих документ, а связывающих "ключевые" слова. Это может привести к большому числу нерелевантных документов, которые будут выданы при поиске, если шаблон попадет на " ***излишнюю информацию*** ": глаголы, служебные слова, местоимения и т.д.

- ***большой объем индекса*** за счет " ***излишней информации***" — следовательно, увеличение пространства необходимого для хранения индекса и увеличение времени поиска.

Как указывается в [8], эти недостатки обусловлены самой концепцией такого индексирования. Действительно, с одной стороны, наличие в индексе всех слов текста гарантирует его нахождение по любому из них, но с другой, может существенно затруднить поиск, если текст содержит некоторые лирические отступления, напрямую не связанные с рассматриваемой темой.

Таким образом, мы возвращаемся к выделению "ключевых" слов из документа, для того чтобы гарантировать валидность результатов поиска. Только в отличие от систем первого поколения, в которых применялось ручное индексирование, данный процесс должен выполняться полностью автоматически в связи со значительно возросшим потоком документов. Кроме того, индексирование "ключевых" слов позволит значительно сократить объем индекса, а значит, и время поиска по нему.

"Ключевые" слова — это слова, определяющие содержание документа, характеризующие его смысл. Согласно [8] все многообразие документов можно разделить на виды с точки зрения их организации:

- ***структурированные документы*** — имеют четкую (известную) организацию содержания информации в документе. Определенные поля данных, их последовательность и положение, например: договора, акты, служебные записки и т.д.;

- ***неструктурированные документы*** — не обладают структурой в разрезе полей данных, например: статьи, книги и т.д.

Первый вид — это хорошо структурированные документы, с обработкой которых нет проблем, поэтому данная группа документов в дальнейшем рассматриваться не будет.

Далее разделим неструктурированные документы на подвиды с точки зрения возможности выделения "ключевых" слов. В качестве предпосылки предполагается исходить из того, что метод определения характерных слов документа должен зависеть от того, что в этом документе важно для контекстного поиска. Таким образом, разделим все неструктурированные документы на следующие группы (подвиды):

- ***контекстно-идентифицируемые*** — описывают конкретные вопросы (статьи, заметки, книги и т.д. на определенную тему или по определенным вопросам);

- ***контекстно-неидентифицируемые*** — не несут информации по конкретным вопросам (например, большинство художественной литературы).

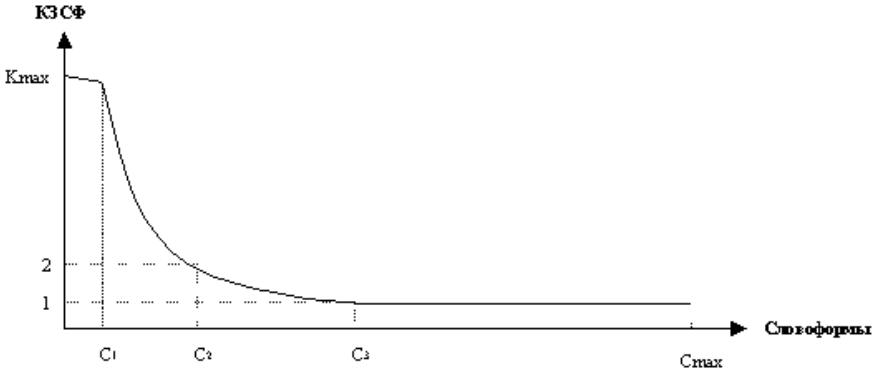
Нас интересует в основном первая группа документов. Она характеризуется тем, что в ней существует явно выделенная тема, о которой идет речь в тексте. Причем описание производится с помощью специальных терминов данной темы и сопроводительных слов, их поясняющих. Анализируя задачи, цели и способы поиска таких документов, можно заметить, что он происходит именно на основе этих самых терминов, которые и будут в данном случае "ключевыми" словами текста.

С учетом вышесказанного рассмотрим две концепции, которые позволяют улучшить условную схему индексирования документа, но предполагают различный подход в выделении ключевых слов.

## Индексирование согласно коэффициенту значимости словоформ (КЗСФ)

Согласно подходу, предложенному в [2], процесс автоматического индексирования выглядит следующим образом. Входной документ преобразовывается в поток слов, из которых выделяются словоформы путем отсечения окончаний и суффиксов, при этом для каждой словоформы необходимо запоминать частоту повторений, которую можно рассматривать как весовой коэффициент, отражающий его значимость. Полученный список сортируется по убыванию коэффициента КЗСФ.

Результаты данной работы представлены на графике.



**Обозначения:**

- $C_{max}$ — число словоформ в тексте документа;
- $K_{max}$ — максимальный КЗСФ;
- $C_1$ — число словоформ с КЗСФ примерно равным Kmax;
- $C_2$ — число словоформ с КЗСФ>2;
- $C_3$ — число словоформ с КЗСФ>1.

**Числовые соотношения:**

- $K_{max} \approx$ 8–15 для одностраничного документа (A4);
  $\approx$ 50–300 для 5–10 страничного документа (A4);
- $C_1 \approx$ 1–5 в зависимости от документа;
- $C_2 \approx$ 20–30% от Cmax;
- $C_3 \approx$ 50% от Cmax;

Представленные результаты интерпретируются автором, следующим образом.

1. Все слова со словоформами, находящимися правее точки $C_3$, не должны попадать в индекс, т.к. они не только никак не характеризуют документ, но зачастую к нему не относятся. Причем чем больше документ, тем это корректнее.

2. Слова со словоформами из зоны $C_2$–$C_3$ можно отнести к “псевдоключевым” только для очень небольших документов (порядка 1–2 страниц). Для больших документов их лучше игнорировать.

3. Слова из начала списка, покрывающие накопительной суммой своих КЗСФ примерно 5–10% от суммы КЗСФ всех слов списка, позволяют однозначно классифицировать документ по теме в случае, если каждой теме будет сопоставлен список характерных слов.

Таким образом, описанный подход выделения “ключевых” слов может быть реализован следующим образом:

– преобразование полученного текстового документа в поток слов;
– преобразование слов в словоформы;
– создание списка словоформ документа, упорядоченного по КЗСФ;

– выбор точки игнорирования по КЗСФ (ТИкзсф) в зависимости от Cmax;
– занесение в индекс словоформ, расположенных левее Тикзсф.

Он позволяет сократить объем индекса по крайней мере в два-четыре раза в зависимости от выбора точки игнорирования для КЗСФ без потери качества поиска по сравнению с полнотекстовым индексированием. Более того, число выдаваемых системой найденных нерелевантных документов значительно сократится за счет очистки индекса от мусора, а время поиска сократится за счет уменьшения индекса.

**Индексирование согласно "предметному указателю"**

В [9] предлагается концепция формирования индекса по принципу формирования предметного указателя. Известно, что предметный указатель эффективно используется для поиска в научно-технической литературе. Предметный указатель — это терминологическая база данного текста. Она включает базовые термины (существительные) и уточненные термины (существительные с определяющими их прилагательными и, возможно, предлогами).

Сама структура такого индекса должна обеспечить не только быстрый, но и релевантный поиск. Для повышения релевантности используется распространенный подход: при формировании терминологической словарной базы конкретного документа сохраняется не только сам термин, но и частота его вхождения в документ. Поэтому при выполнении поиска можно упорядочить его результаты по частоте вхождения искомого термина в документ. Кроме того, можно ввести некоторое пороговое значение f (например, $f > 1$), которое должно использоваться в качестве критерия отбора записей в поисковом запросе. Для формирования терминологического индекса требуется решить следующие задачи.

- Определить часть речи слова в документе (морфологический анализ).
- Выяснить, что является составным термином (синтаксический анализ). Предполагается, что простым термином является существительное. С составным же термином дело обстоит сложнее, поскольку нужен достоверный критерий того, какая последовательность слов является терминологически связанной.
- Определить словоформу.
- Удалить все записи с частотой вхождения ниже некоторого порогового значения. Т.е. предполагается, что термины, которые встре-

тились в документе, скажем, один раз, неадекватно характеризуют его содержание. Пороговое значение предлагается подбирать эмпирически.

Естественно, что при данном подходе не на любые запросы будет получен ответ, например, если образец поиска будет содержать только исключенные из терминологического индекса слова. Но, с другой стороны, поиск документов по словам: “например” или “следовательно” не несет особого смысла. Кроме того, учет таких слов может привести к ошибочному выполнению запроса. В качестве примера рассмотрим поиск по словам “можно” и “термин”. Ясно, что если поиск ведется по вхождению в документ хотя бы одного из двух терминов, то возможно, что он весь будет состоять из документов, содержащих слово “можно” (причем с достаточно высокой частотой вхождения) и не содержащих слово “термин”. Если же критерий поиска построен на вхождении в документ обоих слов, то релевантность такого поиска может вызвать сомнения. Не будет удивительным, если частота вхождения слова “можно” значительно превысит частоту вхождения слова “термин”, в результате чего наверху списка окажутся документы, имеющие меньшую релевантность относительно слова “термин”.

Следует отметить критичность данного алгоритма по отношению к точности определения части речи и правильности исключения "незначимых" слов.

В первую очередь сложность вызывает определение критерия, который позволил бы отличить существительное от прилагательного. Причем дело здесь не только в том, что существительное и прилагательное могут иметь в предложении одинаковые окончания, т.е. морфологический анализ в этом случае не сможет нам помочь, но и в том, что существительное и прилагательное могут быть представлены одним и тем же словом. Так слово “данные” в термине “экспериментальные данные” является существительным, а в словосочетании “данные нам в ощущениях” — прилагательным. Для четкой идентификации части речи потребуется достаточно сложный синтаксический анализ.

## ЗАКЛЮЧЕНИЕ

Рассмотренные подходы дают существенные улучшения индекса, однако эмпирические подборы пороговых значений являются узкими местами обоих концепций. Недостатком постоянного порогового значения является то, что в небольших документах может не оказаться терминов с частотами,

выше порогового значения, но нельзя сказать, что такие документы вообще не несут никакой информации. Использование переменного порогового значения, принимающего некоторое значение в интервале между максимальной и минимальной частотой терминов в данном документе, было бы предпочтительней, но для этого требуется разработка алгоритма его вычисления. Кроме того, можно определить пороговое значение для совокупности документов [8].

Предполагается комбинация и адаптация этих подходов в новой версии ИС ТРАНСФОРМ, для того чтобы пользователи системы имели доступ к актуальной и релевантной информации .

## СПИСОК ЛИТЕРАТУРЫ


1. **Волянская Т., Малинина Ю.В.** Трансформ: интерфейс для ввода информации // Поддержка супервычислений и интернет-ориентированные технологии. — Новосибирск: ИСИ, 2001. — С. 125–139.
2. **Гацко А**. Концепция индексирования по ключевым словам. — computerclub.dore.ru
3. **Ермолаев Д.С.** Компьютерный морфологический разбор слов русского языка. — www.icreator.ru
4. **Игумнов Е**. Основные концепции и подходы при создании контекстно-поисковых систем на основе реляционных баз данных. — www.citforum.ru
5. **Корнеев В.В., Гареев А.Ф., Васютин С.В., Райх В.В**. Базы данных. Интеллектуальная обработка информации. — М.: "Нолидж", 2000. — 352 с.
6. **Красилов Н., Косякин И., Черных Д.** Об одной модели документооборота // Открытые системы. — 1997. — № 1. — www.osp.ru
7. **Малинина Ю.В.** ИС ТРАНСФОРМ: Прототип интерфейса для визуального исследования БД // Проблемы систем информатики и программирования. — Новосибирск: ИСИ, 1999
8. **Марков А.** Концепция построения электронного архива // Открытые системы. — 1997. — №1. — эл. публик. www.osp.ru
9. **Моисеенко В, Майстренко А.** Релевантность полнотекстового поиска: подход на основе построения терминологической базы документов. — www.citforum.ru
10. **Lawrence S., Bollacker K., Lee C.** Indexing and Retrieval of Scientific Literature // Proc. 8th Intern. Conf. on Information and Knowledge Management. — CIKM 1999.
11. **Porter M.F.,** An algorithm for suffix stripping // Readings in Information Retrieval. — 1997. — Vol. 14, N 3. — P. 130–137.