УДК 004.415.2.043

АВРАМЕНКО В.С.

ПАРАМЕТРИЧЕСКИЕ МЕТОДЫ ПОИСКА ИНФОРМАЦИИ

Целью данной работы является создание программных средств для повышения эффективности автоматизированного документального поиска в гипертекстовых ресурсах Интернет. Для этого предлагается метод автоматического индексирования параметрических данных при обработке документов, содержащих числовую информацию с наличием как различных единиц размерностей, так и различных условных обозначений семантически одних и тех же количественных мер.

Ключевые слова: автоматическое индексирование, обработка документов, параметрические данные, числовая информация

An aim hired is creation of programmatic facilities for the increase of efficiency of the automated documentary search in hypertext resources the Internet. For this purpose the method of the automatic code indexing of self-reactance data is offered at treatment of documents, containing numerical information with the presence of both different units of dimensions and different conditional denotations semantically of the same quantitative measures.

Введение

Одним из основных источников информации сегодня являются ресурсы глобальной информационной сети Интернет, и обеспечение доступа к информации обычно связывается с обеспечением доступа именно к ресурсам Интернет. Развитие сети Интернет в наши дни связано в первую очередь с ростом объема информационных ресурсов и повышением качества инфраструктуры физических сетей. В течение следующих нескольких лет основными тенденциями дальнейшего развития будут дальнейший рост объемов ресурсов, накопление информации и увеличение числа пользователей имеющих доступ к глобальной информационной сети.

Развитие исследований в области информационных проблем требует построения таких автоматизированных систем обработки данных, которые извлекают из информационных потоков необходимые сведения. Эти системы, именуемые обычно интеллектуальными роботами, широко применяются в поисковых системах Интернет [4]. Сходная по своей сущности задача возникает при оперативной обработке больших и быстро движущихся потоков текстовой информации, содержащей параметрические данные.

Целью данной работы является повышение эффективности автоматизированного документального поиска в гипертекстовых ресурсах Интернет. Для этого предлагается метод автоматического индексирования параметрических данных при обработке документов, содержащих числовую информацию с наличием как различных единиц размерностей, так и различных условных обозначений семантически одних и тех же количественных мер.

Вспомогательная информация для автоматического индексирования

Метод автоматического индексирования параметрической информации основывается на использовании сведений о внешней (синтаксической) структуре текста и дополнительных сведений, извлекаемых из вспомогательных списков и таблиц.

В процессе исследования текстов и словарей и справочников [3, 5, 6] были получены данные, на основе которых составлены следующие вспомогательные списки и таблицы:

1. Список (словарь) неинформативных терминов (стоп-слов).
2. Таблица слов, входящих в неинформативные выражения.
3. Список окончаний русского языка.
4. Список суффиксов русского языка.
5. Таблица префиксоидов русского языка.
6. Таблица суффиксоидов порядковых и количественных числительных.
7. Таблица порядковых и количественных числительных.
8. Таблица кратных и дольных единиц размерностей физических величин.
9. Таблица единиц размерностей физических величин.

При составлении списков и таблиц были проанализированы словари, образованные из лексики массивов системы АСИОР и МОДИС,

обратный словарь русского языка и другие материалы [1-2].

В словари неинформативных терминов русского и английского языков вошли служебные слова (предлоги, союзы, артикли), сокращения, неинформативные слова и словосочетания. Термины словаря разбиты на группы в зависимости от количества символов, оставшихся после отбрасывания окончаний. Техника отбрасывания окончаний описана ниже (блок лексического анализа). В словарь включена также некоторая информация, необходимая для блока синтаксического анализа.

Списки окончаний и суффиксов разбиты на определенные классы, в состав которых входят положительные и отрицательные элементы (стоп-элементы) с соответствующими морфологическими и семантическими признаками. Во вновь разработанном алгоритме отождествления и отбрасывания окончаний и суффиксов, учитывающем сегментацию суффиксальной последовательности и местоположение отрицательных или положительных элементов этой последовательности, данные списки сведены до минимума. Так, список русских суффиксов содержит только 78 положительных и отрицательных суффиксов длиной от одного до четырех символов, в то время как, например, модифицированный список ИПС АСИОР [1] насчитывал около 400 элементов длиной от 1 до 12 символов.

Списки русских и английских суффиксов разбиты на две группы. К первой группе относятся суффиксы, которые могут быть отброшены, если они занимают только конечную позицию в слове (конечные суффиксы). Во вторую группу отнесены все остальные суффиксы и некоторые конечные суффиксы. Таким образом, в данные списки вошли не сложные суффиксы, а, в основном, неделимые суффиксы (суффиксальные морфы [5]). Это такие суффиксы, при дальнейшем делении которых выделяется хотя бы одна часть, которая не является суффиксом.

При обработке документов, содержащих числовую информацию, возникают затруднения, связанные с наличием как различных единиц размерностей, так и различных условных обозначений семантически одних и тех же количественных мер. Поэтому, для опознавания численной информации, выраженной количественными и порядковыми числительными, а также для перевода единиц измерений в Международную систему единиц СИ (Система Интернациональная) составлены различные таблицы (списки):

– префиксоидов русского языка;

– числительных русского (английского) языка (или их основы);

– единиц измерений всех существующих основных систем (СИ, СГС, МТС, МКГСС) с соответствующими числовыми значениями и коэффициентами пропорциональности.

Список префиксоидов русского языка содержит префиксоиды, являющихся составной частью сложных слов, представляющих числовую информацию.

В таблицу единиц измерений включены также производные единицы размерностей этих систем, имеющие собственные наименования. Производные единицы размерностей, не имеющие собственного наименования и называемые по другим единицам, в данную таблицу не включены, так как такие размерности могут вычисляться алгоритмически на стадии перевода их в систему СИ.

В качестве основных единиц, которые должны входить в любую формулу размерности при ее вычислении, предлагается [3]:

1) семь основных единиц СИ: длина (метр), масса (килограмм), время (секунда), сила электрического тока (ампер), термодинамическая температура (кельвин), сила света (кандела), количество вещества (моль);
2) логарифмическая единица отношения двух величин (десятичный логарифм отношения двух одноименных физических величин, например мощностей, токов, звукового давления), характеризующие интенсивность звука (бел, чаще применяют 0,1 долю бела – децибел);
3) логарифмическая единица, характеризующая единицу количества информации (бит);
4) деньги;
5) относительная величина, представляющая собой безразмерное отношение физической величины к одноименной физической величине (число – штук, сотая доля – процент).

Каждой основной единице присваивается 12-значное 16-ричное число, ненулевой разряд которого в соответствующей позиции закреплен за одной из основных единиц. При составлении формулы размерности каждый разряд (цифра) 16-ричного числа складывается (вычитается) только с соответствующим разрядом по

модулю 5. Отрицательные цифры представляются дополнительным кодом: F (-1), E (-2) и т.д. Цифра соответствующего разряда 16-ричного числа указывает сколько раз входит основная единица в формулу размерности (положительные от 1 до 7 – в числитель, отрицательные от 7 (1001) до 1 (1111) – в знаменатель).

Например, формула размерности для давления, которое в системе СИ измеряется в паскалях и выражается через другие основные единицы (1 Па = 1 Н/м^2) равна

-1 1 -2 0 0 0 0 0 0 0 0 0

(длина-м$^{-1}$, масса-кг, время-сек$^{-2}$), так как Ньютон – производная единица. Исходя из второго закона Ньютона она определяется как сила, изменяющая за 1 с скорость тела массой 1 кг на 1 м/с в направлении действия силы. Таким образом, 1 Н = 1 кг·м/с^2. Поэтому другие единицы размерностей будут иметь следующие коэффициенты пересчета:

1 Паскаль (Pa, Па) = 1 м$^{-1}$·кг·сек$^{-2}$ = 1 Н/м^2 = 10^{-5} Бар (bar, бар) = $10{,}197\cdot10^{-6}$ Техническая атмосфера (at, ат) = $9{,}8692\cdot10^{-6}$ Физическая атмосфера (atm, атм) = $7{,}5006\cdot10^{-3}$ Миллиметр ртутного столба (мм рт. ст., mm Hg, Torr, торр) = $1{,}0197\cdot10^{-4}$ Метр водяного столба (м вод. ст., m H_2O) = $145{,}04\cdot10^{-6}$ Фунт-сила на кв. дюйм (psi)

Лексический анализ

На стадии лексического анализа может использоваться простейший лексический анализатор. Он выполняет вспомогательные функции, и поэтому может быть реализован как самостоятельный блок, или как часть блока морфологического анализа.

Блок лексического анализа принимает исходный текст, который он должен разбить на предложения, или уже готовые предложения. Анализируемое предложение попадает на вход лексического анализатора в виде массива символов, содержащего прописные и строчные буквы русского (английского) алфавита, цифры, знаки пунктуации.

Полученный массив анализатор преобразует в массив лексических единиц. Здесь под этим термином подразумевается слово, число, скобки, знаки препинания. Для каждой лексической единицы формируется отдельная строка, в которую копируются все символы, принадлежащие данной лексической единице, и приписывается тип (класс) лексемы:

– слово естественного языка;

– слово, имеющее в составе русские и английские символы;

– слово, имеющее в составе хотя бы один цифровой символ;

– слово, имеющее в составе различные символы, кроме цифровых символов;

– разделители.

Морфологический и семантический анализ

После лексического анализа термины текста последовательно подвергаются семантическому и морфологическому анализу. На стадии предварительного семантического анализа термины всех типов разделяются на слова естественного языка (русские и английские), числовую информацию и термины, состоящие из смешанных символов.

Числовая информация (числа, числительные) и единицы размерностей приводятся к каноническому виду с учетом соответствующих коэффициентов пропорциональности. Все “слова” естественного языка подвергаются морфологическому анализу, задачей которого является отделение у каждого слова основы и приписывание этой основе морфологической и семантической информации, необходимой для установления текстуальных отношений между терминами на этапе синтаксического анализа. Морфологический анализ выполняется по следующей схеме:

1. Сложные слова разбиваются на самостоятельные единицы: префиксоид(ы) и основу префиксоида. Например, сложное слово “двадцатипятитысячный” разбивается на два префиксоида (“двадцати”, “пяти”) и основу префиксоида (“тысячный”).

2. Различные словоформы одного слова приводятся к одному и тому же парадигматическому коду (основе).

3. Определение порядковых и количественных числительных, неинформативных и полуинформативных слов.

После морфологического анализа входной текст будет представлен последовательностью условных единиц (индексов). Каждая из этих единиц содержит сведения о том, из какой лексической единицы она получена (специальный

знак, неинформативный термин, полуинформативный термин, число, единица размерности, русское слово, английское слово, смешанный термин), а также сведения об этой лексической единицы, извлеченные из вспомогательных таблиц и списков.

Синтаксический анализ

Исходной информацией для этапа синтаксического анализа является цепочка индексов с соответствующей морфологической и семантической информацией. Целью синтаксического анализа является разбиение этой цепочки на определенные конструкции (индексы и сегменты) и установление соответствующих связей между элементами этих конструкций.

Индекс – это набор символов, полученный из значащего термина на стадии лексического анализа. Сегмент – последовательность индексов, ограниченная знаками препинания (запятой, тире, двоеточием, восклицательным знаком, точкой с запятой, точкой, обозначающей конец фразы) или некоторыми служебными словами и другими неинформативными терминами.

Одновременно с сегментацией решаются вопросы лексико-грамматической омонимии, приводится к каноническому виду числовая информация, заданная диапазоном и записанная в сложной (составной) форме.

Лексико-грамматическая омонимия на синтаксическом и морфологическом уровнях распознается с помощью информации, полученной на стадии лексического анализа. Проверке подвергаются только “полуинформативные” термины. Например, при анализе слов “пуст(ой) – пуст(ь)”, “почт(а) – почт(и)”, “вес – вес(ь)” признак “полуинформативности” уничтожится у слов “пустой”, “почта”, “вес”, а признак “неинформативности” будет приписан остальным словам. Для разрешения более сложных омонимов рассматривается левое или правое его окружение, или же левое и правое одновременно

Если при анализе цепочки индексов обнаружатся числовые величины, записанные с указанием интервала (например, “с … по”, “от … до”, “в диапазоне”, “-”), перечислением значений одной и той же величины или в сложной (составной) форме, то они приводятся к каноническому виду с приписыванием признаков нижнего и/или верхнего предела и соответствующей единицы размерности.

На заключительной стадии синтаксического анализа из поискового образа выбрасываются сегменты, которые состоят из неинформативных элементов (индексов). Внутри каждого сегмента устанавливаются связи между его элементами. В частности, для параметрической информации устанавливается атрибутивная связь (число + размерность), дефисная связь (число + число).

Выводы

В отличие от технологий ранних и современных информационно-поисковых систем, базирующихся на технологии текстового поиска и обеспечивающих поиск документов на основе их информационного содержания (набор дескрипторов или значения каких-либо других атрибутов), предлагается дополнить эти технологии автоматическим индексированием параметрических данных при обработке документов, содержащих числовую информацию с наличием как различных единиц размерностей, так и различных условных обозначений семантически одних и тех же количественных мер.

Список литературы

1. Авраменко В.С. Автоматизированная система информационного обеспечения разработок / В.С. Авраменко, В.И. Легоньков В.И., В.Р. Хисамутдинов. – М.: Наука, 1980. – 208 с.
2. Авраменко В.С. Математическое обеспечение диалоговых информационных систем / В.С. Авраменко, В.И. Легоньков В.И., В.Р. Хисамутдинов. – М.: Наука, 1990. – 192 с.
3. Бурдун Г.Д. Справочник по международной системе единиц / Г. Д. Бурдун. – М.: Изд-во стандартов, 1971. – 231 с.
4. Крупник А.Б. Поиск в Интернете: Самоучитель / А.Б. Крупник. – СПб.: Питер, 2006. – 268 с.
5. Зятковская Р.Г. Суффиксальная система современного английского языка / Р.Г. Зятковская. – М.: Высшая школа, 1971. – 186 с.
6. Обратный словарь русского языка. – М.: Советская энциклопедия, 1974. – 944 с.