электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.424.4

ПРИНЦИПЫ ОБРАБОТКИ ТЕКСТОВОЙ ИНФОРМАЦИИ В СИСТЕМАХ СОПОСТАВЛЕНИЯ, ПОИСКА И ИДЕНТИФИКАЦИИ

Пролетарская В.А., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Система обработки информации и управление»

Научный руководитель: Ревунков Г.И.,к.т.н.., доцент
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана
bauman@bmstu.ru

В работе рассматривается подход к работе с предметами и явлениями реального мира и обработки информации о них с целью создания систем поиска, сопоставления и идентификации. Первоначально в рамках работы были описаны рекомендации по анализу предметов и явлений, классификации информации о них и создание объектов в информационном пространстве. Рассмотрены особенности работы с различными видами данных и методы их сопоставления.

После выработки четких стандартов представления атрибутов объектов предметной области, которые бы позволяли упростить задачу сопоставления, идентификации и/или поиска, называемой нормализацией, первым шагом должен быть поиск существенной информации и выделение атрибутов, по которой можно было бы идентифицировать однозначное соответствие объектов. Выборка данной информации основывается на анализе доступных сведений о сущности объекта с учетом структуры данной системы, а также предметной области, для которой производится выверка. При этом следует учитывать, что данные при внесении в систему могут быть искажены или представлены в «неявном» виде: возможны ошибки оператора при внесении данных в Систему, полное и краткое(возможно, аббревиатурное) написание названия атрибуты, наличие в текстовом атрибуте малосодержательной информации (например, организационная форма для компании).

Выделение принципов, по которым искажаются атрибуты объекта, требует глубокого анализа каждого атрибута, его источника, инструмента, при помощи которого он получается, способа хранения и передачи информации. Для частичного решения проблемы искажения информации таким образом используются справочники замен, которые являются центральной частью системы нормирования.

Для каждой смысловой группы разрабатываются отдельные справочники замен. Различия в справочниках основаны в первую очередь на различиях в природе и смысловой нагрузке атрибутов каждой группы.

Справочник можно представит в виде таблицы, столбцами которой являются:

- Заменяемый элемент.
- Идентифицирующая операция.
- Замена.

Справочник работает следующим образом: атрибуты объекта представляются в виде множества «подстрок» - всевозможных частей информации, содержащейся в атрибуте. Каждая «подстрока» последовательно сравнивается на выполнение «Идентифицирующей операции» с «Заменяемыми элементами» справочника и, при положительном результате, замещается «Заменой».

После составления справочника замен для смысловой группы атрибутов, можно использовать функцию нормализации.

Кроме справочников замен при нормализации используется функция приведения к единому регистру. Единый регистр – понятие, позаимствованное из теории обработки печатных текстов. Однако, в данной работе, оно имеет более широкий смысл. В настоящем контексте, единый регистр – это не только регистр символов, но и единые единицы и системы измерения, базы распределения и т.д.

Оперируя этими понятиями, можно составить следующую универсальную формулу нормализации:

$n_j\ (a_{i,j}\ ) = \cup\ (\forall a_{i,j}(x,y) =\ z_j\ \ , a_{i,j}(x,y) = S_j)$,

где $\cup\ (a_{i,j})$– функция приведения атрибута $a_{i,j}$ к общему регистру,

Z_j – множество заменяемых элементов j-го атрибута, а $z_j \in Z_j$ – замена для j-го атрибута,

S_j – множество сокращений для замен Z_j,

$a_{i,j}(x,y), x < y$ - подстрока j-го атрибута i-го объекта системы с символа номер x до символа номер y.

При этом сокращение s_j может быть пустым, что приведет к удалению найденной замены.

После нормализации информации следует шаг сопоставления двух объектов системы. Функции сопоставления должны быть уникальны для каждого класса содержащейся в атрибутах информации. При необходимости можно создать функции сопоставления для каждого атрибута объекта. Сопоставление атрибутов может быть четким или нечетким. Ограничившись допущениями и ограничениями задачи, можно

достаточно быстро выделить метод сопоставления, удовлетворяющий по основным критериям:

- Специфика задачи и сопоставляемых атрибутов.
- Точность.
- Скорость работы.

Зачастую самым лучшим с точки зрения этих критериев является метод на основе самого простого алгоритма(см.Таблица 1). Однако, рассматривая задачи с разной спецификой нельзя вслепую переносить методы и функции сопоставления атрибутов, сколь бы хорошо они не зарекомендовали себя ранее.

Был произведен анализ алгоритмов методом взвешенных сумм. Результаты сравнения приведены в Таблице 1.

Таблица 1

Сравнение алгоритмов методом взвешенных сумм

Алгоритм	**Ресурсоемкость**	**Сложность**	**Скорость**	**Возможность задания пороговой точности**	**Возможность ошибки сопоставления**	**Итоговое значение**
1	**2**	**3**	**4**	**5**	**6**	**7**
Стандартный алгоритм сравнения	0,8	1	0,5	1	0,7	0,84
Алгоритм Кнутта-Мориса-Пратта	1	0,7	0,6	0,5	1	0,72
Простой алгоритм сравнения Бойера-Мура с учетом корреляционного веса	0,9	1	0,9	1	0,9	0,96
Алгоритмы нечеткого сравнения	0,5	0,5	0,7	0,8	0,8	0,67
Оптимизированный алгоритм Бойера-Мура	0,7	0,7	1	1	0,8	0,87
Весовой коэффициент	0,1	0,3	0,1	0,3	0,2	

Из специфики области задачи было решено применить метод нечеткого сопоставления данных с использованием алгоритма поиска наибольшей общей подстроки с учетом корреляции с контролируемым порогом.

Для того чтобы не идентифицирующими атрибутами было возможно описать объект, необходимо сформировать из них идентифицирующий набор. Составление подобного набора можно произвести статистическими, либо аналитическими методами, в зависимости от поставленной задачи и условий решения [1].

Аналитический метод составления идентифицирующих наборов заключается в экспертном определении списка атрибутов, который в рамках задачи системы необходим и достаточен для идентификации объекта. Выбор идентифицирующего набора аналитическим методом напрямую зависит от поставленной задачи.

Статистические методы целесообразно применять при большом количестве не идентифицирующих атрибутов и их групп. В основе метода лежит выборка из большого массива данных искомых атрибутов. В методах составления идентифицирующих наборов остро стоит проблема выбора не идентифицирующих атрибутов. Необходимо выбрать такую совокупность, чтобы исключить лишние атрибуты [1].

Для выделения ключевых атрибутов необходимо провести проверку по всем выделенным значимым атрибутам с использованием составленных алгоритмов нормализации и поиска сходства, а затем проверить результаты на предмет их удовлетворения требованиям по возможному числу ошибок. Для атрибутов, сопоставление по которым не удовлетворяет требованиям точности, необходимо совмещение с другими проверками для понижения удельного числа ошибок.

Завершив классификацию атрибутов и составление функций нормирования и сопоставления, необходимо определить последовательность, в которой будут обрабатываться атрибуты.

Порядок обработки атрибутов необходим для минимизации производимых вычислений и оптимизации использования операций, которые иногда могут быть трудоемки и дорогостоящи. Составление порядка обработки атрибутов основывается, как и многое другое, на понимании задачи и объекта системы [1,2].

Последовательность обработки объектной информации выполняется в соответствии с порядком, представленным в Таблице 2.

Таблица 2

Порядок обработки атрибутов

Порядковый номер операции обработки	**Класс атрибута**	**Ресурсоемкость нормализации и сопоставления**
1	**2**	**3**
1	Идентифицирующий атрибут	Низкая
2	Идентифицирующий набор	Низкая
3	Идентифицирующий атрибут	Высокая
4	Идентифицирующий набор	Высокая
5	Идентифицирующий атрибут, получаемый при помощи операции	Низкая
6	Идентифицирующий атрибут, получаемый при помощи операции	Высокая
7	Идентифицирующий набор, получаемый при помощи операции	Низкая
8	Идентифицирующий набор, получаемый при помощи операции	Высокая

Определившись с последовательностью обработки атрибутов, следует определиться с тем, будет ли производиться предварительная нормализация атрибутов объектов. В рамках предварительной нормализации производится нормализация атрибутов известных объектов и их множеств.

Так в задачах поиска и идентификации может быть проведена предварительная нормализация атрибутов объектов эталонного множества.

Для задачи сопоставления можно провести предварительную нормализацию обоих сопоставляемых множеств.

Целью предварительной нормализации является уменьшение ресурсоемкости последующего сопоставления. Это может быть очень полезно, если задачей системы

подразумевается минимизация времени принятия решения или если предстоит обработка большого количества объектов.

После изучения атрибутов, их особенностей и определения места операций в будущей системе, а также принятия решения о целесообразности предварительной нормализации, должен быть составлен итоговый алгоритм работы системы.

Результатом работы является анализ процесса разработки систем поиска, сопоставления и идентификации, базирующаяся на объектно-ориентированном подходе к работе с информацией. Для конкретной задачи произведено сравнение методов нечеткого сопоставления атрибутов методом взвешенных сумм. Приведены общие принципы составления стандартов в рамках системы для представления атрибутов.

Список литературы

1. Пролетарская В.А. Разработка алгоритмического модуля выверки данных информационной системы// Сборник статей докладов общеуниверситетской научно-технической конференции «Студенческая научная весна-2012», посвященной 165-летию Н.Е. Жуковского. 02 - 29 апреля 2012 г., МГТУ им. Н.Э. Баумана / МТ. 12. Часть 1.- М. .: НТА «АПФН», 2012. (Сер. Профессионал). – С.330-335
2. Пролетарская В.А. Исследование методов анализа объектной информации // Сборник трудов №9 молодых ученых, аспирантов и студентов «Информатика и системы управления в XXI веке» – М.: МГТУ им. Н.Э. Баумана, 2012. – С.77-87.
3. Graham A. Stephen Анализ строк. Перевод М.С.Галкиной под редакцией П.Н. Дубнера, 1992
4. Бойцов Л. М. Использование хеширования по сигнатуре для поиска по сходству // Прикладная математика и информатика. 2000.- N 7.
5. Бойцов Л. М. Поиск по сходству в документальных базах данных: хеширование по сигнатуре оптимальное соотношение скорости поиска, простоты реализации и объема индексного файла. // Программист. - 2001. - N 1.
6. Григорьев Ю. А., Ревунков Г. И. Банки данных: Учеб. для вузов. – М.: Изд-во МГТУ им. Н. Э. Баумана, 2002. – 320 с. (Сер. Информатика в техническом университете).