Procedures for the identification of distorted information
Gusnin S.[1], Petukhov A.[2] (Russian Federation)
Процедуры отождествления искаженной информации
Гуснин С.Ю.[1], Петухов А.Н.[2] (Российская Федерация)

[1]*Гуснин Сергей Юрьевич / Gusnin Sergey – кандидат технических наук, доцент,*
кафедра информационные и сетевые технологии,
институт информационных систем и технологий,
Московский авиационный институт (национальный исследовательский университет), г. Москва
[2]*Петухов Андрей Николаевич / Petukhov Andrey – кандидат технических наук, доцент,*
кафедра информационной безопасности,
Московский институт электронной техники (национальный исследовательский университет), г. Москва

Аннотация: *в статье представлен анализ подходов к построению алгоритмов и технологий отождествления искаженной информации. Рассматриваются достоинства и недостатки различных алгоритмов. Указывается на преимущество посткоординированных процедур перед предкоординированными. Рассматривается связь между распознаванием образов и отождествлением. Выявляются недостаточно учтённые в процедурах факторы. Выделяются задачи перспективных исследований.*

Abstract: *the analysis of approaches to the construction of algorithms and techniques of identification of distorted information. We consider the advantages and disadvantages of various algorithms the article presents an analysis of approaches to the construction of algorithms and techniques of identification of distorted information. We consider the advantages and disadvantages of various algorithms. Points to the benefits postcoordinated procedures before precoordinated. The connection between pattern recognition and identification. Rounds identified inadequate procedures factors are allocated the task of Advanced Studies .*

Ключевые слова: *информационная безопасность, отождествление, коэффициенты корреляции, расстояние, контентная фильтрация, DLP-системы.*

Keywords: *information security, identification , correlation coefficients , distance, content filtering , DLP - system.*

Локализация в информационных потоках и массивах данных конкретного содержания (контента) осуществляется с помощью сопоставления информационных фрагментов естественного языка и принятия решения о степени соответствия этих фрагментов друг другу, т.е. с помощью отождествления информации. Этот базовый элемент любого информационного поиска в базах данных и в сети Интернет нашел свое применение в практике сервисов информационной безопасности сначала в средствах контентной фильтрации, а затем в более развитых инструментах систем защиты от утечки информации (DLP). Славная своим многолетним применением программа гарантированного уничтожения информации TERRIER использует для обнаружения фрагментов данных, подлежащих стиранию, специальные словари ключевых слов, подготовка которых может рассматриваться как некоторый этап отождествления.

Отождествление – это комплекс информационно-логических процессов, направленных на установление отношения между накопленной и поступившей информацией, между сигнатурой и потоком данных, между образцом и массивом. Отношение, о котором идет речь, может трактоваться двояко, что определяет два взаимосвязанных, но различных смысла отождествления. Если имеется в виду отношение тождественности элементов данных, терминов в том виде, в котором они сформулированы, речь идет о тождественности информации по форме, и процесс установления такого отношения ограничивается отождествлением информации. Если же тождественность понимать шире, а именно по отношению к более масштабным интегрированным агрегатам (сообщениям, документам, описаниям, потокам), сведения о которых содержатся в отождествляемых терминах, то отношение тождественности устанавливается по отношению к этим агрегатам. Второй смысл является более широким, но эффективное решение проблем отождествления терминов является первым и существенным шагом в процессе квалификации более крупных информационных конструкций.

Из приведенного определения вытекает, что необходимыми являются, по крайней мере, три элемента: (а) исходный образец (поступившая информация, контентная сигнатура), (б) множество элементов данных (накопленная информация, контролируемый поток, массив данных) и (в) процедура установления отношения. Исходный образец и множество элементов данных в совокупности представляют собой отождествляемую информацию, а аппарат, устанавливающий отношение между образцом и элементом множества, называют процедурой отождествления. Сразу отметим, что эти процедуры призваны не устанавливать «правильные» формы и значения, не выяснять как «должно быть на самом деле», а «всего лишь» определять, насколько можно быть уверенным, считая сопоставляемые данные одинаковыми, тождественными (отсюда происхождение самого термина «отождествление»).

Результатом отождествления в любом случае является некоторая промежуточная информация, которая характеризует отношение между образцом и каждым контентным элементом множества потока или массива

накопленных данных. Эта промежуточная информация может иметь различную форму. Самой простой является бинарная, в этом случае максимально упрощается принятие решения о дальнейшей обработке, т.к. существуют лишь две альтернативы, но решение будет более гибким, если результат отождествления сможет принимать несколько (более двух) значений, а самой общей формой представления промежуточной информация является значение из некоторого непрерывного интервала.

Процесс отождествления интегрированных агрегатов данных может иметь внутреннюю функциональную структуру, в которой выделяют два уровня: отождествление значений отдельных терминов и отождествление агрегата в целом. На первом уровне отождествления используется лишь информация, заключающаяся в самих значениях терминов, а на втором уровне оно проводится на основании уже полученных результатов и может не зависеть от конкретных значений.

В отношении результата отождествления верхнего уровня, который несет в себе информацию о характере отношения между отождествляемыми фрагментами данных, естественно стремление обеспечить наиболее полное представление об этой информации, что соответствует выбору непрерывной формы представления результата. Что касается второго уровня, то такое стремление целесообразно только в случае, когда результат отождествления является лишь одним из нескольких факторов, определяющих принятие решения о квалификации агрегата. Если же отождествление агрегата полностью определяет решение о дальнейшей его обработке, то наиболее привлекательна форма представления результата, обеспечивающая взаимно-однозначное соответствие промежуточной информации и вариантов решения, множество которых принципиально конечно.

Ввиду глубины и обширности темы критериев эффективности процедур отождествления, оставим вопросы их выбора и вычисления для самостоятельного рассмотрения, хотя эти вопросы, в конечном счете, несомненно, являются ключевыми в реальном проектировании. Отметим только, что в отождествляемой информации содержится некоторая доля неопределенности отнесения к конкретному объекту, которая обуславливает принципиально достижимый уровень эффективности средств отождествления, и в общем случае возникновение ошибок отождествления информации обусловлено тремя причинами:

- априорной объективно содержащейся в исходной информации неопределенностью по отношению к соответствующим реальным информационным объектам;
- неполнотой адекватности представлений о количественных и качественных характеристиках факторов, порождающих неопределенность в информации;
- несовершенством алгоритмических средств отождествления реализованных в системе.

Естественная многозначность лексики («один термин - много объектов») требует умения измерять информативность терминов и применять совместное (мажоритарное) отождествление с использованием наиболее информативных терминов. Синонимия («много терминов - один объект»), возникновение лексических дериватов заставляют привлекать развитую словарную службу. Существование парадигматических (типа «род-вид» или «часть-целое») и синтагматических (установленных на уровне конкретного контента) связей делают необходимым включение в систему классификаторов и тезауруса.

Кроме того, в составе связного текста слова приобретают различные морфологические формы, оставаясь при этом основанием для отождествления. Для устранения этого фактора неопределенности применяются различные приемы «морфологической нормализации». Арсенал таких приемов широк - от формального отказа от морфологических признаков (стемминг, лемматизация) до полноценных методов морфологического анализа (установление флективных классов, приведение к каноническим языковым формам) [1].

Все перечисленные направления снижения неопределенности (и, тем самым ошибок отождествления) занимают подобающее им место в ансамбле информационно-алгоритмических инструментов отождествления и заслуживают самостоятельного анализа. В рамках этой работы внимание будет привлечено к фактору, относительно независимому от лингвистических аспектов, а именно к технологическим искажениям, которым подвергается информация до попадания в поле зрения поиска или контентной фильтрации.

Проходя путь от своего источника информация, многократно передается и преобразуется. Неотъемлемым свойством этих преобразований является их стохастический характер. Кроме того, в различные моменты времени одна и та же информация может проходить различные траектории передачи и преобразования, и выбор конкретной траектории, с точки зрения внешнего наблюдателя, происходит случайно.

Специфическая особенность многих каналов поступления информации состоит в определенной трудности апостериорного контроля воспринятых данных путем повторного подтверждения или дополнительной ассоциации их с другими сведениями, достоверность которых известна и достаточна. Речь идет о информационных каналах в той части, которая не поддерживается автоматизированными информационно коммуникационными средствами и не предполагает использования традиционных методов помехоустойчивого кодирования (или эти методы недостаточно эффективны на контентном уровне). Поэтому термин «технологические искажения» прежде всего следует относить к технологиям за пределами информационно-коммуникационных систем (технологиям предсистемной обработки - восприятие на слух, чтение рукописного текста, неформальное преобразование из одного языка в другой с попыткой сохранить

звучание и т.п.). Поэтому, именно существующая структура и методы предсистемной обработки обуславливают наличие в информации случайных отклонений от формы, предоставляемой источником, эти отклонения будем называть искажениями.

Статья представляет собой обзор (не претендующий на исчерпывающий характер) подходов к построению алгоритмов и технологий отождествления искаженной информации.

В простейшем случае средства отождествления обеспечивают определение полного (графического) совпадения данных. Очевидно, что любой фактор, изменяющий такую графическую форму данных, вызывает ошибку отождествления. Попытки уменьшить интенсивность ошибок нашли свое выражение в создании более или менее развитых средств отождествления. Эти средства весьма разнообразны, поэтому необходимо провести некоторую систематизацию множества различных результатов, достигнутых в этой области.

Задача отождествления, как в отношении отдельного термина, так и в отношении агрегата в целом с точки зрения структуры функционирования процедуры может быть сформулирована двояко:

1. Прямая задача отождествления - располагая двумя фрагментами информации (значением термина или агрегата), определить количественную меру их соответствия, трактуемую в дальнейшем как результат вычисления отношения тождественности.

2. Обратная задача отождествления - на основании данного фрагмента информации и известных ограничений, накладываемых на количественную меру соответствия, сформировать новые фрагменты, такие, что при решений по отношению к ним и известному фрагменту прямой задачи, результат отождествления удовлетворял бы заданным ограничениям.

Сравнивая широту возможностей, предоставляемую для дальнейшего принятия решения, можно показать, что обе задачи с точки зрения результатов эквивалентны. Однако, такая эквивалентность существует лишь принципиально, а реализация решения как прямой так и обратной задачи имеет свои особенности, которые определяют их роль при проектировании систем. Эти особенности определяют технологию поиска или фильтрации, в процессе которых решается задача отождествления.

Для решения прямой задачи необходимо располагать двумя сопоставляемыми описаниями, в то время как только одно из них поступило извне. Это обстоятельство приводит к необходимости последовательного просмотра многих элементов данных, что весьма нетехнологично. В случае фильтрации это обстоятельство гармонично вписывается в схему обработки.

Решая обратную задачу с использованием технических средств, обеспечивающих возможность непосредственного доступа к конкретному элементу данных, можно существенно упростить процесс поиска. Это обстоятельство в большинстве случаев определяет технологические преимущества решения обратной задачи отождествления при организации поиска, хотя в некоторых специальных ситуациях решение прямой задачи отождествления является неизбежным.

По месту в общей технологической схеме обработки информации процедуры отождествления можно разделить на две группы:

Предкоординированные процедуры, обрабатывающие хранимую информацию в момент ее поступления в систему (при первоначальном накоплении или текущем пополнении) и обеспечивающие путем преобразования этой информации или специального размещения данных возможность использования при поиске упрощениях методов отождествления, вплоть до сравнения на полное совпадение. Наиболее выраженный характер «предкоординированности» имеет метод создания «регулярных выражений» (шаблонов, масок), предполагающий запись данных с использованием специальных служебных символов, позволяющих при отождествлении допускать установленные виды несовпадений (повторения, выпадения или появления нескольких символов, неразличения символов в пределах одного вида или группы и т.п.). Упомянутые словари программы TERRIER также могут быть подготовлены с учетом вариантности ключевых слов, являя пример предкоординированной процедуры.

Посткоординированные процедуры, реализующие отождествление информации в процессе поиска или фильтрации и представляющие возможность хранить отождествляемые данные в том виде в котором они поступают в систему.

Главное преимущество посткоординированных процедур состоит в том, что они не связывают массивы или потоки данных необходимостью использования специальных форм хранения или представления информации, что делает его более гибким и допускающим возможность адаптации при изменении внешних условий. Поэтому, при прочих равных условиях, предпочтительнее посткоординированные процедуры. Это не исключает возможностей специальной разметки данных аналогично методу «регулярных выражений», но здесь такая разметка производится уже в процессе отождествления. Например, при преобразовании иностранного текста с сохранением фонетического облика (транскрибировании) приходится вводить служебные символы, соответствующие факторам, влияющим на произношение (ударение, открытость слога, начало или конец слова и т.п.).

В процессе объединения (слияния) фрагментарных данных и при использовании критериев эффективности информационного поиска, отличных от традиционных коэффициентов полноты и точности, выбор типа процедуры может оказывать влияние на эффективность, и предпочтительней, могут оказаться предкоординированные процедуры.

Анализ различного алгоритмического содержания разработанных и применяемых методов решения задач отождествления позволяет выделить две обширные группы процедур:

- процедуры аппроксимации степени соответствия;
- процедуры вычисления производных (вторичных) форм представления данных.

Аппроксимация степени соответствия отождествляемых объектов основана на эвристическом выборе меры, характеризующей «сходство» этих объектов между собой. Такая мера должна быть относительно просто вычислима и может быть связана с вероятностью того, что сопоставляемые информационные фрагменты характеризуют единый реальный объект (без учета многозначности лексики) следующим регрессионным соотношением:

$P_X(z,d) = V(x(z,d))+\xi_X$

где x(z,d) - значение эмпирической меры сходства при отождествлении поискового предписания z с поисковым образом d (или их элементов);

V(x(z,d)) - оценка вероятности тождественности реальных объектов, которым соответствует информационные фрагменты z и d

ξ_X – случайная величина, с плотностью распределения $f_x(p)$, зависящей в общем случае от x(z,d).

Средние значения вероятностей ошибочного отождествления определяются из следующих соотношений:

$$P1 = \int_{x_\theta}^{1} \eta(x) \int_{\theta}^{1-V(x)} f_x(p)\, dx\, dp$$

$$P_2 = \int_{x_\theta}^{0} \eta(x) \int_{\theta}^{1-V(x)} f_x(p)\, dx\, dp$$

где P_1 и P_2- вероятности ошибок I и II рода;

Θ - пороговое значение оценки вероятности P_x с помощью которого принимается решение о результате отождествления по следующему правилу:

$P_x > \theta$ - объекты тождественны; $P_x < \theta$ - объекты различны

x_θ - значение эмпирической меры сходства, для которого справедливо: $V(x_\theta)= \Theta$

η(x) - плотность распределения значений эмпирической меры сходства, определяемая частотами встречаемости различных пар информационных фрагментов z и d.

Таким образом, количественные характеристики ошибок отождествления зависят от выбора следующих характеристик:

- вида эмпирической меры «сходства»;
- вида функции V(x(z,d));
- значения порога Θ .

Можно показать, что для любой фиксированной эмпирической меры «сходства» обеспечивается наилучшее соотношение характеристик ошибок, если V(x(z,d)) выбрана таким образом, чтобы оценка P_x была несмещенной, т.е. выполнялось следующее условие:

$$\int_{-\infty}^{+\infty} p\, f_x(p)\, dp = 0$$

Используемые для решения практических задач эмпирические меры «сходства» весьма разнообразны, В ряде работ предлагается следующая классификация таких мер:

- коэффициенты связи, представляющие собой некоторые алгебраические функции, аргументами которых являются количества совпавших и несовпавших символов;
- коэффициенты корреляции, учитывающие взаимную связь различных символов в различных положениях (эти коэффициенты по своей природе являются статистическими);
- коэффициенты, выполняющие функцию расстояния, т.е. величины, для которых справедливы аксиомы, определяющие метрику.

В почтовых автоматизированных системах для отождествления значений некоторых реквизитов почтового отправления используется алгоритм аппроксимации значения меры соответствия, а в качестве эмпирической меры сходства выбран коэффициент связи. С помощью этого алгоритма вычисляется, какую долю составляет количество не совпавших символов по отношению к длине значения реквизита.

Другим примером использования коэффициента связи является расстояние Левенштейна (редакционное расстояние или дистанция редактирования) между двумя строками - это минимальное количество операций вставки одного символа, удаления одного символа и замены одного символа на другой, необходимых для превращения одной строки в другую.

Развитием является расстояние Дамерау-Левенштейна - мера разницы двух строк символов, определяемая как минимальное количество операций вставки, удаления, замены и транспозиции (перестановки двух соседних символов), необходимых для перевода одной строки в другую. К операциям

вставки, удаления и замены символов, определенных в расстоянии Левенштейна, добавлена операция транспозиции (перестановки) символов.

Следует отметить, что процедуры, аппроксимирующие значение меры соответствия, принципиально являются посткоординированными и могут быть использованы только при решении прямой задачи отождествления.

Вычисление производных (вторичных) форм представления данных основано на определении значения словарной функции, которая ставит в соответствие каждому значению признака некоторый символьный код (или группу кодов), являющийся производной формой представления данного значения. Словарная функция подбирается таким образом, чтобы определение степени соответствия между двумя отождествляемыми информационными фрагментами сводилось к вычислению весьма простой функции от значений соответствующих кодов, например, арифметической разности или бинарной функции, принимающей значение 1 в случае полного совпадения кодов и 0 в противном случае. Среди множества разработанных и практически используемых алгоритмов вычисления производных форм представления данных можно выделить следующие группы:

- алгоритмы «хеширования» (как правило, некриптографические, различного вида свертки) и рандомизирующие алгоритмы, позволяющие «сгладить» распределения результатов искажающих факторов в различных элементах информации.
- алгоритмы выделения фрагментов информации, наименее подверженных воздействию искажающих факторов и объединения наиболее искажающихся элементов в группы, характеризующие значительным превышением, вероятностей перехода элементов друг в друга внутри группы по отношению к вероятностям перехода в элементы за пределами группы.

Авторы одной из ранних DLP-систем включили в состав средств отождествления информации специальную рандомизирующую процедуру, использующую равномерно распределенные псевдослучайные числа. Каждое из таких чисел является элементом матрицы W=| Wij|, строки которой соответствуют символу русского алфавита, а столбцы - номеру символа в слове. Каждое слово, в соответствии с этой матрицей преобразуется в последовательность чисел:

$Wi_1 1, Wi_2 2, \ldots Wi_l l$

где l - длина слова

Полученная последовательность рассматривается как совокупность реализаций случайной величины, характеризующейся некоторым распределением. На этой выборке определяется оценка математического ожидания, которая является аргументом функции, определяющей степень соответствия отождествляемых объектов.

Наиболее распространенными в настоящее время в практике функционирования информационных систем является процедуры, в основе которых лежат принципы выделения более искажающихся и менее искажающихся элементов и исключения наиболее неустойчивых в смысле подверженности искажениям фрагментов данных и объединения их в группы [2]. В частности, такие процедуры нашли применение в системах обработки персональных данных, где используются два класса правил преобразований, которым подвергается информация (в основном имена собственные), в процессе отождествления:

- преобразование одних букв в другие, трудно произносимых буквосочетаний, или сочетаний из двух и более букв в более простые или одну букву;
- преобразование сложных и многословных имен собственных, преобразование отдельных их частей, слияние значений имен собственных.

Кроме того используются алгоритмы вычисления вторичных форм представления обеспечивающие выделение относительно устойчивых по отношению к воздействию искажений структур. В распространенной кадровой системе «Босс-кадровик» для формирования кадровых досье был применен алгоритм отождествления искаженной информации, основанный на вычислении фонетического кода. При формировании этого кода наиболее искажающиеся символы (гласные) исключаются, остальные объединяются в группы.

Фонетический код представляет собой четырехсимвольную последовательность из одной буквы (первая буква слова) и трех цифр - номеров соответствующих групп согласных (или нулей). Отождествленными являются те значения, у которых совпадает цифровал часть фонетического кода, а первые буквы относятся к одной группе. В ряде систем функционируют аналогичные алгоритмы, с помощью которых значению признака ставится в соответствие некоторой цифровой код (характеристика фамилии), при этом значения «похожие» в некотором эвристическом смысле, порождают одинаковые коды.

Алгоритм отождествления ключевых слов профилей контентых фильтров, предполагает формирование нескольких вариантов значения признака при загрузке информации. На этом этапе выделяются первые шесть символов и из них составляются шесть пятисимвольных вариантов путем исключения одного из символов. При отождествлении значение признака, поступившего в запросе, также преобразуется в шесть (или менее, в зависимости от длины значения) вариантов и проводится поиск на полное совпадение в массиве вариантов, сформированных при загрузке. Можно показать, что порождаемые множества пятисимвольных вариантов имеют непустое пересечение только в случае, когда расстояние Дамерау-

Левенштейна для выделенных шестисимвольных начал не превышает единицу, т.е искажение «локализовано» в пределах одной позиции.

Все чаще предлагаются алгоритмы решения прямой задачи отождествления с использованием развитых средств представления данных о характере и свойствах искажающих воздействий. Оба отождествляемых значения разбиваются с помощью специальных правил на элементы (контексты), которые наряду с информацией о составе входящих в них символов содержат данные о лингвистических особенностях положения, занимаемого контекстом в слове. Степени соответствия определяется как функция экспериментально измеряемых вероятностных характеристик взаимных переходов контекстов при искажениях, эта функция выбирается таким образом, чтобы минимизировать сумму ошибок первого и второго рода.

Кроме перечисленных алгоритмов отождествления следует указать на широкий спектр методов, который становятся доступен для исследования и практического применения после сведения задач отождествления в форму задач распознавания образов.

Существуют различные точки зрения на возможность применения моделей и методов распознавания образов для решения проблем отождествления. Так в [3] указывается на вероятную несостоятельность аналогий между распознаванием образов и отождествлением в связи с принципиальной невозможностью априорного перечисления и описания классов распознаваемых объектов. С другой стороны, там же говорится о том, что «условия отождествления в полной мере соответствуют исходным посылкам постановки задачи статистического распознавания образов». Трудности, связанные с отсутствием априорных описаний классов, предлагается преодолеть путем введения двуальтернативного разбиения множества всевозможных пар значений отождествляемых данных на те, которые следует считать достаточно соответствующими и те, которые не являются таковыми.

Анализ этих экстремальных концепций позволил выработать некоторую компромиссную точку зрения. Действительно, отсутствие априорных описаний классов, казалось бы, исключает возможность применения основных результатов распознавания образов. Идея разбиения множества пар значений представляется неконструктивной из-за чрезвычайно громоздкой модели. Непосредственное использование методов таксономии и кластерного анализа нецелесообразно, т.к. отождествление направлено не на определение соответствия некоторому внешнему и априорно неизвестному объекту, которым является поступающая информация, а на выявление «близости» объектов распознавания, которыми в данном случае являются элементы данных, друг другу. Однако, как только объект становится известным, классы распознавания вполне определяются, что делает возможной постановку задачи распознавания образов, а выбор метода для ее решения определяется видом сформированных классов. Таким образом, широта аналогий между отождествлением и распознаванием образов ограничивается множеством методов распознавания, инвариантных по отношению к виду распознаваемых классов.

Отождествление информации представляет собой основной элемент функциональной структуры комплексов контентной фильтрации и DLP-систем, определяющий достоверность результатов их работы. Тем не менее, изложенное выше позволяет предполагать, что процедуры отождествления еще далеки от совершенства и имеют ощутимый потенциал развития, по крайней мере, в связи со следующими обстоятельствами:

- мало используются данные о статистических свойствах информационной среды, в которой проявляются искажения и другие факторы неопределенности;
- при разработке средств отождествления применяются недостаточно адекватные представления о характере и свойствах проявления факторов неопределенности;
- использование эвристических приемов при создании средств отождествления не позволяет обеспечивать оптимальные режимы обработки данных;

Поэтому в качестве вывода следует указать на целесообразность постановки и решения следующих задач:

- задачи построения моделей искажений путем глубокой идентификации фактора неопределенности на основании результатов наблюдения его проявлений и с учетом того, что получить использовать априорные сведений о внутренней структуре и характеристиках этого фактора практически крайне трудно;
- задачи разработки принципов оптимизации процедур автоматического отождествления информации в комплексах контентной фильтрации и DLP-системах.

Литература

1. *Толпегин П.В.* Информационные технологии анализа русских естественно-языковых текстов – М., Российская академия наук вычислительный центр им. А.А. Дородницына, 2006;
2. *Воронов Э.С* Методы и алгоритмы анализа естественно-языковых сообщении. VIII Всероссийская научно-практическая конференция «Технологии Microsoft в теории и практике программирования». Национальный исследовательский Томский политехнический университет – Томск, 2007
3. *Фу К.* Лингвистический подход к распознаванию образов. В сб. «Классификация и кластер», М., Мир,

1980