Разработка методов и средств контроля достоверности и актуальности фактографического наполнения информационных систем

© А.С. Серый
Институт Систем Информатики им. А.П. Ершова СО РАН
Новосибирск
Alexey.Seryj@iis.nsk.su

Аннотация

В данном исследовании представлены методы и подходы к автоматизации обработки входящего потока данных в информационной системе, где информация — это множество информационных объектов, соответствующих понятиям и отношениям онтологии предметной области системы. Решаются задачи поиска референциальных связей и идентификация объектов. Кроме того, предлагаются методы получения из трастовых метрик (trust metrics) информационных ресурсов соответствующих метрик для извлекаемых документов и той информации, которая заключена в извлекаемых из документов объектах. Предполагается, что такой подход позволит обеспечить удаление утративших доверие данных, тем самым, снизив долю участия эксперта в процессе проверки информации и уменьшив количество ошибок и противоречий в системе.

1 Введение

В современном мире информационные системы занимают довольно обширную нишу. Глобальная Сеть непрерывно пополняется новой информацией: как текстовой, так и мультимедийной. Пользователю все труднее становится найти то, что было бы для него полезно. Отсюда и появление многочисленных поисковых сервисов, информационных порталов, а также систем, аккумулирующих информацию, относящуюся к некоторой ограниченной области знаний или предметной области.

Результаты исследований двух последних десятилетий привели к активному использованию онтологий в качестве концептуальных схем реляционных баз данных, лежащих в основе информационных систем [2]. В контексте данной работы понятия онтологии и концептуальной схемы используются как равнозначные, онтология предметной области задается в виде базовых понятий, организованных в таксономию, и совокупности связей между ними. Данные, при этом, представляются в виде множества разнотипных информационных объектов — экземпляров понятий и отношений онтологии. В совокупности объекты образуют контент или информационное наполнение системы. Каждый объект определяется понятием или отношением онтологии и, являясь экземпляром класса, имеет заданную им структуру.

Информационная система должна отображать изменения, происходящие в ее предметной области. Очевидно, что накапливаемые в системе факты (свойства или утверждения об объектах) могут оказаться неверными, противоречивыми или некорректными. Поддержание контента в актуальном состоянии повышает эффективность исполнения системой своих функций, позволяет менее расточительно использовать компьютерные ресурсы и снижает вероятность возникновения ошибок.

В данной работе предложены метод автоматической обработки входящего потока информационных объектов и метод оценки достоверности данных в информационной системе.

Входящим потоком данных в нашем случае считается множество информационных объектов, источником которых являются текстовые документы. Основными задачами данного этапа являются поиск референциальных связей объектов и разрешение контекстной омонимии (или идентификация) объектов. Контекстная омонимия зачастую сопровождает процесс автоматической обработки текстов на естественном языке и проявляется в наличии двух и более вариантов отождествления полученных из текста объектов с объектами базы данных информационной системы. Разрешение референции предполагает поиск кореферентных информационных объектов, т.е.

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

описывающих одну и ту же внеязыковую сущность предметной области, называемую референтом. Разработано множество методов поиска референциальных связей языковых выражений в текстах, но, в силу сложности подобных задач вообще и для русскоязычных текстов в частности, они не всегда решаются целиком. Не охваченные в процессе обработки текста случаи могут послужить причиной появления информационных объектов, собранных на основе кореферентных выражений. Наличие подобных объектов во входящем потоке данных нежелательно, т.к. снижает точность идентификации. Разработанный для решения этой задачи подход позволяет абстрагироваться от технологии обработки текста, лишь налагая на формат объектов некоторые требования, определяемые способом описания онтологии предметной области [3].

Задача поддержания актуальности данных ставится разработчиками информационных систем повсеместно, однако используемые методы могут сильно различаться. Универсальным методом можно назвать периодическую сверку с источником. В большей степени это относится к информационно-справочным системам. Перепись населения страны также можно назвать сверкой с источником и на примере такой переписи можно увидеть, что процедура перепроверки данных может быть весьма продолжительной и дорогостоящей; более того, перепроверка данных возможна не всегда. Предлагаемый в данной работе метод позволяет оценить достоверность данных в информационной системе, спроектированной на основе онтологии, отслеживать его изменения и удалять информацию, которой более нельзя доверять.

2 Поиск кореферентных объектов

Поиск кореферентных объектов рассматривается как подготовительный этап процедуры идентификации и включает в себя установление степени сходства объектов, построение множества гипотетических эквивалентов для каждого объекта и объединение кореферентных объектов. Подход, разработанный для решения этой задачи, опирается на результаты работы группы исследователей из университета Стэнфорда [4] по разрешению референции между языковыми выражениями в текстах на английском языке, а также на исследования компании RCO [5] закономерностей использования референции при построении связных предложений на русском языке. Кроме того, концепты предметной области и их экземпляры, представленные в системе, должны удовлетворять ограничениям, описанным в [6].

Разрешение кореферентности объектов представляет собой итерационный процесс, где одной итерации соответствует единичный проход по множеству входных объектов и проверка каждого из них на наличие эквивалента — ближайшего кореферентного объекта. В случае обнаружения для объекта q эквивалентного объекта q' они объединяются в кластер, который в дальнейшем интерпретируется как единый объект q''. В рамках одной итерации для каждого объекта q выполняются действия, описанные в п. 2.1–2.3.

2.1 Вычисление степени сходства q со всеми объектами из его окрестности

Для сравнения объектов вводится коэффициент сходства $SI(q^1, q^2)$ (similarity index), где q^1 и q^2— сравниваемые объекты.

$$SI(q^1, q^2) = \begin{cases} k \cdot SI_C + (1-k) \cdot SI_L;\ SI_c \neq 0 \\ 0;\ SI_c = 0 \end{cases} \quad (1)$$

$$0 \leq k \leq 1$$

Операция вычисления SI не является коммутативной, поэтому будем говорить, что вычисляется степень сходства объекта q^2 с объектом q^1. Объект q^1 при этом называется эталоном, а q^2 — кандидатом. $SI_C = SI_C(q^1, q^2)$ называется *таксономической близостью* объектов q^1 и q^2 и зависит от взаимного расположения соответствующих им классов онтологии в ее иерархическом древе, $SI_L = SI_L(q^1, q^2)$ характеризует близость наборов свойств: атрибутов и связей. Коэффициент k регулирует уровень влияния онтологических и атрибутивно-реляционных факторов на итоговую величину. Его значение определяется экспериментальным путем и может изменяться в зависимости от задачи. Формулы для вычисления выражения (1) и его подвыражений подробно описаны в [6].

2.2 Построение множества потенциальных эквивалентов объекта q

Множество потенциальных эквивалентов объекта q состоит из всех объектов q', удовлетворяющих условиям (2).

$$Pr(q) = \{q' \in Ctx(q) | SI(q', q) > \alpha > 0\} \quad (2)$$

Здесь $Ctx(q)$— это некоторая окрестность объекта q в списке объектов (изначально объекты упорядочены по встречаемости в источнике). Размер окрестности определяется исходя из правил и экспериментальных наблюдений, в частности описанных в компании RCO [5, 6].

2.3 Выбор эквивалента для q из множества его потенциальных эквивалентов

Эквивалентом объекта q считается ближайший к нему объект q' из множества $Pr(q)$ с максимальным либо близким к максимальному значением $SI(q', q)$. Если таковой отсутствует или не является предшествующим объекту q, то говорим, что q упомянут в тексте впервые. Объект, состоящий в кластере и не имеющий эквивалента, т.е. соответствующий самому первому упоминанию, будем называть его глобальным эквивалентом или G-эквивалентом. В случае невозможности выделить единственный эквивалент, говорим, что объект q не имеет эквивалента.

2.4 Условия остановки и результат

Итерации следует повторять до тех пор, пока существует возможность строить новые кластеры

или пополнять уже существующие. Первая итерация, не принесшая новых данных, считается завершающей.

Интерпретация кластеров как обычных объектов позволяет на каждом шаге процесса в полной мере использовать информацию о референциальных связях объектов, добытую на предыдущих шагах. За счет интеграции внутри кластера информации обо всех содержащихся в нем объектах такой подход повышает эффективность всего процесса в целом.

Отношение кореферентности объектов, обозначим его $\Re$, очевидно, является отношением эквивалентности. Множество объектов Q разбивается, таким образом, на непересекающиеся кластеры, представляющие собой классы эквивалентности по отношению $\Re$, а после объединения кореферентных объектов мощность Q совпадет с мощностью соответствующего фактор-множества $Q/\Re$. Ясно, что $|Q/\Re| \le |Q|$. Снижение количества кореферентных объектов призвано повысить эффективность следующего этапа — идентификации.

3 Идентификация информационных объектов

Идентификация заключается в разрешении контекстной омонимии входных объектов, когда входному объекту по его набору атрибутов можно сопоставить несколько объектов из базы данных.

Предлагаемый подход предполагает наличие «стартового» списка идентифицированных объектов, который, может быть получен с помощью процедуры поиска по точному совпадению минимального набора атрибутов, определяющих объект. Если был найден лишь один объект, то входной объект считается идентифицированным, и дальнейший его анализ уже не требуется. В итоге множество информационных объектов разделяется на множество *A* идентифицированных объектов и множество *B*, куда входят те объекты, которые не удалось идентифицировать «сходу». Каждому объекту $q \in B$ сопоставляется множество Q объектов из базы данных — множество похожих объектов. Строится оно путем сравнения *q* с объектами базы данных по различным подмножествам атрибутов. Алгоритм построения множества похожих объектов представлен в листинге 1. Здесь `Intersect(q, i)` возвращает объекты, совпадающие с *q* не менее чем по *i* атрибутам.

```
algorithm ПОИСК ПОХОЖИХ ОБЪЕКТОВ
var q = {a_k|k = 1,…, n} %объект,
включающий n атрибутов
Q результирующее множество объектов
T, i вспомогательные переменные
begin
Q ←∅
i ← 1
while i ≤ n
T ← Intersect(q, i)
if T = ∅ then
return Q
Q ← T
i ←i + 1
end while
return Q
end algorithm
```

Листинг 1. Поиск похожих объектов

Результатом работы алгоритма будет множество Q объектов, совпадающих с заданным объектом *q* по максимальному количеству атрибутов.

Для того чтобы идентифицировать объект, необходимо сузить множество похожих объектов до одного, т.е. снять неопределенность. Алгоритм идентификации описан в листинге 2 (подробнее см. [7]). В описании алгоритма присутствуют следующие вспомогательные функции:

- `Active(q)` возвращает **`true`**, если `q` активен, **`false`** – иначе
- `Activate(q)` присваивает объекту активный статус
- `Deactivate(q)` присваивает объекту неактивный статус
- `Move_Object(q, Q)` переносит объект `q` во м-во `Q`
- `Move_Rel(r, R)` переносит отношение `r` во м-во `R`
- `Filter(Q, R, i)` удаляет из м-ва `Q` объекты, имеющие не более `i` отношений, аналогичных отношениям из `R`.

```
algorithm ИДЕНТИФИКАЦИЯ ОБЪЕКТОВ
var A множество идентифицированных
объектов
    B множество неидентифицированных
    объектов
    F^A, F^B, D^A, D^B, S^b, i
    вспомогательные переменные
begin
 A ← стартовое м-во
 идентифицированных объектов
 B ← м-во неидентифицированных
 объектов
 while B ≠ ∅
  Choose b ∈ B: Active(b) = true
  if ∄b then
   return A
  F^A← связи b с объектами из A
  F^B← связи b с объектами из B
  S^b← ПОИСК_ПОХОЖИХ_ОБЪЕКТОВ(b)
  i ← 1
  while i ≤ |F^I|& S^b ≠ ∅
   Filter(S^b, F^I, i)
   if ∃!q∈S^b then
    Move_Object(b, A) %q – объект БД,
    эквивалентный b
    ∀d∈B: ∃r(b,d) %r – связь объектов
```

```
    d и b
     D^A← связи d с объектами из A
     D^B← связи d с объектами из B
     ∀r ∈ D^B: r(b, d)
      Move_Rel(r, D^A)
     if Active(d) = false then
      Activate(d)
    ВЫХОД ИЗ ЦИКЛА
   i ← i + 1
  end while
  if b ∉ A then
   Deactivate(b)
 end while
 return A
end algorithm
```

Листинг 2. Идентификация объектов

Фактически, для b∈B мы находим непустое подмножество множества S^b, такое, что его элементы имеют наибольшее число связей, аналогичных связям объекта b. Контекстная омонимия для объекта b снимается, если это подмножество содержит единственный элемент.

4 Достоверность как показатель доверия к информации

Для того чтобы оценить полезность факта для информационной системы, необходимо определить его трастовую метрику или *достоверность*. Фактом, в нашем случае, называется минимальное знание об объекте, другими словами это либо значение атрибута объекта, либо его связь с другим объектом. Достоверность (trustworthiness) определяет, до какой степени может доверять данному факту рядовой пользователь информационной системы. Для оценки используются характеристики источников факта, и учитывается время его существования в информационной системе. Данные характеристики описаны ниже.

Пусть F — некоторый факт, d^i — i-й документ, упоминающий F. Обозначим экспертную оценку документа d^i как $x^i \in [-1;1]$. Экспертная оценка характеризует уровень доверия эксперта к информации из документа d^i на основании знаний об источнике этого документа и, возможно, какой-то дополнительной информации, которой располагает эксперт. Границы интервала, в которых заключено значение экспертной оценки, соответствуют предельным случаям: полному доверию при $x^i = 1$ и, соответственно, полному недоверию при $x^i = -1$. Значение $x^i = 0$ соответствует отсутствию информации об источнике у эксперта. Значения по умолчанию в случае отсутствия экспертной оценки x^i вычисляются по формуле $x^i = \frac{N-1}{N}$, где N— количество различных источников, содержащих документ d^i.

Введем характеристику источника, выражающую вероятность получения из него достоверного знания. Она должна быть непрерывным образом связана с экспертной оценкой. Семейство функций вида **(3)** очевидно обладает требуемыми свойствами.

$$f_{\mathfrak{M}}(x) = \left(\frac{x+1}{2}\right)^{\mathfrak{M}} \mathfrak{M} = 1, 2, \ldots \quad \textbf{(3)}$$

Допустим, что нам известно среднее значение допущенных при извлечении фактов ошибок. Пусть φ — это среднее отношение допущенных ошибок к общему числу извлеченных фактов. В простейшем случае мы считаем φ константной величиной, но в общем виде ничто не мешает обозначить ее $\varphi(t)$ и вычислять как функцию от некоторого аргумента. В дальнейшем для определенности будем считать $\varphi = const$.

С помощью функций $f_{\mathfrak{M}}(x)$ и параметра φ породим семейство вероятностных характеристик документа d^i.

$$\boldsymbol{\delta}^{\boldsymbol{i}}_{\mathfrak{M}} = \varphi\left(\frac{x^i+1}{2}\right)^{\mathfrak{M}} \quad \textbf{(4)}$$

Далее, если не указано обратное, значение $\mathfrak{M}$ будет считаться равным единице, поэтому нижний индекс будет опускаться. Если значения x^i вычислялись по умолчанию, то $\delta^i = \varphi\left(1 - \frac{1}{2N}\right)$. При $N = 1$ в системе представлен только один источник документа d^i, и отсутствует какая-либо информация о его свойствах. Занижение значений δ^i по причине неполноты знаний об источниках документов, очевидно, повлияет и на достоверность фактов, в частности, ускоряя потерю актуальности. В предельно неблагоприятном случае (при $N = 1$), получим $\delta^i = \frac{\varphi}{2}$, вместо $\delta^i \approx \varphi$ (при $N \gg 1$).

Информация может со временем стать менее актуальной и, соответственно, менее заслуживающей доверия пользователя. Косвенным признаком утери актуальности факта является длительное отсутствие упоминаний факта в новых документах. Введем следующую функцию $h(t)$, зависящую от времени.

$$h(t) = \frac{1}{1+\ln(\frac{t}{\mathcal{M}}+1)}, \mathcal{M} = const \quad \textbf{(5)}$$

Будем называть $h(t)$ *темпоральным множителем*. Здесь $\mathcal{M}$ — это время, за которое значение достоверности понизится в l раз, из формулы (5) следует, что $l = 1 + ln2 \approx 1.69$. Таким образом, величина $\tau = \frac{t}{\mathcal{M}}$ — это безразмерное время, равное отношению времени существования факта в системе на время, необходимое для понижения его достоверности в l раз. Значение $\mathcal{M}$ подбирается исходя из оценки экспертом скорости устаревания фактов в данной предметной области.

За основу модели достоверности факта был взят неоднородный дискретный марковский случайный процесс, имеющий три состояния $\{E_1, E_2, E_3\}$, определяющих текущий уровень доверия к факту: «недоверие», «неопределенность» и «доверие» соответственно. Вероятность пребывания в третьем состоянии — это вероятность того, что факт заслуживает доверия. Остальные две вероятности

имеют вспомогательный характер. Моментами времени процесса считаем поступление очередного подтверждения факта, т.е. нового документа, упоминающего факт, X_n — случайная величина, выражающая состояние факта в момент времени *n*.

Обозначим через $\bar{\pi}^n$ вектор-строку распределения случайной величины X_n. Начальное распределение процесса задается предварительно: $\bar{\pi}^0 = (\pi_1^0, \pi_2^0, \pi_3^0), \pi_i^0 = P(X_0 = E_i)$. После n шагов вектор $\bar{\pi}^0$ переходит в $\bar{\pi}^n = \bar{\pi}^0 \cdot \mathbb{P}^{(n)}$, где $\mathbb{P}^{(n)} = (p_{jk}(0,n))$ — матрица перехода за n шагов, элементы которой вычисляются при помощи тождества Колмогорова-Чепмена, выполняющегося для любого $m < r < n$, в том числе для $r = n - 1$.

$$p_{jk}(m,n) = \sum_v p_{jv}(m,r)p_{vk}(r,n) \qquad \textbf{(6)}$$

Вероятности вида $p_{vk}(n, n+1)$ определяются матрицей перехода за один шаг или *переходной матрицей*, обозначаемой $\mathbb{P}(n, n+1)$.

$$\mathbb{P}(n,n+1) = \begin{bmatrix} 1-\delta^{n+1} & 0 & \delta^{n+1} \\ 1-\delta^{n+1} & 0 & \delta^{n+1} \\ \frac{1-p_{33}}{2} & \frac{1-p_{33}}{2} & p_{33} \end{bmatrix} \qquad \textbf{(7)}$$

$$p_{33} = \frac{1}{2}(p_{23}(0,n) + \delta^{n+1})$$

Для учета влияния времени существования факта в системе на вектор распределения было построено семейство линейных операторов, обозначаемое $\mathbf{T_t}$. Один из операторов семейства $\mathbf{T_t}$ применяется к вектору распределения, являющемуся результатом последнего, на тот момент, шага случайного процесса. Кроме того, этот же вектор, не претерпевший никаких темпоральных изменений, вовлекается в следующий шаг процесса, при условии, что факт не устарел и не был исключен из системы за прошедшее время. Как любое линейное преобразование операторы $\mathbf{T_t}$ можно записать в виде матрицы:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1-h(t) & 0 & h(t) \end{bmatrix} \qquad \textbf{(8)}$$

4.1 Калибровка параметров

Для проверки и калибровки параметров была проведена серия экспериментов. Напомним, что значения координат вектора распределения $\bar{\pi}$ зависят от двух параметров. Это показатель степени $\mathfrak{M}$ семейства функций (3) величина $\mathcal{M}$ из формулы вычисления темпорального множителя (5).

В общем виде зависимость распределения по времени представляет собой множество точек в четырехмерном евклидовом пространстве. Для удобства визуализации разобьем его на три двумерных зависимости — зависимость каждой из координат вектора распределения по времени.

Эксперимент показал, что при увеличении параметра $\mathfrak{M}$ вектор $\bar{\pi}$ сильнее реагирует на появление ненадежных источников с низкой экспертной оценкой. Однако, нежелательно, чтобы один отдельно взятый источник оказывал сильное влияние на $\bar{\pi}$, т.к. в этом случае ошибка при оценке источника может значительно исказить течение процесса. С учетом этих соображений оптимальным для $\mathfrak{M}$ очевидно является значение $\mathfrak{M} = 1$.

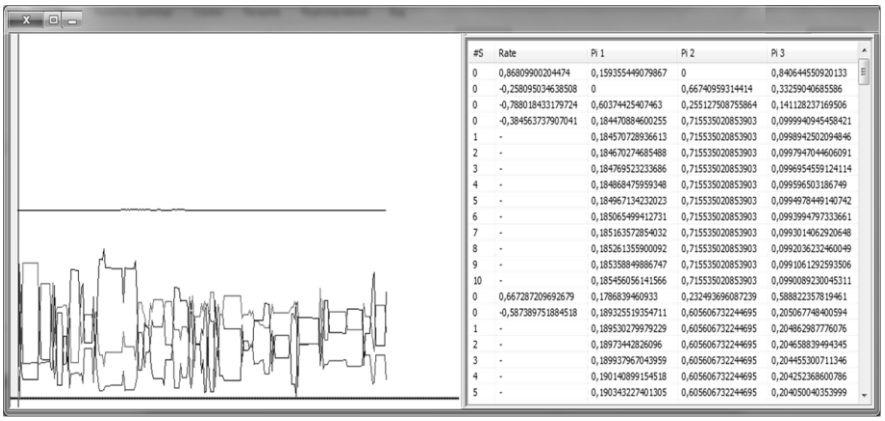

Рис. 1 Экспериментальная калибровка параметров

По результатам экспериментов установлены следующие значения параметров: $\mathfrak{M} = 1$, $\mathcal{M} = 1000$ (предполагается, что время измеряется в сутках). На Рис. 1 показаны графики координат вектора $\bar{\pi}$ для случайного процесса длительностью 200 и вышеуказанными значениями параметров. В силу дискретности процесса каждый график представляет собой множество точек, поэтому для наглядности точки соответствующих графиков были соединены между собой отрезками, сформировав, таким образом, ломаные линии.

Параметр $\mathcal{M}$ не оказывает влияния на случайный процесс, его значение выбирается исходя из предпочитаемой скорости убывания актуальности факта во времени. Для различных видов фактов зависимость от времени, а, следовательно, и значение $\mathcal{M}$, может быть задано индивидуально.

4.2 Критерии удаления ненадежных данных

Изменение достоверности факта *F* в информационной системе описывается цепочкой пар $< j, \pi_j^F >$, где j — момент времени, π_j^F — достоверность в момент времени *j*, т.е. дискретным множеством. Принятие решения относительно факта на основе только текущего значения достоверности

неэффективно вследствие того, что достоверность может опуститься ниже минимально допустимого значения в случае погрешности при оценке, низкого авторитета выбранного источника и других возмущающих факторов, для уменьшения степени влияния подобных возмущений необходимо анализировать окрестность текущей точки. Анализ дискретных окрестностей также оказался неэффективен, т.к. не позволяет принять решение в случае колебаний достоверности вокруг среднего значения. В этом случае мы можем усреднить и оценить значения в промежуточных точках, аппроксимируя или интерполируя имеющееся множество точек, либо его подмножество, гладкой кривой. Отдельные сегменты кривой позволяют оценить уровень доверия в заданной окрестности текущего момента времени без учета влияния более ранней истории. Для решения этой задачи был проведен анализ различных методов аппроксимации и интерполяции кривыми и по его итогам выбран метод аппроксимации B-сплайном [8, 9].

Чтобы принять решение о дальнейшей судьбе факта, доверие к которому на текущий момент времени опустилось ниже минимально допустимого, предлагается выделять общую тенденцию поведения, основываясь при этом на анализе кривой аппроксимации хвоста из нескольких значений. С учетом соображений о вычислительной сложности было решено строить сплайн четвертого порядка без кратных вершин, при этом число опорных точек должно быть не меньше четырех, что и определило минимальный размер хвоста. Выбор хвоста минимально возможной длины для экспериментов обусловлен, во-первых, соображениями производительности, во-вторых — относительно малым количеством возможных видов кривых, и, кроме того, если вероятность погрешности при оценке одного значения достоверности равна p, то вероятность погрешности на четырех значениях составит p^4, что дает нам ~6% вероятности погрешности при $p = \frac{1}{2}$.

Какие тенденции может описывать кривая? Это может быть тенденция к убыванию, что говорит в пользу исключения факта, либо тенденция к возрастанию. Также кривая может не иметь выраженной тенденции. Введенная нумерация кривых отражает их форму и записывается в виде ***a.b.c,*** где $a \in A = \{1,2\}; b \in B = \{1,2,3,4,5,6\}; c \in C = \{1,2\}$.

A:
1. кривая имеет точку перегиба
2. кривая не имеет перегибов

B:
1. кривая возрастает
2. кривая убывает
3. кривая возрастает, а затем убывает
4. кривая убывает, а затем возрастает
5. кривая убывает, возрастает, затем снова убывает
6. кривая возрастает, убывает, затем снова возрастает

C:
1. кривая выпукла вниз в начальной точке
2. кривая выпукла вверх в нач. точке

Согласно такой нумерации кривая с номером, например, **1.2.1** — это кривая, имеющая точку перегиба, выпуклая вниз в некоторой окрестности начальной точки и убывающая на всей области определения. Хотя таким образом можно пронумеровать 24 кривые, всего их 16. Это обусловлено выбранным числом опорных точек и порядком кривой и подтверждено экспериментальной проверкой выборки из ~10 миллионов хвостов, сгенерированных 500 тысячами случайных процессов. Для нас интересны, в первую очередь кривые вида ***.1.*** и ***.2.***, поскольку они показывают общую тенденцию.

Строго убывающая кривая может уничтожить факт, строго возрастающая — предотвратить его удаление. Пусть *MIN* — минимально допустимый уровень доверия, при котором на факт еще можно положиться без какой-либо дополнительной проверки, F — некоторый факт, $E^j = \{e_1^j, e_2^j, e_3^j, e_4^j\}$ — j-й хвост соответствующего ему случайного процесса, e_k^j это последние четыре значения достоверности F. Рассмотрим граничные случаи.

Допустим, что $e_4^j < MIN;\ e_3^j \geq MIN$. Это первый граничный случай, при этом факт удаляется, если кривая выражает тенденцию к убыванию, т.е. имеет вид ***.2.***. Если кривая имеет какой-либо другой вид, факт остается в системе. Точно таким же образом разрешаются промежуточные случаи, когда $e_4^j < MIN;\ e_3^j < MIN, e_2^j \geq MIN$ или $e_4^j < MIN;\ e_3^j < MIN, e_2^j < MIN, e_1^j \geq MIN$. Второй граничный случай наступает, когда $e_k^j < MIN$ $(k = 1,2,3,4)$. Все значения находятся ниже минимального порога. Здесь на первый план выходят кривые вида ***.1.***. Соответственно, если кривая строго возрастает, то факт F все равно *не* будет удален. Если кривая имеет вид, отличный от указанного, то факт исключается из информационной системы как утративший доверие.

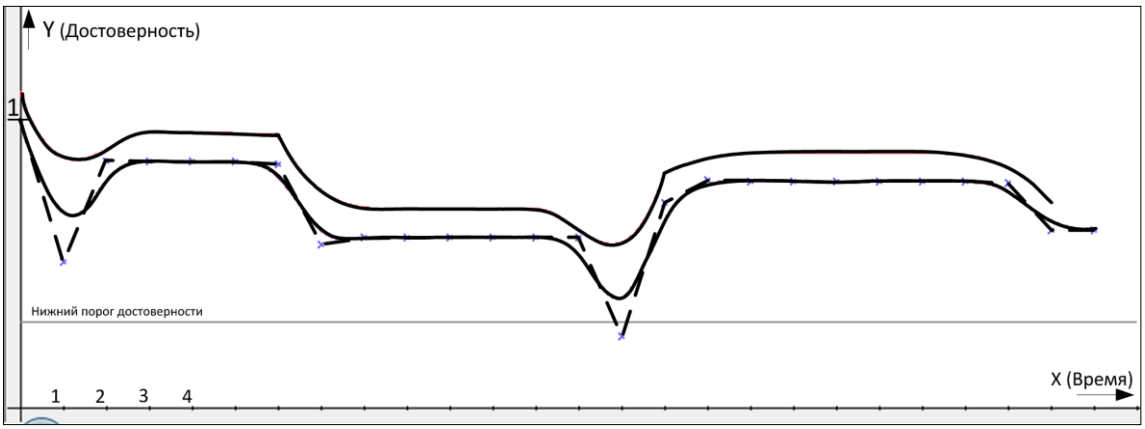

Рис. 2 Различия полной и кусочной аппроксимации

Новые хвосты могут вычисляться со смещением в одну точку. На Рис. 2 показана полная аппроксимирующая кривая в сравнении с кривой, составленной из отдельных хвостов [8, 10]. Пунктиром обозначена ломаная линия, проведенная через опорные точки. Для наглядности сегменты второй кривой построены со смещением в три точки по оси абсцисс, чтобы избежать наложения частей кривых друг на друга, и со смещением 0.1 по оси ординат (единица по шкале О*y* относится к единице шкалы О*x* как 20/3).

5 Заключение

Предлагаемые методы и подходы призваны обеспечить автоматическую обработку входящего потока данных и смоделировать изменение достоверности фактов в информационной системе, предметная область которой формально описана с помощью онтологии, а также описать механизм удаления ненадежной информации из системы.

Поиск референциальных связей между объектами и идентификация помогают отсеять нежелательные данные при пополнении системы, в то время как контроль достоверности и механизм отсеивания утративших доверие фактов способствуют сохранению целостности и непротиворечивости информации, прошедшей эту проверку. В особенности это касается фактов, требующих регулярного подтверждения. Примером таких фактов являются статьи кодексов (уголовного, гражданского, налогового и пр.). Каждое переиздание кодекса может подтвердить силу той или иной статьи, либо скорректировать или удалить ее, вводя новые. Данные, не подтверждаемые в течение долгого времени или подтверждаемые сомнительными источниками, постепенно будут удаляться из информационной системы.

Предлагаемые методы применяются при разработке информационной системы с документально подтверждаемой информацией. Ожидается, что результаты практического применения поспособствуют выявлению ошибок и недочетов, укажут на особенности и дадут опыт настройки процесса под различные виды фактов.

Литература

[1] Когаловский М. Р. Перспективные технологии информационных систем. — М.: ДМК Пресс; М: Компания АйТи, 2003. — 288 с.

[2] Когаловский М.Р. Системы доступа к данным на основе онтологий // Труды Второго симпозиума «Онтологическое моделирование», Казань 2010 – М: ИПИ РАН, 2011. – С. 45–78

[3] Загорулько Ю.А., Боровикова О.И., Кононенко И.С., Сидорова Е.А. Подход к построению предметной онтологии для портала знаний по компьютерной лингвистике. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». М.: РГГУ, 2006.

[4] Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, MihaiSurdeanu and Dan Jurafsky. Stanford’s Multi-Pass Sieve Coreference Resolution System at the CONLL-2011 Shared Task. // In Proc. of the 15th Conference on Computational Natural Language Learning: Shared Task. Portland. Oregon. USA. 2011. P. 28–34.

[5] Ермаков А.Е. Референция обозначений персон и организаций в русскоязычных текстах СМИ: эмпирические закономерности для компьютерного анализа. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2005». М: Наука, 2005. С. 131–135.

[6] Серый А.С., Сидорова Е.А. Поиск референциальных отношений между информационными объектами в процессе автоматического анализа документов // Труды XIV Всероссийской научной конференции RCDL-2012 Электронные библиотеки: перспективные методы и технологии, электронные коллекции. – Переславль-Залесский, 2012. С.206–212

[7] Серый А.С., Сидорова Е.А. Идентификация объектов в задаче автоматической обработки документов. // Компьютерная лингвистика и интеллектуальные технологии: Труды

международной конференции «Диалог 2011». М.: РГГУ, 2011. С. 580-591.

[8] Ли, К. Основы САПР (CAD/CAM/CAE) / Кунву Ли. – СПб. : Питер, 2004. – 560 с.

[9] Роджерс, Д.Ф. Математические основы машинной графики: пер с англ. / Д. Роджерс, Дж. Адамс. – М.:Мир, 2001. – 604 с.

[10] Кокс, Д. Идеалы, многообразия и алгоритмы. Введение в вычислительные аспекты алгебраической геометрии и коммутативной алгебры : пер.с англ. / Д. Кокс, Дж. Литтл, Д. О`Ши ; под ред. В.Л. Попова. – М.: Мир, 2000. – 687 с.

Developing methods for maintaining data reliability in an information system based on facts

Alexey S. Sery

The paper will discuss methods and approaches for automating the process of the incoming data analysis in ontology based information systems where data is presented as a set of information objects. It is proposed how to establish a referential identity or co-reference between objects and how to maintain information reliability, which means defining its trust metric and monitoring up-to-dateness. The former depends on the trust metrics of information sources, the latter — on the lifetime mostly. The proposed trust management technique also includes removing spotted unreliable data from the system data storage, and by doing so reduces expert participation in the data verifying process and number of errors in the system.