ИНЖИНИРИНГ ОНТОЛОГИЙ

УДК 004.853:004.82 *Научная статья* DOI: 10.18287/2223-9537-2023-13-1-113-124

Доверие к данным при пополнении онтологий и графов знаний

© 2023, А.С. Серый

Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия

Аннотация

Рассматривается задача оценки доверия к информации, извлекаемой из текстовых источников для пополнения онтологий или графов знаний. За единицу информации или факт, принимается минимальное знание об экземпляре предметной области, выражаемое единичным *RDF*-триплетом. Приведено описание вероятностной модели оценки доверия, основанной на марковских случайных процессах. При оценке модель строится на основании доступной информации об источниках с учётом ранее извлечённых данных. Предложен метод оценки доверия к информации с параллельным взвешиванием источников. Подобный подход востребован в ситуациях, когда качественные характеристики источников неизвестны или недоступны. В рамках тестирования модели были автоматически сгенерированы наборы численных данных различных объёмов, проведены эксперименты по взвешиванию источников и оценке доверия к извлекаемой из них информации. Результаты экспериментов показали, что в большинстве случаев веса источников, вычисляемые на основе предлагаемой модели, тем больше, чем меньше среднее отклонение предоставленной ими информации от истинной, доверие к фактам увеличивается с уменьшением расстояния до истинных данных. Выполнено сравнение с моделями агрегации данных. В большинстве случаев агрегация, выполненная на основе оценки доверия, демонстрировала наименьшее среднее отклонение от истинных данных среди рассмотренных моделей. Полученные результаты показывают, что предлагаемая модель эффективна в сравнении с другими аналогичными моделями и может применяться в задачах оценки доверия к фактам, представляемым вещественными числами.

Ключевые слова: *онтология, граф знаний, извлечение данных, доверие к информации, марковский процесс.*

Цитирование: *Серый А.С.* Доверие к данным при пополнении онтологий и графов знаний // Онтология проектирования. 2023. Т.13, №1(47). С.113-124. DOI:10.18287/2223-9537-2023-13-1-113-124.

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Введение

Современное глобальное информационное пространство невозможно представить и проанализировать усилиями человека, даже если речь идёт об экспертах в конкретных предметных областях (ПрО). Одним из путей решения данной проблемы стало использование методов автоматического анализа данных, которые широко применяются во всех сферах человеческой деятельности, связанной с обработкой информации. В первую очередь это методы обработки неструктурированных источников, например текстов, изображений и веб-страниц, позволяющие извлекать определённую информацию и представлять её в структурированном виде — в базах данных, онтологиях и графах знаний. Последние являются основным, на данный момент, способом интеграции больших структурированных данных [1]. Для извлечения информации применяется множество различных методов и подходов: от конвейерных

процессов на основе технологии *Apache NiFi* [2] до трансформерных нейросетей типа *BERT* [3]. Полученные графы затем применяются в интеллектуальных информационных системах (ИС) как источники знаний и основа логического вывода [4].

Обработка большого числа источников почти неизбежно приводит к появлению противоречивых знаний, т.е. нескольких альтернативных утверждений относительно одной и той же сущности. Это могут быть, к примеру, разные прогнозы погоды, цены акций, ожидаемое время прибытия авиарейсов, противоречивая информация о местах жительства или работы людей и т.д. Противоречия возникают как вследствие ошибок, так и потому, что информация, предоставленная источником, давно не обновлялась и устарела[1]. Таким образом, требуется не просто извлечь знания из источника, но и оценить их надёжность или уровень доверия к ним. В данной работе предлагается метод оценки доверия к информации, извлекаемой из различных источников для пополнения базы знаний (БЗ) ИС, основанной на онтологии.

1 Обзор предшествующих работ

Проблема оценки надёжности знаний, особенно в тех случаях, когда знания, полученные из разных источников, противоречат друг другу, исследуется давно. В работах [5, 6] надёжность источников и извлекаемых данных оценивались по заранее заданным правилам. Информация зачастую предполагалась статичной, т.е. представленной в виде завершённой таблицы соответствия фактов и источников [7, 8], которая затем не изменяется. В работах [9, 10] рассматриваются ситуации, когда информация из источников поступает последовательно, а истинные знания изменяются со временем. Исследования проводились на численных данных, в качестве примеров были выбраны прогнозы среднесуточной температуры, прогнозы капитализаций на фондовых рынках и время прибытия авиарейсов. Проведённый анализ результатов показал эффективность предлагаемых решений.

Более сложной задачей является оценка текстовых данных. Предметом исследования в данном направлении являются социальные сети: с одной стороны как источник большого количества противоречивой информации, с другой — как средство, оказывающее значительное влияние на образ мыслей и мнение людей. В работах [11, 12] исследовались способы и пути распространения слухов внутри социальных сетей. В [11] рассмотрены механизмы распространения слухов, проанализированы их жизненные циклы и зависимость таких показателей, как уровень поддержки и обсуждаемость, от типов пользователей, вовлекаемых в их распространение. Исследование [12] сосредоточено на верификации слухов. В работе [13] собран набор данных и разработана мультимодальная модель машинного обучения для решения задачи обнаружения и верификации слухов, касающихся девяти различных событий. Каждый элемент набора данных был аннотирован одной из трёх меток в зависимости от степени надёжности: **Правда** (***True***), **Неправда** (***False***) и **Не подтверждено** (***Unverified***). Схожая задача оценки высказываний пользователей решалась в [14] при помощи серии известных методов машинного обучения: наивный байесовский классификатор, логистическая регрессия, метод опорных векторов, деревья решений и др.

Сложность задач исследования в области анализа надёжности информации возрастает с ростом объёмов доступной информации и стремительным распространением ложной информации. Методы глубокого обучения применяются как современное и мощное средство. Разработанный в [12, 13] набор данных применялся для анализа новостей в социальной сети *Twitter* [15]. В [16] использована нейросетевая модель на основе свёрточных и рекуррентных нейронных сетей для распознавания ложной информации.

[1] В современном информационном пространстве особое значение приобретает проблема выявления заведомо ложной или умышленно искажённой информации. *Прим. ред.*

2 Модель доверия

2.1 Факты в онтологии

Пусть БЗ ИС построена на основе онтологии $\mathcal{O}$ некоторой ПрО, где $\mathcal{O} = \{C_O, D_O, Dat_O, Rel_O\}$. Конечное непустое множество C_O представляет совокупность концептов ПрО, конечные непустые множества D_O, Dat_O и Rel_O — соответственно доменов, атрибутов и отношений. Каждый атрибут из Dat_O имеет область значений $d \in D_O$, а элементы множества $Rel_O \subseteq C_O \times C_O$ — это бинарные отношения между концептами из C_O. Объединение $Dat_O \cup Rel_O$ атрибутов и отношений называется множеством свойств онтологии $\mathcal{O}$. Класс можно определить в виде тройки (c, Dat_c, Rel_c), где через c обозначено имя класса, $Dat_c \subseteq Dat_O, Rel_c \subseteq Rel_O$ — его свойства. Каждый атрибут $\alpha^c \in Dat_c$ имеет область значений $d_{\alpha^c} \in D_O$, а каждое отношение $\rho^c \in Rel_c$ связывает класс c некоторым классом $c_{\rho^c} \in C_O$. Класс или множество классов c_{ρ^c} образуют область значений отношения ρ^c.

Пусть $a \in c_a$, если a является экземпляром класса $c_a \in C_O$. Экземпляр представляется тройкой вида $a = (c_a, Dat_a, Rel_a)$ такой, что $Dat_a = \{(\alpha, V_{\alpha_a}) | \alpha \in Dat_{c_a}, V_{\alpha_a} \subseteq d_\alpha \in D_O\}$ — атрибуты экземпляра a, и $Rel_a = \{(\rho, V_{\rho_a}) | \rho \in Rel_{c_a})\}$ — его связи с другими экземплярами. Здесь V_{ρ_a} — множество экземпляров, с которыми a связан отношением ρ.

В данной работе задача пополнения БЗ ИС рассматривается как задача пополнения онтологии, т.е. как добавление, удаление и изменение экземпляров в соответствии с данными, полученными извне. При этом за областью рассмотрения остаётся редактирование ядра онтологии — множества $\mathcal{O}$. В терминах ИС, БЗ которой построена на основе онтологии, единицей информации считается минимальное знание об экземпляре ПрО — значение его атрибута или его связь с другим экземпляром. Можно называть такое знание единичным фактом или просто фактом. Автоматическая обработка текстовых источников позволяет извлекать факты и добавлять их в БЗ ИС. Информация, полученная из разных источников, может оказаться противоречивой, порождая множества конфликтных фактов. Требуется ранжировать эти множества по уровню доверия таким образом, чтобы предоставить пользователям ИС наиболее надёжную информацию.

При оценке доверия к фактам предлагаемая в данной работе модель основывается на доступной информации об источниках, из которых факты были получены. Под источниками здесь понимаются общедоступные электронные ресурсы, из которых извлекаются численные или текстовые данные. Предполагается, что в ИС используются некоторые качественные показатели источников, например рейтинг, если таковые доступны, или создаются собственные оценки, опираясь на всю информацию, доступную на текущий момент. Это означает, что модель доверия учитывает характеристики источников данных, но не включает описание методов их получения. Необходимо только, чтобы данные характеристики принимали значения из множества $\mathbb{R}^+$. В любой момент времени множество источников, из которых извлекаются факты для пополнения БЗ ИС, конечно. Пусть это будет множество S, а для любого источника $s \in S$ искомая качественная характеристика — μ^s. В дальнейшем индекс s в обозначении μ^s будет опускаться в тех случаях, когда не имеет значения, из какого конкретного источника была получена информация.

2.2 Доверие как случайный процесс

В каждый момент времени для факта F, являющегося частью экземпляра a, должна быть определена величина, показывающая насколько надёжным является данный факт по сравнению с другими фактами в ИС, Tr^F- трастовая метрика. БЗ не является статичной, и с появле-

нием новых источников, содержащих другие факты об a, доверие к F может изменяться. История изменений Tr^F представляет собой последовательность, в которой каждый следующий член зависит только от предыдущего, а также от поступившего в обработку источника $s \in S$. ИС может считать или не считать факт F достоверным, т.е. F имеет два потенциальных состояния: **Ненадёжный** (***Unreliable, U***) и **Надёжный** (***Reliable, R***). Информация о том, в каком из состояний на данный момент находится F, эквивалентна его Tr^F. Последовательность значений Tr^F становится эквивалентной последовательности вида $(X_t, t = 0,1, \dots)$ дискретных случайных величин, принимающих значения из бинарного множества состояний $\{U, R\}$. Тогда для последовательности (X_t) выполняется условие $P\big(X_k = x_k \mid X_{k_1} = x_{k_1}, X_{k_2} = x_{k_2}, \dots, X_{k_r} = x_{k_r}\big) = P\big(X_k = x_k \mid X_{k_r} = x_{k_r}\big)$, для любых $k_1 < k_2 < k_3 < \dots < k_r < k$, т.е. она удовлетворяет определению марковских случайных процессов.

Следующий член случайного процесса вычисляется всякий раз, когда в систему поступает новая информация об F. Пусть T - множество моментов времени, соответствующих членам случайного процесса X_t, т.е. $t \in T$. Значение Tr^F в момент t оценивается как вероятность того, что факт F является надёжным, т.е. $Tr^F = P(X_t^F = R)$.

Для любого t величина X_t распределена как (π_U^t, π_R^t), где $\pi_x^t = P(X_t = x)$. Вектор распределения $\bar{\pi} = (\pi_U, \pi_R)$ можно назвать вектором распределения доверия (*Trust Distribution Vector, TDV*). *TDV* показывает вероятность факта оказаться достоверным или недостоверным. Очевидно, что $\pi_U + \pi_R = 1$. Распределение $\bar{\pi} = \left(\frac{1}{2}, \frac{1}{2}\right)$ и близкие к нему соответствуют состоянию неопределённости, когда судить о достоверности F невозможно.

Согласно теории случайных процессов, вектор $\bar{\pi}^{t+1} = (\pi_U^{t+1}, \pi_R^{t+1})$ получается умножением вектора предыдущего шага $\bar{\pi}^t$ на 2×2 стохастическую *матрицу перехода* $\mathrm{P}(t, t+1)$. Элементы $p_{ij}(t, t+1)$ матрицы $\mathrm{P}(t, t+1)$ — это вероятности перехода из i-го состояния в j-е на шаге $(t+1)$, при этом $p_{i1} + p_{i2} = 1, i = 1, 2$. Здесь и далее предполагается, что состояния 1 и 2 — это состояния U и R соответственно. В таких обозначениях $\pi_1 = \pi_U, \pi_2 = \pi_R$.

2.3 Переходная матрица случайного процесса

В рамках модели матрица перехода $P(t, t+1)$ представляется как функция от μ и вектора $\bar{\pi}^t$.

$$P(t, t+1) = softmax\left(\pi^\top \begin{pmatrix} \frac{\pi_1}{\mu} & \mu\pi_2 \end{pmatrix}\right) = \begin{pmatrix} softmax\left(\frac{\pi_1^2}{\mu}, \mu\pi_1\pi_2\right) \\ softmax\left(\frac{\pi_1\pi_2}{\mu}, \mu\pi_2^2\right) \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{1}{1 + e^{\pi_1(\mu\pi_2 - \pi_1/\mu)}} & \frac{1}{1 + e^{\pi_1(\pi_1/\mu - \mu\pi_2)}} \\ \frac{1}{1 + e^{\pi_2(\mu\pi_2 - \pi_1/\mu)}} & \frac{1}{1 + e^{\pi_2(\pi_1/\mu - \mu\pi_2)}} \end{pmatrix}. \qquad (1)$$

Из формулы (1) следует, что увеличение рейтинга μ источника информации ведёт к росту вероятности перехода в состояние R и наоборот — информация из источников с низким рейтингом способствует переходу в состояние U. Функция $softmax(\bar{x}) = \left(\frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}\right)_i, \bar{x} = (x_1, x_2, \dots, x_n)$ необходима для представления категориального распределения по строкам матрицы P.

Матрица P обладает несколькими полезными свойствами как функция $P(\mu, \bar{\pi}^t)$. На основании формулы (1) можно заключить, что с уменьшением рейтинга источника доверие к получаемой из него информации также уменьшается вплоть до нуля. Заведомо ложный источник с $\mu = 0$ приводит любой *TDV*, кроме $(0, 1)$, к вектору $(1, 0)$ за один шаг. Вектор $(0, 1)$,

т.е. такой, где $\pi_1 = 0$, приводит к $P = \begin{pmatrix} 1/2 & 1/2 \\ \frac{1}{1+e^{\mu}} & \frac{1}{1+e^{-\mu}} \end{pmatrix} \xrightarrow{\mu \to 0} \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$. Таким образом, заведомо ложный источник переводит идеальный *TDV* $(0, 1)$ в $(1/2, 1/2)$.

Аналогично, в случае заведомо надёжного источника, при $\mu \to \infty$, матрица P переводит любой *TDV*, кроме $(1, 0)$, в идеальный вектор $(0, 1)$. В случае $\pi_2 = 0$ $P \xrightarrow{\mu \to \infty} \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix}$. Это означает, что заведомо надёжный источник переводит распределение $(1, 0)$ в $(1/2, 1/2)$.

Третьим полезным следствием формулы (1) является тот факт, что для любого распределения $\bar{\pi}$ $(\bar{\pi}P)_2(\mu_1) < (\bar{\pi}P)_2(\mu_2)$ при $\mu_1 < \mu_2$, т.е. доверие Tr^F монотонно как функция от μ.

3 Экспериментальные исследования

3.1 Параметры и обозначения

ИС, реализующая модель доверия для оценки поступающих данных, полагается на качественные показатели источников, из которых они получены. Модель не содержит описания методов оценки источников. При доступности достаточного количества альтернативных фактов существуют методы, позволяющие параллельно оцениванию доверия к информации, «взвесить» её источники. В данной работе предлагается метод параллельного оценивания, основанный на [9, 10]. В этих работах описанные модели использовались для численных данных, таких как прогнозы среднесуточной температуры, количество пешеходов на улице, капитализация компаний на фондовых рынках и т.п. Основная решаемая с использованием моделей задача состоит в вычислении по имеющейся альтернативной информации из источников единственного агрегированного значения, наиболее близкого к истинному. Предлагаемая в данной работе модель предназначена для ранжирования всех полученных альтернативных значений по степени доверия к ним.

Входными параметрами оригинальных моделей [9, 10] являются множество моментов времени T, источников S и множество O объектов, информация о которых извлекается из источников. Источник $s \in S$ имеет вес $w_s \in \mathbb{R}^+$, и в момент времени t из него извлекается информация об $c_t^s \geq 0$ объектах. Решение задачи заключается в минимизации потерь вида (2).

$$L_t = \theta \sum_{s=1}^{S} w_s \sum_{o=1}^{c_t^s} \left(v_{o,t}^s - v_{o,t}^*\right)^2 - \sum_{s=1}^{S} c_t^s \, log(w_s). \qquad (2)$$

Здесь $v_{o,t}^s$ и $v_{o,t}^*$ — это, соответственно, значение, полученное для объекта o из источника s в момент времени t и агрегированное значение, вычисленное для o в момент t. Для минимизации потерь необходимо, с одной стороны, уменьшить вес источников, дающих информацию, сильно отличающуюся от агрегированного значения, с другой — избегать приближения весов к нулю. Значение параметра θ задаётся заранее.

В данном случае объектами выступают факты, т.е. единицы информации об экземплярах концептов онтологии ПрО. Количество источников заранее неизвестно, однако в любой момент времени оно конечно, поэтому, без ограничения общности, можно считать множество источников аналогичным таковому в оригинальной модели в ситуации, когда источник s может не содержать информации о конкретном объекте в момент времени t. Это означает, что в момент t учитываются только те источники, из которых удалось извлечь требуемую информацию. Множеством, аналогичным множеству объектов O, будет $\mathcal{F} = \{f | f = (a, \alpha), a \in c_a \in C_O, \alpha \in Dat_{c_a} \cup Rel_{c_a}\}$ — множество пар экземпляр/свойство, которое можно

назвать множеством типов извлекаемых фактов. В этом случае $v_{f,t}^{s}$ дополняет пару f до законченного триплета, т.е. факта. Время T при этом соответствует определению, данному в разделе 2.2. Полный список обозначений приведён в таблице 1.

Таблица 1 – Принятые обозначения

Обозначение	Определение
S	Количество источников, из которых извлекается информация. $S < +\infty$.
w_s	Вес источника s.
$\mathcal{F}$	Множество типов извлекаемых фактов.
T	Множество моментов времени.
$v_{f,t}^{s}$	Факт типа $f \in \mathcal{F}$, полученный из источника s в момент t.
$v_{f,t}^{*}$	Агрегированное значение f в момент t.
c_t^s	Количество фактов, полученных из источника s в момент t.
D	База данных ИС.
$D_{f,t}$	Количество альтернативных значений f в базе данных в момент t.
v_f^j	j-е альтернативное значение f в D.
$e_{f,t}^{s}$	Ошибка источника s в момент t на факте типа f.

В работе [9] рассмотрены ситуации, когда истинное знание постоянно меняется, то есть истинное значение факта зависит от t. Это соответствует задачам оценки прогнозов погоды или количества товаров на складе. Существуют задачи, где истинное знание не изменяется длительное время, но могут появляться источники, распространяющие неточную информацию и порождающие новые альтернативные значения. В качестве примера можно привести количество сотрудников в организации, данные статистических исследований, место работы или жительства персоны и т.п. В подобных ситуациях альтернативных значений сравнительно немного, они распространяются разными источниками и поэтому могут быть извлечены ИС многократно. Для таких случаев предлагается использовать функцию потерь (3), которая, наряду с новой информацией, учитывает и ту, что уже содержится в БЗ ИС.

$$L_t = \theta\left(\sum_{s=1}^{S} w_s \sum_{f=1}^{c_t^s}\left(v_{f,t}^{s} - v_{f,t}^{*}\right)^2 + \lambda \sum_{f\in\mathcal{F}}\sum_{j=1}^{D_{f,t}} T\, r\left(v_f^j\right)\left(v_f^j - v_{f,t}^{*}\right)^2\right) - \sum_{s=1}^{S} c_t^s \log(w_s). \quad (3)$$

В формуле (3) гиперпараметры θ и λ также должны быть заданы предварительно. Агрегированные значения в рамках построенной модели доверия носят вспомогательный характер и необходимы только для взвешивания источников и вычисления рейтинговых характеристик $\mu_{f,t}^{s} = {}^{1}/_{e_{f,t}^{s}}$, вычисляемых как величина, обратная ошибке $v_{f,t}^{s}$ по отношению к $v_{f,t}^{*}$. Величина $\mu_{f,t}^{s}$ используется при оценке доверия к $v_{f,t}^{s}$. В зависимости от вида функции потерь агрегированные значения $v_{f,t}^{*}$ вычисляются по одной из формул (4):

$$v_{f,t}^{*} = \frac{\sum_{s=1}^{S} w_s \cdot v_{f,t}^{s}}{\sum_{s=1}^{S} w_s};$$
$$v_{f,t}^{*} = \frac{\sum_{s=1}^{S} w_s \cdot v_{f,t}^{s} + \lambda \sum_{j=1}^{D_{f,t}} Tr\left(v_f^j\right) v_f^j}{\sum_{s=1}^{S} w_s + \lambda \sum_{j=1}^{D_{f,t}} Tr\left(v_f^j\right)}. \quad (4)$$

Вид формулы для $v_{f,t}^{*}$ не влияет на способ взвешивания источников. Их веса вычисляются по формуле (5), применяемой в работе [9].

$$w_s = \frac{2\alpha - 2 + \sum_{t=1}^{T} c_t^s}{2\beta + \theta \sum_{t=1}^{T}\sum_{f=1}^{c_t^s}\left(e_{f,t}^{s}\right)^2}. \quad (5)$$

В [9] была доказана сходимость процесса взвешивания, т.е. веса источников при такой оценке сходятся к определённым значениям. Начальные веса определяются случайно и подчиняются гамма-распределению с параметрами α и β.

3.2 Результаты на наборах численных данных

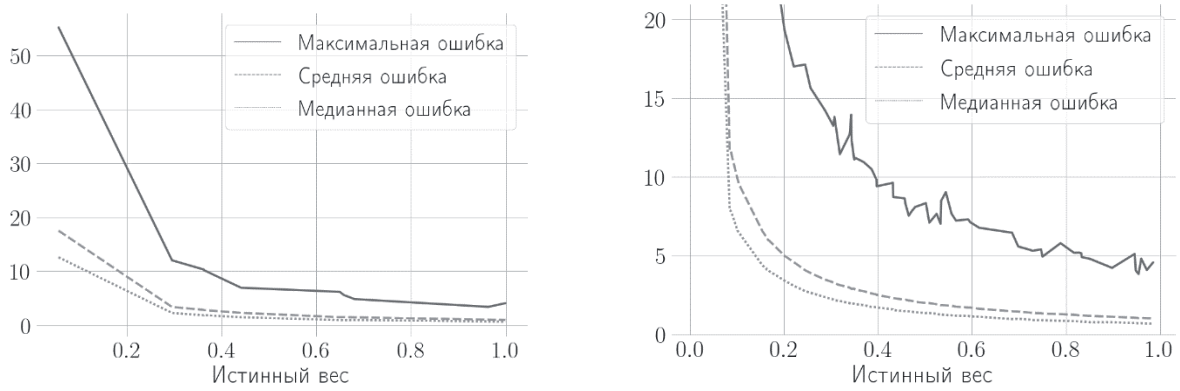
Оригинальные модели предназначены для работы с численными данными, соответственно, все $v_{f,t}^{s}$ — это целые или вещественные числа. Предлагаемая модель оценки доверия, обозначенная *Markov Trust Evaluation model* (*MarkTE*), сравнивалась с моделью агрегации [9] *DYNAmic Truth Discovery* (*DynaTD*) на случайно сгенерированных массивах данных разных размеров, представленных в таблице 2.

Таблица 2 – Искусственные наборы данных

	S	T	$\mathcal{F}$
Small	10	15	50
Medium	65	45	250
Large	150	100	500

Приведены результаты решения двух задач, обозначенных как *NF* и *FX*. Задача *NF* (*Not Fixed truth*) соответствует ситуации, когда из каждого источника в каждый момент времени извлекается значение $v_{f,t}^{s}$, а истинные значения зависят от t. Это означает, что для каждого $f \in \mathcal{F}$ и каждого t существует отдельное истинное значение $v_{f,t}^{a}$. Условия задачи *NF* соответствуют таковым в экспериментах, приведённых в работах [9, 10]. Так как истинные значения постоянно меняются, то в функции (3) не имеет смысла учитывать информацию, полученную ранее, поэтому в задаче *NF* в формуле (4) принята $\lambda = 0$. Задача *FX* (*FiXed truth*) соответствует ситуации, когда значение вида $v_{f,t}^{s}$ может быть извлечено из источника s не гарантированно, но с некоторой вероятностью, распределённой как $U_{(0.4,0.6)}$, а истинные значения v_{f}^{a} фиксированы и не зависят от времени. При этом количество альтернатив для f также фиксировано, а одни и те же значения могут встречаться в разных источниках в разное время. Данная ситуация схожа с теми, на которые ориентирована модель *MarkTE*, и при оценке учитывалась информация из БЗ с $\lambda > 0$. Путём экспериментов в условиях задач *NF* и *FX*, предложенная модель доверия была протестирована в разных условиях.

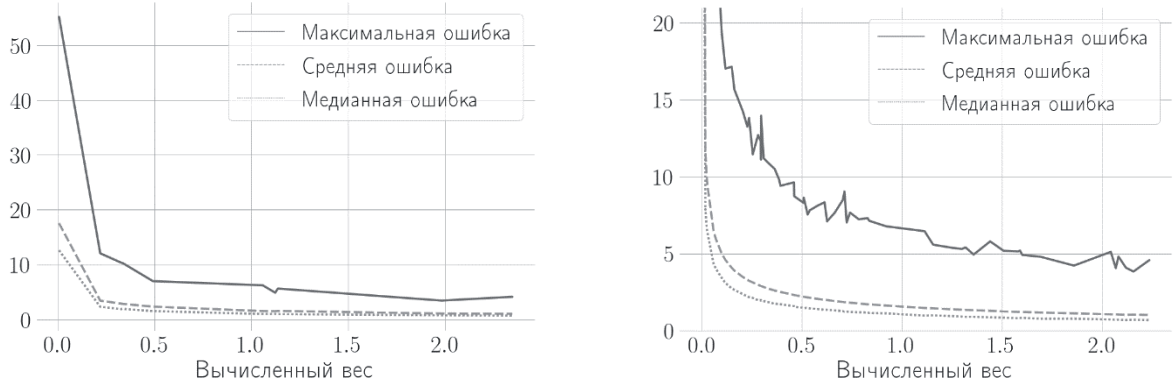
Для каждой задачи были сгенерированы отдельные наборы данных. Каждый источник $s \in S$ получил истинный вес w_{s}^{a}, и для каждого $f \in \mathcal{F}$ были определены истинные значения. Вероятность того, что источник предоставил информацию, не соответствующую действительности, обратно пропорциональна его весу. Как показано на рисунке 1, максимальная, средняя и медианная ошибки источников тем больше, чем меньше назначенный им истинный вес. Ошибки вычислялись как $\left|v_{f,t}^{s} - v_{f,t}^{a}\right|$ в задаче NF и $\left|v_{f,t}^{s} - v_{f}^{a}\right|$ в задаче FX.



(а) Small + NF (б) Medium + NF

Рисунок 1 – Ошибки источников в зависимости от их истинного веса

Характеристики $\mu_{f,t}^{s}$ для переходных матриц (1) вычислялись как $\mu_{f,t}^{s} = {}^{1}/_{e_{f,t}^{s}}$, где $e_{f,t}^{s} = \left|v_{f,t}^{s} - v_{f,t}^{*}\right|$. Зная вектор распределения доверия $\bar{\pi}^{t-1}$ значения $v_{f,t}^{s}$, распределение на шаге t вычислялось как $\bar{\pi}^{t} = \bar{\pi}^{t-1} \cdot P\left(\mu_{f,t}^{s}\right)$. Веса источников последовательно пересчитывались по формуле (5). Зависимость весов, вычисленных на основе модели *MarkTE*, от ошибок источника на истинных значениях показана на рисунке 2.



(а) Small + NF (б) Medium + NF

Рисунок 2 – Ошибки источников в зависимости от их вычисляемого веса

Сходство графиков на рисунках 1 и 2 говорит о том, что веса источников, предоставивших информацию, близкую к истинной, сходятся к величинам бóльшим, нежели те, к которым сходятся веса источников, содержащих информацию, далекую от истины. Таким образом, источники были взвешены корректно, и более надёжные получили бóльшие веса по сравнению с менее надёжными.

Модель оценки доверия служит для ранжирования полученных знаний, тогда как модель *DynaTD* выполняет агрегацию. Для сравнения полученных результатов была введена операция агрегации (6) для всех s, таких что $v_{f,t}^{s}$ существует, выполняемая на основе оценок доверия полученных знаний, а не их источников, как в (5).

$$v_{f,t}^{m} = \frac{\sum_{s=1}^{S} v_{f,t}^{s} \cdot Tr\left(v_{f,t}^{s}\right)}{\sum_{s=1}^{S} Tr\left(v_{f,t}^{s}\right)}. \qquad (6)$$

В качестве показателей эффективности были использованы результаты сравнения полученных агрегированных значений с истинными: средняя абсолютная ошибка (*Mean Absolute Error, MAE*) и средняя квадратичная ошибка (*Root Mean Squared Error, RMSE*).

$$V = \left\{(t,s,f) \in T \times S \times \mathcal{F} \middle| \exists v_{f,t}^{s}\right\}$$

$$MAE = \frac{\sum_{(t,s,f)\in V}\left|v_{f,t}^{s} - v_{f,t}^{a}\right|}{|V|}$$

$$RMSE = \sqrt{\frac{\sum_{(t,s,f)\in V}\left(v_{f,t}^{s} - v_{f,t}^{a}\right)^{2}}{|V|}}.$$

Базисный уровень был реализован двумя дополнительными моделями *Mean* и *Median*. Агрегированные значения $v_{f,t}^{*}$ в модели *Mean* вычисляются как среднее по всем $v_{f,t}^{s}$, в модели *Median* агрегацию выполняет функция медианы.

Результаты оценки *MAE* и *RMSE* для всех моделей и наборов данных на каждой задаче приведены в таблице 3. Видно, что агрегированные значения, полученные по формуле (6) на основе оценок доверия в соответствии с моделью *MarkTE*, в большинстве случаев оказались ближе к истинным. Отсюда можно заключить, что эффективность модели *MarkTE* находится на конкурентном уровне по сравнению с моделью *DynaTD* и другими, сравнения с которыми

были проведены в [9]. Полученные оценки доверия оказались более точными весовыми коэффициентами при агрегации.

Таблица 3 – Сравнение результатов моделей *MarkTE, DynaTD, Mean* и *Median*

Задача		**Мера**	*MarkTE*	*DynaTD*	*Mean*	*Median*
Small	*NF*	*MAE*	**0,4367**	0,5448	2,8096	0,5601
		RMSE	**0,5765**	0,9925	3,5978	0,7068
	FX	*MAE*	**0,3841**	0,5306	0,5449	0,6093
		RMSE	**0,542**	0,673	0,6888	0,8243
Medium	*NF*	*MAE*	0,2093	**0,2087**	1,2755	0,2507
		RMSE	**0,2674**	0,3603	1,5951	0,3155
	FX	*MAE*	**0,3297**	0,6205	0,6377	0,8149
		RMSE	**0,4289**	0,7759	0,7969	1,0263
Large	*NF*	*MAE*	0,1389	**0,132**	1,9844	0,1688
		RMSE	**0,1814**	0,2727	2,4887	0,2117
	FX	*MAE*	**0,2791**	0,582	0,6018	0,7967
		RMSE	**0,3599**	0,7325	0,757	1,0124

Каждая единица информации, выраженная значением $v^s_{f,t}$, получила оценку доверия $Tr(v^s_{f,t})$. На рисунке 3 показана зависимость $Tr(v^s_{f,t})$ от абсолютной ошибки $|v^s_{f,t} - v^a_{f,t}|$ на примере задачи *Medium+NF*.

Набор данных *Medium* (см. таблицу 2) содержал $65 \cdot 45 \cdot 250 = 731250$ значений. В целях наглядности на рисунке 3 представлены результаты для случайной выборки из 100 значений. На графике видна очевидная тенденция уменьшения доверия с ростом ошибки, из чего можно заключить, что предлагаемая модель *MarkTE* корректно оценивает доверие к поступающим в ИС данным, назначая наибольшие показатели доверия значениям с минимальной ошибкой.

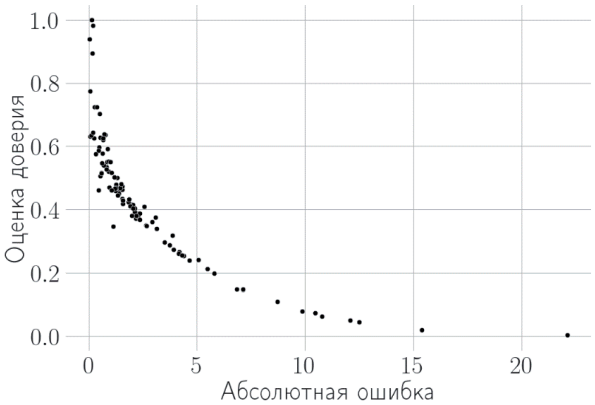


Рисунок 3 – Оценка доверия к фактам в зависимости от абсолютной ошибки

Заключение

Предлагаемая модель оценки доверия к информации, извлекаемой из внешних источников для пополнения БЗ ИС, построенной на основе онтологии некоторой ПрО, способна демонстрировать эффективные показатели по сравнению с другими моделями на задачах оценки численных данных. В качестве данных могут выступать показатели, например, стоимость акций и капитализация компаний, информация о наличии товаров на складе и др. В общем случае модель *MarkTE* способна оценить доверие к текстовым данным или данным, представленным в виде *RDF*-триплетов, что соответствует их представлению в онтологиях.

Список источников

[1] ***Баклавски К.*** Онтологический Саммит 2020. Коммюнике: Графы знаний / К. Баклавски, М. Беннет, Г. Берг-Кросс, Т. Шнайдер, Р. Шарма, Д. Сингер. Перевод с англ. Д. Боргест // Онтология проектирования. 2020. Т.10, №4(38). С.540–555. DOI: 10.18287/2223-9537-2020-10-4-540-555.

[2] ***Simsek U., Umbrich J., Fensel D.*** Towards a Knowledge Graph Lifecycle: A pipeline for the population of a commercial Knowledge Graph. In: A. Paschke, C. Neudecker, G. Rehm, J.A. Qundus and L. Pintscher (eds): Proceedings of the Conference on Digital Curation Technologies Qurator-2020 (Berlin, Germany, 2020, January 20-21). CEUR Workshop Proceedings, vol. 2535, CEUR-WS.org. https://ceur-ws.org/Vol-2535/paper_10.pdf.

[3] ***Fernàndez-Cañellas D. et al***. Enhancing Online Knowledge Graph Population with Semantic Knowledge. In: The Semantic Web ISWC 2020. Lecture Notes in Computer Science, vol 12506. Springer, Cham. 2020. P.183–200. DOI: 10.1007/978-3-030-62419-4_11.

[4] ***Cimmino A., García-Castro R.*** Helio: a framework for implementing the life cycle of knowledge graphs. Semantic Web. Preprint 2022. P.1–27. DOI: 10.3233/SW-233224.

[5] ***Galland A., Abiteboul S., Marian A., and Senellart P.*** Corroborating information from disagreeing views. In: Proceedings of the third ACM international conference on Web search and data mining WSDM-2010. (New York, USA, 2010, February 4–6). 2010. P.131–140. DOI: 10.1145/1718487.1718504.

[6] ***Li X., Dong X.L., Lyons K.B., Meng W., Srivastava D.*** Truth finding on the deep web: Is the problem solved? In: Proceedings of the VLDB Endowment. vol. 6(2). 2012. P.97–108. DOI: 10.14778/2535568.2448943.

[7] ***Pochampally R. et al.*** Fusing data with correlations. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data SIGMOD-2014 (Snowbird, Utah, USA, 2014, June 22–27). 2014. P.433–444. DOI: 10.1145/2588555.2593674.

[8] ***Dong X.L., Gabrilovich E., Murphy K., Dang V., Horn W., Lugaresi C., Sun S., Zhang W.*** Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. In: Proceedings of the VLDB Endowment. vol. 8, 2015. P.938–949. DOI: 10.14778/2777598.2777603.

[9] ***Li Y. et al.*** On the discovery of evolving truth. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD-2015 (Sydney, NSW, Australia, 2015, August 10–13). 2015. P.675–684. DOI: 10.1145/2783258.2783277.

[10] ***Yao L. et al.*** Online truth discovery on time series data. In: Proceedings of the 2018 SIAM international Conference on Data Mining SDM-2018 (San Diego, USA, 2018, October 6–13). 2018. Society for Industrial and Applied Mathematics. P.162–170. DOI: 10.1137/1.9781611975321.19.

[11] ***Zubiaga A., Liakata M., Procter R., Wong Sak Hoi G., Tolmie P.*** Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE 2016. 11(3): e0150989.DOI: 10.1371/journal.pone.0150989.

[12] ***Kochkina E., Liakata M., Zubiaga A.*** All-in-one: Multi-task learning for rumour verification. In: Proceedings of 27th International Conference on Computational Linguistics COLING-2018 (Santa Fe, New-Mexico, USA, 2018, August 20–26). Association for Computational Linguistics (ACL). 2018. P.3402–3413. DOI: 10.48550/arXiv.1806.03713.

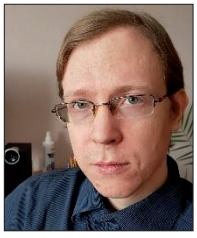
[13] PHEME dataset for Rumour Detection and Veracity Classification. https://www.kaggle.com/datasets/usharengaraju/pheme-dataset.

[14] ***Chen X., Yuan Y., Lu L., Yang J.*** A multidimensional trust evaluation framework for online social networks based on machine learning. IEEE Access. vol. 7, 2019. P.175499–175513. DOI: 10.1109/ACCESS.2019.2957779.

[15] ***Vyas P., El-Gayar O.*** Credibility analysis of news on twitter using LSTM: An exploratory study. In: Proceedings of 26th Americas Conference on Information Systems AMCIS 2020 (Virtual conference, 2020, August 10–14). Association for Information Systems. https://scholar.dsu.edu/cgi/viewcontent.cgi?article=1150&context=bispapers.

[16] ***Hirlekar V.V., Kumar A.*** Tweet Credibility Detection for COVID-19 Tweets using Text and User Content Features. International Journal of Advanced Computer Science and Applications, 13(4), 2022. P.430–439. DOI: 10.14569/IJACSA.2022.0130451.

Сведения об авторе



Серый Алексей Сергеевич, 1987 г. рождения. Окончил Новосибирский государственный университет в 2010 г. Младший научный сотрудник лаборатории искусственного интеллекта Института систем информатики им. А.П. Ершова (Новосибирск). В списке научных трудов более двух десятков работ в области представления знаний и компьютерной лингвистики. Author ID (RSCI): 714554; ORCID: 0000-0001-8275-4700; Author ID (Scopus): 56403204900; Researcher ID (WoS): K-1557-2018. *alexey.seryj@iis.nsk.su.* ✉.

Поступила в редакцию 10.01.2023, после рецензирования 31.01.2023. Принята к публикации 11.02.2023.

Scientific article DOI: 10.18287/2223-9537-2023-13-1-113-124

Data credibility when populating ontologies and knowledge graphs

© 2023, A.S. Sery

A.P. Ershov Institute of Informatics Systems of Siberian Branch of RAS, Novosibirsk, Russia

Abstract

The problem of assessing trust in the information extracted from textual sources to populate ontologies or knowledge graphs is considered. For a unit of information or a fact, the minimum knowledge about an instance of the subject area, expressed by a single RDF triplet, is taken. The paper provides a description of a probabilistic trust evaluation model based on Markov random processes. When assessing, the model is built on the basis of available information about sources, taking into account previously extracted data. A method for assessing the credibility of information with parallel weighting of sources is also provided. The proposed approach is in demand when the quality of the data sources is unknown or unavailable. As part of testing the model, sets of numerical data of various sizes were automatically generated, experiments were carried out to weigh the sources and assess trust in the information extracted from them. It was shown that in most cases the weights of the sources calculated on the basis of the proposed model are the greater, the smaller the average deviation of the information they provide from the true one, and the confidence in facts increases with decreasing distance to the true data. Comparison with data aggregation models is made. In most cases, the aggregation based on the trust score showed the smallest average deviation from the true data among the considered models. The obtained results show that the proposed model is effective in comparison with other similar models and can be used in problems of assessing trust in facts represented by real numbers.

Key words: *ontology, knowledge graph, data extraction, information trustworthiness, Markov Process.*

For citation: *Sery AS.* Data credibility when populating ontologies and knowledge graphs [In Russian]. *Ontology of designing*. 2023; 13(1): 113-124. DOI:10.18287/2223-9537-2023-13-1-113-124.

Conflict of interest: The author declares no conflict of interest.

List of figures and tables

References

[1] ***Baclawski K, Bennett M, Berg-Cross G, Schneider T, Sharma R, Singer J, Sriram, R.D.*** Ontology summit 2020 communiqué: Knowledge graphs. Applied Ontology. 2021; 16(2): 229–247. DOI: 10.18287/2223-9537-2020-10-4-540-555.

[2] ***Simsek U, Umbrich J, Fensel D.*** Towards a Knowledge Graph Lifecycle: A pipeline for the population of a commercial Knowledge Graph. In: A. Paschke, C. Neudecker, G. Rehm, J.A. Qundus and L. Pintscher (eds): Proceedings of the Conference on Digital Curation Technologies Qurator-2020 (Berlin, Germany, 2020, January 20-21). CEUR Workshop Proceedings, vol. 2535, CEUR-WS.org, https://ceur-ws.org/Vol-2535/paper_10.pdf.

[3] ***Fernàndez-Cañellas D. et al***. Enhancing Online Knowledge Graph Population with Semantic Knowledge. In: The Semantic Web ISWC 2020. Lecture Notes in Computer Science, vol 12506. Springer, Cham. 2020. 183–200. DOI: 10.1007/978-3-030-62419-4_11.

[4] ***Cimmino A, García-Castro R.*** Helio: a framework for implementing the life cycle of knowledge graphs. Semantic Web. Preprint 2022. 1–27. DOI: 10.3233/SW-233224.

[5] ***Galland A, Abiteboul S, Marian A, Senellart P.*** Corroborating information from disagreeing views. In: Proceedings of the third ACM international conference on Web search and data mining WSDM-2010. (New York, USA, 2010, February 4–6). 2010. 131–140. DOI: 10.1145/1718487.1718504.

[6] ***Li X, Dong XL, Lyons KB, Meng W, Srivastava D.*** Truth finding on the deep web: Is the problem solved? In: Proceedings of the VLDB Endowment. 2012; 6(2): 97–108. DOI: 10.14778/2535568.2448943.

[7] ***Pochampally R. et al.*** Fusing data with correlations. In: Proceedings of the 2014 ACM SIGMOD international conference on Management of data SIGMOD-2014 (Snowbird, Utah, USA, 2014, June 22–27). 2014. 433–444. DOI: 10.1145/2588555.2593674.

[8] ***Dong XL, Gabrilovich E, Murphy K, Dang V, Horn W, Lugaresi C, Sun S, Zhang W.*** Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. In: Proceedings of the VLDB Endowment. 2015; 8: 938–949. DOI: 10.14778/2777598.2777603.

[9] ***Li Y. et al.*** On the discovery of evolving truth. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD-2015 (Sydney, NSW, Australia, 2015, August 10–13). 2015. 675–684. DOI: 10.1145/2783258.2783277.

[10] ***Yao L. et al.*** Online truth discovery on time series data. In: Proceedings of the 2018 SIAM international Conference on Data Mining SDM-2018 (San Diego, USA, 2018, October 6–13). 2018. Society for Industrial and Applied Mathematics. 162–170. DOI: 10.1137/1.9781611975321.19.

[11] ***Zubiaga A, Liakata M, Procter R, Wong Sak Hoi G, Tolmie P.*** Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE 2016. 11(3): e0150989. DOI: 10.1371/journal.pone.0150989.

[12] ***Kochkina E, Liakata M, Zubiaga A.*** All-in-one: Multi-task learning for rumour verification. In: Proceedings of 27th International Conference on Computational Linguistics COLING-2018 (Santa Fe, New-Mexico, USA, 2018, August 20–26). Association for Computational Linguistics (ACL). 2018. 3402–3413. DOI: 10.48550/arXiv.1806.03713.

[13] PHEME dataset for Rumour Detection and Veracity Classification. https://www.kaggle.com/datasets/usharengaraju/pheme-dataset.

[14] ***Chen X, Yuan Y, Lu L, Yang J.*** A multidimensional trust evaluation framework for online social networks based on machine learning. IEEE Access. 2019; 7: 175499–175513. DOI: 10.1109/ACCESS.2019.2957779.

[15] ***Vyas P, El-Gayar O.*** Credibility analysis of news on twitter using LSTM: An exploratory study. In: Proceedings of 26th Americas Conference on Information Systems AMCIS 2020 (Virtual conference, 2020, August 10–14). Association for Information Systems. https://scholar.dsu.edu/cgi/viewcontent.cgi?article=1150&context=bispapers.

[16] ***Hirlekar VV, Kumar A.*** Tweet Credibility Detection for COVID-19 Tweets using Text and User Content Features. International Journal of Advanced Computer Science and Applications, 2022; 13(4): 430–439. DOI: 10.14569/IJACSA.2022.0130451.

About the author

Alexey Sergeevich Sery (b.1987) holds a master's degree in mathematics from Novosibirsk State University (2010) and the position of Junior Researcher at the A.P. Ershov Institute of Informatics Systems of Siberian Branch of RAS. He is the author of more than 20 papers in the fields of NLP systems and Knowledge Representation. Author ID (RSCI): 714554; ORCID: 0000-0001-8275-4700; Author ID (Scopus): 56403204900; Researcher ID (WoS): K-1557-2018. *alexey.seryj@iis.nsk.su.* ✉.

Received January 10, 2023. Revised January 31, 2023. Accepted February 11, 2023.