Math-Net.Ru
Общероссийский математический портал

И. М. Адамович, О. И. Волков, Иерархическая форма представления биографического факта, *Системы и средства информ.*, 2016, том 26, выпуск 2, 108–122

DOI: 10.14357/08696527160207

Использование Общероссийского математического портала Math-Net.Ru подразумевает, что вы прочитали и согласны с пользовательским соглашением
http://www.mathnet.ru/rus/agreement

Параметры загрузки:
IP: 18.220.196.68
26 декабря 2024 г., 09:39:33

ИЕРАРХИЧЕСКАЯ ФОРМА ПРЕДСТАВЛЕНИЯ БИОГРАФИЧЕСКОГО ФАКТА

И. М. Адамович[1], *О. И. Волков*[2]

Аннотация: Данная статья посвящена факту в конкретно-историческом исследовании. Показана специфика биографического факта (БФ), проявляющаяся в его двойственности, отражающей его связи как с реальным миром, так и с информационным пространством (ИП). Описана иерархическая форма представления БФ, использующаяся в системе Т-парсер для автоматического извлечения фактов из текстов. Доказано соответствие данного представления формальным свойствам БФ. Определены и описаны операции логики фактов для данного представления. Также продемонстрирована связь иерархического представления фактов с онтологией предметной области и вытекающая из этого перспективность ее использования в автоматизированных процедурах обработки фактов. Показана возможность частичной формализации фактов на этапе их выделения. Описана технология обработки данных в биографическом исследовании, автоматизирующая как этап сбора информации, так и этап сопоставления выявленных фактов с целью взаимоувязывания и разрешения противоречий, включающая полную формализацию фактов на базе онтологии. Показаны проблемы этого подхода и намечены пути их решения.

Ключевые слова: биографический факт; иерархическая форма факта; автоматизированная технология обработки фактов; онтология; логика фактов

DOI: 10.14357/08696527160207

1 Введение

Специфика биографического исследования состоит в том, что в центре внимания исследователя находится конкретная личность и все без исключения стороны (социальные, экономические, политические, этнические, художественные и т. п.) ее реальной жизни [1]. Соответственно БФ, выявляемые в результате исследования, отличаются большим многообразием как по составу характеристик, так и по связям между ними.

Уже разработаны и используются средства автоматизации обработки БФ (сравнения, нормализации, выявления связей, разрешения противоречий, вывода новых фактов). Примером может служить разработанная в ИПИ ФИЦ ИУ

[1]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, Adam@amsd.com

[2]Институт проблем информатики Федерального исследовательского центра «Информатика и управление» Российской академии наук, Volkov@amsd.com

РАН система «Фактограф» [2], обеспечивающая формализацию, хранение, интеграцию, аналитико-синтетическую обработку БФ, получаемых из гетерогенных источников. Но выявление БФ остается очень трудоемким процессом и отнимает у исследователя много сил и времени [3], поэтому очень востребованными являются системы автоматического извлечения фактов из текстов на естественном языке (ЕЯ), в том числе биографической и художественно-исторической направленности.

Работы по интеллектуальной обработке ЕЯ-текстов и извлечению из них структур знаний давно ведутся в лаборатории Компьютерной лингвистики и когнитивных технологий обработки текстов ИПИ ФИЦ ИУ РАН. В основе этих работ лежит созданная И. П. Кузнецовым концепция формализма представления знаний, получившего название «расширенные семантические сети» (РСС), а также разработанный им язык ДЕКЛ, обеспечивающий гибкое преобразование структур РСС, включая поиск и порождение новых знаний [4–8]. В рамках дальнейшего развития этого подхода была разработана система Т-парсер [9] для извлечения структурированных данных из текста при помощи контекстно-свободных грамматик. Т-парсер опирается на иерархическое (древовидное) представление факта. Принципиальная адекватность такого представления сути БФ была наглядно показана в [9], но, вообще говоря, утверждение о строгом соответствии данного представления формальным признакам БФ требует строгого обоснования, что и является основной целью данной статьи.

Также показана связь данной структуры с онтологией предметной области и вытекающая из этого перспективность ее использования в автоматизированных процедурах обработки.

2 Специфика биографического факта

Существуют различные определения факта в научном смысле этого слова (научного факта). Но все они сводятся к тому, что научный факт отражает объективные свойства реальных объектов и процессов. Даже в областях, где невозможно непосредственное взаимодействие с объектом (например, в палеонтологии), исследователь имеет дело с объективными свидетельствами (окаменелостями, ископаемыми останками). Совсем другое дело — факт в конкретно-историческом исследовании (частным случаем которого является биографическое исследование). Исследователь, как правило, имеет дело с субъективными свидетельствами и оценками. Даже в ситуации, когда, казалось бы, можно опереться на объективные данные, невозможно исключить субъективную составляющую. Например, информацию о росте персоны — объекта исследования можно получить объективным измерением, если сохранилась его одежда. Но факт принадлежности этой одежды искомой персоне устанавливается исключительно по семейным преданиям, чьим-то свидетельствам и т. п., т. е. на основании субъективной информации.

Информация, зафиксированная документально, образует ИП — бумажные и другие твердые носители, а в последние десятилетия и электронные ресур-

сы [10]. Информационное пространство формирует образ реального мира, в котором реальные объекты (РО) имеют свое отражение. Информационный образ (ИО) при этом не тождественен РО. Следовательно, факты, содержащиеся в документах, адекватно описывают образы РО, но не сами объекты. Более того, один РО может иметь несколько ИО в ИП. Цель же исследователя — установление научных фактов, получение объективной информации о РО исследования (насколько это вообще возможно в научном исследовании). Поэтому биограф работает с фактами двух типов:

(1) факты, описывающие ИО (i-факты). Обладают свойством фрагментарности и противоречивости. Не существуют в отрыве от своих метаданных, т. е. информации о документе-источнике.

Получение этих фактов — необходимая и наиболее трудоемкая стадия биографического исследования, состоящая в работе в архивах и просмотре огромного числа документов. Автоматизация этого этапа является важной и актуальной задачей в связи со все увеличивающимся общественным интересом к семейной истории;

(2) факты, описывающие РО (r-факты), т. е. объективное знание об объекте исследования. Работа исследователя на втором этапе исследования состоит в сопоставлении i-фактов между собой, а также с ранее выявленными r-фактами, относящимися к объекту исследования, и с нормалями (наборами правил, регламентирующих зависимости между фактами) [11] с целью взаимоувязывания и разрешения противоречий. При этом r-факты после публикации результатов исследования могут выступать i-фактами в каком-либо другом исследовании.

3 Иерархическая форма представления биографического факта

Формы представления информации могут быть самыми различными, но не все они подходят к БФ в силу их специфики: так, семантическая сеть, представляющая собой узлы, соответствующие объектам предметной области, и дуги, связывающие узлы и описывающие отношения между ними, вполне пригодна для хранения информации [12], но i-факты принципиально фрагментарны, а отношения, задаваемые ими, нечетки, неоднозначны и взаимно противоречивы. Так, при изучении даже таких официальных документов, как метрические книги и данные переписи населения, часто приходится сталкиваться с разным указанием даты рождения одного и того же человека в разных документах. Даже степень родства может быть указана по-разному. Что же касается такой информации, как причина смерти в разделе «Об умерших», то ее достоверность крайне низка, поскольку она устанавливалась без участия медицинских работников.

Образовать четкую сеть могут только r-факты, что, собственно, и является целью исследования, но это их свойство никак не может быть использовано в процессе его проведения.

Также существуют достаточно универсальные формы представления, в принципе подходящие и для БФ. К примеру, в Томита-парсере компании Яндекс — средстве для извлечения структурированных данных из текста на ЕЯ [13] — используется табличная форма представления фактов. Такая форма позволяет эффективно описывать структуру факта, хотя и требует несколько громоздкого механизма объединения таблиц для сведения фактов в факты более высокого уровня, необходимость в чем обусловливается многопроходной технологией извлечения информации. В случае же биографического исследования сама структура извлекаемого из источника факта не всегда может быть описана заранее. Следовательно, структура БФ должна быть адаптирована:

- к специфике биографической информации;
- к технологии поиска фактов;
- к технологии дальнейшей работы с ними с учетом возможности их автоматизированной обработки.

В [3] предложена иерархическая форма представления факта и показано ее соответствие специфике биографической информации и удобство ее использования для описанной технологии автоматического извлечения фактов из биографических и художественно-исторических текстов с помощью системы Т-парсер. Также было показано принципиальное соответствие данной структуры самому понятию БФ. Так, в соответствии с [10], под БФ понимается утверждение, что некая характеристика некоторого объекта принимает некоторое конкретное значение. Применительно к Т-парсеру характеристикой, подразумевающей наличие своего объекта, является правило, а значением — выделенный этим правилом фрагмент разбираемого предложения. Так, правило «фамилия», выделившее в тексте

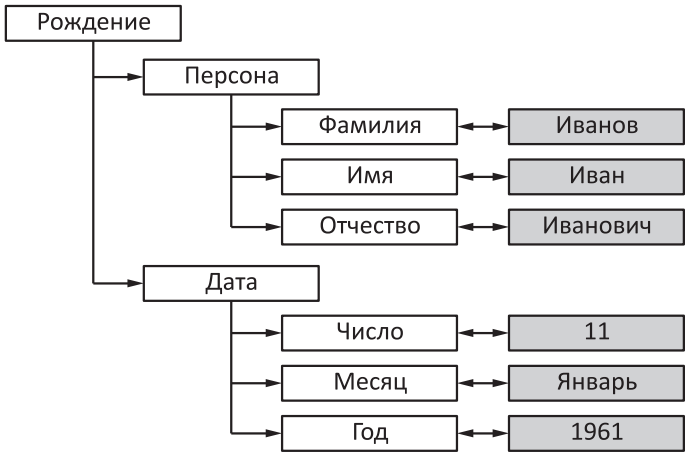

Рис. 1 Пример факта «рождение»

фрагмент «Иванов», сформирует факт, что некто имеет фамилию Иванов, т. е. существует объект, у которого характеристика «фамилия» принимает значение «Иванов».

В соответствии с логикой работы Т-парсера результат применения правила, в свою очередь, может быть выделен неким правилом, что формирует иерархию фактов, т. е. под фактом в Т-парсере понимается иерархическая структура именованных понятий, терминальные узлы которой содержат выделенные из фразы слова или словосочетания, которые интерпретируются как их значения (рис. 1).

Вообще говоря, для иерархического представления факта не является принципиальным, что значения имеют только терминальные узлы. Для общности следует считать, что значения есть у каждого узла и значением по умолчанию является факт его наличия. Так, если в приведенном примере убрать значение «Иванович» у узла «Отчество», то это следует интерпретировать как факт наличия отчества у персоны.

4 Соответствие основным свойствам биографического факта

Как показано в [10], БФ должен подчиняться правилам логики фактов, т. е. для БФ должны быть определены конъюнкция, дизъюнкция, импликация и отрицание. Также для БФ должны быть справедливы основные свойства: симметрия, транзитивность и вариативность. Покажем, что представление факта как иерархической структуры не противоречит этим требованиям.

Конечно же, существует тривиальная трактовка этого требования: любой факт, независимо от формы его представления, можно трактовать как высказывание, могущее быть как истинным, так и ложным (i-факт вполне может быть ложным, если объект ИП, который он описывает, неадекватно отражает соответствующий объект реального мира), а поэтому к ним применимы все логические операции булевой алгебры. Но такой подход не является конструктивным для БФ, поскольку истинность i-факта заранее не известна и ее только требуется установить в процессе исследования. Поэтому следует рассмотреть более конструктивные, специфические для выбранной формы представления фактов, аналоги основных логических операций.

4.1 Формальное описание факта

Введем ряд обозначений и понятий. Для этого опишем иерархический факт Tree как совокупность узлов, представляющих собой совокупность именованного понятия и значения, которое, вообще говоря, может быть пусто, рекуррентной формулой

$$\mathrm{Tree} = \{\mathrm{Node}, \{\mathrm{Tree}_i | i = 1, \ldots, m\}\} \ ,$$

т. е. дерево представляет собой корневой узел и множество его поддеревьев в количестве m штук. Для терминальных узлов множество поддеревьев пусто, т. е. $m = 0$. Корневой узел любого (в том числе являющегося частью другого

(*а*) (*б*)

Рис. 2 Два эквивалентных дерева: (*а*) Tree, (*б*) $\mathrm{Tree}' = \mathrm{Eq}\,(\mathrm{Tree})_{\mathrm{Node}_2}$

дерева) дерева будем обозначать как Rt, т. е. $\mathrm{Node} = \mathrm{Rt}(\mathrm{Tree})$. Обратную операцию обозначим как Tr, т. е. $\mathrm{Tree} \equiv \mathrm{Tr}(\mathrm{Rt}(\mathrm{Tree}))$.

Для любого узла Node_n, удаленного от корневого на расстояние n, существует последовательность $\{\mathrm{Node}_0, \ldots, \mathrm{Node}_{n-1}\}$ вышестоящих узлов, где Node_0 — корневой узел. Тогда иерархический факт можно описать формулой:

$$\begin{aligned}\mathrm{Tree} = \big\{\mathrm{Node}_0, \big\{\mathrm{Tree}_i^0 | \mathrm{Rt}\left(\mathrm{Tree}_i^0\right) \neq \mathrm{Node}_1\big\}, \\ \mathrm{Node}_1, \big\{\mathrm{Tree}_i^1 | \mathrm{Rt}\left(\mathrm{Tree}_i^1\right) \neq \mathrm{Node}_2\big\}, \ldots \\ \ldots, \mathrm{Node}_{n-1}, \big\{\mathrm{Tree}_i^{n-1} | \mathrm{Rt}\left(\mathrm{Tree}_i^{n-1}\right) \neq \mathrm{Node}_n\big\}, \mathrm{Node}_n, \{\mathrm{Tree}_i^n\}\big\}\,.\end{aligned}$$

Будем говорить, что дерево Tree′ эквивалентно дереву Tree относительно узла Node_n, если

$$\begin{aligned}\mathrm{Tree}' = \big\{\mathrm{Node}_n, \{\mathrm{Tree}_i^n\}, \mathrm{Node}_{n-1}, \big\{\mathrm{Tree}_i^{n-1} | \mathrm{Rt}\left(\mathrm{Tree}_i^{n-1}\right) \neq \mathrm{Node}_n\big\}, \ldots \\ \ldots, \mathrm{Node}_1, \big\{\mathrm{Tree}_i^1 | \mathrm{Rt}\left(\mathrm{Tree}_i^1\right) \neq \mathrm{Node}_2\big\}, \\ \mathrm{Node}_0, \big\{\mathrm{Tree}_i^0 | \mathrm{Rt}\left(\mathrm{Tree}_i^0\right) \neq \mathrm{Node}_1\big\}\big\}\,,\end{aligned}$$

т. е. состав узлов в Tree′ совпадает с Tree, порядок узлов для цепочки $\{\mathrm{Node}_0, \ldots, \mathrm{Node}_n\}$ меняется на обратный, а для остальных сохраняется. Корневым узлом в Tree′ становится Node_n (рис. 2). Кратко будем записывать этот факт как $\mathrm{Tree}' = \mathrm{Eq}\,(\mathrm{Tree})_{\mathrm{Node}_n}$.

Далее факты будем обозначать заглавными латинскими буквами. Для i-фактов, не существующих в отрыве от своей метаинформации, вводим соответствующее обозначение: факт A с метаинформацией M будем обозначать как $A|_M$.

4.2 Операции над фактами

Теперь рассмотрим аналоги основных логических операций применительно к БФ в иерархической форме, для которых введем соответствующие обозначения:

1. F-конъюнкция ($\&_F$).
   Определим F-конъюнкцию следующим образом:

   $$\exists\, N_1 \in A|_{M_1},\ N_2 \in B|_{M_2} :\ \mathrm{Tr}\,(N_1) = \mathrm{Tr}\,(N_2) \Rightarrow A|_{M_1} \&_F B|_{M_2} = \mathrm{Tr}\,(N_1)\ ,$$

   т. е. операция определена и имеет смысл только для i-фактов, причем с различной метаинформацией, что соответствует фактам, полученным из разных источников. Если эти факты имеют совпадающее поддерево (т. е. общий подфакт), то этот подфакт можно считать имеющим независимое подтверждение из разных источников и, следовательно, достоверным. Соответственно, данный подфакт можно считать новым r-фактом, являющимся результатом F-конъюнкции исходных i-фактов. Таким образом, операцию F-конъюнкции можно назвать операцией сверки.
2. F-дизъюнкция ($\vee_F$).
   Определим F-дизъюнкцию следующим образом:

   $$\exists\, N_1 \in A,\ N_2 \in B :\ \mathrm{Tr}\,(N_1) = \mathrm{Tr}\,(N_2) \Rightarrow A \vee_F B = \mathrm{Eq}\,(A)_{N_1} \cup \mathrm{Eq}\,(B)_{N_2}\ .$$

   Операция представляет собой объединение двух фактов в один. Следовательно, она определена и имеет смысл только для фактов с одинаковой метаинформацией, что соответствует фактам, полученным из одного источника, или для r-фактов. Факты приводятся к эквивалентному виду относительно корневых узлов общего поддерева и объединяются в единое дерево.
3. F-импликация ($\Rightarrow_F$).
   Определим F-импликацию следующим образом:

   $$A \Rightarrow_F B = B|_A\,.$$

   Это неформальная операция — один из важных элементов этапа обработки фактов в биографическом исследовании. Как показано в [3], поиск информации в биографическом исследовании проходит итерационно и направление поиска на последующих этапах зависит от результатов поиска на предыдущем этапе. Так, операцию выявления нового направления в исследовании задает именно F-импликация, т. е. формирование важного с точки зрения исследования факта, логически зависящего (как правило, на основании некоторых нормалей) от истинности некоторого i-факта, который сам по себе может казаться малозначимым. Так, факт службы в армии некоторой персоны в определенное время может означать невозможность его участия в это время в некотором важном событии. С позиций описываемой структуры это соответствует новому i-факту, в качестве метаданных которого принимается этот малозначимый факт.

4. F-отрицание ($\neg_N$).

   F-отрицание является обычным булевским отрицанием и поэтому не требует специального значка F в обозначении, но, поскольку применяется к конкретному узлу факта, обозначение включает узел. Так, $B = \neg_N A$ означает, что факт B совпадает с фактом A, кроме значения узла N, которое заменено на его логическое отрицание.

4.3 Свойства фактов

В соответствии с [10] БФ должен обладать следующими свойствами:

1. Симметрия.

   Для большинства двуместных отношений факту $\beta(p, q)$, где p и q — объекты, однозначно соответствует факт $\beta'(q, p)$.

   Пример: *МестоРаботы*(«Иванов», «Контора») $\Leftrightarrow$ *Сотрудник*(«Контора», «Иванов»).

   Данное свойство для БФ в иерархическом представлении безусловно выполняется. Так, если β есть факт с иерархией ($\mathit{ФИО} = \text{«}\mathit{Иванов}\text{»}) \rightarrow (\mathit{Работа} = \text{«}\mathit{Контора}\text{»})$, то $\beta' = \mathrm{Eq}\,(\beta)_{\text{Работа}}$.

2. Транзитивность.

   Факты, основанные на таких характеристиках, как иерархия, обладают свойством транзитивности. Пример: *Работа* («Иванов», «Контора») $\wedge$ *Место* («Контора», «Москва») $\Rightarrow$ *Место* («Иванов», «Москва»).

   Иерархическое представление БФ автоматически поддерживает свойства, основанные на иерархии понятий. Так, из иерархии понятий, приведенных на рис. 3, выделяются цепочки:

   а) Иванов–Контора: $(\mathit{Фамилия} = \text{«}\mathit{Иванов}\text{»}) \leftarrow (\mathit{Персона}) \rightarrow (\mathit{Работа}) \rightarrow \rightarrow (\mathit{Название} = \text{«}\mathit{Контора}\text{»})$;

   б) Иванов–Москва: $(\mathit{Фамилия} = \text{«}\mathit{Иванов}\text{»}) \leftarrow (\mathit{Персона}) \rightarrow (\mathit{Работа}) \rightarrow \rightarrow (\mathit{Место} = \text{«}\mathit{Москва}\text{»})$.

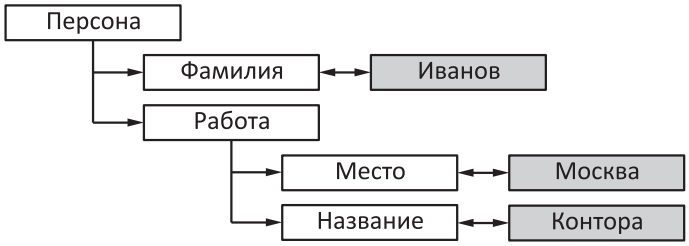

Рис. 3 Пример иерархии понятий

3. Вариативность. Для вариативных характеристик, например именования, значения разбиваются на классы эквивалентности, определяемые нормалями, а формула (речь идет о «нормализованной» записи факта $(\forall\ t \in \Delta t)\beta(p, q, t) = a$, означающей, что некая характеристика объекта или объектов в некий момент времени принимает определенное значение) принимает вид принадлежности данному классу.

Сама по себе иерархическая форма представления факта не накладывает никаких дополнительных ограничений на форму представления значений характеристик, но, как будет показано далее, легко может быть увязана с онтологией, необходимой для дальнейшей обработки фактов, что решает проблему вариативности за счет формализации понятий предметной области.

5 Связь иерархической формы факта с онтологией

Специфика i-фактов такова, что их превращение в r-факты, т. е. в факты в научном понимании этого слова, представляет собой весьма нетривиальную и не формализуемую на 100% процедуру. Но такая ее важная составляющая, как процедура сопоставления i-фактов с целью выявления информационных лакун и противоречий, вполне может быть автоматизирована. Существует специализированное программное обеспечение, разработанное специально для сопоставления биографической информации [2]. Но для любой автоматизированной процедуры обработки данные (т. е. факты) должны быть формализованы. А поскольку БФ формируются из текстов на ЕЯ, для успешного определения значений слов и правильных связей между ними необходимо знание мира, который описывается в тексте. Такое знание предоставляется онтологией. Онтология представляет собой формализацию знаний о взаимосвязях объектов и целых классов объектов реального мира, которая позволяет компьютеру использовать эти знания и даже дополнять информацию об отдельных объектах с помощью логического вывода [14]. Подходы к задачам выделения фактов из текстов на ЕЯ с опорой на онтологии активно разрабатываются в последнее время [15–17]. Более того, одно из определений факта звучит как «содержащаяся в тексте и метаданных документа совокупность связей между сущностями, описываемыми в онтологии информационной системы» [18].

Частичная формализация на основе онтологии составляющих БФ в иерархическом представлении вполне возможна и вытекает из технологии формирования таких фактов. Так, на рис. 4 схематически представлена технология формирования БФ, позволяющая применять автоматизацию как на этапе формирования i-фактов, так и на этапе формирования r-фактов с использованием онтологии. Технология предполагает выполнение следующих этапов:

1. Формируются правила определения фактов. Имена правил привязываются к онтологии.

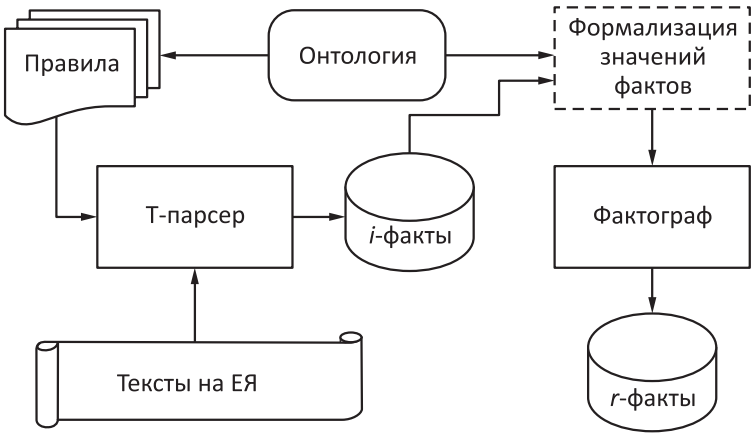

Рис. 4 Автоматизированная технология формирования r-фактов

2. Посредством Т-парсера с использованием подготовленных правил из текстов на ЕЯ вычленяются i-факты в иерархической форме. Имена узлов иерархии при этом формализованы за счет привязки извлекающих правил к онтологии. Тем самым, поскольку имена характеристик для большинства узлов и определяют значения типа «факт наличия», часть значений становится формализованной уже на этом этапе.
3. Формализуются непустые значения узлов. В данный момент это осуществляется вручную, поскольку применение «Фактографа» осуществлялось для обработки информации, полученной из уже частично формализованных источников — метрических книг и исповедных ведомостей. В дальнейшем при использовании большого объема данных, полученных из текстов на ЕЯ, потребуется автоматизация и этого процесса.
4. С использованием модуля «Фактограф» формируются r-факты. В перспективе возможно также использование иных модулей, отражающих специфику изучаемых объектов и предметной области.

6 Выводы

Описано и обосновано иерархическое представление БФ. Описана автоматизированная технология формирования БФ из текстов на ЕЯ. Намечены пути развития данной технологии. При этом следует отметить и проблемы, присущие предложенному подходу:

– сложность создания онтологии с учетом широты предметной области биографического исследования. В качестве решения этой проблемы могут быть предложены следующие подходы:

(а) децентрализация создания онтологии [19];

(б) автоматизация создания онтологии [20];

– отсутствие автоматизированной процедуры формализации непустых значений узлов. Данная проблема непосредственно связана с проблемой создания онтологии и может быть легко решена при наличии таковой. Видится целесообразным объединение процедуры формализации на основе онтологии с процедурой парсинга текста на ЕЯ, т. е. дооснащение Т-парсера средствами взаимодействия с онтологией.

Поскольку элементы описанной технологии уже применяются и показали свою эффективность для ряда задач биографического поиска, следует считать выбранную форму представления БФ приемлемой, а предложенную технологию — перспективной.

Литература

1. *Иконникова С. Н.* Биографика как часть исторической культурологии // Вестник СПбГУКИ, 2012. № 2(11). С. 6–10.
2. *Адамович И. М.* Методы и средства справочно-поисковой поддержки научных и социально-культурных проектов на основе интеграции данных разнородных биографических источников: Отчет о НИР. — М.: ИПИ РАН, 2012. С. 67–84.
3. *Адамович И. М., Волков О. И.* Средства поддержки интернет-поиска при проведении биографических исследований // Системы и средства информатики, 2014. Т. 24. № 2. С. 178–192.
4. *Kuznetsov I. P., Kozerenko E. B., Kuznetsov K. I., Timonina N. O.* Intelligent System for Entities Extraction (ISEE) from natural language texts // Workshop (International) on Conceptual Structures for Extracting Natural Language Semantics (SENSE'09) at the 17th Conference (International) on Conceptual Structures (ICCS'09) Proceedings / Eds. U. Priss, G. Angelova. — Moscow, Russia: University Higher School of Economics, 2009. P. 17–25.
5. *Kuznetsov I. P., Kozerenko E. B., Matskevich A. G.* Intelligent extraction of knowledge structures from natural language texts // 2011 IEEE/WIC/ACM Joint Conferences (International) on Web Intelligence and Intelligent Agent Technology (WI-IAT 2011) Proceedings, 2011. P. 269–272.
6. *Kozerenko E. B., Ermakov P. V.* The strategies of syntactic analysis based on head-driven grammars and the methods of their implementation in information systems // Информатика и её применения, 2011. Т. 5. Вып. 4. С. 107–113.
7. *Кузнецов И. П., Сомин Н. В.* Выявление имплицитной информации из текстов на естественном языке: проблемы и методы // Информатика и её применения, 2012. Т. 6. Вып. 1. С. 49–58.
8. *Шарнин М. М., Кузнецов И. П.* Особенности семантического поиска информационных объектов на основе технологии баз знаний // Информатика и её применения, 2012. Т. 6. Вып. 2. С. 113–121.
9. *Адамович И. М., Волков О. И.* Система извлечения биографических фактов из текстов исторической направленности // Системы и средства информатики, 2015. Вып. 25. № 3. С. 235–250.

10. *Маркова Н. А.* Логика биографических фактов // Информатика и ее применения, 2012. Т. 6. Вып. 2. С. 87–96.

11. *Маркова Н. А.* Электронная коллекция биографических фактов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Тр. XIV Всеросс. науч. конф. RCDL'2012. — Переславль-Залесский: Университет города Переславля, 2012. С. 287–293.

12. *Мизарбеков А. М.* Алгоритм хранения информации в семантической сети // Известия ЮФУ. Технические науки, 2012. Т. 130. Вып. 5. С. 25–28.

13. *Огородник Р. В., Серебреная Л. В.* Обработка текста с помощью Томита-парсера // Информационные технологии и системы 2014 (ИТС-2014): Мат-лы междунар. науч. конф. — Минск: БГУИР, 2014. С. 230–231.

14. *Диконов В. Г., Богуславский И. М., Тимошенко С. П.* Онтология для поддержки задач извлечения смысла из текста на естественном языке // Информационные технологии и системы (ИТиС'12): Сб. тр. 35-й Конф. молодых ученых и специалистов ИППИ РАН. — М.: ИППИ РАН, 2012. С. 152–161.

15. *Сидорова Е. А.* Инструментальные средства фактографического анализа документов в информационных системах, основанных на онтологиях // Вестник НГУ. Сер. Информационные технологии, 2008. Т. 6. Вып. 3. С. 126–134.

16. *Святогор Л. А., Гладун В. П.* Машинное понимание текстов естественного языка: онтологическая парадигма // Искусственный интеллект, 2010. № 3. С. 249–258.

17. *Оробинская Е. А., Дорошенко А. Ю.* Использование онтологий для автоматической обработки текстов на естественном языке // Вестник НТУ ХПИ: Сб. научных тр., 2011. № 30. С. 101–106.

18. *Барахнин В. Б., Федотов А. М.* Построение модели фактографического поиска // Вестник НГУ. Сер. Информационные технологии, 2013. Т. 11. Вып. 4. С. 16–27.

19. *Слободюк А. А., Маторин С. И., Четвериков С. Н.* О подходе к созданию онтологий на основе системно-объектных моделей предметной области // Научные ведомости БелГУ, 2013. № 22(165). Вып. 28/1. С. 159–167.

20. *Найханова Л. В.* Технология создания методов автоматического построения онтологий с применением генетического и автоматного программирования. — Улан-Удэ: БНЦ СО РАН, 2008. 244 с.

Поступила в редакцию 18.12.15

HIERARCHICAL FORMAT OF A BIOGRAPHICAL FACT

I. M. Adamovich and O. I. Volkov

Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilova Str., Moscow 119333, Russian Federation

Abstract: This article focuses on the facts in the specific historical studies. The specific duality of a biographical fact, which is evidenced in its connections with

the real world and with the information space, is mentioned. The hierarchical format of a biographical fact, which is used in the "T-parser" system of automatic extraction of facts from texts in natural language, is described in detail. The accordance of this format with the formal properties of a biographical fact is proven. The logic of biographical facts operators for this format is explored and described. The relation of the hierarchical format of facts with the domain ontology and the prospects of its usage in automated systems of facts processing are also demonstrated. The possibility of partial formalization at the stage of fact extraction is analyzed. The technology of biographical data processing, which automates the fact extraction stage and the stage of facts comparison with the purpose of integration and conflict resolution, including full fact formalization on the basis of an ontology, is proposed and analyzed. The problems of this approach are described and the possible ways of their solution are proposed.

Keywords: biographical fact; hierarchical format of fact; automated technology of facts processing; ontology; logic of facts

DOI: 10.14357/08696527160207

References

1. Ikonnikova, S. N. 2012. Biografika kak chast' istoricheskoy kul'turologii [Biografical studies as part of the historical cultural studies]. *Vestnik SPbGUKI* [Bulletin of Saint-Petersburg State University of Culture and Art] 2(11):6–10.
2. Adamovich, I. M. 2012. Metody i sredstva spravochno-poiskovoy podderzhki nauchnykh i sotsial'no-kul'turnykh proektov na osnove integratsii dannykh raznorodnykh biograficheskikh istochnikov [Methods and tools of information support of scientific and sociocultural projects based on the integration of heterogeneous data of different biographical sources]. Research Report. Moscow: IPI FRC CSC RAS. 67–84.
3. Adamovich, I. M., and O. I. Volkov. 2014. Sredstva podderzhki internet-poiska pri provedenii biograficheskikh issledovaniy [The technology of Internet-searching as the part of the biographic investigation]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 2(24):178–192.
4. Kuznetsov, I. P., E. B. Kozerenko, K. I. Kuznetsov, and N. O. Timonina. 2009. Intelligent System for Entities Extraction (ISEE) from natural language texts. *Workshop (International) on Conceptual Structures for Extracting Natural Language Semantics (SENSE'09) at the 17th Conference (International) on Conceptual Structures (ICCS'09) Proceedings*. Eds. U. Priss and G. Angelova. Moscow, Russia: University Higher School of Economics. 17–25.
5. Kuznetsov, I. P., E. B. Kozerenko, and A. Matskevich. 2011. Intelligent extraction of knowledge structures from natural language texts. *2011 IEEE/WIC/ACM Joint Conferences (International) on Web Intelligence and Intelligent Agent Technology (WI-IAT 2011) Proceedings*. 269–272.
6. Kozerenko, E. B., and P. V. Ermakov. 2011. The strategies of syntactic analysis based on head-driven grammars and the methods of their implementation in information systems. *Informatika i ee Primeneniya — Inform. Appl.* 5(4):107–113.
7. Kuznetsov, I. P., and N. V. Somin. 2012. Vyyavlenie implitsitnoy informatsii iz tekstov na estestvennom yazyke: Problemy i metody [Extraction of implicit information from

the texts in natural language: Problems and methods]. *Informatika i ee Primeneniya — Inform. Appl.* 6(1):49–58.

8. Sharnin, M. M., and I. P. Kuznetsov. 2012 Osobennosti semanticheskogo poiska informatsionnykh ob"ektov na osnove tekhnologii baz znaniy [Semantic search of natural language information on the basis of knowledge base technology]. *Informatika i ee Primeneniya — Inform. Appl.* 6(2):113–121.
9. Adamovich, I. M., and O. I. Volkov. 2015. Sistema izvlecheniya biograficheskikh faktov iz tekstov istoricheskoy napravlennosti [The system of facts extraction from historical texts]. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 25(3):235–250.
10. Markova, N. A. 2012. Logika biograficheskih faktov [A logic of biographical facts]. *Informatika i ee Primeneniya — Inform. Appl.* 6(2):49–58.
11. Markova, N. A. 2012. Elektronnaya kollektsiya biograficheskikh faktov [Digital collection of biographic facts]. *14th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" Proceedings*. Pereslavl-Zalessky, Russia. 287–293.
12. Mirzabekov, Ya. M. 2012. Algoritm khraneniya informatsii v semanticheskoy seti [Algorithm for storing information in semantic networks]. *Izvestiya UFU. Tekhnicheskie nauki* [Herald of SFedU. Engineering Sciences] 130(5):25–28.
13. Ogorodnik, R. V., and L. V. Serebrenaya. 2014. Obrabotka teksta s pomoshch'yu Tomita-parsera [Text processing using Tomita-parser]. *Scientific Symposium (International) "Information Technology and Systems 2014" Proceedings*. Minsk. 220–231.
14. Dikonov, V. G., I. M. Boguslavsky, and S. P. Timoshenko. 2013. Ontologiya dlya podderzhki zadach izvlecheniya smysla iz teksta na estestvennom yazyke [Ontology to support semantic analysis of natural language texts]. *35th Conference "Information Technologies and Systems" (ITiS'12) Proceedings*. Moscow: IPPI RAN. 152–161.
15. Sidorova, E. A. 2008. Instrumental'nye sredstva faktograficheskogo analiza dokumentov v informatsionnykh sistemakh, osnovannykh na ontologiyakh [Factographic text analysis tools in information systems based on ontologies]. *Novosibirsk State University J. Information Technologies* 6(3):126–134.
16. Sviatogor, L. A., and V. P. Gladun. 2010. Mashinnoe ponimanie tekstov estestvennogo yazyka: Ontologicheskaya paradigma [Machine understanding of natural language texts: An ontological paradigm]. *Artificial Intelligence* 3:249–258.
17. Orobinska, E. A., and A. Yu. Doroshenko. 2011. Ispol'zovanie ontologiy dlya avtomaticheskoy obrabotki tekstov na estestvennom yazyke [Ontologies using for natural language texts automatic processing]. *Sb. nauch. tr."Vestnik NTU "KhPI"* [NTU "KhPI" Proceedings]. Kharkov. 30:101–106.
18. Barakhnin, V. B., and A. M. Fedotov. 2013. Postroenie modeli faktograficheskogo poiska [A model of factographic retrieval]. *Novosibirsk State University J. Information Technologies* 11(4):16–27.
19. Slobodyuk, A. A., S. I. Matorin, and S. N. Chetverikov. 2013. O podkhode k sozdaniyu ontologiy na osnove sistemno-ob"ektnykh modeley predmetnoy oblasti [About approach for building ontologies based on UFO domain models]. *Nauchnye Vedomosti BelGU* [Belgorod State University Scientific Bulletin] 22(28/1):159–167.
20. Naikhanova, L. V. 2008. *Tekhnologiya sozdaniya metodov avtomaticheskogo postroeniya ontologiy s primeneniem geneticheskogo i avtomatnogo programmirovaniya* [The

technology of automatic ontology building with genetic and automate programming methods creation]. Ulan-Ude: The Buryat Scientific Center of SB RAS. 244 p.

Received December 18, 2015

Contributors

Adamovich Igor M. (b. 1934) — Candidate of Science (PhD) in technology, Head of Department, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; Adam@amsd.com

Volkov Oleg I. (b. 1964) — leading programmer, Institute of Informatics Problems, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 44-2 Vavilov Str., Moscow 119333, Russian Federation; Volkov@amsd.com