Имитационная модель компьютерного анализа фактов

Г. Л. Топровер, С. Л. Киселев

В статье представлена формальная модель алгоритмического анализа фактов на базе имитационного подхода с применением методов теории распознавания образов.

Ключевые слова: фактология, анализ фактов, фактография, управление знаниями.

В настоящей статье рассмотрен один из способов формализации процесса компьютерного анализа фактов и обоснована возможность построения *фактоаналитической программной системы* (ФПС), по эффективности не уступающей профессиональному Эксперту-аналитику с учетом ограничений, обусловленных несовершенством современных лингвистических и аналитических алгоритмов. Описываемая модель применяется в качестве основы математического обеспечения системы управления фактографической информацией «XFiles» компании «Ай-Теко» [9].

Для удобства ссылок материал всех статей содержит сквозную алфавитно-цифровую нумерацию глав с вложенной нумерацией разделов.

A. Вводные замечания

В современном Интернете, включая электронные СМИ, социальные сети, коммерческие и частные сайты, ежедневно публикуются сообщения, содержащие сотни тысяч *фактов* самой разной значимости: от глобального (*Катастрофа в Японии...*), государственного (*Президент России высказал...*) и регионального (*Сегодня в Москве...*) до корпоративного (*Фирма X объявила о...*), клубного (*«Вышел новый*

альбом. . . ») и личного (*«Я провел день. . . »*). При этом даже невооруженным глазом заметна высокая комплиментарность этого массива данных, когда представленные факты синергетически подтверждают, дополняют и развивают друг друга. К сожалению, естественноязыковая природа публикаций и, главным образом, отсутствие единой интерпретационной модели всерьез затрудняют консолидацию отдельных фактов и их дальнейшую интеграцию в структуры знаний более высокого порядка.

Подавляющее большинство современных работ в области компьютерной обработки фактов ([1, 2, 5, 6, 8] и другие) посвящены вопросам эффективного поиска, фильтрации, верификации и организации фактов, а также методам извлечения релевантной информации из текстовой или иной формы их представления. Однако скурпулезный сбор фактических данных является лишь предварительным этапом решения гораздо более общей *задачи анализа фактов*.

В самом деле посмотрим на новостные заголовки: «Мэрия Самары разместила заказ. . . » или «Президент Медведев выступил. . . » — это факты. Полноценный анализ фактов подразумевает их лингвистическую, статистическую и онтологическую обработку, сопоставление друг с другом и с внешними источниками знаний, выявление связей и зависимостей между ними, многоуровневую редукцию и т. д. Результатом этой сложной работы становятся обобщающие утверждения, имеющие прямую практическую ценность для пользователя. Достоверность и полнота таких утверждений является критерием качества аналитического механизма.

С учетом технических условий и опыта эксплуатации ранних версий ФПС «XFiles», нам представляется целесообразным использовать *имитационный подход* [3] с применением методов ориентированной лингвистики и теории распознавания образов.

Б. Фактографические принципы

Б1. Вне зависимости от значения слова «факт» в философии или в бытовом обиходе, в нашей системе *факт* по существу информационный объект с идентификатором и набором системных дескрипторов, как технологических (например, «время регистрации факта в

системе»), так и смысловых («субъект», «объект», «аттрибут» и т. п.) Значения смысловых дескрипторов извлекаются из исходного информационного материала (текста, записей БД и пр.), а их совокупность уникально определяет факт в пространстве характеристик.

Б2. ***Характеристика*** представляет собой ограниченную дискретную величину F, принимающую значения из перечислимого и, в общем случае, неупорядоченного множества $\{F\} = \{f_1, \ldots, f_{N_F}\}$. Отметим, что характеристика может принимать и ко́мплексные значения — например, в случае характеристики «объект» для факта с множественными объектами.

Б3. Характеристики факта, извлекаемые из источника этого факта и сохраняемые системой в качестве смысловых дескрипторов, будем называть *базовыми* (или *фактографическими*) в противоположность *производным* характеристикам, которые будут определены ниже (см. раздел В1). Множество всех базовых характеристик $\{\Phi\} = \{F^1, \ldots, F^{M_\Phi}\}$ образует ***фактографическое пространство*** Φ, которое является ограниченным в силу ограниченности областей значений составляющих его характеристик и, в общем случае, неполным по причине их возможной взаимозависимости.

Б4. ***Факт***, таким образом, определяется как точка в фактографическом пространстве:

$$\varphi \in \Phi : \ \varphi = (f^1, \ldots, f^M), \text{ где } f^i \in \{F^i\}, \ \ 1 \leqslant i \leqslant M_\Phi.$$

Все факты системы образуют множество $\{\varphi\}$, которое мы будем называть *коллекцией фактов*.

Естественно было бы определить в пространстве Φ *отношение тождественности*

$$\varphi_1 \equiv \varphi_2 : \ f_1^i = f_2^i, \ 1 \leqslant i \leqslant M_\Phi,$$

то есть равенство значений всех характеристик фактов определяет идентичность этих фактов.

Б5. Интуитивно ясно, что фактографические пространства разных ФПС (или разных версий одной ФПС) отличаются своей описательной силой, то есть степенью детализации представления факта.

В дальнейших построениях мы будем полагать, что фактографическое пространство Φ нашей системы имеет *адекватную описательную силу*, то есть равнозначные исходные описания отображаются в одну точку пространства Φ, в то время как разнозначные исходные описания отображаются в разные точки этого пространства. Построение *априорно адекватного* фактографического пространства является нетривиальной задачей, если вообще разрешимой — в практическом плане приходится полагаться на здравый смысл, сходимость итеративного процесса совершенствования системы и стационарность сред, производящих исходные информационные материалы.

Б6. Отметим, что фактографическое пространство само по себе не дает нам формализма для аргументированного аналитического обобщения фактов, поэтому далее мы рассмотрим способы преобразования этого пространства в более удобную для анализа форму.

В. Фактологические принципы

В1. Возьмем произвольную функцию H, определенную на фактографическом пространстве: $h = H(\varphi)$, $\varphi \in \Phi$. Эта функция задает дискретную величину H со множеством значений $\{H\} = \{h_1, \ldots, h_{N_H}\}$, причем N_H не превышает мощности множества допустимых значений пространства Φ. Такую величину H мы будем называть производной (или *фактологической*) характеристикой.

В2. Произвольная совокупность фактологических характеристик $\{\Psi\} = \{H^1, \ldots, H^{M_\Psi}\}$ образует ***фактологическое пространство*** Ψ, которое является ограниченным и, в общем случае, неполным по причинам, указанным в разделе Б3. Отношение тождественности точек ψ_1 и ψ_2 в пространстве Ψ определяется также, как в разделе Б4, то есть $\psi_1 \equiv \psi_2 :\ h_1^i = h_2^i$, $1 \leqslant i \leqslant M_\Psi$.

При этом соответствующая совокупности $\{\Psi\}$ комбинация фактологических функций вида $H(\varphi)$ определяет, очевидно, *однозначное* ***фактологическое преобразование*** пространств $\Psi :\ \Phi \to \Psi$, то есть любому *базовому факту* $\varphi \in \Phi$ соответствует ровно один *производный факт* $\psi = \Psi(\varphi)$, $\psi \in \Psi$. Обратное не всегда верно: два нетож-

дественных факта $\varphi_1 \neq \varphi_2$ могут отображаться в один производный факт $\Psi(\varphi_1) = \Psi(\varphi_2)$.

В3. Базовые факты φ_1 и φ_2 такие, что $\Psi(\varphi_1) = \Psi(\varphi_2)$ будем далее называть ***подобными*** *по фактологическому преобразованию* Ψ, обозначая это отношение $\varphi_1 \sim_\Psi \varphi_2$. С учетом однозначности отображения $\Psi(\varphi)$, коллекция фактов $\{\varphi\}$ разбивается на *классы подобия* $\Omega_1, \ldots, \Omega_{K_\Psi}$, так что $\bigcup \Omega_i = \{\varphi\}$, $1 \leqslant i \leqslant K_\Psi$ и $\Omega_i \cap \Omega_j = \varnothing$, $\forall 1 \leqslant i, j \leqslant K_\Psi$, то есть любой базовый факт $\varphi \in \Phi$ принадлежит ровно к одному классу Ω_i. Множество всех классов подобия $\{\Omega\}_\Psi = \{\Omega_1, \ldots, \Omega_{K_\Psi}\}$ назовем ***классификацией*** *коллекции фактов* $\{\varphi\}$ *по преобразованию* Ψ. Для удобства обозначений введем *классифицирующий оператор* $\Omega(\Psi, \{\varphi\}) : \{\varphi\} \to \{\Omega\}_\Psi$. Две классификации $\Omega(\Psi_1, \{\varphi\})$ и $\Omega(\Psi_2, \{\varphi\})$, образованные преобразованиями $\Psi_1(\varphi)$ и $\Psi_2(\varphi)$ соответственно, будем полагать *тождественными* при условии равенства $\{\Omega\}_{\Psi_1} = \{\Omega\}_{\Psi_2}$ в обычном теоретико-множественном смысле. Понятно, что $\Omega(\Psi_1, \{\varphi\}_1) = \Omega(\Psi_2, \{\varphi\}_2)$ возможно только при $\{\varphi\}_1 = \{\varphi\}_2$ и, кроме того, $\Omega(\Psi, \{\varphi\}_1) = \Omega(\Psi, \{\varphi\}_2) \Leftrightarrow \{\varphi\}_1 = \{\varphi\}_2$. Учитывая, что любое преобразование $\Psi(\varphi)$ задает единственную классификацию на заданной коллекции фактов, условие $\Omega(\Psi_1, \{\varphi\}) = \Omega(\Psi_2, \{\varphi\})$, $\forall\{\varphi\}$ определяет *тождественность преобразований* $\Psi_1(\varphi)$ и $\Psi_2(\varphi)$.

В4. Дополним понятие фактологической функции так, чтобы областью ее определения могло быть не только базовое фактографическое пространство Φ, но и любое фактологическое пространство Ψ, образованное другими фактологическими функциями. Таким образом, функция $g = G(\psi)$, $\psi \in \Psi$, определенная на фактологическом пространстве $\Psi : \Phi \to \Psi$ задает дискретную величину $G = \{g_1, \ldots, g_{N_G}\}$, которую уместно называть *производной характеристикой второго порядка*. Соответственно определяются фактологическое пространство Γ и фактологическое преобразование $\Gamma : \Psi \to \Gamma$ второго порядка, и далее более высоких порядков.

В5. Сформулируем одно важное свойство фактологических классификаций, которое потребуется нам в дальнейших рассуждениях. Пусть на фактографическом пространстве Φ определены несколько (пусть P) произвольных фактологических преобразований $\Psi^1, \ldots, \Psi^P$, образующих P фактологических пространств Ψ^1 :

$\Phi \to \Psi^1, \ldots, \Psi^P : \Phi \to \Psi^P$, которые в свою очередь задают P классификаций заданной коллекции фактов: $\Omega(\Psi^1, \{\varphi\}), \ldots, \Omega(\Psi^P, \{\varphi\})$. Нетрудно доказать, что пространство $\Psi = \Psi^1 \cup \ldots \cup \Psi^P$, получаемое объединением этих пространств, определяет классификацию $\Omega(\Psi, \{\varphi\}) = \Omega(\Psi^1, \{\varphi\}) \times \ldots \times \Omega(\Psi^P, \{\varphi\})$, которая является комбинаторной композицией их классификаций. По существу это означает, что редукция классов в фактологических классификациях имеет последовательно-итеративный характер — методический и практический смысл этого свойства мы рассмотрим в разделе Д5.

Г. Аналитические тезисы и алгоритмы

Г1. Проведем мысленный эксперимент. Допустим, в нашем распоряжении имеется коллекция фактов $\{\varphi\}$ и мы передаем ее для анализа авторитетному Эксперту-аналитику (или целому агенству). Результатом работы Эксперта станут *аргументированные умозаключения* о тенденциях, отношениях и явлениях, которые не обозначены в фактах эксплицитно, но четко проявляются при их группировании и сопоставлении определенным образом, например (в зависимости от темы нашей коллекции): «Руководство РФ проявляет повышенный интерес к...», «Банк X скоро усилит свои позиции в...» или даже «Этот новый пылесос удобен, но не долговечен» — при том, что с точки зрения непрофессионала в исходных фактах не было даже намека на подобные утверждения! Вместе с тем, высокий профессиональный уровень Эксперта, то есть сложнейший агрегат его опыта, образования и эрудиции вкупе с глубоким знанием реалий заданной области, заставляет нас относиться к этим умозаключениям с должным уважением и руководствоваться ими в принятии решений. В соответствии с логикой имитационного подхода, целью и главным критерием успеха разработки ФПС является способность системы делать подобные выводы автоматически, без участия Эксперта.

Разберемся, что представляют из себя умозаключения Эксперта, которые мы в дальнейшем будем называть ***тезисами***. Для иллюстрации наших рассуждений возьмем общепонятный пример: пусть это будет система сбора и анализа фактов в предметной области бытовой техники, где роль источников фактической информации иг-

рают описания товаров, данные продаж, отзывы покупателей, журнальные обзоры и т. п.

Во-первых, можно считать, что тезис носит характер онтологического предиката, то есть является *логическим утверждением* о наличии качественных свойств объектов предметной области и отношений между ними Так, в системе анализа бытовой техники непременно существует тезис о долговечности товара. Заявление Эксперта «По имеющимся данным пылесос X долговечен» реализует этот тезис как истинный по отношению к пылесосу X, а «Отзывы указывают на недолговечность пылесоса X» — как ложный.

Во-вторых, умозаключения Эксперта врядли будут безапелляционными — оценка степени надежности анализа является важной частью аналитической работы и имеет существенное значение для заказчика. Именно этим объясняется вероятностная модальность реальных аналитических выводов: «Факты *отчетливо указывают* на...», «Факты *дают некоторые основания* думать, что...» и т. д. Таким образом, правильнее говорить о тезисе, как об утверждении нечеткой логики [4], где в роли показателя истинности выступает та самая оценка достоверности, измеряемая обычно в диапазоне $[0, 1]$.

В-третьих, грамотный Эксперт не позволит себе делать «голословные» утверждения, а всегда подкрепляет их подборкой подтверждающих и/или опровергающих фактов: «Факты №№... дают основания полагать, что...» и тому подобное. В нашей модели такая подборка будет ничем иным как *классификацией* фактов (см. раздел В3).

В-четвертых, естественно предположить, что профессиональный Эксперт сначала производит классификацию фактов по критериям тезиса (гипотезы), а затем использует ее для расчета истиности этого тезиса, но не наоборот — сначала высказывается об истиности или ложности тезиса, а потом подгоняет факты под этот вывод. Это означает, что показатель истиности тезиса на заданной коллекции фактов следует считать функцией классификации фактов для данного тезиса.

Итак, в рамках настоящего исследования постановим, что любой тезис $T(\{\varphi\})$, высказываемый Экспертом на коллекции фактов $\{\varphi\}$, представляет собой комбинацию трех элементов:

1) *онтологического предиката* τ;

2) *аргументирующей классификации* фактов $\{\Omega\}_\tau(\{\varphi\})$;

3) *показателя истинности* $D_\tau(\{\Omega\})$,

что можно суммарно записать в следующем виде: $T(\{\varphi\}) : [\tau, \{\Omega_\tau\}(\{\varphi\}), D_\tau(\{\Omega\})]$.

Г2. Аналитический тезис по определению происходит от Эксперта-человека, а следовательно является неформализуемым концептом. По этой причине важные для нас свойства тезисов придется представить в виде постулатов.

1) Множество тезисных предикатов $\{\tau\}$ *конечно и не зависит от конкретной коллекции фактов:* $\forall\{\varphi\}^* \supset \{\varphi\}$: $\{\tau\}^* = \{\tau\}$.

Действительно, любое умозаключение τ, высказанное на коллекции фактов $\{\varphi\}$, может быть высказано и для более широкой коллекции $\{\varphi\}^* \supset \{\varphi\}$, хотя возможно и с другим значением истинности. Иными словами, каждый вновь прибывающий факт может влиять на меру истинности тезисного предиката τ, но не делает его бессмысленным. Это означает, что тезисы являются принадлежностью фактографического пространства Φ а не существующей в нем коллекции фактов $\{\varphi\}$. В нашем примере, предикат долговечности пылесоса X имеет смысл (то есть *исчислим*), даже если в системе нет пока ни одного отзыва об этом пылесосе. В этом случае наш Эксперт делает вывод, что долговечность имеет значение «неопределено». Другое дело, что в реальном аналитическом отчете такие заявления опускаются для экономии места, но Эксперт готов к вопросу клиента «Что вы думаете насчет долговечности пылесоса X?».

2) Для каждого тезисного предиката τ *существует фактологическое преобразование* $\Psi_\tau(\varphi)$*, задающее классификацию любой коллекции фактов* $\{\varphi\}$ *тождественную тезису* $T(\{\varphi\})$*:*

$$[\tau, \{\Omega_\tau\}(\{\varphi\}), D_\tau(\{\Omega\})] : \ \forall\tau\ \exists\Psi_\tau(\varphi) : \ \Omega(\Psi_\tau\{\varphi\}) = \{\Omega_\tau\}(\{\varphi\}),\ \forall\{\varphi\},$$

и хотя бы одна оценочная функция $\Delta_\tau(\{\Omega\})$ *тождественная показателю истинности* $D_\tau(\{\Omega\})$*:*

$$\forall\tau\ \exists\Delta_\tau(\{\Omega\}) : \ \Delta_\tau(\{\Omega\}) = D_\tau(\{\Omega\}),\ \forall\{\Omega\}.$$

Этим постулатом заявляется вполне очевидная алгоритмическая представимость тезиса, то есть существование алгоритма, тождественного тезису $T(\{\varphi\})$ в смысле соответствия «данные-результат». Иначе пришлось бы признать, что Эксперт не копит личный опыт и строит способы классификации и/или оценки истиности тезиса *ad hoc* для каждой коллекции фактов $\{\varphi\}$.

Действительно, процесс исчисления Экспертом-человеком отнологического предиката в тезис $T(\{\varphi\})$ хоть и не может быть эксплицирован в общем случае, но полностью удовлетворяет всем основным критериям алгоритмичности:

1) Детерминированность — одинаковые коллекции фактов Эксперт прокомментирует одинаково: $\{\varphi\}^* = \{\varphi\} \Rightarrow T(\{\varphi\}) = T(\{\varphi\}^*)$.
2) Результативность — тезис всегда имеет результатное значение, пусть даже «неопределено».
3) Универсальность — тезис исчислим на любой коллекции фактов, включая пустую коллекцию.

Итак, для любого тезиса $T(\{\varphi\}) : \ [\tau\{\Omega_\tau\}(\{\varphi\}), D_\tau(\{\Omega\})]$ существует *алгоритм* $\Theta(\{\varphi\}) : \ [\tau, \Psi_\tau(\varphi), \Delta_\tau(\{\Omega\})]$, который на произвольной коллекции фактов $\{\varphi\}$ вычисляет классификацию фактов и показатель истинности предиката τ, идентичные данному тезису $T(\{\varphi\}) = \Theta(\{\varphi\})$:

$$\forall\{\varphi\} : \ \Omega(\Psi_\tau\{\varphi\}) = \Omega_\tau(\{\varphi\}) \wedge \Delta_\tau(\Omega(\Psi_\tau\{\varphi\})) = D_\tau(\Omega_\tau(\{\varphi\})).$$

Вообще говоря, таких алгоритмов может быть несколько, но по всей вероятности, только один будет соответствовать «человеческой» процедуре. При этом совсем необязательно, что именно он будет самым оптимальным для практической реализации.

Г3. Постулаты предыдущего раздела позволяют заявить, что для произвольного заданного фактографического пространства Φ можно построить конечное множество алгоритмов $\{\Theta\}_\Phi$ такое, что совокупный результат вычисления этих алгоритмов $\{\Theta\}_\Phi(\{\varphi\})$ на любой заданной коллекции фактов $\{\varphi\} \in \Phi$ будет неотличим от полного множества тезисов $\{T\}_\Phi(\{\varphi\})$, представленных Экспертом на той же коллекции фактов.

Задача анализа фактов состоит в построении *информативных обобщений* имеющихся фактов, то есть *тезисов* в нашей терминологии. Эксперт справляется с этой задачей наилучшим образом по определению, а значит важнейшим показателем эффективности ФПС становится совпадение результатов ее работы с результатами Эксперта. В итоге, разработка системы сводится к построению множества алгоритмов $\{\Theta\}$, оптимального в смысле интегрального критерия вида $Q(\{\Theta\}) = \int_{\{\varphi\}} \Sigma_{\{\tau\}} W_\tau(\Delta_\tau(\{\varphi\}) - D_\tau(\{\varphi\}))$, где $W_\tau(x)$ — функция стоимости ошибки вычисления тезиса τ.

Поясним вышесказанное. Дело в том, что клиенту-заказчику по большому счету все равно, кто делает анализ фактов и формулирует тезисы: человек-аналитик, автомат или их комбинация — его живо интересует только соотношение «скорость-цена-качество». На сегодняшний день, мы полагаем, ответственную факто-аналитическую работу на заказ выполняют все-таки люди: специалисты-аналитики, частные детективы и иногда даже секретари — все зависит от запросов и кошелька клиента. Получается либо «хорошо, но долго и дорого», либо «дешево и быстро, но плохо» и т. п. Наша цель заключается в создании компьютерных средств, способных взять на себя если не всю, то хотя бы часть работы и тем самым улучшить упомянутое соотношение «скорость-цена-качество». Материал разделов Г1–Г3 позволяет думать, что эта цель теоретически достижима.

Г4. Тут неизбежно возникает вопрос: раз для оценки качества тезисов системы на каждой коллекции фактов все равно не обойтись без присутствия Эксперта (причем, самого лучшего), не проще ли сразу поручить эту работу ему? Однако, постулат 1 (раздел Г2) дает основания считать, что множество исчислимых тезисов $\{\tau\}$ универсально в рамках предметной области. Следовательно, мы можем ожидать, что система алгоритмов $\{\Theta\}$, построенная и отлаженная при участии Эксперта (в ТРО — «обученная») для одного заказчика, окажется работоспособной (с минимальной доводкой) в среде другого заказчика, оперирующего в той же предметной области[1] Более того, пе-

[1]В качестве гипотезы выскажем следующее, пока не доказанное утверждение: *Пусть в операционной среде* Λ *определено фактографическое пространство* Φ *с описательной силой* $\Im_\Lambda(\Phi)$*, на котором построена система аналитических алгоритмов* $\{\Theta\}$*, дающая в среде* Λ *качество* $Q_\Lambda(\{\Theta\})$*. Если в другой операционной*

ренос системы в технически отличную, но родственную предметную область окажется не столь болезненным, как ее построение с нуля. Таким образом, один раз подготовленная система оправдает расходы на привлечение Эксперта многократной установкой в разных средах.

Например, нам удалось построить эффективную систему мониторинга дебиторов для банка А с фактографическими характеристиками вроде «декларируемый доход», «сумма выплат», «места отдыха», «марка автомобиля», а также тезисами типа «дебитор скрывает доходы» и «дебитор приближается к границе риска невыплаты». Очевидно, что такая система сможет работать и в другом банке Б, причем перенастройка системы сведется, по большому счету, к замене множества значений характеристик «имя» и т. п. Причина такой универсальности в том, что операционные и административные схемы современных организаций не создаются ad hoc, а наоборот, сильно стандартизированы и отображаются друг на друга с точностью до названий департаментов, стоимости услуг и фамилий руководства.

Итак, все сказанное в разделах Г1–Г4 утверждает *теоретическую возможность* построения эффективной ФПС, но не предлагает конструктивных путей решения этой задачи. Этому вопросу посвящена следующая глава нашей работы.

Д. Практическая задача

Д1. Признаем, что выдвинутая в разделе Г3 заявка на создание *автоматической* ФПС, полностью заменяющей Эксперта-аналитика, кажется на данном этапе преждевременной. Эту общую задачу мы обозначим как стратегическую, но пока неосуществимую и переведем рассмотрение в более практическую плоскость. Как обсуждалось ранее, аналитическая работа с фактами имеет целью построение полез-

---

среде Λ^* *пространство* Φ *обладает описательной силой* $\Im_{\Lambda^*}(\Phi) \geqslant \Im_{\Lambda}(\Phi)$*, то качество этой системы алгоритмов в новой среде* $Q_{\Lambda^*}(\{\Theta\}) \geqslant Q_{\Lambda}(\{\Theta\})$. Понятие описательной силы применительно к фактографическим пространствам обсуждалось в разделе Б5, но вопрос о ее количественной мере был отложен. Термин «оперативная среда» заимствован из ориентированной лингвистики, здесь его можно понимать как «множество всех возможных фактов с приписанным к нему Экспертом».

ных для клиента умозаключений-тезисов, для каждого из которых требуется выполнить два основных действия:

1) выделить релевантные тезису классы фактов,
2) интерпретировать эту классификацию в аналитическую гипотезу.

Мы предлагаем на текущем этапе сосредоточиться на компьютеризации первого из этих действий, поскольку именно подбор, сопоставление и сортировка фактов в больших массивах содержат много утомительной рутины, которую было бы правильно переложить на компьютер. Таким образом, назначение разрабатываемой нами программной системы формулируется как *инструмент построения фактологических классификаций*, то есть не автономный анализатор, а техническое средство, которое снимает с оператора-аналитика «грязную работу» и позволяет ему сфокусироваться на творческой деятельности.

Надо заметить, что такое «сужение» целей совсем не означает, что мы отказываемся от более амбициозной задачи построения полностью автоматического анализатора фактов, как обсуждалось в разделе Г3. Наоборот, как мы увидим ниже (см. шаг № 6 в схеме следующего раздела), этот подход позволяет набрать опыт и подготовиться к полной автоматизации.

Д2. Итак, в свете новой формулировки задачи рабочий цикл нашей аналитической системы выглядит так.

1. Система получает документы или другие источники фактов, выделяет в них индивидуальные факты и вычисляет их базовые характеристики $\{\Phi\} = \{F^1, \ldots, F^{M_\Phi}\}$. Тем самым, каждый исходный факт транслируется в точку φ фактографического пространства Φ, а все они образуют коллекцию фактов $\{\varphi\}$.
2. Система оборудована набором заранее подготовленных аналитических алгоритмов $\{\Theta\}$ (их происхождение обсуждается в следующем разделе). Каждый такой алгоритм $\Theta \in \{\Theta\}$ состоит из классификатора фактов $\Psi_\Theta(\{\varphi\})$ и критериальной функции $\Delta_\Theta(\{\Omega\})$.

3. Система поочередно запускает алгоритмы из множества $\{\Theta\}$ и для каждого Θ:

   3.1. Классификатор Ψ_Θ вычисляет одну или более функций вида $H(\varphi)$ — фактологических характеристик — и, следовательно, реализует некое фактологическое преобразование $\Psi_\Theta : \Phi \to \Psi_\Theta$. Результатом его работы становится некая классификация фактов $\Omega(\Psi_\Theta\{\varphi\})$.

   3.2. Критерий Δ_Θ обрабатывает полученное множество классов $\{\Omega\} = \Omega(\Psi_\Theta\{\varphi\})$ и выносит свое решение об информативности этой классификации.

   3.3. В случае положительного решения $\Delta_\Theta(\Omega(\Psi_\Theta\{\varphi\}))$ данная классификация $\{\Omega\}$ предъявляется Эксперту-оператору в удобной форме, графической или табличной.

   3.4. Эксперт-оператор оценивает предъявленную классификацию и либо

      3.4.1. Признает ее информативной и использует для построения аналитического тезиса $[\tau, \{\Omega\}, D]$. Алгоритм Θ в этом случае помечается как качественный на коллекции фактов $\{\varphi\}$, $Q(\Theta\{\varphi\}) = 1$, а тезис регистрируется системой. Либо

      3.4.2. Признает ее нерелевантной или ошибочной и отбрасывает. Тогда алгоритм Θ помечается как некачественный для коллекции $\{\varphi\}$, $Q(\Theta\{\varphi\}) = 0$, а разработчик получает сигнал об этом.

4. Эксперт-оператор использует презентационные средства системы (диаграммы, таблицы, графы и пр.) для классификации и анализа фактов в ручном режиме. При этом, возможно, он выстроит хотя бы одну новую классификацию фактов $\{\Omega\}$, упущенную системой на шаге № 3 и аргументирующую некий тезис $[\tau, \{\Omega\}, D]$. В этом случае разработчик получает соответствующий сигнал с полной информацией о новом тезисе.

5. Эксперт-оператор переносит все тезисы $\{T\}$ выстроенные на данной коллекции фактов в окончательный аналитический отчет для клиента.

6. Разработчик совершенствует набор алгоритмов $\{\Theta\}$ с учетом сигналов, полученных от системы на шагах № 3.4.2 и № 4. Кроме того, по информации с шагов № 3.4.1 и № 4 разработчик устанавливает соответствия между множеством алгоритмов $\{\Theta\}$ и множеством тезисов $\{T\}$ — эти данные будут потом использованы для автоматизации построения тезисов.
7. Цикл повторяется либо
   - 7.1. с шага № 1 на другой коллекции фактов $\{\varphi\}^*$, расширенной $\{\varphi\} \subset \{\varphi\}^*$ или новой $\{\varphi\} \not\subset \{\varphi\}^*$, либо
   - 7.2. с шага № 2 с обновленным на шаге № 6 набором алгоритмов $\{\Theta\}^*$.

Теоретические принципы, изложенные в разделах Г1–Г3, позволяют нам надеяться, что итеративный процесс совершенствования системы будет сходящимся, так что со временем шаг № 6 станет лишним и система будет готова к автономной эксплуатации.

Д3. Вернемся к схеме предыдущего раздела в момент, когда множество аналитических алгоритмов $\{\Theta\}$ еще пусто. Здесь особое значение приобретает шаг № 4, где мы должны предоставить Эксперту-оператору удобные и эффективные средства для построения классификации фактов. Если пока оставить в стороне технические и пользовательские аспекты «удобства» и «эффективности», то речь идет о предъявлении оператору одной или нескольких классификаций-кандидатов, из которых он выберет действительно значащие.

Д4. Возьмем фактологическое преобразование $\Psi : \Phi \to \Psi$, задающее на коллекции фактов $\{\varphi\}$ классификацию $\Omega(\Psi, \{\varphi\}) = \{\Omega_1, \ldots, \Omega_N\}$. Допустим, известны *априорные вероятности* $P_i = P(\varphi \in \Omega_i)$, $1 \leqslant i \leqslant N$ того, что произвольный факт φ попадает в класс Ω_i. С другой стороны, на достаточной большой конкретной коллекции $\{\varphi\}$ вычислимы *апостериорные вероятности* $P_i^* = P(\varphi \in \Omega_i/\{\varphi\})$ того же события. Как правило, значимое отклонение P_i^* от P_i является индикатором информативной классификации, что формально выражается интегральным критерием такого, например, вида: $\Sigma_i[(P_i^* - P_i)/P_i(1 - P_i)]^2/N \geqslant \varepsilon$, $1 \leqslant i \leqslant N$.

Д5. Предложенный выше способ оценки информативности классификации $\Omega(\Psi, \{\varphi\})$, а значит и задающего ее фактологического преобразования $\Psi : \Phi \to \Psi$, требует информации об априорных вероятностях $P_i = P(\varphi \in \Omega_i)$, что далеко не всегда возможно на практике для преобразований высших порядков (раздел В4). Однако, это требование выглядит вполне реалистичным для фактологических преобразований первого порядка.

Рассмотрим некоторую фактологическую функцию $H(\varphi)$, $\varphi \in \Phi$ со множеством значений $\{H\} = \{h_1, \ldots, h_N\}$. Часто (хотя не всегда), существует возможность оценить априорные вероятности $P(h_i)$, $1 \leqslant i \leqslant N$ того, что функция $H(\varphi)$ принимает значение h_i на произвольном факте φ. Утверждение, сделанное в разделе В5, дает основания полагать, что отклонение апостериорных вероятностей $P(h_i/\{\varphi\})$ на конкретной коллекции $\{\varphi\}$ от априорных $P(h_i)$ является индикатором информативной классификации вокруг соответвующих значений h_i.

Д6. Итак, можно предложить следующую процедуру выдвижения классификаций-кандидатов на шаге № 4 схемы раздела Д2. Пусть заранее определено множество «базовых» фактологических функций $\Re$ таких, что для любой такой функции $H(\varphi) \in \Re$ известны априорные вероятности $P(H(\varphi) = h_i)$ получения значения h_i из полного множества значений $\{H\} = \{h_1, \ldots, h_{N^H}\}$ на произвольном факте φ. Для простоты ограничимся только функциями с булевым множеством значений $\{H\} = \{1, 0\}$, где ситуация $H(\varphi) = 1$ имеет смысл истинности некого онтологического условия, приписанного данной функции.

4.1. Система производит перебор функций $H(\varphi) \in \Re$ на данной коллекции фактов $\{\varphi\}$ объемом $M(\{\varphi\})$. Для каждой такой функции:

4.1.1. Вычисляются значения $H(\varphi)$ для всех фактов $\varphi \in \{\varphi\}$. Строится множество фактов $\{\psi\} \subset \{\varphi\}$, получивших значение 1: $H(\psi) = 1$, $\forall \psi \in \{\psi\}$.

4.1.2. Вычисляется критерий $\Sigma(H, \{\varphi\}) = [(P^*_H - P_H)/P_H(1 - P_H)]^2$, P_H — априорная вероятность $P(H(\varphi) = 1)$, P^*_H — отношение $M(\{\psi\})/M(\{\varphi\})$.

4.1.3. Полученное значение $\Sigma(H, \{\varphi\})$ сравнивается с некоторым эмпирическим порогом δ_H.

4.1.4. В случае $\Sigma(H, \{\varphi\}) < \delta_H$ система переходит к шагу 4.1.1 со следующей $H(\varphi)$.

4.1.5. В случае $\Sigma(H, \{\varphi\}) \geqslant \delta_H$ система переходит на *следующий уровень рекурсии*, а именно выполняет шаги 4.1.1–4.1.5 с перебором других функций из $\Re$ на выделенном подмножестве фактов $\{\psi\}$.

Конкретно, система начинает перебор функций $J(\varphi) \in \Re$, $J(\varphi) \neq H(\varphi)$. Для каждой такой функции:

4.1.5.1. Вычисляются значения $J(\psi)$ для всех фактов $\psi \in \{\psi\}$. Строится множество фактов $\{\chi\} \subset \{\psi\}$, получивших значение 1: $H(\chi) = 1$, $\forall \chi \in \{\chi\}$.

4.1.5.2. Вычисляется критерий $\Sigma(J, \{\psi\}) = [(P_J^* - P_J)/P_J(1 - P_J)]^2$, P_J — априорная вероятность $P(J(\psi) = 1)$, P_J^* — отношение $M(\{\chi\})/M(\{\psi\})$.

4.1.5.3. Полученное значение $\Sigma(J, \{\psi\})$ сравнивается с некоторым эмпирическим порогом δ_J.

4.1.5.4. В случае $\Sigma(J, \{\psi\}) < \delta_J$ система переходит к шагу 4.1.5.1 со следующей $J(\varphi)$.

4.1.5.5. В случае $\Sigma(J, \{\psi\}) \geqslant \delta_J$ система переходит на *следующий уровень рекурсии.*

.....

4.1.5.6. Система переходит к шагу 4.5.1 со следующей функцией $J(\varphi)$.

4.1.6. Система переходит к шагу 4.1.1 со следующей функцией $H(\varphi)$.

4.2. Все обнаруженные системой «статистически значимые» комбинации базовых фактологических функций из множества $\Re$ вида $H \times J \times \ldots$ вместе с соответствующими классами фактов вида $\{\varphi\}^* \subset \{\varphi\}$, $H(\varphi) = 1 \wedge J(\varphi) = 1 \wedge \ldots$, $\forall \varphi \in \{\varphi\}^*$ предъявляются Эксперту-оператору через наглядные презентационные средства: диаграммы, графики, таблицы и пр. Далее, для каждой такой комбинации:

4.2.1. Если Эксперт не признает данную классификацию $\Omega = \{\varphi\}^*$ информативной, она отбрасывается.

4.2.2. Если Эксперт использует данную классификацию $\Omega = \{\varphi\}^*$ для построения некоторого аналитического тезиса $[\tau, \{\Omega\}, D]$ — либо напрямую, либо в качестве «базы» для более тонкого анализа — информация об этом тезисе передается разработчику для дальнейшего усовершенствования системы.

Д7. В предыдущем разделе предполагалось существование множества $\Re$ базовых фактологических функций с известным распределением априорных вероятностей. Мы полагаем, что универсального рецепта построения таких функций не существует, но интуитивно ясно, что в качестве строительного материала для них логично использовать онтологические данные предметной области, которые дают *обоснованное обобщение значений характеристик* пространства Ф. Сюда относятся географические, социологические, исторические, административные, политические, инженерные и подобные *иерархии*, а также особые области пространства Ф, имеющие самостоятельное онтологическое значение. Вот несколько очевидных примеров:

- Города → регионы → страны (география)
- имена → место рождения / место прежней службы / состав семьи (социология)
- время действия → довоенный / послевоенный (история)
- имена → должности (администрация)
- имена → партии (политика),
- технические объекты → отрасль промышленности (инженерия)

Имеет смысл обратить особое внимание на характеристику «время факта». В силу своей количественной природы, эта характеристика может служить хорошей стартовой точкой для поиска информативных композитных фактологических преобразований (см. пример в разделе Д6). В следующей главе мы рассмотрим эту характеристику более подробно, используя ее как иллюстрацию нашей формальной модели анализа фактов.

Е. Хронологическая регулярность фактов

Поиск хронологических регулярностей в массивах фактов занимают особое место в контексте общей задачи анализа фактов как в силу своей прямой прагматической ценности, так и в качестве отправной точки композитного анализа (см. раздел Д7).

Е1. Пусть в нашем фактографическом пространстве Φ определена базовая характеристика t «время действия факта», $t \in \{\Phi\}$. Помимо дискретности и ограниченности, множество значений этой характеристики $T = \{t_1, \ldots, t_{N^t}\}$ естественно обладает свойством упорядоченности $t_n \leqslant t_{n+1}$. На характеристике t определим ***хронологическую функцию*** $\chi = X(t)$, $t \in T$, с возможными значениями $\{X\} = \{\chi_0, \ldots, \chi_{N^X}\}$. Очевидно, что хронологическая функция $X(t)$ может быть однозначно доопределена до фактологической функции $X(\varphi)$, $\varphi \in \Phi$, а значит она сама по себе выражает преобразование $\Psi_X : \Phi \to \Psi_X$, задающее на коллекции фактов $\{\varphi\}$ классификацию $\Omega(\Psi_X, \{\varphi\})$. Иными словами, функция $X(\varphi)$ опирается только на ось времени и игнорирует все остальные базовые характеристики: например, $\{X\}$ состоит из названий месяцев и $X(\varphi)$ отображает факт φ в название месяца даты этого факта.

Рассмотрим произвольную фактологическую функцию $h = H(\varphi)$, $\varphi \in \Phi$, и соответствующее ей фактологическое преобразование $\Psi_H : \Phi \to \Psi_H$. Объединение функций $H(\varphi)$ и $X(\varphi)$ образует новое пространство $\Psi_{H\times X} : \Phi \to \Psi_{H\times X}$, задающее на коллекции фактов $\{\varphi\}$ комбинаторную классификацию $\Omega(\Psi_{H\times X}, \{\varphi\})$ — см. раздел В5. При этом точное совпадение этой классификации $\Omega(\Psi_{H\times X}, \{\varphi\})$ с исходной хронологической классификацией $\Omega(\Psi_X, \{\varphi\})$ имеет важный смысл:

Фактологическое преобразование Ψ_H назовем ***регулярным*** *по отношению к хронологической функции $X(\varphi)$ на коллекции фактов $\{\varphi\}$, если $\Omega(\Psi_X, \{\varphi\}) = \Omega(\Psi_{H\times X}, \{\varphi\})$.* Соответственно, хронофункция $X(\varphi)$ будет в этом случае называться *регуляризующей* преобразование Ψ_H на коллекции фактов $\{\varphi\}$.

Е2. Общее определение хронологической регулярности может быть наглядно конкретизировано для случая фактологических преобразований булевого типа, задающих только два класса — «знача-

щий» и «нулевой». Положим, что фактологическая функция $H(\varphi)$ выражает какое-либо онтологическое условие, то есть $H(\varphi) = 1$ если факт φ удовлетворяет условию, и $H(\varphi) = 0$ в противном случае. «Значащий» класс $\Omega_H^1 = \{\varphi\}_H \subset \{\varphi\}$ представляет собой подмножество *всех* фактов, удовлетворяющих заданному условию: $H(\varphi) = 1$, $\forall\varphi \in \{\varphi\}_H \wedge H(\varphi) = 0$, $\forall\varphi \notin \{\varphi\}_H$. Например, функция $H(\varphi)$ может выражать условие «факт φ сообщает о выступлении Президента», тогда $H(\varphi) = 1$ для всех фактов, представляющих выступления Президента, и $H(\varphi) = 0$ для всех прочих фактов. Такая функция $H(\varphi)$ — точнее ее фактологическое преобразование Ψ_H — разбивает полную коллекцию фактов $\{\varphi\}$ на два класса Ω_H^1 и Ω_H^0. Подмножество $\{\varphi\}_H$ включает в себя все факты о выступлениях Президента, вне зависимости от темы, даты и времени.

Пусть хронофункция $X(\varphi)$ также имеет булевый результат, то есть $X(\varphi) = 1$, если факт φ удовлетворяет какому-то закону на оси времени, и $X(\varphi) = 0$ в противном случае. Вычисление функции $X(\varphi)$ — точнее, ее Ф-дополнения $X(\varphi)$ — на всех фактах множества $\{\varphi\}_H$, очевидно, разбивает класс Ω_H^1 на два подкласса $\Omega_{H^1\times X^1}$ и $\Omega_{H^1\times X^0}$. Например, $X(\varphi) = 1$ для событий по понедельникам и $X(\varphi) = 0$ для прочих дней недели. После применения этой функции к подмножеству $\{\varphi\}_H$ у нас получатся классы «выступления Президента по понедельникам» и «выступления Президента по другим дням недели».

Если при этом выяснится, что $\Omega_{H^1\times X^1} = \Omega_{H^1}$ и $\Omega_{H^1\times X^0} = \varnothing$, то условие регулярности (раздел Е1) выполнено. Если, например, все факты из класса «выступления Президента» попали в класс «выступления Президента по понедельникам», а класс «выступления Президента по другим дням недели» остался пустым, то из этого можно сделать серьезные выводы о графике работы Президента и, с известной долей достоверности, о его здоровье.

Е3. Для целей практической реализации условие хронологической регулярности фактов можно сформулировать следующим образом:

Подмножество фактов $\{\psi\} \subset \{\varphi\}$ и, соответственно, определяющее его фактологическое преобразование Ψ, является регулярным относительно хронофункции $X(\varphi) = \{0, 1\}$, если значения $X(\varphi)$ на

подмножестве $\{\psi\}$ *образуют дельта-распределение:* $P(X(\varphi) = 0/\varphi \in \{\psi\}) < \sigma$, *где* σ — *допустимое статистическое отклонение.*

E4. Рассмотрим несколько практически значимых видов хронологической регулярности, но для удобства изложения сделаем сначала несколько предварительных замечаний

Введем в наше рассмотрение функцию $T(\varphi)$, $\varphi \in \Phi$, отображающую факт φ в значение характеристики «время действия факта» $t \in \{\Phi\}$. По предположению раздела E1 эта функция определена для всех возможных фактов φ в фактографическом пространстве Φ, а множество ее значений зависит от выбранной единицы измерения времени. Граничные значения $T(\varphi)$ обозначим $T_{\min}$ и $T_{\max}$: $T_{\min} \leqslant T(\varphi) \leqslant T_{\max}$.

Учтем, что выбор единицы измерения может сделать функцию $T(\varphi)$ вырождающей по отношению к базовой характеристике t. В самом деле, в системном репозитории ФПС время действия факта хранится в виде системной даты с точностью до дня — хотя источник не факта не всегда допускает такую точность («. . . выступил в августе прошлого года. . . »). Однако, в практическом анализе даты часто огрубляют до недель, декад, месяцев, лет и даже веков («События второй половины 17 века доказывают, что. . . »), а также их частей («в конце июля»). В этом случае разные исходные даты будут отображаться в одну аналитическую дату по $T(\varphi)$.

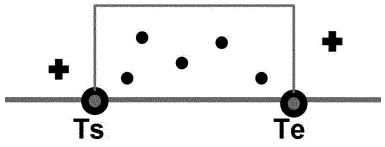
В силу упорядоченности значений характеристики t, функция $T(\varphi)$ задает отношение порядка на произвольной коллекции фактов $\{\varphi\} = \{\varphi_1, \varphi_2, \ldots, \varphi_N\}$ так, что $T(\varphi_n) \leqslant T(\varphi_{n+1})$, $1 \leqslant n \leqslant N-1$. При этом будем считать, что в случае $T(\varphi_n) = T(\varphi_{n+1})$ порядок следования фактов задается любым удобным детерминированным способом.

Очевидно, что «огрубление масштаба» функцией $T(\varphi)$ не должно сказываться на порядке фактов, поэтому самым первым способом разрешения проблемы порядка в случае $T(\varphi_n) = T(\varphi_{n+1})$ должна быть техническая дата действия факта. Если даже технические даты фактов φ_n и φ_{n+1} совпадают, то можно использовать регистрационный факт в системе.

Вычисление хронологической регулярности во всех нижеперечисленных случаях производится согласно схеме раздела E2 и по определению раздела E3. Именно, для всех фактов коллекции $\{\varphi\} =$

$\{\varphi_1, \varphi_2, \ldots, \varphi_N\}$ вычисляется булевое условие $X(\varphi) = \{0, 1\}$, $1 \leqslant n \leqslant N$ и определяется количество фактов $N_{X=1}$, для которых данное условие $X(\varphi)$ оказалось истинным. Далее, мы рассчитываем статистический показатель ложности $P(X(\varphi) = 0)$ условия $X(\varphi)$ и сравниваем его с допустимым отклонением: $1 - N_{X=1}/N = P(X(\varphi) = 0) < \sigma$.

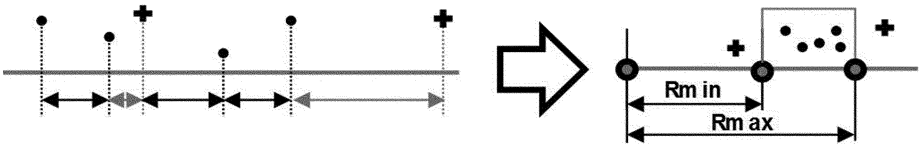
а. Ситуационная регулярность обобщает все факты коллекции $\{\varphi\}$, имевшие место на некотором временно́м отрезке, границы которого обусловлены некоторой *опорной ситуацией*. Например, «до войны» — «во время войны» — «после войны», или «после покупки компании», или «до смены руководства».



Хронофункция: $X(\varphi_n) = [T_s \leqslant T(\varphi_n) \leqslant T_e]$.

Здесь $T_s \geqslant T_{\min}$ и $T_e \leqslant T_{\max}$ — время начала и окончания «опорной ситуации» соответственно. При этом случай $T_s = T_{\min}$ определяет вариант «до ситуации», а $T_e = T_{\max}$ - «после ситуации».

б. Интервальная регулярность формализует обобщение последовательности фактов, следующих друг за другом через определенный промежуток времени.

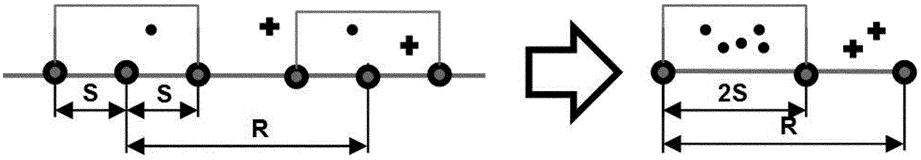


Хронофункция: $X(\varphi_n) = [R_{\min} \leqslant T(\varphi_n) - T(\varphi_{n-1}) \leqslant R_{\max}]$.

Значения $R_{\min}$ и $R_{\max}$ задают допустимый интервал следования фактов.

Этот вид хронологической регулярности весьма эффективно описывает ситуации, связанные с потреблением и восполнением каких-то ресурсов.

в. Периодическая регулярность выявляет общность последовательности фактов, привязанных к некоторым периодическим отметкам на оси времени.

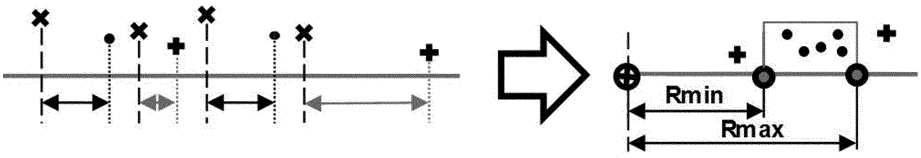


Хронофункция: $X(\varphi_n) = [(T(\varphi_n) - T)\%R \leqslant S] \wedge [(T(\varphi_n) - T)/R \geqslant (T(\varphi_{n-1}) - T)/R]$.

Здесь символ «/» означает целочисленное деление, а символ «%» — остаток от такого деления. Значение T задает момент начала наблюдаемой последовательности, величина R определяет период регулярности, а S выражает максимально допустимое отклонение от регулярности.

Периодическая регулярность, по-видимому, наиболее свойствена хронологическим тенденциям организационного и социального характера вроде «ежемесячно по 10-м числам плюс-минус 2 дня», «еженедельно строго по вторникам», «в последнюю декаду года» и пр.

г. Корреляционная регулярность обнаруживает аналитически-значимую зависимость одной группы фактов $\{\varphi\}$ от другой группы фактов $\{\psi\}$, причем $\{\psi\} \cap \{\varphi\} = \varnothing$.



Хронофункция: $X(\varphi_n) = [R_{\min} \leqslant T(\varphi_n) - T(\psi_m) \leqslant R_{\max}]$.

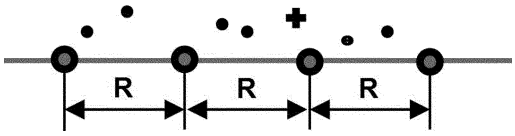
Значения $R_{\min}$ и $R_{\max}$ задают, соответственно, минимальный и максимальный допустимый интервал отстояния фактов группы $\{\varphi\}$ от фактов группы $\{\psi\}$.

В своей более строгой форме, корреляционная регулярность выражается как:

$X(\varphi_n) = [R_{\min} \leqslant T(\varphi_n) - T(\psi_m) \leqslant R_{\max}] \wedge [T(\varphi_n) - T(\psi_m) \leqslant T(\varphi_n) - T(\varphi_{n-1})]$, где вторая часть требует соответствия фактов группы $\{\varphi\}$ фактам группы $\{\psi\}$ без дублирования.

Во многих случаях корреляционная регулярность служит индикатором причинно-следственной связи между явлением, представленным группой фактов $\{\psi\}$ (причина) и явлением, представляемым группой фактов $\{\varphi\}$ (следствие).

д. Групповая регулярность указывает на то, что явление, представляемое фактами коллекции $\{\varphi\}$, имеет свойство проявляться определенное количество раз K в заданный периодический отрезок времени R:



$$X(\varphi_n) = \bigvee_{K-S \leqslant M \leqslant K+S} \Big\{ \bigvee_{1 \leqslant i \leqslant M} \{[T(\varphi_{n-i})/R \neq T(\varphi_{n+1-i})/R] \wedge$$

$$\wedge [T(\varphi_{n+1-i})/R = T(\varphi_{n+M-i})/R] \wedge [T(\varphi_{n+M-i})/R \neq T(\varphi_{n+M+1-i})/R]\} \Big\}.$$

Сложный вид хронофункции $X(\varphi_n)$ обусловлен тем, что она допускает возможность случайного отклонения количества повторений K на S раз в каждую сторону ($S < K$). Легко видеть, что $X(\varphi_n) = 1$ только для фактов, образуюших в течении периода R *группы* размером $K \pm S$.

Групповая регулярность соответствует ситуациям типа «ровно три раза в год», «пару раз в неделю» или «обычно четыре раза в месяц, но иногда получается только два».

Ж. Заключение

Представленная формальная имитационная модель алгоритмического анализа фактов далека от завершения — как и все прочее в

пока еще не сформировавшейся окончательно отрасли компьютерной фактологии. Мы продолжаем развивать вышеописанные схемы и формализмы, охватывая новые предметные области, типы фактов и аналитических связей между ними. Кроме того, в настоящее время исследовательская группа компании «Ай-Теко» реализует масштабную экспериментальную программу верификации основных положений данной модели на обширном фактическом материале, накопленном системой «XFiles». Мы планируем сообщать о результатах этих экспериментов в наших дальнейших публикациях.

Список литературы

[1] Banko M. et al. Open Information Extraction From The Web. University of Washington, DCSE, IJCAI, 2007.

[2] Carlson A. et al. Active Learning for Information Extraction via Bootstrapping. Carnegie Mellon University, 2010.

[3] Fain V. S., Rubanov L. I. Activity And Understanding: Structure Of Action And Orientated Linguistics // World Scientific. 1998.

[4] Zadeh L. A. Fuzzy Logic And Approximate Reasoning. Synthese, 1975.

[5] Ильин Н., Киселёв С., Рябышкин В., Танков С. Технологии извлечения знаний из текста // Открытые системы. 2006. № 6. http://www.osp.ru/text/302/2700556/.

[6] Киселев С. Модель информационной системы бизнес-разведки // Открытые системы. 2005. № 5–6.

[7] Киселев С. Л., Ермаков А. Е., Плешко В. В. Поиск фактов в тексте естественного языка на основе сетевых описаний // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. М.: Наука, 2004. С. 282–285.

[8] Топровер Г., Киселев С. Алгоритмический анализ фактов // Открытые системы. 2011. № 5.

[9] Киселев С. Л. Системы «Аналитический курьер» и «X-Files» — основа технологии извлечения знаний текстов из произвольных источников // Бизнес и безопасность в России. 2007. № 48. С. 102–106.