ЛОГИКО-ЛИНГВИСТИЧЕСКАЯ МОДЕЛЬ ГЕНЕРАЦИИ ФАКТОВ ИЗ ТЕКСТОВЫХ ПОТОКОВ ИНФОРМАЦИОННОЙ КОРПОРАТИВНОЙ СИСТЕМЫ

Нина Хайрова, Наталья Шаронова, Аджит Пратап Сингх Гаутам

***Аннотация**: Подсистема накопления и генерации фактов представляет основу для принятия решений и проведения бизнес-разведки интегрированной корпоративной системы. Причина относительно малого количества систем генерации фактов из слабоструктурированной текстовой информации заключается в отсутствии четких алгоритмов извлечения фактов из текста, проверки их на непротиворечивость и невозможности семантической интерпретации полученных результатов, что не позволяет объединить их в общее единое пространство фактографической информации. В работе предлагается логико-лингвистическая модель идентификации и экстракции фактов, позволяющая получить пространство фактов, динамически наполняемое из англоязычного текстового контента интегрированной корпоративной системы. Факт записывается в виде триплета: Subject - Predicate - Object, в котором предикат представляет отношение, а субъект и объект определяют два предмета или понятия. Такой факт записывается в виде двухместного предиката в логике первого порядка. Выделяются два типа фактов: факты, описывающие связь двух сущностей, одна из которых определяется как субъект, а вторая как объект предикатного действия, и факты, фиксирующие значение заранее определенного свойства. Математическая модель, связывающая информацию, содержащуюся в определении смысловых связей, с элементами поверхностной структуры предложений английского языка базируется на формальном аппарате алгебры конечных предикатов. Семантические связи между извлеченными понятиями текста, выражающие тот или иной факт, определяются через предикат, связывающий категории наличия предлога после глагола, существование апострофа, определяющего притяжательный падеж, расположения понятия в предложении, связи которого определяются, наличия глагола to be и формы основного глагола. В статье рассмотрен вид фактов, представляющий утверждение о некотором обладании, приобретении (или наличии) у некоторой сущности субъекта некоторой сущности объекта, и выделены связанные с ним факты второго типа, определяющие атрибут времени, места, способа действия и т.д. Разработана программная имплементация полученной модели, представляющая собой веб-приложение, на вход, которого поступают текстовые потоки*

разнородных источников информационной системы, а на выходе формируется базовое пространство фактов интегрированной корпоративной системы

Ключевые слова: *система генерации фактов, автоматическая экстракция фактов, англоязычный текст, семантические связи, алгебра конечных предикатов.*

ACM Classification Keywords*: H.3.3 .Information Search and Retrieval, I.2.4. Knowledge Representation Formalisms and Methods*

Введение

Центральной составляющей современной интегрированной корпоративной системы является база знаний, которая должна включать в себя единое информационное пространство взаимосвязанных фактов или гипотез вне зависимости от типа источника получения информации.

Сегодня система извлечения фактов является одним из наиболее эффективных инструментов выделения нужной для принятия решений информации и для проведения аналитической бизнес-разведки [Киселев, 2005], практически заменяя обычный поиск информации. Факт о некоторой сущности представляет собой структурированную экстракцию из предложения текста документа в виде значения факта: его суть, время и место совершения, его участники [Andersen, 1994].

Основная причина относительно малого объема рынка систем извлечения знаний состоит в том, что практически ни одна система текстовой аналитики, не выполняет формально семантической интерпретации полученных результатов. Если система и включает извлечение фактов, то результаты работы программы, как правило, интерпретируются экспертом-аналитиком, сохраняя в результате добытые знания-факты в головах экспертов. Такое сохранение фактов в разных документах и базах данных, не позволяет объединить их в общее единое пространство фактографической информации для повторного использования. Что в свою очередь приводит к потере ценности полученных сведений [Ландэ, 2009].

Еще одной насущной проблемой обработки фактографической информации является оценка достоверности автоматически определяемой фактографической информации, что особенно важно в связи с все более увеличивающейся плотностью потока текстовой информации в средствах масс-медиа и различного рода социальных сетях, форумах и блогах. Множественность значений факта обусловлена возможностью разной интерпретации одного и того же явления, а также противоречивостью, неточностью или нечеткостью поступающих из внешних источников сведений.

Общая постановка задачи

Целью работы является разработка подсистемы идентификации и экстракции фактов, позволяющая получить пространство фактов, динамически наполняемое из текстового-контента интегрированной корпоративной системы. На вход подсистемы поступают текстовые потоки разнородных источников информационной системы, на выходе формируется базовое пространство фактов интегрированной корпоративной системы.

Помимо количественных показателей основной формой представления фактов являются триплеты. Базовое пространство полученных фактов доступно для аналитической обработки, формирования гипотез, генерирования выводов и интеграции в единое информационное пространство фактов интегрированной корпоративной системы.

Факт представляет собой знание в форме утверждения, достоверность которого строго установлена [Andersen, 1994]. В сфере информационных технологий и теории обработки знаний под фактом, обычно, понимают зафиксированное и произошедшее событие, сопровождаемое временными и географическими метками, аргументирующей информацией, ссылками на источник и т. д.

Факт может быть извлечен из текстовой информации (как слабо структурированной, так и не структурированной) и может определять как свойства объекта, так и связь объекта с другими объектами. При этом под объектом мы понимаем сущность, информация о которой накапливается в системе и может быть в ней само идентифицирована [Ландэ, 2009]. Для извлечения и структурирования фактографической информации в тексте выделяются сущности, и используется структурированное представление семантики факта в терминах предикатных операций. Факты выделяются из предложений, содержащих упоминание сущности или анафорические ссылки на нее. В свою очередь, фактографическую информацию можно разделить на хорошо структурированную и плохо структурированную (Рис. 1).

К хорошо структурированным сведениям (так называемая параметрическая информация) относятся, прежде всего, сведения количественного характера, а так же качественные сведения, имеющие хорошо регламентированную форму. К плохо структурированной фактографической информации относятся сведения, представленные различными нерегламентированными словесными конструкциями, представленными на естественном языке [Барахнин,1980].

Алгоритмы фактографического анализа зависят, в свою очередь, от степени структурированности конкретного документа. По степени структурированности данные документа можно разделить, подобно общей классификации степени формализации информации, на табличные данные, отображенные в виде фактов; массивы однородных слабоструктурированных

текстовых документов, обычно описывающие конкретную предметную область и документы произвольного слабоструктурированного типа [Fader, 2011].

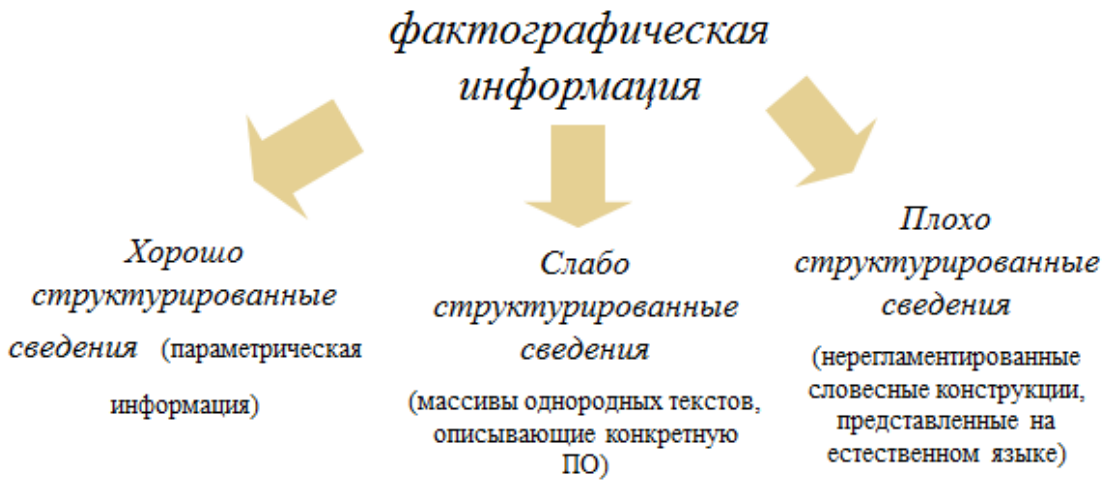

Рис. 1. Общая схема представления фактографической информации

Структурное описание модели

Выделение фактов из слабоструктурированной текстовой информации включает следующие этапы [Baeza-Yates, 1999]:

- Entity Extraction – извлечение слов или словосочетаний, важных для описания смысла текста (списки терминов предметной области, персоналий, организаций, географических названий и т.д);
- Feature Association Extraction – исследование связей между извлеченными понятиями;
- Event and Fact Extraction – извлечение сущностей, распознавание фактов и действий.

Для реализации первого этапа выделения понятий используется стандартный лингвистический процессор [Ritter, 2011], включающий графемный, морфологический, синтаксический и контекстный этапы обработки, с добавлением специализированных методов и алгоритмов обработки документов. Так как очень часто в задачах по извлечению фактографической информации нужно найти в тексте упоминания лиц, компаний, правительственных организаций и местоположений, и другие подобные типы сущностей, то для их выделения используются специальные формализмы графемного анализа. На этапе морфологического анализа используются декларативный и алгоритмический методы. Каждый неправильный глагол английского языка представлен в базе данных во всех его формах, то есть глагол write имеет

формы write-writes-wrote-written-writing, формы правильных глаголов определяются алгоритмически.

Факт представляет собой триплет: *Subject ->Predicate ->Object*, в котором предикат представляет собой отношение, а субъект и объект указывают на два предмета. Практическое запись такого факта осуществляется строкой в таблице реляционной базы данных, поля которой представляют субъект и объект триплета, а имя таблицы соответствует отношению между предметами или предикатом триплета. Кроме того можно использовать представление в виде двухместного предиката в логике первого порядка.

Следующим этапом после выделения слов или словосочетаний, представляющих узлы триплета факта, является выделение отношений, устанавливаемых данным фактом между словами.

Выделяем два типа фактов: факты описывающие связь двух сущностей, при этом одна из сущностей будет определяться как субъект, а вторая как объект предикатного действия. Например, "*the company had revenue*" (субъект: *company*, объект: *revenue*, предикат: *had*). Если второго участника связи в базе нет, то он создается автоматически.

Второй вид факта представляет собой триплет: *предмет – атрибут – значение*, где предмет – это объект, о котором фиксируется факт, атрибут – некоторое именованное, заранее определенное свойство, а значение представляет собой некоторое значение, область определения которого может быть в некоторых случаях известна. Например, это могут быть факты атрибутов места и времени осуществления некоторого действия.

Для выделения излечения связей между определенными понятиями в тексте необходимо выделить семантические (или понятийные) связи в предложении. Для чего необходимо разработать строгую модель, связывающую информацию, содержащуюся в определении смысловых связей с элементами поверхностной структуры предложений естественного языка.

Такой подход рассматривается в рамках падежной грамматики и основывается на понятии глубинных падежей, введенных Ч. Филлмором, выделившим пропозицию, или основной смысл предложения, как предикат, выражаемый в поверхностной структуре чаще глаголом, связанным с помощью определенных глубинных падежей с участниками данной ситуации, или партиципантами [Филлмор,1981]. Семантические падежи в различных естественных языках имеют разные формы формального выражения, которые необходимо четко определить для автоматической идентификации и экстракции фактов из текстов. Например, в русском и украинском языках, семантическая информация партиципантов кодируется, в основном, грамматическими поверхностными падежами, тогда как в английском — она передается сочетанием с предлогом, порядком слов в предложении.

Описание математической модели

Введем на универсуме U, включающем все возможные понятия и объекты анализа сложной языковой системы [Хайрова, 2012], множество грамматических характеристик синтаксической сочетаемости слов английских предложений, влияющих на понятийные связи, $M = \{m_1, \dots, m_n\}$, где n – количество характеристик системы. Используя формальный аппарат алгебры конечных предикатов [Бондаренко, 2007].

Отношения между характеристиками можно представить в виде $m_i * m_j * \dots * m_k$, где $m_i, m_j, \dots, m_k \in M$, а знак * – обозначает, что конъюнкция данных характеристик соответствует некоторой семантической функции или некоторому глубинному смысловому отношению между словами, грамматические характеристики которых выражаются $m_i, m_j, \dots, m_k$.

На множестве M введем систему предикатов S так, чтобы любой предикат $P(q_m) \in S$, обращался в 1 на множестве слов с грамматической информацией, соответствующей определенной семантической функции, и был равен 0 в противном случае. Таким образом, множество предикатов S можно сопоставить с множеством грамматических характеристик приписанных словам предложения, называющим сущности триплета факта.

Для формализации семантических функций предложений английского языка и их явного представления средствами поверхностной структуры были выделены и описаны следующие синтаксические и морфологические категории:

$$z^{to} \vee z^{by} \vee z^{with} \vee z^{about} \vee z^{of} \vee z^{on} \vee z^{at} \vee z^{in} \vee z^{out} = 1,\ y^{ap} \vee y^{aps} \vee y^{out} = 1,\ x^{f} \vee x^{l} \vee x^{kos} = 1,$$

$$m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were} \vee m^{out} = 1,\ p^{III} \vee p^{ed} \vee p^{l} \vee p^{ing} \vee p^{II} = 1,$$

где использованы предметные переменные, характеризующие следующие категории:

- наличие предлога *to*, *by*, *with*, *about*, *of* после предиката триплета или его отсутствие – z^{to}, z^{by}, z^{with}, z^{about}, z^{of}, z^{at}, z^{on}; z^{in}, z^{out};
- наличие или отсутствие апострофа в конце слова, определяющего притяжательный падеж у субъекта триплета – y^{ap}, y^{aps}, y^{out};
- расположение существительного, определяющего сущность, перед глаголом в личной форме, после глагола в личной форме или после косвенного дополнения – x^{f}, x^{l}, x^{kos};
- наличие или отсутствие любой формы глагол *to be* – m^{is}, m^{are}, m^{havb}, m^{hasb}, m^{hadb}, m^{was}, m^{were}, m^{out};
- первая, вторая/третья и четвертая форма основного правильного глагола , и вторая, третья формы неправильного основного глагола – p^{l}, p^{ed}, p^{ing}, p^{II}, ^{III}.

Семантические связи между извлеченными понятиями текста определяются через предикат P, связывающие категории наличия предлога после предиката, существование апострофа,

определяющего притяжательный падеж, расположения понятия, факт связи которого определяется, наличия глагола *to be* и формы основного глагола:

$$P(x, y, z, m, p) \rightarrow P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p). \quad (1)$$

Зададим на декартовом квадрате множества $S * S$ предикат $\gamma(x_n, y_n, z_n, m_n, p_n)$, принимающий значение 1, если комплекс выбранных категорий для фразы n формирует некоторые семантические связи понятий триплета, т.е. формирует некий факт, и значение 0 в противном случае.

Таким образом, отношения грамматических элементов английского предложения, идентифицирующих некоторый факт, можно задать формулой:

$$P(x, y, z, m, p) = \gamma_k(x, y, z, m, p) \wedge P(x) \wedge P(y) \wedge P(z) \wedge P(m) \wedge P(p). \quad (2)$$

Практически никогда подмножество согласующихся категорий информации, выражающей факты, не совпадает с декартовым произведением на множестве грамматических признаков. Грамматические категории, которые в своей конъюнкции не формируют семантические связи и соответственно факты, исключаются из формулы (1) множителем $\gamma_k(x_n, y_n, z_n, m_n, p_n)$, $k \in [1;h]$, где h — количество, принятых к рассмотрению в системе типов фактов.

В процессе реализации модели был определен набор глаголов, соответствующих центральной части триплета идентифицируемых типов фактов. Одним из рассмотренных типов фактов является утверждение об обладании, приобретении (или наличии) у некоторой сущности субъекта некоторой сущности объекта. Такое утверждение в англоязычных текстах будет определяться предикатами (глаголами), заранее определенными в базе данных: *have*, *purchase*, *buy*, *acquire*, *get*, *gain*, *obtain*.

В соответствии с формулой (2) семантическая связь, выделяющая субъект триады данного утверждения будет определяться следующим предикатом:

$$\begin{aligned}\gamma_1(x_n, y_n, z_n, m_n, p_n) = {} & z^{out} y^{out} x^{f} m^{out} p^{I} \vee z^{out} y^{out} x^{f} m^{out} p^{II} \vee z^{out} y^{out} x^{f} m^{out} p^{ed} \vee \\ & \vee z^{by} y^{out} x^{I} p^{ed}(m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \vee \\ & \vee z^{by} y^{out} x^{I} p^{III}(m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee \vee m^{was} \vee m^{were}).\end{aligned} \quad (3)$$

Объект данного факта будет явным образом выделен из предложения с помощью предиката, соответствующего конъюнкции предметных переменных грамматических категорий членов предложения:

$$\begin{aligned}\gamma_2(x_n, y_n, z_n, m_n, p_n) = {} & z^{out} y^{out} x^{I} m^{out} p^{I} \vee z^{out} y^{out} x^{I} m^{out} p^{ed} \vee z^{out} y^{out} x^{I} m^{out} p^{II} \vee z^{out} y^{out} x^{f} p^{III} \\ & (m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}) \vee z^{out} y^{out} x^{f} p^{ed}\end{aligned} \quad (4)$$

$$(m^{is} \vee m^{are} \vee m^{havb} \vee m^{hasb} \vee m^{hadb} \vee m^{was} \vee m^{were}).$$

Ко второму виду фактов, связанных с теми же глаголами, можно отнести определение атрибутов времени, места, способа действия и т.д. Например, факт времени осуществленного действия выделяется из предложения с помощью предиката.

$$\gamma_3 (x_n, y_n, z_n, m_n, p_n) = z^{on} x^{kos} y^{out} m^{out} \vee z^{in} x^{kos} y^{out} m^{out} \vee z^{at} x^{kos} y^{out} m^{out}. \quad (5)$$

Дополнительным лингвистическим условием выражения семантических связей, определяющим атрибутивный факт места осуществления действия, является представление объекта триплета факта именем собственным (обычно графически выражаемым с большой буквы), так как в данном факте интерес представляет именно населенный пункт, а не местоположение, как, например, *in mansion*.

Факт принадлежности, или собственности, объекта некоторому субъекту выделяется из предложений с вышеперечисленными глаголами, но определяется следующим предикатом

$$\gamma_3 (x_n, y_n, z_n, m_n, p_n) = z^{out} x^{f} (y^{ap} \vee y^{aps}). \quad (6)$$

Программная имплементация модели

Программная имплементация модели представляет собой веб-приложение, анализирующие текст или список анализируемых текстовых файлов. Извлеченная системой фактографическая информация представляется пользователю форме диалогового окна (Рис. 2).

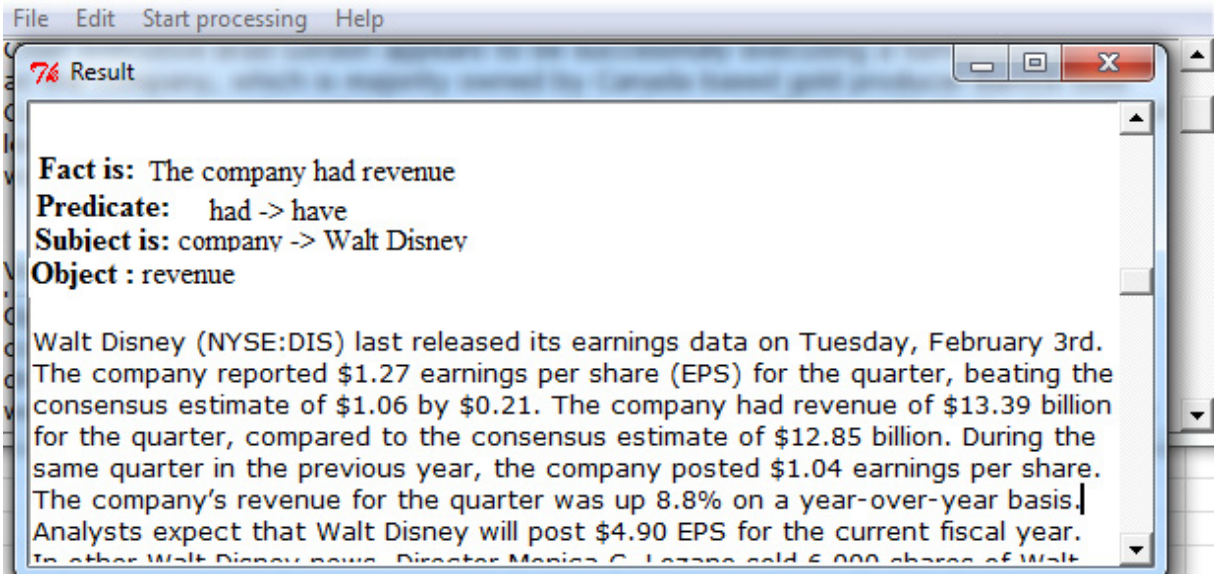

Рис. 2. Общая схема представления фактографической информации.

Программа отображает извлеченную фактографическую информацию в виде факта и первичных предложений, из которых данный факт был извлечен. В поле "Fact is" представлено извлеченное утверждение в виде триплета; в поле "Predicate" представлен извлеченный глагол и его каноническая форма; в полях "Subject" и "Object" соответственно извлеченные субъект и объект триплета. Кроме того, извлеченные названия сущностей анализируются по базе данных гипонемических отношений экономических терминов и подвергаются анализу, устанавливающему анафорические ссылки.

Идентифицированные факты записывается последовательно, перед абзацем текста, из которого он извлекается. Факты деятельности располагаются в порядке значимости, определенной системой.

Выводы

Результатом данного исследования является разработка логико-лингвистической модели извлечения фактов из слабоструктурированных текстов на английском языке. Используемая технология идентификации и экстракции фактов, основывающаяся на использовании специальных семантико-лингвистических методов, включающих специализированный лингвистический процессор, учитывающий как анафорические ссылки, так и словоизменительные формы, позволяют получить полноту и точность получаемого фактографического пространства, сравнимую с экспертными оценками.

Литература:

[Andersen, 1994] Andersen, P. M., Huettner A. K. Knowledge engineering for the JASPER fact extraction system. / Intrgated Computer-Aided Enginrering. – 1 (6), 1994. – P. 473–493.

[Baeza-Yates, 1999] Baeza-Yates R., Ribeiro-Neto B. Modern Information Retrieval. / Addison-Wesley, 1999. 340 p.

[Fader, 2011] Fader, S. Soderland, O. Etzioni.Identifying relations for open information extraction. / Conference on Empirical Methods in Natural Language Processing. Edinburgh, Scotland, 2011. – P. 1535 – 1545.

[Ritter, 2011] Ritter A., Clark S., Mausam K., Etzioni O. Named entity recognition in tweets: an experimental study / Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. – Edinburgh/Scotland, P. 1524–1534.

[Барахнин, 1980]. Барахнин В. Б., Федотов А. М.. Проблемы разработки технологии фактографического поиска – М.: Институт вычислительных технологий СО РАН, 1980. – 150 с.

[Бондаренко, 2007] Бондаренко М. Ф., Шабанов-Кушнаренко Ю. П. Теория интеллекта / Харьков: Комп. СМИТ, 2007. 576 с.

[Киселев, 2005] Киселев С. Модель информационной системы бизнес-разведи / Открытые системы #05-06/2005. Режим доступа: http://www.osp.ru/os/2005/05-06/185595/

[Ландэ, 2009] Ландэ Д. В., Снарский А. А., Безсуднов. И. В. Интернетика: Навигация в сложных сетях: модели и алгоритмы.– М.: Либроком (Editorial URSS), 2009. – 264 с.

[Филлмор, 1981] Филлмор Ч. Дело о падеже открывается вновь // Новое в зарубежной лингвистике– М.: Изд. иностр. лит., 1981, вып. 10. – С. 496-530.

[Хайрова, 2012] Хайрова Н. Ф. Використання логіко-алгебраїчної моделі семантичних відмінків для семантичного аналізу речення / Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2012.- Вип. № 38. – С. 239 – 245.

Authors' Information

Нина Хайрова *– professor kafedry intellektualnykh kompyuternykh system Natsionalnogo technicheskogo universiteta „Kharkovskiy politechnicheskiy institut", ul. Frunze, 21, Kharkov, 61002, Ukraina; e-mail: nina_khajrova@yahoo.com*

Nauchnye interesy: iskusstvennyy intellekt, identifikatsiya znaniy iz tekstov, Text Mining, Opinion Mining, Web Mining, Natural language processing

Наталья Шаронова *– professor, zaveduyushchiy kafedroy intellektualnykh kompyuternykh system Natsionalnogo technicheskogo universiteta „Kharkovskiy politechnicheskiy institut", ul. Frunze, 21, Kharkov, 61002, Ukraina; e-mail: nvsharonova@mail.ru*

Nauchnye interesy: iskusstvennyy intellekt, matematicheskoe modelirovanie, avtomatizirovannye bibliotechnye sistemy, prikladnaya lingvistika

Аджит Пратап Сингх Гаутам *– aspirant kafedry intellektualnykh kompyuternykh system Natsionalnogo technicheskogo universiteta „Kharkovskiy politechnicheskiy institut", ul. Frunze, 21, Kharkov, 61002, Ukraina; e-mail: apsgautam@gmail.com*

Nauchnye interesy: integrirovannye korporativnye sistemy, informatsionnye technologii, modeli predstavleniya znaniy

Logic-linguistic model of fact generation from text streams of corporate information system

Nina Khairova, Nataliya Sharonova, Ajit Pratap Singh Gautam

Abstract: *This paper proposes a logical-linguistic model extracting semi-structured facts in English texts. To identify the fact some entities expressed by lexical units as well as semantic relations between them are defined in the text. The semantic relations are expressed by semantic functions of sentence participants. A fact is written in form of a triplet: Subject - Predicate - Object, in which the Predicate represents the relations and Subject and Object define the subjects, objects or concepts. Two types of the facts are defined. The first type is fact that describes relation between two entities; the second one is fact that fixes the value of a predetermined attribute. The functions are described by predicates of algebra of finite predicates. The mathematical model allows associating meaning relations of concepts of a sentence with elements of the syntactic and morphological structure of the English sentence. The model is applied to the semantic stage of linguistic processor of information subsystem for facts identification, which are essential for business analysis, in the framework a semi-structured texts. Software implementation of the model is designed. The input subsystem receives text streams disparate sources of information of the integrated corporate system, basic facts of space of the system are output. The accuracy and completeness extracted facts from texts in English by the subsystem are compatible with extracted facts by an expert.*

Keywords: *the system of facts generation, the automatic facts extraction, semantic relations, the algebra of finite predicates, natural language processing.*