УДК 004.415

Н. Д. Елагина, М. Г. Пантелеев
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В. И. Ульянова (Ленина)

Извлечение тематических фактов из неструктурированных текстов и базовых знаний

Рассматривается подход к получению требуемых фактов с использованием базовых знаний и извлекаемых из документов фактов, а также его реализация в рамках системы FactE, предназначенной для анализа инновационных технологий. Подход иллюстрируется на примере вывода фактов о предприятиях, являющихся потенциальными потребителями заданной инновационной технологии. Приводятся примеры онтологического представления соответствующих базовых знаний и правил вывода. Представлена архитектура и алгоритмы функционирования системы. Обсуждаются возможные направления ее дальнейшего развития.

Извлечение фактов, извлечение информации, основанное на онтологии, базовые знания

Автоматизированное извлечение информации из текстов является важной и актуальной задачей, учитывая скорость пополнения объемов текстовых данных в сети Интернет. Поскольку эта задача имеет множество практических применений, разработка соответствующих систем вызывает интерес исследователей. Извлечение информации является очень широкой задачей, и в настоящее время предложено несколько подходов и методов для ее решения [1]–[3]. В частности, в последние годы в связи с развитием концепции семантического web большое внимание уделяется подходу к извлечению информации, основанному на онтологиях [4]–[8]. Существующие на данный момент в области OBIE системы (KIM [9], PANKOW [7], SOBA [10], Text-To-Onto [11] и др.) ориентированы в основном на извлечение требуемой категории фактов, непосредственно содержащихся в тексте анализируемых документов. Тем не менее, во многих случаях значимые для пользователя ИАС факты не содержатся в тексте документов в явной форме, однако могут быть выведены посредством рассуждений на основе содержащихся в текстах фактов и неких общих знаний о предметной области.

В данной статье рассматривается задача извлечения тематических фактов, решаемая в контексте разработки информационно-аналитической системы оценки инновационных технологий, и предлагается подход, позволяющий повысить полноту извлечения фактов за счет использования онтологий для извлечения фактов из текста и вывода фактов, не содержащихся в текстах явно. Вывод явно не упомянутых фактов осуществляется на основе приобретенных в процессе анализа фактов и базовых знаний онтологии. Также предлагаемый подход предусматривает использование явных онтологических знаний об особенностях первичных источников информации, позволяющих существенно повысить эффективность извлечения из них тематических фактов. В настоящее время на web доступно значительное число ресурсов, имеющих специфическую тематическую направленность, структуру документов и т. п. Эти особенности целесообразно учитывать при обработке соответствующих документов на предмет извлечения фактов определенного типа. Так, например, ресурсы, хранящие информацию о патентах, имеют хорошую структуризацию по техническим областям (в соответствии с классификаторами) и четкую структуризацию самих документов (описаний патентов).

В статье детально рассматриваются особенности предлагаемого подхода с иллюстрацией на конкретном примере; представлены фрагменты онтологических баз знаний и правил вывода, используемых в рамках данного подхода; представлена архитектура и обобщенные алгоритмы функционирования системы; приведены примеры ре-

ализации вывода фактов и определены направления дальнейшей работы.

Подход к получению тематических фактов. Под *тематическим фактом* (ТФ) (thematic fact, TF) понимается утверждение, характеризующее некоторую сущность (субъект факта, S) в заданном аспекте. Аспект, в котором характеризуется субъект факта, определяет базовое отношение R, связывающее S с другой сущностью – объектом факта (O). Таким образом, тематический факт формально может быть представлен как

$$TF = R\ (S, O).$$

Отношение R определяет *категорию факта* (КФ).

Предлагаемый подход реализуется в рамках ИАС оценки инновационных технологий, поэтому при рассмотрении примеров в качестве субъекта фактов выступает некоторая инновационная технология. Вместе с тем, предлагаемые решения являются универсальными и могут использоваться в ИАС другого функционального назначения.

Интересующие пользователя-эксперта аспекты технологии определяют категории фактов, которые должны извлекаться системой из неструктурированных текстов. К таким категориям относятся, например, предприятия-разработчики технологии, уровень готовности технологии, предприятия – потенциальные потребители технологии и т. п. Перечень интересующих категорий фактов задан априори и используется в качестве исходных данных при разработке информационного обеспечения ИАС.

Предлагаемый подход направлен на повышение полноты извлечения фактов посредством решения двух проблем, идентифицированных в процессе разработки и опытной эксплуатации ИАС:

1. Пропуск содержащихся в текстах релевантных фактов вследствие многообразия и сложности возможных формулировок. Факты могут быть не извлечены вследствие:

– *лексического* многообразия структурных элементов ТФ: субъектов, отношений и объектов. Пример синонимии базового отношения факта: «Предприятие E *разрабатывает* технологию T» и «Предприятие E *работает над созданием* технологии T»;

– *синтаксического* многообразия выражения ТФ. Порядок следования структурных элементов ТФ в предложении не является жестко фиксированным. Например: «Предприятие E разрабатывает технологию T» и «Технология T разрабатывается предприятием E».

2. Отсутствие интересующих фактов в текстах *в явном виде*, при том что такие факты могут быть логически выведены из найденных в текстах фактов с использованием некоторых общих знаний.

Далее рассмотрены аспекты предлагаемого подхода, соответствующие выделенным проблемам.

Использование онтологий при извлечении ТФ из текстов. Извлечение из текстов фактов различных категорий выполняется с использованием соответствующих шаблонов. Для повышения полноты извлечения элементы шаблонов, соответствующих разным категориям фактов, связываются с элементами *лексической онтологии*. Это позволяет максимально полно использовать возможные лексические формы выражения субъекта, объекта и отношения ТФ, а также однозначно определить их смысловую принадлежность.

Для каждой заданной категории ТФ разработан набор шаблонов извлечения. Эти шаблоны специфицируют в общем виде специфичные для данной категории субъект, объект и отношение как элементы лексической онтологии. Например, шаблон извлечения для категории фактов о предприятиях-потребителях заданной технологии может быть определен в общем виде следующим образом:

элемент_определяющий_Заинтересованность (элемент_определяющий_Технологию, элемент_определяющий_Предприятие) (1)

Проблема синтаксического многообразия ТФ решается с помощью онтологического описания структуры шаблонов посредством перечисления его элементов без строгой фиксации их порядка.

Подход OBIE позволяет использовать лексическую онтологию для управления процессом извлечения фактов. Для этого разработанные шаблоны извлечения определены в онтологии с помощью соответствующих отношений шаблона. *Отношения шаблона* представляют собой бинарные отношения, связывающие сущность, представляющую шаблон категории фактов, с сущностями, представляющими каждый элемент шаблона. Таких отношений для каждого шаблона 3:

является_Субъектом_Шаблона(шаблон_категории_факта, субъект_шаблона);

является_Объектом_Шаблона(шаблон_категории_факта, объект_шаблона);

является_Отношением_Шаблона(шаблон_категории_факта, отношение_шаблона).

Таким образом, в соответствии с предложенным подходом в лексической онтологии определены:

– сущности, соответствующие элементам шаблонов извлечения для каждой категории фактов;

– родовидовые отношения и отношения лексической синонимии;

– отношения шаблонов для каждого определенного в ИАС шаблона извлечения.

Лексическая онтология в соответствии с предложенным подходом имеет следующую структуру. Сущностями верхнего уровня абстракции в данной онтологии являются: Категория_Факта, Субъект_Факта, Отношение, Объект_Факта. При разработке шаблонов извлечения у этих классов определяются конкретные подклассы. Для данных подклассов определяются шаблонные отношения, связывающие их друг с другом. В качестве элементов шаблона извлечения используются описательные экземпляры классов онтологии, соответствующие определяемому отношению, его субъекту и объекту.

Под *описательным экземпляром* (ОЭ) понимается экземпляр класса онтологии, реализующий его объектные отношения. ОЭ позволяют формализовать знания лексической онтологии. Они специфицируют набор характерных объектных отношений, их домен и область допустимых значений на уровне классов и позволяют реализовать эти отношения для конкретных экземпляров. Для ОЭ может быть определен набор лексических экземпляров, принадлежащих тому же классу.

Лексический экземпляр (ЛЭ) – это экземпляр класса онтологии, выражающий лексическую словоформу данного класса. Все ЛЭ, определенные для одного ОЭ, семантически эквивалентны друг другу, т. е. являются лексическими синонимами.

Рассмотрим пример, иллюстрирующий концепции ОЭ и ЛЭ. Пусть имеется общий класс «Категория_Факта», для которого определено отношение «имеет_Отношение» с областью допустимых значений в виде класса «Отношение_Факта». Рассмотрим категорию фактов о предприятиях-потребителях некоторой технологии, имеющей шаблон (1). На основе структуры (1) для класса «Категория_Факта» создается подкласс «Предприятия_Потребители», а для класса «Отношение_Факта» – подкласс «Заинтересованность». Связать эти классы объектным отношением «имеет_Отношение» не представляется возможным, поэтому для каждого из них создаются ОЭ, которые связываются этим отношением. Далее, для классов «Предприятия_Потребители» и «Заинтересованность» требуется определить их лексическое наполнение, т. е. определить набор ЛЭ. Чтобы привязать набор ЛЭ к шаблону, вводится отношение «имеет_Лексическое_выражение», связывающее каждый ЛЭ с ОЭ соответствующего класса. Например, для класса «Заинтересованность» могут существовать ЛЭ «заинтересован», «планирует приобрести». Наличие одного или нескольких ЛЭ для всех структурных компонентов шаблона (в произвольном порядке следования) в структурном элементе документа является основанием для извлечения этого элемента как тематического факта.

На рис. 1 приведен фрагмент онтологии, описывающий знания о предприятиях-потребителях искомой технологии. Шаблон извлечения для данной категории фактов на основании введенных определений специфицирован следующим образом:

ОЭ_класса_Заинтересованность

(ОЭ_класса_Технология, ОЭ_класса_Предприятие)

На рис. 1 классы верхнего уровня и основные подклассы, участвующие в процессе извлечения фактов, показаны в толстых рамках, их конкретные экземпляры – в пунктирных рамках. Отношение наследования («is_subclass_of») показано (*), отношение инстанцирования («is_a») показано пунктирной стрелой. Отношение лексической принадлежности «hasLexicalForm» показано (**). Под цифрой 1 – отношение, определяющее объект шаблона «hasObject»; под цифрой 2 – отношение шаблона «hasRelationship»; под цифрой 3 – отношение, определяющее субъект шаблона «hasSubject».

Лексическая онтология расширяется в процессе анализа текстов за счет извлеченных ТФ (ИФ). Каждый ТФ приводится к формализованной форме и добавляется в виде утверждения (тройки), состоящего из *объектного отношения*, определенного для данной категории фактов и связывающего обнаруженные в тексте лексические формы субъекта и объекта ТФ. Например, для рассматриваемой категории предприятий-потребителей технологии предметная онтология

может расшириться утверждением «*Заинтересован* (полимерные нанокомпозиты, Вертолеты России)» (рис. 1).

Логический вывод фактов на основе базовых знаний и фактов, извлеченных в процессе анализа документов. Рассмотрение подхода к логическому выводу релевантных фактов, не представленных явно в анализируемых текстах (*логически выводимые факты, ЛВФ*), начнем с примера. Предположим, что эксперту требуется получить факты о предприятиях, являющихся потенциальными потребителями некоторой (заданной) технологии Т.

Полагаем, что для некоторых классов изделий в онтологической БЗ хранятся *базовые знания* (т. е. знания, специфичные для данной предметной области) двух типов:

1) знания о структуре изделий в виде иерархии составляющих данный класс изделий подсистем, компонентов, элементов (используются подсвойства свойства «hasPart» (и инверсного «partOf»): «hasSubsystem», hasComponent»,…);

2) знания о материалах, используемых при изготовлении определенных подсистем (компонентов), используется свойство «useMaterial» (инверсное свойство «isUsedIn»).

Пусть в процессе обработки текстов удается выделить факт о том, что рассматриваемая технология Т является перспективной для производства материала М. Кроме того, из текстов был извлечен факт о том, что некое предприятие планирует производить изделие определенного типа Р. Привлекая базовые знания о том, что изделия данного типа содержат компоненты, в изготовлении которых используется материал М, можно заключить, что данное предприятие является потенциальным потребителем технологии Т. Правило вывода таких фактов имеет вид:

hasSubsystem (?Product, ? Subsystem) ^
hasComponent (?Subsystem, ?Component) ^
isUsedIn (?Material, ?Component) ^
isUsedFor (?Technology, ?Material) ^
plansToProduce (?Enterprise, ?Product)
=>
isPotentialProducer (?Technology, ?Enterprise) (2)

Таким образом, подход к логическому выводу тематических фактов использует следующие компоненты знаний:

ВК – *базовые знания о предметной области;*
IR – *правила вывода неявных фактов;*

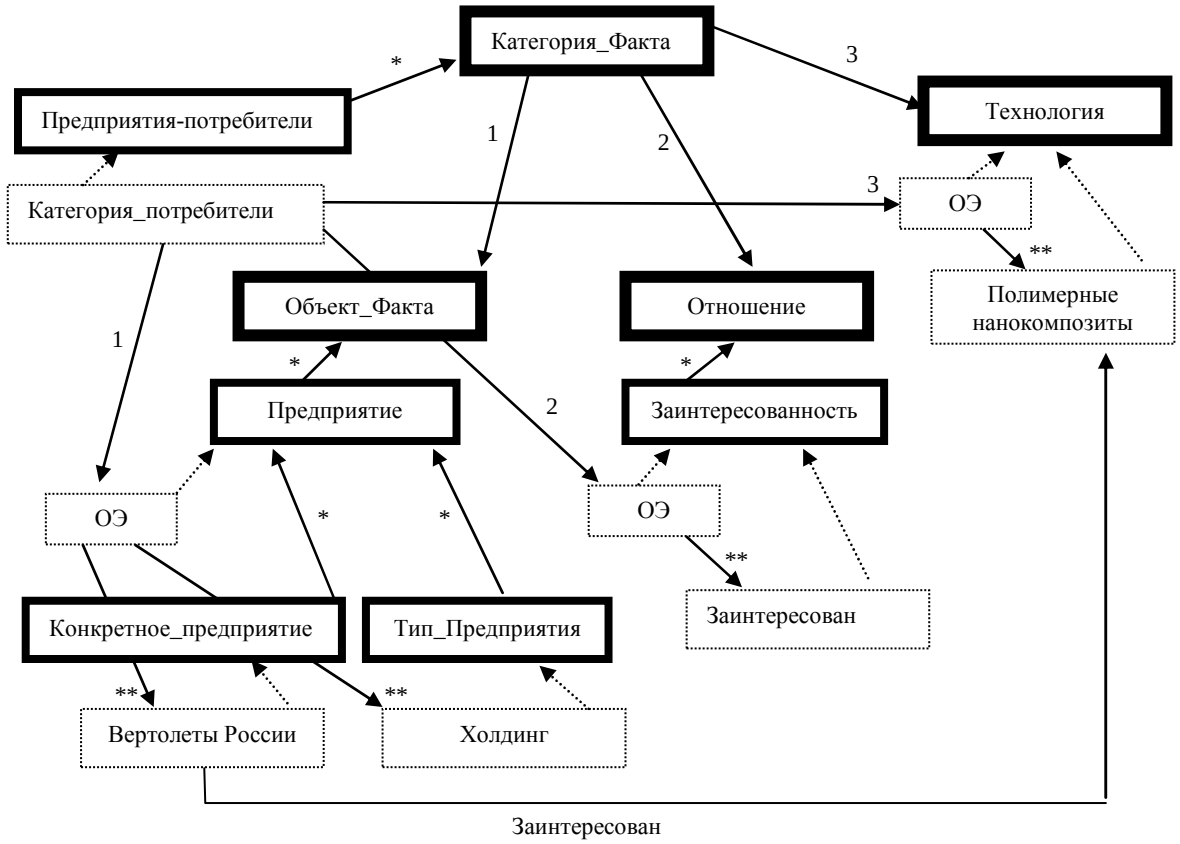

Рис. 1

EF – *извлекаемые из текстов факты;*

DF – *логически выводимые факты*.

Тогда формально вывод новых фактов можно представить в виде

$$(BK, EF) \models_{IR} DF$$

Реализация. Рассматриваемый подход к извлечению тематических фактов был реализован во фреймворке FactE. Фреймворк является ИАС, на вход принимает пользовательский запрос, содержаший название инновационной технологии, а также массив текстов, которые необходимо обработать. Считается, что список категорий фактов задан заранее, равно как и специфицированы правила вывода (ПВ) новых фактов. В соответствии с общими концепциями построения систем OBIE и с учетом предложенного подхода укрупненный алгоритм работы FactE включает в себя следующие шаги:

• Извлечение явно упомянутых в текстах тематических фактов. В рассматриваемом подходе данный процесс управляется лексической онтологией, хранящей шаблоны извлечения для категорий фактов. Каждой категории фактов ставится в соответствие имя описательного экземпляра соответствующего класса лексической онтологии. По имени этого ОЭ можно получить шаблон извлечения со всеми лексическими словоформами для отношения и объекта шаблона (субъект уже задан во входном запросе в виде названия инновационной технологии). Основные шаги этого этапа можно представить в виде псевдокода:

1) получить описательный экземпляр, представляющий для рассматриваемого набора категорий субъект фактов (Ontology[], *Technology_Name_Instance*)

2) зафиксировать в предметной онтологии субъект текущего запроса, внедрив отношение HasLexicalForm для описательного экземпляра объекта фактов (Ontology[], query, Technology_Name_Instance, *Query_Instance*)

for each document in doc_array[] do

{

//для каждого документа в текущей серии обработки

1) получить чистый текст (document, *plain_text*)

2) разбить текст на предложения (plain_text, *sentences[]*)

3) найти такие предложения, в которых есть упоминание об искомой технологии

(query, sentences[], fact_sentences[])

for each fact in fact_sentences[] do

{

//для каждого потенциального тематического факта

for each category in fact_categories[] do

{

1) получить описательный экземпляр класса онтологии, соответствующего данной категории фактов (Ontology[], category_instance_name, *Category_Instance*)

2) следуя отношению HasObject полученного описательного экземпляра категории, выявить описательный экземпляр объекта данной категории фактов (Ontology[], Category_Instance, *Category_Object*)

3) следуя отношению HasRelation полученного описательного экземпляра категории, выявить описательный экземпляр отношения данной категории фактов (Ontology[], Category_Instance, *Category_Relation*)

4) следуя отношению HasLexicalForm описательного экземпляра объекта категории, получить список его возможных лексических форм (Ontology[], Category_Object, *object_lexical_forms[]*)

5) следуя отношению HasLexicalForm описательного экземпляра отношения категории, получить список его возможных лексических форм (Ontology[],Category_Relation, *relation_lexical_forms[]*)

6) сформировать лексический паттерн, состоящий из субъекта и отношения факта (subject_lexical_forms[], relation_lexical_forms[], *pattern*)

7) проверить рассматриваемое предложение на соответствие полученному паттерну (fact, pattern, *is_Match*, *matched_object_lexical_form*)

if is_Match then

{

//если предложение соответствует шаблону, т. е. в нем были найдены //совпадения для обоих списков лексических форм

1) найти экземпляр, соответствующий найденной в тексте лексической форме объекта факта (Ontology[], matched_object_lexical_form, *Concrete_Object_Instance*)

2) получить отношение, характерное для рассматриваемой категории (Ontology[], category_instance_name, *Category_Relation*)

3) создать в лексической онтологии новый факт:

Category_Relation <Concrete_Object_Instance, Query_Instance>:(Ontology[],Category_Relation,

Concrete_Object_Instance, Query_Instance, *ontological_fact*)

4) присвоить рассматриваемому *предложению* рассматриваемую категорию (fact, category_tag)

5) сохранить факт в базу ИФ extracted_facts[] <- fact

}

end if

}

end for

}

end for

}

end for

• Логический вывод новых фактов на основе базовых знаний онтологии и ИФ. Основные шаги этого этапа можно представить в виде псевдокода:

for each inferred_fact_category in inferred_fact_categories[] do

{

1) получить правило вывода фактов, использующее базовые знания онтологии и приобретенные в результате анализа документов знания (inferred_fact_category, *inference_rule*)

2) следуя метаданным, определяющим источник для каждого предиката правила, определить массив извлеченных фактов, к которым будет применено правило (extracted_facts[], inference_rule, *required_facts[]*)

3) следуя метаданным, определяющим источник для каждого предиката правила, задействовать требуемые базовые знания и применить полученное правило вывода; получить список новых утверждений (required_facts[], basic_Knowledge[], inference_rule, *is_Match*, *derived_facts[]*)

if is_Match then

{

//если удалось найти новые факты для рассматриваемой категории

for each derived_fact in derived_facts[] do

{

1) сформировать текстовый факт из утверждения (derived_fact, *text_fact*)

2) присвоить текстовому факту рассматриваемую категорию (text_fact, inferred_fact_category_tag)

3) сохранить факт в базу ЛВФ derived_facts[]

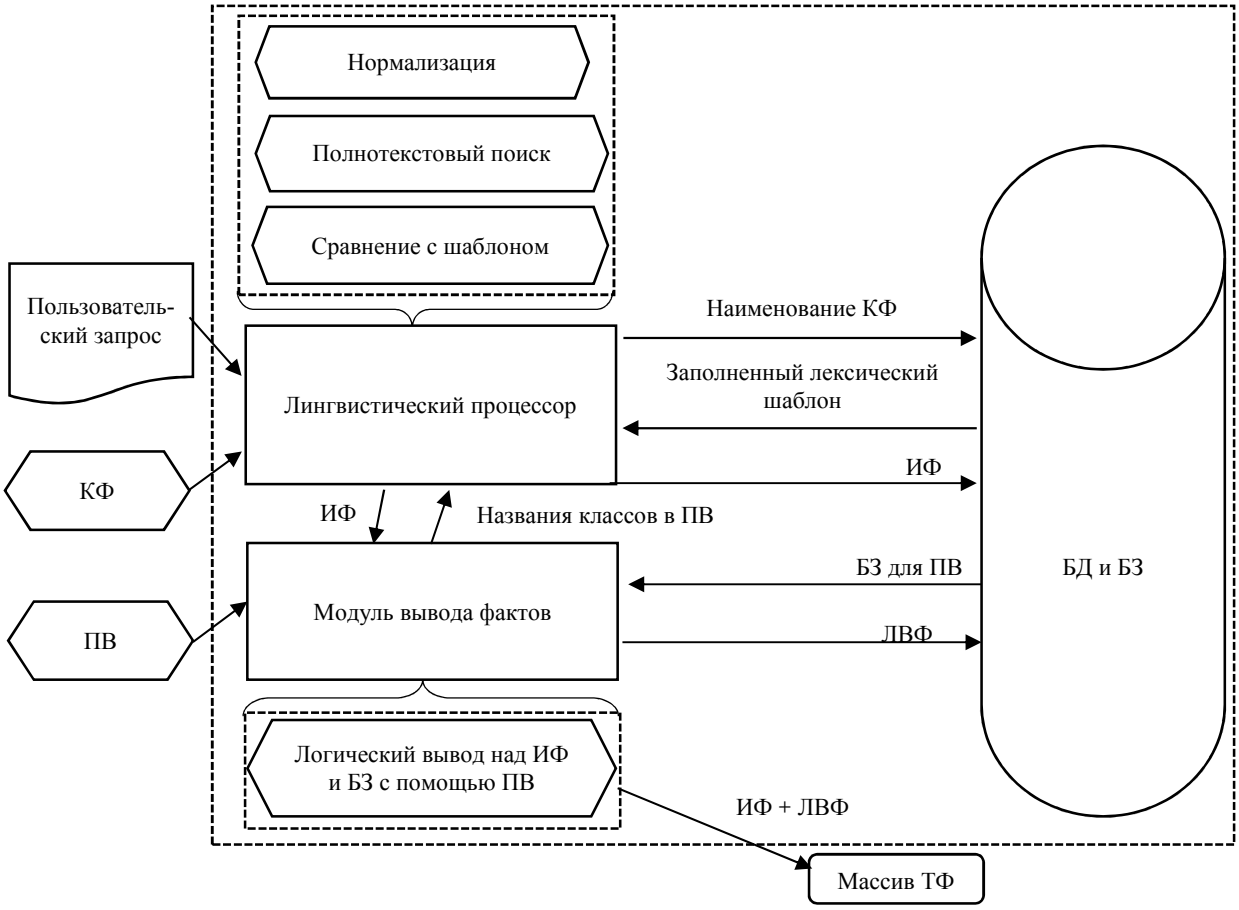

Рис. 2

<- fact
}
end for
}
end if
}
end for

На основании общего алгоритма была разработана архитектура фреймворка FactE, представленная на рис. 2. Фреймворк имеет 2 функциональных модуля:

– лингвистический процессор, решающий задачу извлечения фактов из текстов на основе предоставляемых лексической онтологией шаблонов извлечения;

– модуль вывода фактов, предназначенный для организации логического вывода новых фактов.

Работа модулей контролируется предметной онтологией, включающей как знания лексической онтологии (управляющей процессом извлечения явных знаний), так и базовые знания (использующиеся для логического вывода новых фактов).

При реализации фреймворка FactE для решения частных задач были задействованы следующие программные средства: Apache HttpClient, Apache Tika, GATE, Apache Open NLP Sentence Splitter, Apache Lucene, Apache Jena Core, Apache Jena SDB.

Экспериментальные результаты. Ниже представлены результаты тестирования предложенного подхода на практике. В качестве пользовательского запроса была выбрана технология производства полимерных нанокомпозитов.

Базовые знания предметной области FactE включают следующие утверждения:

Включает (вертолет, несущий винт) (3)

Содержит (несущий винт, лопасти) (4)

Используется для изготовления (углепластик, лопасти) (5)

Может использоваться для изготовления материала (полимерные нанокомпозиты, углепластик) (6)

Из [12] FactE был извлечен следующий тематический факт:

Планирует производить («Вертолеты России», вертолет) (7)

Тогда на основании явного факта (7), базовых знаний (3)–(6) и общего правила (2) может быть выведен следующий факт, относящийся к категории «Предприятия – потенциальные потребители технологии Т», где Т – заданная технология, в данном случае – полимерные нанокомпозиты:

Является потенциальным потребителем («Вертолеты России», полимерные нанокомпозиты).

Заключение и дальнейшее направление исследования. В статье был предложен подход к извлечению тематических фактов, явно указанных в документах или же выведенных на основе базовых знаний онтологии и ее приобретенных знаний. Рассмотрен фреймворк FactE, реализующий данный подход, его архитектура и основной алгоритм работы. Результаты показывают, что использование логического вывода на знаниях может давать существенное преимущество при решении задачи извлечения информации: некоторые релевантные и полезные знания, не указанные в документах явно и не доступные при применении традиционных средств анализа текста, могут быть обнаружены.

На следующем этапе разработки планируется улучшить качество обработки документов посредством включения более сложных лингвистических средств, позволяющих разрешать контекстные связи между предложениями, а также реализовать разрешение кореферентности, что позволит обнаружить некоторые явные знания в тексте, не доступные для идентификации в настоящий момент. Также планируется организовать проверку онтологии на консистентность при обнаружении и добавлении в базу новых фактов, что может позволить предотвратить появление заведомо ложных для данной предметной области фактов.

СПИСОК ЛИТЕРАТУРЫ

1. A Survey of Web Information Extraction Systems / C. H. Chang, M. Kayed, M.R. Girgis et al. // Knowledge and Data Engineering, IEEE Transactions. Institute of Electrical and Electronics Engineers. 2006. Vol. 18, is. 10. P. 1411–1428.

2. Toman M. Comparison of Approaches for Information Extraction from the Web // Proc. of the 9th Intern. PhD Workshop on Systems and Control: Young Generation Viewpoint, Slovenia, 2008. URL: http://dsc.ijs.si/files/papers/S202%20Toman.pdf.

3. Appelt D., Israel D. J. Introduction to Information Extraction Technology // Artificial Intelligence Communications. IOS Press. 1999. Vol. 12, is. 3. P. 161–172.

4. Wimalasuriya D. C., Dou D. Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches // J. of Information Science archive. CA,

USA: Sage Publications, Inc. Thousand Oaks. 2010. Vol. 36, is. 3. P. 306–323.

5. Wu F., Weld D. S. Autonomously semantifying Wikipedia // Proc. of the Sixteenth ACM Conf. on Information and Knowledge Management (CIKM'07). New York: ACM Press, 2007. P. 41–50.

6. Kietz J., Maedche A., Volz R. A method for semi-automatic ontology acquisition from a corporate intranet // Proc. of the EKAW'00 Workshop on Ontologies and Text, 2000. URL: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol51/Maedche.pdf

7. Cimiano P., Handschuh S., Staab S. Towards the self-annotating web // Proc. of the 13th Intern. Conf. on World Wide Web. New York: ACM Press, 2004. P. 462–471.

8. Maynard D., Peters W., Li Y. Metrics for evaluation of ontology-based information extraction // Proc. of the 4th Workshop on Evaluating Ontologies for the Web (EON2006), Edinburgh, Scotland, 2006. URL: https://km.aifb.kit.edu/ws/eon2006/eon2006maynardetal.pdf

9. KIM – semantic annotation platform / B. Popov, A. Kiryakov, A. Kirilov et al. // Proc. of the 2nd Intern. Semantic Web Conf. (ISWC2003). Heidelberg: Springer, 2003. Vol. 2870 . P. 834–849.

10. Buitelaar P., Siegel M. Ontology-based Information Extraction with SOBA // Proc. of the Fifth Intern. Conf. on Language Resources and Evaluation, Genoa, Italy, 2006. URL: http://www.lrec-conf.org/proceedings/lrec2006/

11. Maedche A., Staab S. The Text-To-Onto Ontology Learning Environment // Proc. of the Eighth Intern. Conf. on Conceptual Structures. Berlin: Springer, 2000. P. 14–18.

12. Перспективный скоростной вертолет (ПСВ) В-37. URL: http://bastion-karpenko.ru/v-37_psv/.

N. D. Yelagina, M. G. Panteleyev
Saint-Petersburg state electrotechnical university «LETI»

DERIVING OF THEMATIC FACTS FROM UNSTRUCTURED TEXTS AND BACKGROUND KNOWLEDGE

The article considers the approach to obtaining all relevant facts with the use of basic knowledge and extracted from the documents facts, as well as its implementation in the framework of the system FactE, designed for the analysis of innovative technologies. The approach is illustrated by the example output facts about the enterprises, which are potential consumers of a given innovation. Examples ontological views of relevant basic knowledge and inference rules. Presents the architecture and algorithms of the system. Discuss possible directions of its carry out further development.

Fact extraction, ontology-based information extraction, background knowledge