Итерационное извлечение шаблонов описания событий по новостным кластерам

© Д. С. Котельников
Московский государственный университет
имени М.В. Ломоносова

© Н. В. Лукашевич
Научно-исследовательский
вычислительный центр МГУ
имени М.В. Ломоносова

Москва

info@dmitriu.com

louk@mail.cir.ru

Аннотация

В статье описывается метод итерационного формирования шаблонов описания событий по новостным кластерам. Небольшое количество размеченных примеров используется для построения базовых шаблонов, которые обогащаются за счет вариативности описания события в новостных сообщениях близкой тематики. Проведены эксперименты, в которых показана возможность формирования шаблонов для различных типов отношений.

1 Введение

В связи с ростом объемов информации и развитием сети Интернет, задача автоматической обработки текста на естественном языке и извлечения структурированной информации приобретает все большую актуальность. Подобные системы обрабатывают огромное количество текстов, составляя различные базы знаний.

Традиционными подзадачами систем извлечения информации из текста являются:

1. выделение именованных сущностей, например названий организаций, людей, географических объектов, временные и денежные обозначения и др.;
2. нахождение различных обозначений одного и того же объекта в тексте;
3. извлечение фактов или отношений между несколькими сущностями.

Разработано большое количество систем выделения фактов из текстов на русском языке [8, 9, 10, 11, 12], которые используют так называемый инженерный подход, когда шаблоны для извлечения информации описываются вручную экспертами. Для такого рода систем характерна высокая трудоемкость создания и низкая полнота извлекаемой информации. Проблема состоит в том, что эксперту трудно предусмотреть все способы описания некоторого события в тексте.

Использование итерационного метода извлечения шаблонов [2] позволяет обойти основной недостаток инженерного подхода — необходимость участия человека в процессе написания новых шаблонов. Идея метода состоит в повторении двух итераций:

1. поиск документов, которые содержат участников уже установленных фактов и формирование новых шаблонов;
2. извлечение новых фактов, использующих шаблоны, полученные на первой итерации.

Для обучения алгоритму достаточно нескольких размеченных примеров. В предыдущих работах [1, 2] применимость метода исследовалась для извлечения бинарных отношений, но особенно сложной задачей для систем извлечения информации из текста является выделение ситуаций, в которых задействовано несколько участников.

Таким образом, актуальной является задача автоматического формирования шаблонов описания для произвольного события с использованием небольшого количества размеченных примеров.

В работе описывается итерационный подход к автоматическому формированию шаблонов извлечения фактов с использованием нескольких близких по смыслу новостных сообщений, объединенных в кластер. В новостном кластере часто оказывается достаточное количество предложений, в которых некоторое событие распознано вполне успешно, так и предложения, в которых содержатся те же участники, но событие не распознано вовсе. Именно эту вторую группу предложений можно использовать для получения новых шаблонов описания событий.

Небольшое количество новостных кластеров с указанием фактов, которые в них содержатся, используются для формирования базовых шаблонов. Получившиеся шаблоны применяются на всей коллекции для извлечения новых фактов, которые опять используются для формирования шаблонов. Таким образом, несколько новостных кластеров, описывающих различные случаи упоминания однотипного события, позволяют обнаружить дополнительную лексическую информацию для отражения в

Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2012, Переславль-Залесский, Россия, 15-18 октября 2012 г.

шаблонах описания события. Повторение несколько итераций позволяет достичь наилучшего результата.

Оценка качества предложенного метода исследовалась на размеченной коллекции для нескольких видов событий. Приводится сравнение полученных результатов с результатами системы, основанной на инженерном подходе.

Дальнейшее изложение статьи организовано следующим образом: в разделе 2 приводится обзор ключевых исследований итерационного метода и применения новостных кластеров для извлечения информации из текста, в разделе 3 описывается метод выделения сущностей, в разделе 4 описан наш итерационный подход к автоматическому построению шаблонов, в разделе 5 представлены результаты экспериментов и оценка предложенного подхода.

2 Обзор работ по исследуемой тематике

Впервые итерационный метод формирования шаблонов для извлечения отношений (книга, автор) из частично структурированных HTML документов описан в статье [2]. Обучение начинается с небольшого количества фактов, составленных человеком. Для извлечения шаблонов система использует контексты в виде трех подстрок (левый, правый и средний) около упоминаний сущностей. Поисковая система используется для нахождения сайтов, на страницах которых содержатся соответствия книги и её автора.

Развитием этого подхода для текстов на естественном языке является работа [1], в которой шаблоны представлены тремя векторами лемм с весами, отражающими левый, средний и правый контексты между извлекаемыми сущностями. На каждой итерации производится оценка качества получившихся шаблонов.

В работе [5] используются кластеры близких по мере PMI слов [4], полученные на коллекции из миллиона новостных сообщений:

$$PMI(w_0, w) = -\log \frac{T(w, w_0)}{T(w)T(w_0)},$$

Шаблоны учитывают порядок слов в предложении. Кластеры близких по смыслу слов используются для обогащения конкретных значений лемм в шаблонах, а так же для проверки получившихся фактов.

Использование новостных кластеров для формирования шаблонов описано в статьях [6, 7]. Для каждого кластера формируются граф участников некоторого события, после чего конкретные значения обобщаются до понятий с помощью тезауруса WordNet [3]. В получившихся графах производится поиск общих поддеревьев, которые образуют шаблоны для извлечения информации из текста.

В статье [15] приводится формальное описание правил извлечения фрагментов текста и метод автоматического построения правил, которые используются для классификации географических объектов по типам. Алгоритмы классификации строятся с помощью дискреционных процедур распознавания по прецедентам.

3 Выделение сущностей

Одной из важных подзадач извлечения информации из текста является выделение конкретных сущностей, упомянутых в тексте. Под термином сущность мы будем понимать объект определенного типа, имеющий имя; обозначения даты или времени; числовые выражения.

Тексты новостей предварительно обрабатываются морфологическим анализатором. Морфологическая омонимия частично снимается за счет согласования прилагательного и существительного, который к нему относится. Омонимия на падеж существительного частично разрешается предлогом.

3.1 Описание метода выделения сущностей

Для выделения даты используется небольшой словарь, содержащий названия месяцев, дней недели и времен года, а так же шаблоны, которые позволяют выделять дату из нескольких подряд идущих чисел с разделителями или комбинаций чисел и слов из словаря. Например, «05.07.1988», «5 августа 2006 года».

Числовые выражения начинаются с числительного или числа (нескольких чисел, разделенных точкой или запятой) и включают стоящие после него существительные. Например, «5,5 миллиардов долларов», «10 млрд. долл.», «пятьдесят процентов»

Для выделения именованных сущностей мы использовали упрощенный алгоритм, в котором написание слов с заглавной буквы является одним из основных маркеров имени сущности. Выделение именованных сущностей производится в два этапа.

На первом этапе из предложений выделяются последовательности **[P]** из одного или нескольких идущих подряд слов, которые могут содержать одну или несколько именованных сущностей, по следующему алгоритму:

1. Производится поиск слова **S** написанного с заглавной буквы или аббревиатуры, **S** добавляется в **[P]**;
2. Если перед словом **S** стоит существительное, то оно так же включается в **[P]**, вместе с прилагательными, которые к нему относятся.
3. В **[P]** включаются все аббревиатуры и слова, написанные с заглавной буквы, которые следуют непосредственно за **S**.
4. Если последовательность **[P]** заканчивается прилагательным, например «Международный», то в **[P]** также включается и существительное, к которому оно относится.

При формировании последовательностей учитываются разделители, кроме одинарных и двойных кавычек.

Например, из предложения «Россия готова участвовать в кредите ЕС Киеву на энергоцели - Путин», будут извлечены следующие последовательности: «Россия», «ЕС Киеву», «Путин». Из примера видно, что могут получаться последовательности, содержащие сразу несколько именованных сущностей, например, «ЕС Киеву».

На втором этапе производится анализ совместной встречаемости нескольких последовательностей в одном предложении, которые полностью содержатся в некоторой другой последовательности.

Для этого в кластере производится поиск нескольких сущностей из одного предложения, не граничащих друг с другом, объединение которых дает другую полную последовательность. При разделении последовательности на несколько именованных сущностей делаются дополнительные проверки на корректность получившихся сущностей.

Так в кластере, содержащем предложения:

1. Россия готова участвовать в кредите **ЕС Киеву** на энергоцели - заявил Путин;
2. В ответ на жесткую позицию России, которая заявила, что не будет выполнять подписанный **Киевом** протокол по транзиту газа из-за внесенных в него оговорок, **ЕС** и Украина говорят, что согласны рассмотреть претензии Москвы»

Последовательность «ЕС Киеву» будет разбита на две именованные сущности «ЕС» и «Киев».

3.2 Поиск синонимичных сущностей

В тексте новостного кластера один и тот же объект может описываться различными выражениями. Например, в новостном кластере про получение кредита Белоруссией от Международного валютного фонда, встречаются следующие обозначения кредитора: «МВФ», «Совет директоров МВФ», «Исполнительный совет Международного валютного фонда», «Международный валютный фонд» и др.

Для поиска различных вариантов названий одной сущности используются контексты. Если для двух сущностей полностью совпадают контексты длины два по обе стороны, то две сущности считаются синонимами. Например, для предложений:

1. Заместитель главы Газпрома Александр Медведев заявил, что поставки российского газа в **Евросоюз** через территорию Украины могут быть возобновлены.
2. Поставки российского газа в **Европу** через территорию Украины могут быть возобновлены.

Получаются синонимичные сущности: «Евросоюз» и «Европа», которые используются только в рамках одного кластера.

Второй вид контекстов, который используется для поиска различных названий одного объекта — контексты через глагол. В тексте производится поиск глаголов, и выделяется пара сущностей расположенная непосредственно по разные стороны от глагола. Если по одну сторону от глагола в двух предложениях стоят одинаковые сущности, то сущности, расположенные по другую сторону от глагола, так же считаются синонимичными. При этом предлоги, которые расположены рядом с сущностями, так же должны совпадать. Например, для предложений:

1. *Президент Дмитрий Медведев* **поручил** правительству Российской Федерации не выполнять протокол до тех пор, пока в нем не будут сняты противоречия.
2. *Президент России Дмитрий Медведев* **поручил** правительству Российской Федерации не выполнять протокол до тех пор, пока в нем не будут сняты противоречия.

будет установлена синонимичность следующих сущностей: «Президент Дмитрий Медведев», «Президент России Дмитрий Медведев».

3.3 Значимые слова

Предыдущее исследование различных описаний событий [13] показало, что присутствие некоторых лемм в предложении может указывать на наличие в нем извлекаемого события. В данной работе экспертом задается одно значимое слово, после чего значимые слова обогащаются за счет нахождения синонимичных слов. Например, для факта покупки значимыми являются слова: «купить», «приобрести» и др.

4 Описание работы системы

В качестве исходных данных для работы системы использовалось значимое слово и новостной кластер, в котором указаны основные участники и их роль в извлекаемом событии. Для всех фактов достаточно указать только один вариант названия именованной сущности.

Новостной архив Google используется для поиска кластеров, которые потенциально могут содержать извлекаемое событие. В качестве ключевого слова при поиске используются значимые слова. Из новостного архива извлекаются ссылки на документы с полным описанием новости и с сайтов новостных изданий скачиваются HTML страницы, и извлекаются тексты, которые в них содержатся.

Размеченный новостной кластер обрабатывается морфологическим анализатором и модулем выделения сущностей. Среди получившихся именованных сущностей производится поиск полных совпадений со слотами фрейма, и извлекаются шаблоны.

4.1 Шаблоны описания событий

Для извлечения информации из текста используются шаблоны, которые учитывают порядок слов в предложении и могут содержать сразу несколько участников события. При формировании и сопостав-

лении шаблонов не учитываются обозначения даты и времени.

В шаблонах используются следующие конструкции, которые позволяют обобщить конкретные значения слотов или некоторых сущностей в предложении:

1. [Number] – соответствует произвольному числовому выражению;
2. [Entity:Debtor: [Дт, Рд]] – именованная сущность, соответствующая участнику события, на которую накладывается ограничение по падежу (дательный или родительный). Если некоторое предложение сопоставится с шаблоном, то именованной сущности будет присвоена роль «Debtor».

Шаблоны строятся для предложений, в которых найдено не менее двух участников, следующим образом:

1. слоты целевого фрейма в предложении, заменяются на конструкцию [Entity:Роль] и добавляется ограничение на падеж;
2. для построения шаблона выделяется непустая подстрока лемм из исходного предложения между двумя разными слотами или глаголом и одним из слотов;
3. шаблон обязательно должен содержать глагол и значимое слово.

Например, из предложения:

«МВФ **предоставит** Белоруссии кредит на сумму $2,46 млрд»

будет извлечен следующий шаблон:

[Entity:Creditor:[Им]] <ПРЕДОСТАВИТЬ> [Entity: Debtor:[Дт]] {КРЕДИТ} [Number:Amount].

так как лемма «кредит» — значимое слово, а «предоставит» — глагол.

Получившиеся шаблоны используются для построения конечного автомата, который используется для сопоставления шаблонов с предложениями и извлечения значений слотов из текста.

Для построения шаблонов описания событий используется следующая схема работы итерационного метода:

1. {E} ← Базовые примеры
2. {O} ← FindOccurences({E}, {D})

   Поиск предложений, в которых входят значения слотов {E} фреймов в новостные документы {D}
3. {P} ← GeneratePatterns({O})

   Построение шаблонов {P} из предложений, в которых удалось найти несколько фактов
4. {E} ← ApplyPatterns({P}, D)

   Извлечение новых фактов {E} из новостных документов с использованием шаблонов
5. Если количество уникальных шаблонов увеличилось, переходим на шаг 2, иначе останавливаемся.

Наилучший результат в среднем достигается после 5-6 циклов. На каждой итерации происходит обогащение шаблонов новыми вариантами описания события.

5 Эксперименты

Экспериментальные исследования проводились на коллекциях новостных кластеров, собранных из архива Google [14]. Рассматривались события выдачи кредита и покупки, которые различаются по способу описания.

Факт получения кредита является уточнением более общего факта передачи некоторого объекта от одного участника другому, поэтому в предложении обязательно должно присутствовать значимое слово — существительное, которое уточняет объект передачи и условия. Например, «кредит», «транш», «займ», «кредитная линия» и другие.

В факте покупки нет жестких ограничений на объект, который участвует в событии, а ограничение накладывается только на совершаемое действие, поэтому значимое слово является глаголом.

Исходные данные для работы программы приведены в таблице 1. Из новостного архива было собрано несколько тысяч новостных кластеров для каждого из фактов.

Таблица 1

Исходная информация для работы системы

Факт	Покупка	Кредит
Значимое слово	«Купить»	«Кредит»
Факт	**Buyer:**«Microsoft» **Goods:** «Yahoo» **Amount:** «44,6»	**Debtor**: «Белоруссия» **Creditor**: «МВФ» **Amount**: 4

Оценка качества работы системы проводилась на коллекции из 84 кластеров, которые были размечены экспертом. Для каждого кластера эксперт выделил факты, которые в нем содержатся, а так же список возможных значений для каждого слота. Пример ручной разметки для некоторого кластера содержащего сразу два факта получения кредита:

Creditor: «ВЭБа», «Внешэкономбанка», «Внешэкономбанка ВЭБ», «ВЭБ»;

Amount: «10,2 млрд рублей»;

Debtor: «ОАО Альфа-Банк Москва», «Альфа-Банк»

Creditor: «акционеры Альфа-Банка»

Amount: «370 млн долл»;

Для оценки качества работы системы использовалась перекрёстная проверка. Множество размеченных примеров разбивалось на 4 блока, обучение производилось на ¾ выборки, тестирование на ¼.

Факт считался правильно извлеченным, если извлеченное значение для каждого слота содержится в значениях, указанных экспертом и неправильно извлеченным, если хотя бы одно значение слота извлечено неправильно. Учитывались только уникальные в пределах одного кластера факты.

В таблице 2 приведены результаты работы метода на различных разбиениях размеченной выборки.

Таблица 2

Результаты оценки метода

№	Точность	Полнота	F-мера
1	0,97	0,41	0,57
2	0,95	0,55	0,69
3	0,98	0,35	0,51
4	0,92	0,71	0,8
Среднее	**0,95**	**0,5**	**0,65**

В таблице 3 приведены результаты оценки извлеченных фактов для каждой итерации при обучении на всей размеченной коллекции. После 5 итерации новые шаблоны больше не формируются. В таблице 4 приведены результаты оценки отдельных предложений из обучающей выборки, в которых был извлечен факт. Если из двух предложений извлекался один и тот же факт, то оба предложения учитывались в результирующей оценке.

В конце таблицы приведены результаты работы одной из систем извлечения информации из текстов на русском языке, основанной на инженерном подходе (ИП). Шаблоны для выделения фактов получения кредита и покупки уже были описаны в данной системе.

Таблица 3

Результаты оценки фактов для кластеров по итерациям

№	Шаблонов	Точность	Полнота	F-мера
1	12	1	0,07	0,13
2	183	0,97	0,34	0,5
3	316	0,94	0,65	0,78
4	325	0,94	0,65	0,78
5	330	0,94	0.65	0,78
ИП	**20**	**0,95**	**0,24**	**0,38**

Таблица 4

Результаты оценки фактов для предложений по итерациям

№	Количество предложений с фактами	Количество ошибок
2	278	3
3	643	16
4	802	18
5	817	18
6	819	18
ИП	**178**	**5**

Примеры наиболее частотных шаблонов для факта получения кредита:

1. [Entity:Creditor:[Им]] <ПРЕДОСТАВИТЬ> [Entity:Debtor:[Рд, Дт, Пр]] {КРЕДИТ}
2. [Entity:Creditor:[Вн, Им]] <ВЫДЕЛИТЬ> [Entity:Debtor:[Им, Рд]] {КРЕДИТ}
3. [Entity:Debtor:[Им]] <ПОЛУЧИТЬ> {КРЕДИТ} НА [Number:Amount]
4. [Entity:Creditor:[Им]] <ВЫДЕЛИТЬ> [Entity:Debtor:[Рд, Дт, Пр]] {КРЕДИТ} В [Number:Amount]
5. [Entity:Creditor:[Им]] <ПРЕДОСТАВИТЬ> [Entity:Debtor:[Рд, Дт]] {КРЕДИТ} В [Number:Amount]
6. [Entity:Creditor:[Им]] <ДАТЬ> [Entity:Debtor:[Рд, Дт, Пр]] {КРЕДИТ}
7. [Entity:Creditor:[Им]] <ОДОБРИТЬ> ВЫДЕЛЕНИЕ [Entity:Debtor:[Рд, Дт, Пр]] {КРЕДИТ}
8. [Entity:Creditor:[Им]] <ВЫДАТЬ> {КРЕДИТ} [Entity:Debtor:[Рд]]
9. [Entity:Creditor:[Им, Вн, Пр]] <ПРЕДОСТАВИТЬ> [Entity:Debtor:[Дт, Рд]] {КРЕДИТ} В РАЗМЕР [Number:Amount]
10. [Entity:Debtor:[Им, Вн]] <ПОЛУЧИТЬ>{КРЕДИТ} ОТ [Entity:Creditor:[Рд]]

и для факта покупки:

1. [Entity:Buyer:[Им]] {КУПИТЬ} [Entity:Goods:[Рд]]
2. [Entity:Buyer:[IM]] {ПОКУПАТЬ} [Entity:Goods:[Рд]]
3. [Entity:Buyer:[Им]] <ХОТЕТЬ> {КУПИТЬ} [Entity:Goods:[Рд]]
4. [Entity:Buyer:[Им]] <МОЧЬ> {КУПИТЬ} [Entity: Goods:[Вн, Рд]]
5. [Entity:Goods:[Им]] <РЕШИТЬ> {КУПИТЬ} [Entity:Buyer:[Им, Вн, Рд]]
6. [Entity:Buyer:[Им]] {КУПИТЬ} АКЦИЯ [Entity:Goods:[Им, Вн, Рд]]
7. <РАЗРЕШИТЬ> [Entity:Buyer:[Рд]] {КУПИТЬ} [Entity:Goods:[Им]]
8. [Entity:Buyer:[Им, Дт]] <НАМЕРИТЬ> {ПРИОБРЕСТИ} [Entity:Goods:[Рд]]
9. [Entity:Buyer:[Рд]] <СОГЛАСИТЬСЯ> {КУПИТЬ} [Entity:Goods:[Им, Вн, Рд]]
10. [Entity:Buyer:[Им]] {ПРИОБРЕСТИ} {Date} [Entity:Goods:[Рд]]

В таблице 5 приведены получившиеся синонимы для значимых слов.

Таблица 5
Результаты нахождения значимых слов

Факт получения кредита	Факт покупки
АВТОКРЕДИТ	ВЫКУПИТЬ
БРИДЖ-КРЕДИТ	ДОКУПИТЬ
ГОСКРЕДИТ	ЗАКУПИТЬ
КРЕДИТ	КУПИТЬ
МИКРОКРЕДИТ	НАКУПИТЬ
СТАБКРЕДИТ	НАПОКУПАТЬ
ТРАНШ	ПЕРЕКУПИТЬ
ЭКСПРЕСС-КРЕДИТ	ПЕРЕПРОДАТЬ
	ПОДКУПИТЬ
	ПОКУПАТЬ
	ПОНАПОКУПАТЬ
	ПОПОКУПАТЬ
	ПОСТАВИТЬ
	ПРИКУПИТЬ
	ПРИОБРЕСТИ
	ПРОДАТЬ
	РАСКУПИТЬ
	РАСПРОДАТЬ
	СКУПИТЬ

Таким образом, итерационный метод автоматического построения шаблонов позволяет значительно улучшить полноту извлекаемой информации без значимого снижения точности. Экспериментальные исследования предложенного подхода показали, что обучение даже на небольшом количестве размеченных примеров, позволяет превзойти результаты работы системы, основанной на инженерном подходе.

6 Заключение

В данной работе описан итерационный метод извлечения шаблонов описания событий по новостным кластерам. Новостные кластеры используются как источник разнообразных описаний событий. Метод основан на нахождении в новостном кластере нескольких предложений с одинаковыми участниками, в одном из которых удалось обнаружить извлекаемое событие. Итерационный метод позволяет существенно сократить количество обучающих примеров и необходимость участия человека в процессе получения новых шаблонов описания событий.

Оценка предложенного подхода производилась на двух фактах получения кредита и покупки методом перекрестной проверки. Эксперименты показали применимость метода для автоматического формирования шаблонов системы извлечения информации из текста.

Литература

[1] Agichtein E., Gravano L. Snowball: extracting relations from large plain-text collections. Proceedings of the Fifth ACM Int. Conference on Digital Libraries, p. 85-94, New York, 2000.

[2] Brin S. Extracting patterns and relations from the World Wide Web. Proceedings of the 1998 Int. Workshop on the Web and Databases, p. 172–183, New York, 1998

[3] G. Miller. Wordnet: A lexical database for English. CACM, 38(11), p. 39–41, 1995.

[4] Lin D. Automatic retrieval and clustering of similar words. In Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL-98), p. 768–774, 1998.

[5] Pasca M., Lin D., Bigham J., Lifchits A., Jain A. NamesAnd Similarities On The Web: Fact Extraction In The Fast Lane. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, p. 809–816, 2006.

[6] Trampus M., Mladenić D. Constructing Event Templates from Written News. Web Intelligence and Intelligent Agent Technologies, p. 507–510, Milan, 2009.

[7] Trampus M., Mladenić D. Learning Event Patterns from Text. Informatica, Volume 35, Number 1, March 2011.

[8] Ермаков А.Е. Извлечение знаний из текста и их обработка: состояние и перспективы. Информационные технологии № 7, с. 50 – 55, М: Вид, 2009.

[9] Ефименко И., Леонтьева Н., Хорошевский В. Семантическое аннотирование под управлением предметных онтологий в проекте OntosMiner. Труды 9-й Конференции по Искусственному Интеллекту, КИИ-2004, Тверь, 2004.

[10] Ефименко И.В., Жалыбин П.П., Минор С.А., Старостин А.С., Хорошевский В.Ф. Проект OntosMiner: воспоминания о будущем. Труды 12-й Конференции по Искусственному Интеллекту, КИИ-2010, Тверь, М.: 2010.

[11] Киселев С.Л., Ермаков А.Е., Плешко В.В. Поиск фактов в тексте естественного языка на основе сетевых описаний. Труды международной конференции «Диалог 2004»: Компьютерная лингвистика и интеллектуальные технологии, с. 282-285, 2004.

[12] Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В. Извлечение информации из текста в системе ИСИДА-Т. Труды 11-й всероссийской научной конференции «Элек-

тронные библиотеки: перспективные методы и технологии, электронные коллекции», RCDL'2009, с. 247–253, Петрозаводск, 2009.

[13] Котельников Д.С., Лукашевич Н.В. Автоматизированное пополнение шаблонов для системы извлечения информации из текста. Труды 12-й всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», RCDL'2010, с. 101–107, Казань, 2010.

[14] Сайт новостного архива Google. http://news.google.ru/news/advanced_news_search?as_drrb=a

[15] Прокофьев П.А., Васильев В. Г., Извлечение информации из текста с автоматическим построением правил. Труды 13-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL'2011, Воронеж, 2011

Iterative Pattern Extraction Using News Clusters

Dmitry Kotelnikov, Natalia Loukachevitch

In this article, we describe an iterative pattern extraction approach. The extraction starts from a few original facts and improves the coverage by using duplicate information concerning the same event from news clusters. Experiments show that our approach can be used to extract various types of facts.