УДК 004.5:004.7

С. В. Прищепа
Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина

Обзор методов экстрагирования событий «из потока новостей»

Рассмотрены проблемы и актуальность экстрагирования событий. Проанализированы основные методы экстрагирования событий и сделана их классификация. Приведены примеры и описание основных методов экстрагирования. Выявлена и обусловлена необходимость совместного использования как методов, основанных на правилах, так и методов, основанных на статистике.

Ключевые слова: *экстрагирование событий, классификация методов экстрагирования, обзор методов экстрагирования событий.*

Актуальность

Ежегодно растущий поток информации и шума в сети Интернет требует от государственных органов и бизнеса огромную скорость ее обработки. Ведь выявить и проанализировать из всего потока данных ценную, правдивую и необходимую под конкретную задачу информацию — отнюдь не простое задание. Большое количество источников информации, резкое увеличение объемов новостных данных и необходимость их быстрой обработки вызвали потребность в создании систем автоматизированного анализа новостного потока. Одним из самых важных и ценных моментов в анализе новостного потока является выявление новых событий.

Задачи, связанные с обработкой новостного потока, имеют довольно специфические особенности, обусловленные природой новостных данных. Сообщения пользователей в социальных сетях и новостные сообщения являются откликами на события реального мира, и поэтому кроме текстового наполнения сюжет объединяют причинно-следственные, временные и другие факторы.

Анализ событий направлен на обработку информации, которая показывает, кто говорит или делает, что говорит или делает, по отношению к кому и когда говорит или делает, что случилось с тем или иным объектом, и множество других событий — от поглощений и банкротств до назначений на должности и рейдерских захватов. Одной из важнейших задач автоматизированной обработки новостного потока в глобальной сети Интернет можно назвать выявление новых событий и

разбиение документов (сообщения в соцсетях, новостные статьи и посты в блогах) на сюжеты/категории.

Государству и бизнесу нужны средства, способные производить поиск тех данных, которые действительно необходимы, а также трансформировать найденные данные в удобно воспринимаемую информацию, необходимую конкретному потребителю. Экстрагирование информации позволяет построить базу данных заданных (интересующих) связей или событий из новостей, форумов, финансовых отчетов, социальных сетей или других источников.

Данная тема имеет множество не решенных проблем и путей усовершенствования уже найденных методов экстрагирования событий.

Как изучалась предметная область

Конкретным типом знаний, которые могут быть извлечены из текста, посредством text mining, являются события. Они могут быть представлены в виде сложных комбинаций отношений (взаимосвязей), связанных с набором эмпирических наблюдений из текстов.

В ходе изучения был проведен поиск и анализ публикаций в сервисе Google Scholar по теме экстрагирования событий и выявлено более пятидесяти актуальных и релевантных теме экстрагирования событий научных материалов (трудов) за прошедшие два года. Изначально библиография релевантных авторов была сформирована автоматически, с помощью зондирования сети Google Scholar по теме «экстрагирования событий из текстов» [12], затем, в ходе ее аналитики, она была исправлена и дополнена. Некоторые материалы из сформированной библиографии несколько отходят от темы непосредственно экстрагирования событий, но в целом, являются хорошим источником знаний по теме «экстрагирования событий из текстов» или решают смежные с данной задачей вопросы, к примеру, выявление сущностей, объектов, лиц, экстрагирования информации в целом.

Из наиболее актуальных и релевантных авторов по теме экстрагирования событий система зондирования сети Google Scholar определила таких как: Oren Etzioni, Alan Ritter, Harith Alani, Alan Ritter, Daniel Weld, Gregor Leban, Michael Strube и Hristo Tanev. Исходя из списка авторов, был произведен поиск и изучение актуальных материалов по теме. Также, с помощью зондирования сети Google Scholar, были построены взаимосвязи и найдены связанные с тематикой темы и ключевые слова, что изображено на рис. 1.

Классификация методов экстрагирования информации

К основным направлениям экстрагирования, которые описываются в научных статьях и патентах, можно отнести 3 вида:

— экстракция событий, основанная на знаниях (правилах);

— экстракция событий, основанная на статистике;

— гибридная экстракция событий.

Как и всегда, у всех методов есть свои плюсы и минусы. Для качественной экстракции событий на практике, как правило, используют гибридную модель добычи событий, а уже она может быть в большей или меньшей степени основана на знаниях или статистике соответственно.

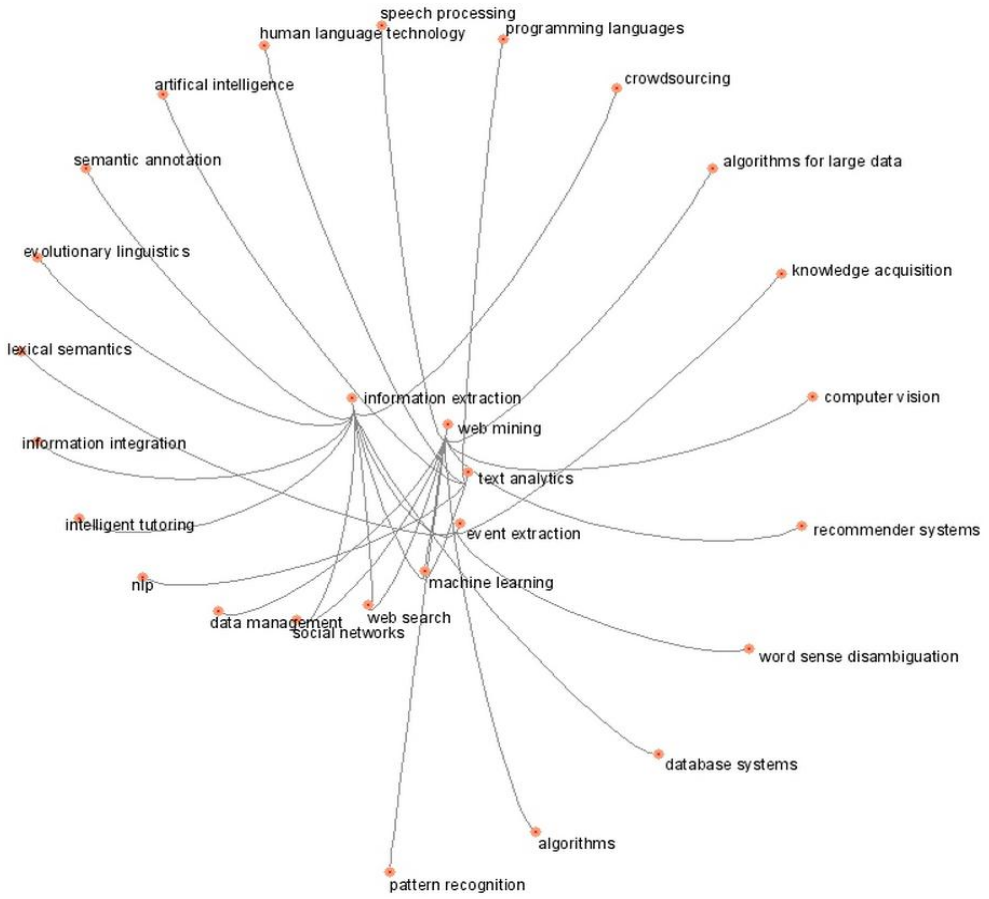

Рис. 1. Граф связанных с темой «Экстрагирования событий из текстов» понятий

Важно отметить, что некоторые из этих методов можно разделить на различные подметоды. На рис. 2 приведен неполный перечень классификации методов экстрагирования информации, на самом деле их количество в разы больше, на рисунке изображены лишь основные.

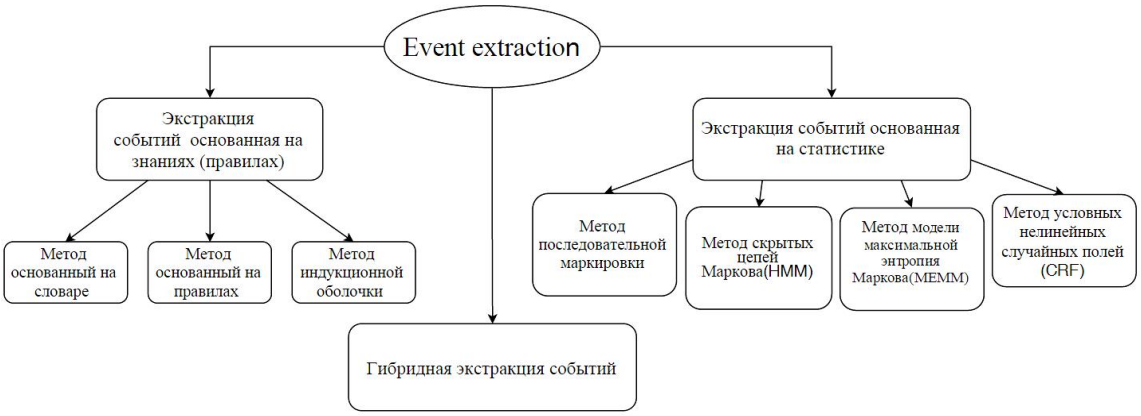

Рис. 2. Структурная схема классификации методов экстрагирования событий

Экстрагирование событий, основанное на знаниях (правилах)

Метод экстрагирования событий, основывающийся на знаниях, или как его еще часто называют «правилах», базируется на моделях, выражающих правила, представляющие экспертные знания. Они построены на основе лингвистического и лексикографического знания, а также человеческого познания относительно текста, который должен быть обработан. Это облегчает проблемы со статистическими методами относительно смысла текста. Информация в данном подходе добывается из корпусов текста с помощью предопределенных или обнаруженных языковых закономерностей, которые могут быть либо лексико-синтаксическими, либо лексико-семантическими шаблонами.

Примером такого подхода может быть фреймворк для определения событий из Твиттера, опубликованный в научной статье Аленом Риттером и другими [1]. Данный фреймворк нацелен на изучение и применение лексико-семантических шаблонов. Экстрагированная информация используется для наполнения базы знаний. Учеными института была представлена методология, нацеленная на экстракцию событий для автоматической системы раннего предупреждения. Авторы применяют лексико-семантические шаблоны для соответствия понятий, а зависимость цепочки усиливается с помощью словарей (списков слов). Соответственно, эти понятия соответствуют всем синтаксически связанным цепям выражений, передавая их составляющие понятия, что встречаются с такими же предложениями.

Извлечение событий на основе правил можно поделить на следующие методы:

1) метод, основанный на словаре;
2) метод, основанный на правилах;
3) метод индукционной оболочки.

Метод, основанный на словаре

Традиционные системы добычи информации, в первую очередь, строят шаблон-словарь, а затем используют его для извлечения новых данных из текстов. Такие системы экстракции называются методами, основанными на словарях или на шаблонах (pattern based systems). Ключевым моментом в таких системах является, как создать и улучшить словарь шаблонов, что может быть использован для идентификации релевантной информации в текстах.

Метод, основанный на правилах

В отличие от предыдущего метода, метод, основанный на правилах, использует несколько основных правил вместо словаря. Такие системы в основном используются для экстракции информации с полуструктурированных веб-страниц [2].

Обычно такой метод сводится к изучению синтаксических/семантических ограничений и разделителей, что связывали текст для извлечения, чтобы распознать правила построения границ целевого текста.

Метод индукционной оболочки

Метод индукционной оболочки — это еще один тип методов, основанных на правилах. Он направлен на структурированные и полуструктурированные документы типа веб-страниц [3]. Данный метод экстракции состоит из правил экс-

тракции и набора программных кодов, которые должны применять данные правила. Метод индукционной оболочки — техника для автоматического изучения оболочек, в которых размещены данные. Получая множество обучающих данных, алгоритм изучает оболочку для экстрагирования целевой информации из распознанных шаблонов. К примеру, в структурированной HTML-странице это могут быть теги <B>*</B> <I>*</I> и т.д, где * — любая информация, что подвергается или может подвергаться экстрагированию и обработке.

Экстрагирование событий, основанное на статистике

Такой метод извлечения событий в глобальных сетях направлен на преобразование данных в знания через использование статистики, машинного обучения и линейной алгебры**.**

Такие подходы опираются исключительно на количественные методы для обнаружения взаимосвязи в событиях. Данные методы нуждаются в больших объемах данных (текстовых корпусов), чтобы разработать модели, работающие с языковыми феноменами. У данного метода есть различные подходы, но все они основаны на определении статистических связей.

Использование данного метода на практике можно найти в литературе. К примеру, в 2009 году Масаюки Окамото и Масааки Кикучи для выявления нерегулярных или локальных событий разработали фреймворк, использующий иерархические техники кластеризации [4]. В то время как сама кластеризация может уже давать перспективные результаты, авторы использовали комбинацию взвешенных неориентированных двудольных графов и кластеризацию для извлечения ключевых понятий и существенных событий из ежедневных веб-новостей.

Техники кластеризации также можно найти у Христо Танев и других соавторов [5]. Они нацелились на экстракцию событий в реальном времени, но сфокусировались сугубо на насильственных событиях и катастрофах. Авторы использовали автоматическую маркировку слов, и представленный ими фреймворк построен для автоматического обучения шаблонов изучаемых событий.

Недостатком данного метода является то, что они обнаруживают отношение корпусов текста без учета семантики, то есть не проводится анализ, связаны ли они по смыслу.

Добыча событий, основанная на статистике, может быть реализована огромным количеством методов, среди которых рассмотрим следующие:

— метод последовательной маркировки;
— метод скрытых моделей Маркова;
— метод модели максимальной энтропии Маркова;
— метод условных случайных полей.

Метод последовательной маркировки

Информационное извлечение может приводиться как задача последовательного маркирования. В последовательной маркировке документ рассматривается как последовательность маркеров и последовательность меток, предназначенных каждому маркеру в зависимости от формы слов.

На рис. 3 продемонстрирован пример разбора предложения, взятого из статьи, опубликованной на одном из новостных сайтов: «Младший сержант киевской патрульной полиции Юрий Зозуля возглавил новую патрульную полицию во Львове». Здесь: [Субъект: Юрий Зозуля] [Основная метка: возглавил] [Объект: патрульную полицию] [Локация: во Львове] [Дата: 20.08.2015]

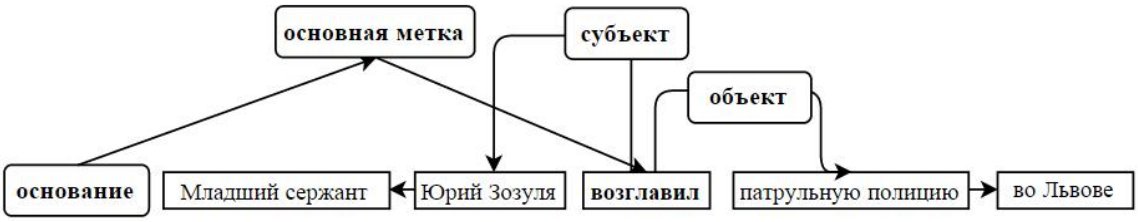

Рис. 3. Пример разбора и маркировки предложения методом последовательной маркировки

В методах извлечения на основе последовательной маркировки набор меток должен быть определен в соответствии с заданием экстракции. При добыче метаданных в научно-исследовательских работах такие метки как название, автор, e-mail, автореферат — заранее определены. Документ рассматривается в качестве последовательности наблюдения X [6]:

$$X = (x_1, x_2, ...x_n).$$

Единицей наблюдения может быть слово, текстовая строка или любой другой блок. Задача данного метода формально состоит в том, чтобы отыскать последовательность $y^* = (y_1, y_2, ..., y_n)$, максимизирующую условную вероятность $P(Y \mid X)$, где, $y^* = \arg\max_y P(y \mid x)$.

Отличие данного метода от методов, основанных на правилах, состоит в том, что данный метод позволяет описывать зависимости между целевой информацией. А затем данные зависимости могут быть использованы для повышения точности извлечения.

Метод скрытых моделей Маркова

Метод скрытых моделей Маркова (СММ) является одной из наиболее распространенных генеративных моделей, используемых в настоящее время. В данном методе каждая наблюдаемая последовательность рассматривается как сформированная последовательность переходов между состояниями, начиная с некоторого начального состояния и заканчивая, когда некоторое, заранее назначенное, конечное состояние достигнуто.

В каждом состоянии элемент наблюдения последовательности генерируется стохастически, прежде чем перейти к следующему состоянию. Каждое состояние ассоциируется со специальным тегом-меткой. Тег, связанный с любым словом, затем можно использовать для объяснения этого слова в некотором роде. Следовательно, можно найти последовательность тегов, что будет наилучшей для каждого данного предложения путем определения последовательности состояний.

Состояния в СММ рассматриваются как скрытые из-за двойной стохастической природы процесса, описываемого данной моделью. При этом скрытая от наблюдателя последовательность может быть просмотрена с учетом стохастических свойств, определяемых наблюдаемой последовательностью. Наилучший принцип выявления лучшей последовательности в наблюдаемых последовательностях основывается на использовании моделей для маркировки последовательных данных с конечным числом состояний.

Формально, метод скрытых моделей Маркова полностью определяется такими параметрами как [7]:

— конечное множество состояний Y;

— конечный выходной алфавит X;

— условное распределение $p(y' \mid y)$, представляющее вероятность перехода от состояния y к состоянию y', где $y, y' \in Y$. Распределение вероятностей наблюдения $p(x \mid y)$ представляется как вероятность наблюдения x, где $x \in X$, $y \in Y$;

— первоначальное состояние распределения $p(y), y \in Y$.

Генеративные модели, к которым относят и скрытые модели Маркова, определяют совместное распределение вероятностей $P(X,Y)$, где X и Y являются случайными величинами, соответственно проходит последовательность наблюдений и соответствующая ей метка последовательности. Для того чтобы рассчитать условную вероятность $P(X \mid Y)$ применяется правило Байеса [7]:

$$y^* = \arg\max_y p(y \mid x) = \arg\max_y \frac{p(x,y)}{p(x)}.$$

Метод модели максимальной энтропии Маркова

Модели максимальной энтропии Маркова являются формой дискриминационных моделей для маркировки последовательных данных. Вместо моделирования совместного распределения вероятностей на наблюдательные метки и метки последовательностей дискриминационные модели определяют условное распределение $P(X \mid Y)$ на наблюдения и метки последовательности. Это означает, что при определении наиболее вероятной последовательности метки для данного наблюдения последовательности, дискриминационные модели используют условное непосредственное распределение, не делая какой-либо зависимости предположений по наблюдениям или перечисляя все возможные последовательности наблюдений для расчета предельной вероятности $P(X)$.

Метод модели максимальной энтропии Маркова рассматривает наблюдаемую последовательность скорее как восстановленную, чем сгенерированную последовательностью меток. Таким образом, вместо определения двух типов распределения, данный метод имеет только один набор по отдельности подготовленных распределений такого вида [8]:

$$p(y' \mid x)(y' \mid y, x),$$

что представляют собой вероятность перехода от состояния y к состоянию y' по наблюдению за x.

Условные случайные поля

Условные случайные поля (Conditional Random Field, CRF) [8, 9] в настоящее время широко используются при извлечении информации, это метод классификации, отличием которого является возможность учета контекста объекта. Достоинством модели является то, что она не требует моделировать вероятностные зависимости между наблюдаемыми переменными. Модель применяется для решения задач машинного обучения, где прецедентом является последовательность случайных величин с поставленными им в соответствие метками, т.е. задач разметки и сегментации последовательностей. Это обуславливает применение CRF в задачах обработки естественного языка.

Метод скрытых моделей Маркова, метод модели максимальной энтропии Маркова и линейные цепи условных случайных полей могут только моделировать зависимости между соседними метками. Но иногда важно моделировать определенные виды далеко отдаленных зависимостей между объектами.

Подход к обнаружению значимых событий и их распределение, основанное на алгоритме латентного размещения

В экстрагировании и классификации событий с коротких текстовых документов нашла свое применение модель на основне латетного размещения Дирихле (LDA). Алгоритм латентного размещения Дирихле предназначен для описания текстов с точки зрения их тематик. Основное предположение модели состоит в том, что каждый документ имеет несколько тематик, смешанных в некоторой пропорции. В данной модели специальный алгоритм генерирует историю данных, связанных с типами событий как скрытыми переменными, в то время как Байесовские методы вывода используются для инвертирования порождающего процесса и вывода соответствующего набора типов для описания наблюдаемых событий. Для заключения используются выборки Гиббса в виде темы, основных фраз события и объектов (лиц), принимающих участие в событии [10, 11].

Одной из проблем коротких текстов в Интернете можно назвать огромное количество неважных для нашего внимания событий, что повседневно публикуются пользователями и могут быть интересны лишь их узкому кругу друзей и родственников. Подход к обнаружению значимых событий и их распределение, основанное на латентной переменной модели, решает данную проблему.

Латентное размещение Дирихле — это вероятностная модель порождения текста, обучение которой позволяет выявить для каждого документа вероятностное распределение по тематикам, что в свою очередь, позволяет проводить экстрагирование и распределение событий на более высоком и качественном уровне. Это иерархическая байесовская модель, состоящая из двух уровней (рис. 4):

1) смесь, компоненты которой соответствуют «темам»;

2) мультиноминальная переменная с априорным распределением Дирихле, которое задает «распределение тем» в документе.

В отличие от обычной кластеризации с априорным распределением Дирихле или обычного «наивного» байесовского классификатора, в данном методе не выбирается кластер один раз, а затем добавляются слова из этого кластера, а для каждого слова сначала выбирается по распределению θ тема, а уже затем добавляется это слово по данной теме.

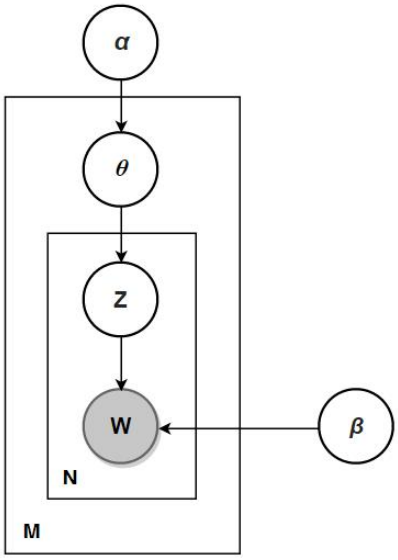

Рис. 4. Граф иерархической Байесовской модели, где: M — количество документов; N — количество слов в документе; W — темы слов; Z — темы слов в документах; θ — распределение тем в документе; α — параметр распределения Дирихле тем в документе; β — параметр распределения Дирихле слов в документе

На выходе после обучения модели LDA получаются векторы θ, показывающие, как распределены темы в каждом документе, и распределения β, показывающие, какие слова более вероятны в тех или иных темах. Таким образом, из результатов LDA легко получить для каждого документа список встречающихся в нем тем, а для каждой темы — список характерных для нее слов, т.е. фактически описание темы. Система обучается сама, обучение проходит автоматически, также большим плюсом такого подхода является то, что размечать набор текстов предварительно не нужно!

Гибридное экстрагирование событий

Несмотря на достоинства обеих подходов в экстрагировании событий, на практике очень трудно оставаться в рамках одного из них. Учитывая, что у обоих подходов, описанных выше, есть свои недостатки, то их совмещение может дать лучший результат.

На практике, большинство систем основаны на «извлечении из текстов информации о событиях на основе имеющихся правил и знаний», которые, в свою очередь, опираются на методы управления данными. Такой гибридный подход позволяет решить проблему нехватки экспертных знаний или проводить быструю самонастройку.

В гибридной системе экстракции событий в большей степени используется подход управления данным. Количество необходимых данных увеличивается, но, как правило, необходимое количество данных в таком подходе меньше чем в случае использования только подхода управления данными. По сравнению с добычей событий, основанной на знаниях, сложность увеличивается за счет комбинации нескольких методов. С другой стороны, количество экспертных знаний, необходимых для эффективного обнаружения событий в целом — меньше, чем для основанных на шаблонах методов, из-за того, что отсутствие предметной области знаний может быть компенсировано использованием статистических методов.

Что касается интерпретируемости, то приписывание результатов к конкретным частям извлеченных событий является более тяжелым, по сравнению с подходом управления данным. Все же, интерпретируемость все еще проводится с использованием семантики.

Недостатки гибридных подходов, в основном, связаны с междисциплинарными аспектами гибридных систем.

Выводы

Новые методы и улучшение уже используемых методов экстрагирования событий — одна из важнейших задач, как в военной и политической сфере, так и в бизнесе. Если раньше количество информации позволяло обрабатывать ее вручную, то сейчас с ростом количества доступной информации и форматов ее представления в сети, большим количеством фейковой информации и спама, без автоматизации и специальных методов экстрагирования событий и другой информации — никуда.

В ходе работы была изучена предметная область, сформирована сеть понятий, получен библиографический список авторитетных авторов и их работ по данной тематике в Google Scholar.

В результате анализа литературы, составлен список основных направлений и методов экстрагирования, которые описываются в научных статьях и патентах или применяются на практике, сделана классификация основных методов экстрагирования событий, рассмотрены методы и научные подходы к выявлению новых событий.

1. *Ritter A.* Named Entity Recognition in Tweets: An Experimental Sudy // A. Ritter, O. Etzioni // Proc. of EMNLP. — 2011. — P. 1524–1534.

2. *Saif H.* Semantic sentiment analysis of twitter / H. Saif, H. Alani // The Semantic Web-ISWC. — 2012.

3. *Freitag D.* Boosted wrapper induction // D. Freitag, N. Kushmerick // Proc. of 17-th National Conf. on Artificial Intelligence. — 2000. — P. 577–583.

4. *Masayuki O.* Discovering Volatile Events in Your Neighborhood: Local-Area Topic Extraction from Blog Entries / O. Masayuki, K. Masaaki // Proc. of 5-th Asia Information Retrieval Symposium, AIRS 2009. — Sapporo (Japan). — 2009, October 21–23.

5. *Tanev H.* Enhancing Event Descriptions through Twitter Mining / H. Tanev, M. Ehrmann, J. Piskorski // Proc.of ICWSM. — 2012.

6. *Qi L.* Joint Event Extraction via Structured Prediction with Global Features / L. Qi, J. Heng, H. Liang // Proc. of 51-st Annual Meeting of the Association for Computational Linguistics (ACL). — 2013.

7. *Collins M.* Discriminative training methods for Hidden Markov models: theory and experiments with Perceptron algorithms / M. Collins // Proc. of the Conf. on Empirical Methods in NLP. — 2002.

8. *McCallum A.* Maximum Entropy Markov Models for information extraction and segmentation / A. McCallum, D. Freitag, F. Pereira // Proc. of the 17-th Internation. Conf. on Machine Learning. — 2000. — P. 591–598.

9. *Lafferty J.* Conditional Random Fields: 34 Probabilistic models for segmenting and labeling sequence data / J. Lafferty, A. McCallum, F. Pereira // Proc. of the 18-th Internation. Conf. on Machine Learning (ICML'01). — 2001. — P. 282–289.

10. *Finkel J.* Incorporating non-local information into information extraction systems by gibbs sampling / J. Finkel, T. Grenager, C. Manning // Proc. of the 43-rd Annual Meeting of the Association for Computational Linguistics. — 2005. — P. 363–370.

11. *Pinoli P.* Latent Dirichlet Allocation based on Gibbs Sampling for gene function prediction. Proc. of the Internation. Conf. on Computational Intelligence in Bioinformatcs and Computational Biology / P. Pinoli, D. Chicco, D. Masseroli // IEEE Computer Society. — 2014. — P. 1–7.

12. *Ланде Д.В.* Автоматична побудова термінологічної мережі як моделі предметної області / Д.В. Ланде, С.В. Прищепа, Т.В. Синькова // Реєстрація, зберігання і оброб. даних — 2015. — Т. 17, № 3. — С. 22–29.

Поступила в редакцию 23.11.2015