

УДК 378; 004.89

Подход к автоматизированному мониторингу и прогнозированию развития инновационных образовательных технологий

Андреев А.М.¹, Березкин Д.В.¹,
Козлов И.А.^{1,*}

[*kozlovilya89@yandex.ru](mailto:kozlovilya89@yandex.ru)

¹МГТУ им. Н.Э. Баумана, Москва, Россия

В рамках международного научного конгресса "Наука и инженерное образование. SEE-2016", II международная научно-методическая конференция «Управление качеством инженерного образования. Возможности вузов и потребности промышленности» (23-25 июня 2016 г., МГТУ им. Н.Э. Баумана, Москва, Россия).

В статье представлен подход к автоматизированному мониторингу и прогнозированию развития инновационных образовательных технологий на основе анализа потока текстовых документов. Предложенный подход включает обнаружение в текстовом потоке событий, относящихся к теме образовательных технологий, формирование ситуаций, определение возможных сценариев их дальнейшего развития и подготовку предложений по действиям, которые необходимо предпринять для успешного внедрения выявленных инновационных технологий в учебный процесс ВУЗа. Приведена многокритериальная модель события, предложен метод обнаружения событий на основе инкрементальной кластеризации. Представлен метод формирования сценариев на основе принципа исторической аналогии. Проведена экспериментальная оценка разработанных моделей и методов.

Ключевые слова: обнаружение событий, сценарный анализ, инкрементальная кластеризация

Введение

В настоящее время в образовании появляются новые методы и технологии обучения, ориентированные на формирование компетенций, востребованных современным обществом. Чтобы соответствовать современным стандартам качества образования, вузам требуется внедрять эти методы в учебный процесс. При этом внедрение нововведений связано с риском, поскольку невозможно гарантировать успешность инновационных педагогических подходов.

В связи с этим существует необходимость обнаружения новых образовательных методов и технологий, отслеживания их развития с течением времени и прогнозирования их возможного дальнейшего развития в будущем с целью определения их перспективности и целесообразности внедрения в учебный процесс ВУЗа. С этой целью может выполняться периодический сбор и последующий анализ текстовых документов из открытых Интер-

нет-источников, а также из специализированных ресурсов. При этом выполнение экспертом анализа этой информации в исходном виде затруднительно ввиду огромного количества документов, генерируемых источниками.

В данной статье рассматривается решение задачи автоматизированного мониторинга и прогнозирования развития образовательных технологий на основе анализа потока текстовых документов. При этом понятие «образовательные технологии» мы трактуем широко, включая в эту категорию не только технологии обучения, но и новые перспективные направления в науке и технике, которые необходимо учитывать при планировании и проведении учебного процесса в ВУЗе.

1. Постановка задачи

В качестве основы для подхода к мониторингу развития новых научных направлений и инновационных образовательных технологий предлагается концепция “4С+П” (сообщения – события – ситуации – сценарии + предложения), которая предполагает выполнение следующих этапов анализа текстового потока:

1. Обнаружение в потоке событий, релевантных теме образовательных технологий.

Под событием будем понимать любое происшествие или явление, связанное с наукой и образованием и отраженное в текстовых документах (научных статьях, новостных сообщениях СМИ, нормативных документах, связанных с образованием). Выделение событий позволяет существенно сократить затраты времени на мониторинг интересующей пользователя темы (поскольку ему не приходится просматривать множество документов, посвященных одному и тому же событию).

В [1-3] рассматриваются различные характеристики событий, такие как место действия, время, множество действующих лиц, условия, результат. Эти характеристики представляют собой аспекты событий, учитываемые человеком при ручном анализе потока документов. Следовательно, при выполнении автоматического обнаружения и интерпретации событий эти аспекты также должны приниматься во внимание.

При ручной работе с текстовым потоком перечисленные аспекты учитываются по-разному в зависимости от целей анализа. Так, при выявлении событий, связанных с проведением научных конференций, важную роль играет пространственный аспект (место проведения), а при отслеживании достижений некоторых научных школ – множество действующих лиц (ученые, представляющие эти школы). В связи с этим разрабатываемый метод автоматического обнаружения событий должен иметь возможность гибкой настройки в зависимости от предметной области и особенностей решаемой задачи.

Кроме того, для пользователя важно иметь возможность ознакомиться со всей имеющейся информацией о наступившем событии, поэтому помимо обнаружения события в текстовом потоке необходимо формировать набор документов, описывающих его.

2. Отслеживание развития ситуаций на основе обнаруженных событий.

Под ситуацией будем понимать цепочку взаимосвязанных событий, отражающую развитие некоторых научных направлений, образовательных технологий, методов обуче-

ния, потребностей организаций в специалистах с требуемым уровнем компетенции и т.п. При анализе текстового потока обнаруженные события необходимо объединять в такие цепочки. При этом нужно учитывать нелинейность развития анализируемых процессов: каждое событие может быть включено во множество цепочек. Так, событие “Круглый стол на тему гибридных моделей организации учебного процесса в вузе” включено как в ситуацию, отражающую ход конференции об электронном обучении, так и в ситуацию, отражающую развитие гибридных моделей организации учебного процесса в России.

3. Определение возможных сценариев дальнейшего развития ситуаций.

Формирование сценариев состоит в построении цепочек событий, являющихся потенциальными продолжениями текущей ситуации. Для обеспечения эффективного использования результатов прогнозирования в целях управления учебным процессом вуза из всех вариантов необходимо выделять сценарии, представляющие наибольший интерес для лиц, принимающих решения (ЛПР) в выработке стратегии развития вуза и формировании образовательных программ – оптимистический, пессимистический и наиболее вероятный.

4. Подготовка предложений.

На основе сформированных сценариев должны быть подготовлены рекомендации по действиям, которые необходимо предпринять для успешного внедрения выявленных инновационных технологий в учебный процесс вуза.

2. Существующие подходы к мониторингу и прогнозированию развития ситуаций

В настоящий момент отсутствует комплексный подход к анализу текстового потока, позволяющий выполнять автоматический мониторинг и прогнозирование развития ситуаций. Существующие подходы [4] позволяют прогнозировать наступление отдельных событий, но не дают возможности получать сценарии дальнейшего развития ситуаций.

При этом существует множество работ, посвященных задаче обнаружения событий в потоке текстовых документов. Представленные в них методы можно разделить на два класса в зависимости от постановки задачи: методы обнаружения новых событий (New Event Detection, NED) и методы разделения анализируемых документов на группы, соответствующие различным событиям.

Методы первого класса обнаруживают появление в текстовом потоке важной новой информации. Момент возникновения нового события определяется

- появлением документа с большим весом, отражающим его новизну и значимость [5];
- изменением закона распределения документов по кластерам [6];
- появлением документа, не содержащего информации о событиях, уже описанных в ранее загруженных сообщениях [7-9].

Такие методы имеют общий недостаток с точки зрения требований, представленных в п. 1: они не способны агрегировать сообщения, описывающие новое происшествие.

Методы второго класса выполняют кластеризацию потока текстовых документов, определяя каждое из сообщений в группу, соответствующую некоторому событию.

Метод, предложенный в [10], предполагает разделение документов по темам с последующей группировкой сообщений в рамках каждой темы по темпоральному принципу.

В [11] предложено использовать иерархическую кластеризацию для выделения событий в статической коллекции документов. Этот метод позволяет осуществлять лишь ретроспективный анализ и не может быть применен для обработки потока сообщений.

В [12,13] предложены методы кластеризации текстового потока на основе генеративных моделей. Хотя предложенные модели отражают основные аспекты события, они не имеют возможности настройки в соответствии с различными предметными областями.

Одним из наиболее популярных подходов, относящихся ко второму классу, является разбиение текстового потока на события с помощью инкрементальной кластеризации [14]. Она позволяет относить документы к тем или иным событиям сразу после их загрузки. Этот подход наиболее близок к поставленным требованиям, однако существующие методы, основанные на инкрементальной кластеризации, используют для анализа документов лишь состав их слов, не учитывая другие важные аспекты событий.

3. Предлагаемый подход к решению задачи

В соответствии с концепцией “4С+П” и вышеописанными требованиями предложен следующий подход к автоматизированному мониторингу и прогнозированию развития инновационных технологий в образовании.

В потоке текстовых документов выявляются события ε_i , релевантные теме образовательных технологий. В основе обнаружения событий лежит разбиение текстового потока на кластеры таким образом, что каждый кластер содержит документы, которые описывают некоторое событие. Для выполнения распределения документов по кластерам, каждое загружаемое сообщение d_j сопоставляется с ранее обнаруженными событиями, причем сравнение выполняется на основании множества критериев, соответствующих различным аспектам события. На основе многокритериального сравнения рассчитывается значение близости между документом и событием и принимается решение о соответствии документа событию.

Из множества обнаруженных событий выделяются пары взаимосвязанных событий $p_{ij} = (\varepsilon_i, \varepsilon_j)$, потенциально принадлежащих одной ситуации. Для их выделения выполняется попарное сравнение событий с точки зрения всех аспектов. На основе формирования таких пар выполняется построение ситуационного графа $G = (E, P)$. В этом графе узлы $E = \{\varepsilon_i\}$ соответствуют событиям, а ребра $P = \{p_{ij}\}$ – выделенным парам. Любой путь в этом графе является потенциальной ситуацией $s = (\varepsilon_s^1, \varepsilon_s^2, \dots, \varepsilon_s^n)$.

Для текущей ситуации s_c выполняется построение сценариев её возможного дальнейшего развития: $\Xi = \{\xi_i\}$, $\xi_i = \langle s_{\xi_i}, rec_{\xi_i}, P_{\xi_i} \rangle$, где

$s_{\xi_i} = (\varepsilon_{\xi_i}^1, \varepsilon_{\xi_i}^2, \dots, \varepsilon_{\xi_i}^n)$ – гипотетическая последовательность событий, которые могут наступить в будущем;

$rec_{\xi_i} = \langle action_{\xi_i}, actor_{\xi_i}, period_{\xi_i} \rangle$ – предложения по действиям $action_{\xi_i}$, которые должны быть предприняты лицом $actor_{\xi_i}$ в срок $period_{\xi_i}$ для содействия или противодействия развитию текущей ситуации по сценарию ξ_i .

P_{ξ_i} – вероятность развития ситуации по сценарию ξ_i .

Генерация сценариев основана на принципе исторической аналогии: текущая ситуация s_c подвергается сравнению с эталонными ситуациями $s_e \in S_e$ из подготовленной экспертами базы эталонов S_e . На основе сопоставления событий, составляющих цепочки, оценивается вероятность того, что текущая ситуация аналогична эталонной. Если вероятность превышает пороговое значение, эталонная ситуация считается потенциальным сценарием дальнейшего развития текущей.

После построения сценариев выполняется определение их приоритетности путем сравнения сценариев на основе множества критериев. Наиболее приоритетный сценарий считается оптимистическим, наименее приоритетный – пессимистическим.

Предложения rec_{ξ_i} определяются эталонной ситуацией s_e , на основе которой был сформирован сценарий ξ_i , и зависят от того, какому событию этой эталонной ситуации соответствует s_c . В связи с этим каждое событие $\varepsilon_{s_e}^k$ эталонной ситуации s_e должно содержать собственную рекомендацию $rec_{\varepsilon_{s_e}^k} = \langle action_{\varepsilon_{s_e}^k}, actor_{\varepsilon_{s_e}^k}, period_{\varepsilon_{s_e}^k} \rangle$.

4. Модель события и документа

Для обеспечения возможности многокритериального сопоставления событий и сообщений, модель события должна содержать компоненты, отражающие его различные аспекты:

$$\varepsilon_i = \langle \varepsilon_i^d, \varepsilon_i^w, \varepsilon_i^{tw}, \varepsilon_i^c, \varepsilon_i^p, \varepsilon_i^n, \varepsilon_i^{dt}, \varepsilon_i^e, \varepsilon_i^g, \varepsilon_i^f \rangle. \quad (1)$$

Модель включает следующие компоненты:

- множество сообщений, описывающих событие: $\varepsilon_i^d = \{d_i^1, d_i^2, \dots, d_i^{N_i^d}\}$. На основе этих документов формируются остальные компоненты модели события;
- вектор слов документов, описывающих событие: $\varepsilon_i^w = (w_i^1, w_i^2, \dots, w_i^{N_i^w})$;
- вектор слов заголовков документов, описывающих событие: $\varepsilon_i^{tw} = (tw_i^1, tw_i^2, \dots, tw_i^{N_i^{tw}})$;
- множество предложений документов, описывающих событие: $\varepsilon_i^c = \{c_i^1, c_i^2, \dots, c_i^{N_i^c}\}$.

Каждое предложение описывается вектором слов $c_i^k = (w_{c_i^k}^1, w_{c_i^k}^2, \dots, w_{c_i^k}^{N_i^w})$;

- множество абзацев документов, описывающих событие: $\varepsilon_i^p = \{p_i^1, p_i^2, \dots, p_i^{N_i^p}\}$. Каждый абзац описывается вектором слов $p_i^k = (w_{p_i^k}^1, w_{p_i^k}^2, \dots, w_{p_i^k}^{N_i^w})$;

- множество числовых значений, извлеченных из текстов документов, описывающих событие: $\varepsilon_i^n = \{n_i^1, n_i^2, \dots, n_i^{N_i^n}\}$;

- множество имён персон и организаций, связанных с событием: $\varepsilon_i^e = \{e_i^1, e_i^2, \dots, e_i^{N_i^e}\}$;

- множество географических объектов, связанных с событием: $\varepsilon_i^g = \{g_i^1, g_i^2, \dots, g_i^{N_i^g}\}$;

- временной интервал, в течение которого происходило событие: $\varepsilon_i^{dt} = (\tau_i^{begin}, \tau_i^{end})$.

Также для более качественного выявления событий, относящихся к темам, подвергающимся мониторингу, модель содержит информацию о темах, которым релевантно событие: $\varepsilon_i^t = (t_i^1, t_i^2, \dots, t_i^{N_i^t})$. Тема задается экспертом в виде формализованного запроса, и значение каждого элемента вектора ε_i^t отражает релевантность события соответствующему запросу.

Для определения близости между событиями и сообщениями каждое сообщение также должно быть представлено аналогичной многокритериальной моделью:

$$d_j = \langle d_j^w, d_j^{tw}, d_j^c, d_j^p, d_j^n, d_j^{dt}, d_j^e, d_j^g, d_j^t \rangle. \quad (2)$$

5. Метод обнаружения событий

5.1. Определение расстояния между событием и документом

При сравнении сообщения d_j и события ε_i выполняется определение расстояния между соответствующими компонентами их моделей:

- для сравнения компонентов, представленных векторами (слова документов, слова заголовков, темы) используется косинусная мера:

$$\rho_{i,j}^w = 1 - \sum_{k=1}^{N^w} [w_i^k w_j^k] / \sqrt{\sum_{k=1}^{N^w} (w_i^k)^2} \sqrt{\sum_{k=1}^{N^w} (w_j^k)^2}. \quad (3)$$

- для сравнения компонентов, представленных множествами, используется мера включения:

$$\rho_{i,j}^n = 1 - |\varepsilon_i^n \cap d_j^n| / |d_j^n|. \quad (4)$$

Если элементы множеств являются взвешенными (именованные сущности, географические наименования, абзацы и предложения), вместо количества элементов учитывается их суммарный вес:

$$\rho_{i,j}^e = 1 - \sum_{e \in \varepsilon_i^e \cap d_j^e} e / \sum_{e \in d_j^e} e. \quad (5)$$

При этом вес предложений и абзацев рассчитывается как суммарный вес их слов.

- при определении расстояния между документом и событием с точки зрения времени берется средний момент между началом и концом события:

$$\rho_{i,j}^{dt} = |d_j^{dt} - (\tau_i^{begin} + \tau_i^{end}) / 2|. \quad (6)$$

Таким образом, результатом сравнения документа d_j и события ε_i является вектор, каждый элемент которого отражает расстояние между d_j и ε_i по некоторому признаку:

$$\rho_{i,j} = \langle \rho_{i,j}^w, \rho_{i,j}^{tw}, \rho_{i,j}^c, \rho_{i,j}^p, \rho_{i,j}^n, \rho_{i,j}^{dt}, \rho_{i,j}^e, \rho_{i,j}^g, \rho_{i,j}^t \rangle. \quad (7)$$

На основании $\rho_{i,j}$ необходимо определить значение расстояния между d_j и ε_i . Для этого применяется машина опорных векторов (SVM), в качестве примеров для обучения которой используются:

- пары вида “событие – документ, относящийся к этому событию” (позитивные примеры);
- пары вида “событие – документ, не относящийся к этому событию” (негативные примеры).

При обучении машины выполняется построение гиперплоскости, разделяющей эти примеры. Расстояние $Dist(\varepsilon_i, d_j)$ между документом d_j и событием ε_i определяется как нормализованное расстояние между вектором $\rho_{i,j}$ и гиперплоскостью.

5.2. Кластеризация документов

Для распределения поступающих документов по кластерам, соответствующим событиям, используется метод инкрементальной кластеризации. Каждое новое сообщение либо относится к одному из ранее сформированных событий, либо используется для формирования нового события. Алгоритм инкрементальной кластеризации состоит из следующих шагов:

1. Новый документ d_j сравнивается с каждым из ранее обнаруженных событий ε_i с определением значения расстояния $Dist(\varepsilon_i, d_j)$.
2. Выбирается наиболее близкое к документу событие: $\varepsilon_o = \arg \min_{\varepsilon_i} [Dist(\varepsilon_i, d_j)]$.
3. Если расстояние между сообщением d_j и событием ε_o не превосходит пороговое значение – сообщение относится к этому событию.
4. Если расстояние больше порогового значения – создается новое событие на основе этого документа.

6. Метод формирования сценариев

При построении ситуационного графа в нём находятся события, аналогичные событиям, составляющим эталонные ситуации. Далее в графе выделяются цепочки (ситуации), содержащие такие события-аналоги. Каждая из выделенных цепочек s_c сравнивается с эталонными ситуациями $s_e \in S_e$. Сравнение ситуаций рассматривается как задача логистической регрессии. Для этого вводится переменная y , такая что

$$y = \begin{cases} 1, & \text{если цепочки не являются аналогами} \\ 0, & \text{если цепочки являются аналогами} \end{cases} \quad (8)$$

Делается предположение, что вероятность наступления события $y = 1$ задана логистической функцией:

$$P(y = 1 | s_e, s_c) = \frac{1}{1 + e^{-\theta W^T}}. \quad (9)$$

$W = (W_{del}, W_{add}, W_{rep}, W_{trep})$ - вектор, отражающий различие между цепочками. Различие определяется весом операций, необходимых для превращения одной цепочки в другую:

- W_{del} - суммарный вес операций удаления события из эталонной ситуации;
- W_{add} - суммарный вес операций добавления события в текущую ситуацию;
- W_{rep} - суммарный вес операций замены события на его аналог;
- W_{trep} - суммарный вес операций изменения временного интервала между событиями.

$\theta = (\theta_{del}, \theta_{add}, \theta_{rep}, \theta_{trep})$ - вектор параметров. Параметры подбираются методом максимального правдоподобия.

Вероятность того, что текущая ситуация является аналогом эталонной рассчитывается как $P_{an}(s_c, s_e) = P(y = 0 | s_c, s_e) = 1 - P(y = 1 | s_c, s_e)$. При $P_{an}(s_c, s_e) \geq 0.5$ делается заключение о том, что ситуации аналогичны. Также при сравнении ситуаций определяется начальная часть $st(s_e, s_c)$ ситуации s_e , которая соответствует текущей ситуации s_c . Заключительная часть $fin(s_e, s_c)$ эталонной ситуации является возможным сценарием дальнейшего развития текущей ситуации. При этом значение $P_{an}(s_e, s_c)$ можно рассматривать как вероятность того, что текущая ситуация будет далее развиваться по соответствующему сценарию.

Заключительные части эталонных ситуаций, признанных аналогичными текущей, являются возможными сценариями её дальнейшего развития. Среди всех сценариев особый интерес представляют три: оптимистический, пессимистический и наиболее вероятный. Наиболее вероятным сценарием является заключительная часть эталонной цепочки s_e^{pr} , для которой вероятность аналогичности текущей ситуации максимальна: $s_e^{pr} = \arg \max_{s_e} P_{an}(s_e, s_c)$.

Для выделения оптимистического и пессимистического сценариев необходимо выполнить ранжирование сценариев по оптимальности. Для этого используется метод анализа иерархий, позволяющий вычислить приоритет каждого из сценариев на основе набора критериев, в качестве которых могут использоваться уровень знаний обучающихся, длительность сценария, педагогическая эффективность, экономическая эффективность (затраты на реализацию сценария в вузе) и т.д. Значения критериев для каждой эталонной ситуации определяются экспертами при формировании базы эталонов.

Приоритетность критериев относительно цели (определения наиболее оптимального сценария) вычисляется на основе попарных сравнений, выполняемых экспертами на этапе обучения. Приоритетность сценариев относительно каждого критерия вычисляется автоматически на основе характеристик сценариев. Таким образом, при анализе сценариев,

сформированных для текущей ситуации, их приоритетность относительно цели вычисляется автоматически. Сценарий, обладающий наибольшим приоритетом, признается оптимистическим, сценарий с наименьшим приоритетом - пессимистическим.

В качестве предложения по содействию оптимальному развитию текущей ситуации используется рекомендация, соответствующая последнему событию цепочки $st(s_e, s_c)$.

7. Система автоматизированного мониторинга и прогнозирования развития образовательных технологий

На основе предложенного подхода разработана система автоматизированного мониторинга и прогнозирования развития ситуаций. Обучение подсистемы обнаружения событий производится экспертом на основе подготовленных эталонных событий, связанных с наукой и образованием. Обученная подсистема выполняет автоматический анализ потока текстовых документов и предоставляет пользователю списки обнаруженных событий и ситуаций. Пользователь имеет возможность выделить заинтересовавшую его ситуацию и добавить её в базу эталонов для обучения подсистемы формирования сценариев. На рис. 1 представлен пример ситуации, состоящей из четырёх событий, выделенной системой.

Ситуация: Конференция «Современные информационные технологии в образовании»

Круглый стол на тему «Мобильное образование: за и против»			
Начало события: 20.06.2016 13:07:22 Конец события: 24.06.2016 11:15:00			
Имя документа	Дата публикации	Время публикации	Источник
Круглый стол на тему «Мобильное образование: за и против»	20.06.2016	13:07:22	ForSMI
Следить урок: живым и интересным помогают мобильные устройства	21.06.2016	21:48:00	Газета "Вечерняя Москва"
Плюсы и минусы мобильного образования обсудили в Москве	24.06.2016	11:15:00	Учительская Газета

Утверждены эксперты конференции «Современные информационные технологии в образовании»			
Начало события: 22.06.2016 17:14:00 Конец события: 22.06.2016 17:14:00			
Имя документа	Дата публикации	Время публикации	Источник
Утверждены эксперты конференции «Современные информационные технологии в образовании»	22.06.2016	17:14:00	Газета "Новые Окрута"

На международной конференции в Троицке обсудили технологии в образовании			
Начало события: 28.06.2016 16:03:00 Конец события: 30.06.2016 13:53:00			
Имя документа	Дата публикации	Время публикации	Источник
На международной конференции в Троицке обсудили технологии в образовании	28.06.2016	16:03:00	Газета "Новые Окрута"
Ищем новые формы обучения	29.06.2016	22:28:00	Газета "Вечерняя Москва"
Тенденции применения IT в образовании обсудили на конференции в Троицке	30.06.2016	13:53:00	Информационный Центр Правительства Москвы

На селекторном совещании в Москве подвели итоги конференции «Современные информационные технологии в образовании»			
Начало события: 06.07.2016 16:13:00 Конец события: 06.07.2016 16:13:00			
Имя документа	Дата публикации	Время публикации	Источник
На селекторном совещании в Москве подвели итоги конференции «Современные информационные технологии в образовании»	06.07.2016	16:13:00	Mangazia

Рис. 1. Пример выявления событий и ситуаций

Обучение подсистемы формирования сценариев выполняется на основе эталонных ситуаций, отражающих историю развития инновационных технологий и их внедрения в учебный процесс в прошлом (рис. 2). Эксперты снабжают каждое событие эталонной ситуации предложениями по действиям, которые необходимо предпринять при наступлении такого события для оптимального развития ситуации в дальнейшем.

Путем сопоставления текущих ситуаций с эталонными подсистема формирует вероятные сценарии их дальнейшего развития и вырабатывает рекомендации.

Название события	Временной промежуток	Актеры и их действия	Рекомендации
Появление первых публикаций о новой технологии	2000-2002	Исследователи: публикация статей о новой технологии	<i>actor</i> Преподаватели ВУЗа
			<i>action</i> Ознакомление с новой технологией, отслеживание новых публикаций
			<i>period</i> 1 год
Появление международных конференций, посвященных технологии	2005-2006	Исследователи и разработчики: проведение конференций	<i>actor</i> Преподаватели ВУЗа
			<i>action</i> Изучение материалов конференций, анализ целесообразности внедрения технологии в учебный процесс
			<i>period</i> 1 год
Появление первых коммерческих разработок	2008	Разработчики: коммерческая реализация инновационной технологии	<i>actor</i> Руководство ВУЗа
			<i>action</i> Подготовка к внедрению технологии в учебный процесс
			<i>period</i> 1 год
Появление спроса компаний на специалистов, владеющих технологиями	2011	Руководство компаний: поиск специалистов, владеющих технологиями	<i>actor</i> Руководство ВУЗа
			<i>action</i> Внедрение технологии в учебную программу
			<i>period</i> 1 год
Появление в учебных программах ведущих ВУЗов курсов, посвященных технологии	2014-2015	Руководство ВУЗов: внедрение новых курсов в учебную программу	<i>actor</i> Руководство ВУЗа
			<i>action</i> Внедрение технологии в учебную программу
			<i>period</i> 6 месяцев

Рис. 2. Пример эталонной ситуации и предложений

8. Экспериментальная проверка методов

Для анализа качества предложенных моделей и методов был проведен эксперимент, целью которого являлось определение зависимости показателей качества (точности, полноты и F-меры) обнаружения событий от мощности обучающей выборки. Полученные зависимости приведены на рис. 3.

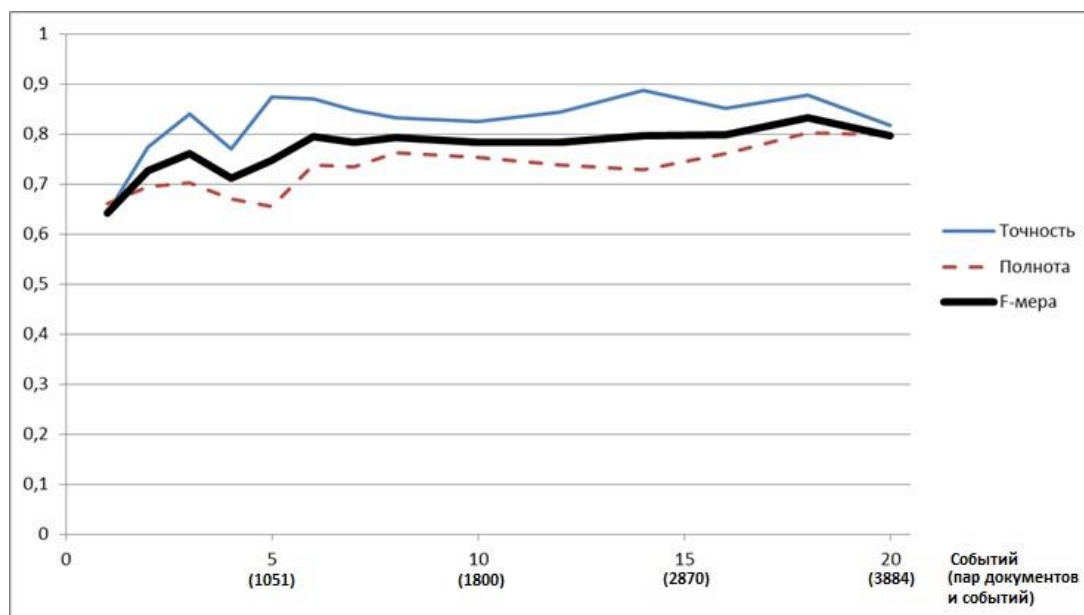


Рис. 3. Зависимость точности (тонкая сплошная линия), полноты (тонкая пунктирная линия) и F-меры (жирная линия) от мощности обучающей выборки

Также анализировалась зависимость вышеуказанных показателей от используемых критериев сравнения документов и событий. В частности, анализировалось качество при анализе близости составов слов документов, при анализе близости составов слов документов и их заголовков и при использовании всех критериев. Результаты эксперимента приведены в табл. 1.

Таблица 1. Зависимость показателей качества обнаружения событий от используемых критериев

Используемые критерии	Точность	Полнота	F-мера
Слова документов	65,8%	65%	65,2%
Слова документов и заголовков	72,2%	63%	67,2%
9 критериев	85,2%	76%	79,8%

В результате проведения эксперимента оказалось, что для обучения системы достаточно 1300 пар документов и событий, генерируемых на основе 6 обучающих событий, с дальнейшим же увеличением обучающей выборки качество работы метода не улучшается. Проведенный эксперимент также доказывает целесообразность многокритериального сравнения документов и событий: при использовании всех критериев достигаются наиболее высокие показатели качества.

Заключение

В работе предложен подход к автоматизированному мониторингу и прогнозированию развития ситуаций, связанных с развитием тех или иных образовательных технологий. В его основе лежит последовательное выполнение обнаружения событий в текстовом потоке, формирования ситуаций и построения сценариев их дальнейшего развития.

Представленная модель события включает компоненты, отражающие его основные аспекты. Предложенный метод обнаружения событий имеет возможность гибкой настройки в соответствии с особенностями образовательной деятельности и потребностями пользователей благодаря использованию методов машинного обучения, в частности, метода опорных векторов.

Представленный метод формирования сценариев дальнейшего развития ситуаций позволяет оценить вероятность полученных сценариев благодаря использованию метода логистической регрессии. При помощи метода анализа иерархий обеспечивается выделение наиболее оптимистического и пессимистического сценариев. На основе сгенерированных сценариев формируются предложения по действиям, необходимым для обеспечения развития ситуации по наиболее благоприятному сценарию.

Предлагаемый подход позволяет отследить появление инновационных технологий в образовании, дать прогноз их вероятного развития и выработать рекомендации, позволяющие обеспечить их внедрение в учебный процесс вуза.

Список литературы

- [1]. Raymond Y., Abdallah S. The Event Ontology. 2007. Режим доступа: <http://motools.sourceforge.net/event/event.html> (дата обращения 29.07.2016).
- [2]. Добров Б.В., Павлов А.М. Исследование качества базовых методов кластеризации новостного потока в суточном временном окне // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XII-ой Всероссийской научной конференции RCDL'2010. (Казань: Казанский университет, 13-17 октября 2010 г.). 2010. С. 287-295.

- [3]. Kumaran G., Allan J. Using names and topics for new event detection // Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver: Association for Computational Linguistics. 2005. P. 121-128.
- [4]. Radinsky K., Horvitz E. Mining the web to predict future events // Proceedings of the sixth ACM international conference on Web search and data mining. New York, NY, USA: ACM. 2013. С. 255-264. DOI:[10.1145/2433396.2433431](https://doi.org/10.1145/2433396.2433431)
- [5]. Ландэ Д.В., Брайчевский С.М., Григорьев А.Н., Дармохвал А.Т., Радецкий А.Б. Выявление новых событий из потока новостей // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». (Бекасово, 30 мая - 3 июня 2007 г.) / Под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. М.: Изд-во РГГУ. 2007. 658 с. С. 349–352.
- [6]. Aggarwal C.C., Subbian K. Event Detection in Social Streams // Proceedings of the 2012 SIAM International Conference on Data Mining. (Disney's Paradise Pier Hotel, Anaheim, California, USA, April 26-28, 2012). SDM. SIAM / Omnipress. 2012. Vol. 12. P. 624-635. DOI: <http://dx.doi.org/10.1137/1.9781611972825.54>
- [7]. Кондратьев М.Е. Анализ методов кластеризации новостного потока // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Восьмой Всероссийской научной конференции (RCDL'2006). (Суздаль, 17-19 октября 2006 г.). Ярославль: Ярославский государственный университет им. П.Г. Демидова. 2006. С. 108-114.
- [8]. Brants T., Chen F., Farahat A. A system for new event detection // Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM. 2003. С. 330-337. DOI:[10.1145/860435.860495](https://doi.org/10.1145/860435.860495)
- [9]. Yang Y., Zhang J., Carbonell J.G., Jin C. Topic-conditioned novelty detection // Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (July 23-26, 2002, Edmonton, Alberta, Canada). ACM. 2002. P. 688-693. DOI: [10.1145/775047.775150](https://doi.org/10.1145/775047.775150)
- [10]. Zhao Q., Mitra P., Chen B. Temporal and information flow based event detection from social text streams // AAAI'07 Proceedings of the 22nd national conference on Artificial intelligence. Vancouver, British Columbia. AAAI-07, AAAI Press. 2007. Vol. 2. P. 1501-1506.
- [11]. Yang Y., Pierce T., Carbonell J. A study of retrospective and on-line event detection // SIGIR '98 Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM. 1998. P. 28-36. DOI: [10.1145/290941.290953](https://doi.org/10.1145/290941.290953)
- [12]. Li Z., Wang B., Li M., Ma W-Y. A probabilistic model for retrospective news event detection // SIGIR '05 Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM. 2005. P. 106-113. DOI: [10.1145/1076034.1076055](https://doi.org/10.1145/1076034.1076055)

- [13]. Ahmed A., Ho Q., Eisenstein J., Xing E., Smola A.J., Teo C.H. Unified analysis of streaming news // WWW'11. Proceedings of the 20th international conference on World wide web. (March 28–April 1, 2011, Hyderabad, India). New York, NY, USA: ACM. 2011. P. 267-276. DOI: [10.1145/1963405.1963445](https://doi.org/10.1145/1963405.1963445)
- [14]. Aggarwal C.C., Philip S.Y. On clustering massive text and categorical data streams // Knowledge and information systems. 2010. Vol. 24. № 2. P. 171-196. DOI: [10.1007/s10115-009-0241-z](https://doi.org/10.1007/s10115-009-0241-z)