Применение временных рядов в задаче фоновой идентификации пользователей на основе анализа их работы с текстовыми данными[1]

В.Ю. Королёв <bruce27@yandex.ru>
А.Ю. Корчагин <proton.ru@gmail.com>
И.В. Машечкин <mash@cs.msu.su>
М.И. Петровский <michael@cs.msu.su>
Д.В. Царёв <tsarev@cs.msu.su>
Факультет вычислительной математики и кибернетики,
Московский государственный университет имени М.В. Ломоносова,
119991, Россия, Москва, Ленинские горы, д. 1с52.

Аннотация. В статье представлен новый подход идентификации пользователя на основе анализа его поведения при работе с текстовой информацией. Для описания поведения пользователя предлагается использовать содержимое текстовых документов, к которым он обращался. Структурированное представление рассматриваемой поведенческой информации осуществляется на основе отображения содержимого электронных документов в тематическое пространство пользователя, формируемое с использованием неотрицательной матричной факторизации. Веса выделенных тематик в документе характеризуют тематическую направленность пользователя во время работы с данным документом. Изменение значений весов тематик во времени формирует многомерный временной ряд, описывающий историю поведения пользователя при работе с текстовыми данными. Построение прогноза такого временного ряда позволит осуществлять идентификацию данного пользователя на основе оценки отклонений наблюдаемой тематической направленности пользователя от спрогнозированных значений. В рамках предложенного подхода был разработан собственный оригинальный метод прогнозирования временных рядов, основанный на ортонормированной неотрицательной матричной факторизации (ОНМФ). Важно отметить, что ранее методы неотрицательной матричной факторизации не использовался для решения задачи прогнозирования временных рядов. Проведённое экспериментальное исследование на примере реальной корпоративной переписки пользователей, сформированной из набора данных Enron, показало применимость предложенного подхода идентификации пользователя. Кроме того, эксперименты с применением других популярных на сегодняшний день методами прогнозирования показали превосходство разработанного метода на основе ОНМФ по качеству классификации тематических характеристик пользователя. Также в работе исследовались два различных подхода оценки отклонений: абсолютная оценка и оценка p-значения. Эксперименты показали, что оба рассмотренные подхода расчёта оценки отклонения временной точки от прогноза применимы в предложенном подходе идентификации пользователя.

Ключевые слова: компьютерная безопасность; идентификация пользователя; тематическое моделирование; ортонормированная неотрицательная матричная факторизация; прогнозирование временных рядов

DOI: 10.15514/ISPRAS-2015-27(1)-8

Для цитирования: Королёв В.Ю., Корчагин А.Ю., Машечкин И.В., Петровский М.И., Царёв Д.В. Применение временных рядов в задаче фоновой идентификации пользователей на основе анализа их работы с текстовыми данными. Труды ИСП РАН, том 27, вып. 1, 2015 г., стр. 151-172. DOI: 10.15514/ISPRAS-2015-27(1)-8.

[1] Работы выполнены при финансовой поддержке Минобрнауки России (Соглашение № 14.604.21.0056 о предоставлении субсидии, Уникальный идентификатор прикладных научных исследований RFMEFI60414X0056).

1. Введение

Актуальность проблемы защиты компьютерной информации в настоящее время не вызывает сомнений. Использование стандартных средств защиты информации, основанных на разграничении прав доступа, контроле целостности, аутентификации пользователей с использованием паролей, ключей или цифровых подписей, а также применение систем контроля работы пользователей, основанных на предопределенных регламентах, политиках, правилах и использовании сигнатурных методов обнаружения вторжений, не дают надежной защиты.

По мнению ведущих специалистов по компьютерной безопасности, перспективным инструментом в настоящей предметной области являются подходы на основе анализа поведенческой биометрии пользователей с использованием статистических методов и методов машинного обучения [1]. В общем случае биометрия включает совокупность процедур и методов распознавания людей по одной или нескольким физиологическим или поведенческим чертам. Поведенческие признаки относятся к индивидуальным особенностям поведения человека, сформированным в результате его уникального опыта, навыков и знаний, которые не являются секретной информацией, но их невозможно абсолютно точно передать или скопировать, а также достаточно тяжело подделать. Например, графическая подпись, стилистические и морфологические особенности устной и письменной речи, динамика работы с устройствами ввода-вывода и т.д. К поведенческим признакам также относятся особенности потребляемой и создаваемой пользователем текстовой информации (документы, электронные сообщения, почта).

Настоящая статья посвящена одной из важнейших задач защиты компьютерной информации — задаче идентификации пользователей, т.е. постоянной (или периодической) оценки достоверности того, что пользователь, работающий с защищаемой компьютерной системой, является действительно тем, от имени кого он авторизовался. В отличие от задачи аутентификации при идентификации не подразумевается явных процедур проверки, требующих интерактивных действий от пользователя.

В статье рассматриваются методы машинного обучения и математической статистики для построения и применения поведенческих моделей с целью решения задач постоянной фоновой идентификации пользователей на основе его работы с текстовыми данными. Идея предлагаемого подхода состоит в тематическом анализе сложившихся в прошлом тенденций работы (поведения) пользователя с текстовым контентом различных (в том числе конфиденциальных) категорий и прогнозировании его дальнейшего поведения. Тематический анализ работы пользователя предполагает определение основных тематик его текстового контента и расчёт соответствующих им весов в заданные интервалы времени. На основе отклонений поведения в работе пользователя с контентом от прогноза можно выявить интервалы времени, когда:

- велась работа с документами несвойственных категорий;
- работа с документами той или иной категории отличается от обычной (исторической).

Настоящая статья имеет следующую структуру. В разделе 2 приведено описание процедуры выделения тематических характеристик из текстовых данных пользователя и формирования тематических временных рядов. В Разделе 3 рассматриваются популярные подходы прогнозирования временных рядов и представлен собственный оригинальный метод прогнозирования на основе ортонормированной неотрицательной матричной факторизации. Раздел 4 посвящен экспериментальному исследованию предложенного подхода идентификации пользователя на примере реальной корпоративной переписки, сформированной из набора электронных писем Enron. Также в данном разделе исследуются два различных подхода оценки отклонений: абсолютная оценка по всем тематикам и оценка p-значения критерия согласия Хи-квадрат. В Разделе 5 делаются основные выводы и приводится заключение.

2. Выделение тематических характеристик из текстовых данных пользователя

В основе предлагаемого подхода идентификации пользователей лежит тематическое моделирование текстовых данных, с которыми работал

пользователь за заданное модельное время. Модельное время разбивается на последовательно измеренные через некоторые (зачастую равные) промежутки времени интервалы, например, в качестве промежутка времени (шага) может быть выбран час, день или время, за которое происходит заданное число событий. (рис. 1) [2]. С помощью тематического моделирования выделяются основные тематики текстового контента пользователя и соответствующие им веса в каждом временном интервале модельного времени.

Рис. 1. Формирование временных рядов тематической направленности пользователя.

Веса тематик во временном интервале характеризуют тематическую направленность пользователя, на их основе формируются временные ряды изменения его тематической направленности для каждой из тематик. Далее по сформированным временным рядам строятся прогнозы (рис. 1). На основе значений отклонений тематической направленности от спрогнозированных данных определяются временные интервалы с несвойственной активностью пользователя с контентом.

Исходя из предыдущих работ авторов [3-7] в качестве методов тематического моделирования были выбраны методы, основанные на неотрицательной матричной факторизации. Методы неотрицательной матричной факторизации работают с векторным представлением текста типа «мешок слов» (англ. «bag-of-words») [8]. В нашем случае в качестве текстов выступают текстовые данные каждого временного интервала. Далее под термином временной интервал в зависимости от контекста будет пониматься либо совокупность текстовых данных анализируемого пользователя за рассматриваемый временной интервал, либо непосредственно интервал времени.

Формально опишем модель построения тематических временных рядов для n временных интервалов модельного времени. Каждый временной интервал j ($1 \le j \le n$) отображается в числовой вектор $A_j=[a_{1,j}, a_{2,j}, \ldots, a_{m,j}]^T$ фиксированной размерности m, где m — число признаков текстовых данный за модельное

время, а *i*-ая компонента вектора определяет вес *i*-го признака в *j*-ом временном интервале.

В качестве признаков в модели «мешка слов» используются лексемы, входящие в текст, а размерность признакового пространства равна размерности словаря лексем. Под лексемами в общем случае понимаются все различные слова текста. Однако обычно применяются некоторые меры по предварительной обработке текста с целью получения более «информативного» признакового пространства [8]: удаление стоп-слов, приведение слов к нормализованной форме (стемминг) и т.д. Цель предварительной обработки текста — оставить только те признаки, которые наиболее информативны, т.е. наиболее сильно характеризуют текст. К тому же уменьшение анализируемых признаков приводит к уменьшению использования вычислительных ресурсов. В интеллектуальном анализе текстовых данных для обозначения признака текста принято использовать термин «терм».

Вес *i*-го терма в векторном представлении *j*-го временного интервала определяется как $a_{i,j}=L_{i,j}G_i$. $L_{i,j}$ — локальный вес терма *i* во временном интервале *j*, G_i — глобальный вес терма *i* во всех временных интервалах. Т.к. для вычисления отклонений от спрогнозированных значений будут использоваться новые временные интервалы, не вошедшие в модельное время, то заранее определить использование того или иного терма в будущих временных интервалах невозможно, поэтому использование глобального веса исключается. В ходе экспериментов, проводимых в Разделе 4, наилучшие результаты были получены при использовании логарифмического веса в качестве локального: $L_{i,j} = 1 + \log(t_{i,j})$, где $t_{i,j}$ — число появлений терма *i* во временном интервале *j* [7, 8].

Таким образом, текстовый контент пользователя за модельное время представляется в виде числовой матрицы, строки которой соответствуют термам, а столбцы текстам каждого временного интервала. Объединение термов в тематики и представление временных интервалов в пространстве тематик осуществляется путём применения к данной матрице неотрицательной матричной факторизации.

Матрица модельных временных интервалов $A \in \Re^{m \times n}$, где *m* — число различных термов, *n* — число временных интервалов. Элементы матрицы *A* принимают неотрицательные значения, т.к. являются весами соответствующих термов во временных интервалах. Тогда цель неотрицательной матричной факторизации состоит в нахождении матриц $W_k \in \Re^{m \times k}$ и $H_k \in \Re^{k \times n}$ с неотрицательными элементами, которые минимизируют целевую функцию [9]:

$$f(W_k, H_k) = \frac{1}{2}\left\|A - W_k H_k\right\|_F^2, k << min(m,n).$$

Матрица $W_k=[w_{ij}]$ задает отображение пространства *k* тематик в пространство *m* термов, матрица $H_k=[h_{ij}]$ соответствует представлению временных интервалов в пространстве тематик, т.е. элемент h_{ij} соответствует представлению *i*-ой тематики в *j*-ом временном интервале. В связи с тем, что элементы матрицы H_k неотрицательны, то их можно рассматривать как вклад(вес) тематики во временной интервал. Чем больше значение элемента h_{ij} по сравнению с другими элементами *j*-го временного интервала, тем более характерна *i*-ая тематика для текста данного временного интервала. На этом свойстве основаны алгоритмы кластеризации, использующие неотрицательную матричную факторизацию [10, 11]. Аналогично и для матрицы W_k, чем больше значение элемента w_{ij} по сравнению с другими элементами *j*-го столбца (*j*-ой тематике), тем более характерен *i*-ый терм для данной тематики [12].

Исходя из описанных свойств неотрицательной матричной факторизации, временной ряд изменения каждой из выделенных *k* тематик формируется из значений элементов соответствующей строки матрицы H_k (рис. 2).

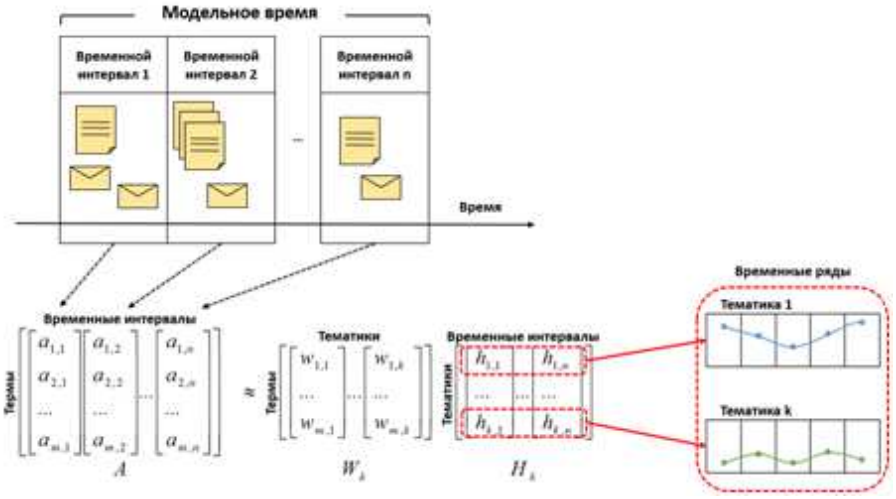

Рис. 2. Построение временных рядов на основе неотрицательной матричной факторизации.

Существуют различные методы реализации неотрицательной матричной факторизации [9, 11, 12, 13]. Однако для реализации предлагаемого подхода, необходимо иметь возможность отображать новые, не вошедшие в модельное время, временные интервалы в уже построенное пространство тематик. Для реализации данного функционала необходимо наложить дополнительное условие ортонормированности матрицы W_k: $W_k^T \cdot W_k = I$. Тогда для отображения матрицы временных интервалов времени прогноза A_{new} в

пространство тематик модельного времени достаточно A_{new} слева умножить на $W_k^T : W_k^T \cdot A_{new} = H_{k\_new}$.

Авторы статьи исследовали применение множества популярных алгоритмов, реализующих ортонормированную неотрицательную матричную факторизацию [9, 11, 13], для использования в предлагаемом подходе. На основе экспериментальных исследований, оставшихся за рамками данной статьи, был выбран алгоритм минимизации целевой функции $f(W_k, H_k) = \frac{1}{2}\|A - W_k H_k\|_F^2 + \frac{\alpha}{2}\|W_k^T W_k - I\|_F^2$, описанный в [9] и позволяющий задавать баланс между точностью приближения исходной матрицы и ортонормированностью получаемых тематик с помощью параметра α.

1. Элементы матриц $W^1 \in \Re_+^{m \times k}$ и $H^1 \in \Re_+^{k \times n}$ инициализируются случайными неотрицательными числами;
2. В цикле p раз выполняются итерационные формулы для вычисления матриц W и H:

   a. $H_{b,j}^{p+1} = H_{b,j}^p \frac{((W^p)^T A)_{b,j}}{((W^p)^T W^p H^p)_{b,j}}, \forall b, j : 1 \le b \le k, 1 \le j \le n$,

   b. $W_{i,a}^{p+1} = W_{i,a}^p \frac{(A(H^{p+1})^T + \alpha W^p)_{i,a}}{(W^p H^{p+1}(H^{p+1})^T + \alpha W^p (W^p)^T W^p)_{i,a}}, \forall i, a : 1 \le i \le m, 1 \le a \le k$.

Для построения прогнозов k тематических временных рядов, заданных матрицей H_k, в предлагаемом подходе идентификации пользователя применялись методы, описанные в Разделе 3. После построения прогнозов тематических рядов вычисляются отклонения тематической направленности пользователя за время прогноза от спрогнозированных значений. Вычисленные отклонения используются для идентификации временных интервалов с несвойственной тематической направленностью пользователя (Рисунок 3).

Отдельно отметим, что матрица W_k, построенная по модельным временным интервалам пользователя, описывает основные тематики пользовательского контента и служит для отображения любых текстовых данных (не обязательно временных интервалов) в пространство тематик данного пользователя, т.е. матрица W_k «характеризует» пользователя с точки зрения его тематических предпочтений в контенте. Поэтому для обозначения матрицы W_k нами также будет использоваться термин тематический «портрет» пользователя. Получаем, что в общем случае временные интервалы одного пользователя можно проецировать в тематическое пространство другого пользователя, и на основе полученных представлений также анализировать временные ряды. Данная возможность позволяет определять временные интервалы, когда один пользователь активно интересовался материалами характерными для другого

пользователя. Также возможно сформировать подобную матрицу W_k вообще не на основе временных интервалов пользователей, а на основе заранее сформированного набора документов организации (тренировочный набор), чьи тематики представляют наибольший интерес для анализа работы пользователей, получив таким образом тематический «портрет» набора документов и уже на его основе строить временные ряды для пользователя.

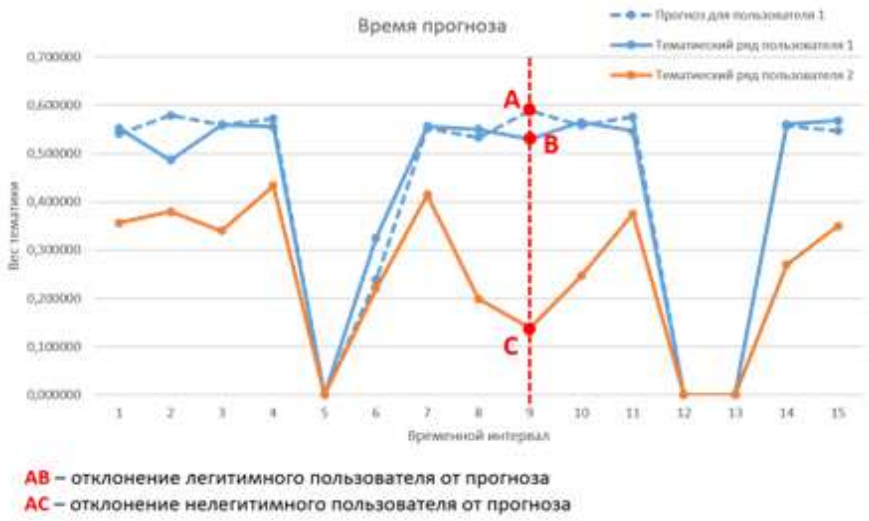

Рис. 3. Пример отклонений тематических временных рядов от прогноза.

3. Прогнозирование временных рядов

Прогнозирование временных рядов заключается в построении модели для предсказания будущих событий, основываясь на известных событиях прошлого [1]. Для построения прогноза временных рядов в статье использовались следующие модели: линейная модель авторегрессии, авторегрессионная модель дерева и предложенная оригинальная модель на основе ортонормированной неотрицательной матричной факторизации.

3.1 Линейная модель авторегрессии и авторегрессионная модель дерева

В линейной модели авторегрессии временных рядов значение временного ряда в данный момент линейно зависят от предыдущих значений этого же ряда. Формально авторегрессионную модель порядка p, которую обычно обозначают, как AR(p), определяют следующим образом:

1. $X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$, где φ_i — параметры авторегрессионной модели, c — константа (для простоты константу, как правило, опускают), ε_t — белый шум;
2. Форма записи с помощью оператора задержки L ($Lx_t = x_{t-1}$): $c + \varepsilon_t = \left(1 - \sum_{i=1}^{p} \varphi_i L_i\right) X_t$.

Авторегрессионная модель дерева (англ. AutoRegressive Tree Model, ART) — модель дерева принятия решений, в «листьях» которого располагаются авторегрессионные модели (AR). Для реализации данной модели в статье используется алгоритм ARTXP, разработанный Microsoft, который основан на их реализации алгоритма дерева принятия решений. Алгоритм ARTXP устанавливает соотношение между переменным количеством предыдущих элементов и каждым текущим элементом, для которого выполняется прогноз [14].

Отметим основные особенности алгоритма Microsoft ARTXP [15]:

- алгоритм ARTXP поддерживает учёт корреляций между несколькими анализируемыми рядами, т.е. поддерживает перекрестное прогнозирование;
- алгоритм ARTXP используется для прогнозирования ближайших значений временного ряда (порядка 5-7 временных шагов) и даёт существенно менее точный долгосрочный прогноз.

3.2 Модель прогнозирования на основе ортонормированной неотрицательной матричной факторизации

Анализируемый временной ряд можно представить в виде вектора, элементами которого являются рассчитанные значения в соответствующих временных точках, например, строки матрицы H_k (Раздел 2). Для того чтобы представить вектор временного ряда в матричной форме введём понятие порядка модели, т.е. количество подряд идущих значений временного ряда по которым определяются основные взаимосвязи, аналогично линейной модели авторегрессии. Тогда вектор временного ряда размерности n при заданном порядке модели p можно представить в виде матрицы размерности $p \times (n\text{-}p+1)$, чьи столбцы соответствуют всевозможным подпоследовательностям длины p подряд идущих временных точек анализируемого ряда (рис. 4).

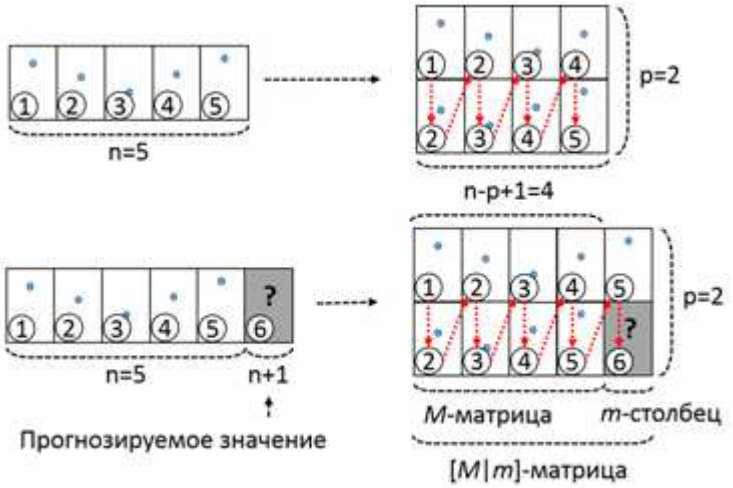

Рис. 4. Отображение вектора временного ряда в матрицу.

Задача прогнозирования следующего (n+1)-го значения (рис. 4) сводится к задаче аппроксимации пропущенных значений в матрице временного ряда (матрица $[M|m]$) на основе уже заполненных значений (матрица M) — т.н. задача подстановки пропущенных значений (англ. Missing Value Imputation) [16-18].

Наиболее распространенным методами матричного разложения к решению задачи подстановки пропущенных значений являются подходы на основе сингулярного разложения (Singular Value Decomposition, SVD) [16-18]. Однако, авторами была исследована возможность применения методов неотрицательной матричной факторизации для решения задачи подстановки пропущенных значений [18], на основе полученных результатов в рамках настоящей статьи будет описан алгоритм подстановки пропущенных значений с использованием ортонормированной неотрицательной матричной факторизации.

Рассмотрим ортонормированную неотрицательную матричную факторизацию матрицы M: $M \approx M_k = Wm_k \cdot Hm_k$, $Wm_k^T \cdot Wm_k = I$, где M_k является аппроксимацией исходной матрицы M, k — число «латентных» признаков, при этом $k << min(p, n\text{-}p+1)$ [6, 18, 19]. Матрица Wm_k задает отображение между пространством «латентных» признаков размерности k и пространством позиций элементов (от 1 до p) в сформированных подпоследовательностях длины p. Таким образом, выделенные «латентные» признаки содержат только наиболее значимую информацию о взаимосвязях между позициями элементов среди всех подпоследовательностей (см Раздел 2). Каждый столбец матрицы Hm_k описывает соответствующую подпоследовательность в виде вектор-столбца с весами соответствующих базисных «латентных» признаков.

Было предложено вычисление пропущенных значений в матрице $[M|m]$ на основе значений матрицы M следующим итерационным способом (по аналогии с решением в [16]):

- *Шаг 0*. Рассчитать ортонормированную неотрицательную матричную факторизацию матрицы M (с изначально заполненными элементами): $M \approx M_k = Wm_k \cdot Hm_k$, $Wm_k^T \cdot Wm_k = I$.
- *Шаг 1*. Инициализировать пропущенное значения в столбце m. Традиционно в литературе пропущенные значения инициализируются средним значением по столбцу или всему вектору временного ряда. Если известны данные о сезонности временного ряда, то при расчёте средних значений можно учитывать и шаг сезонности. На выходе получается полностью заполненный столбец m^i, где $i = 0$.
- *Шаг 2*. Рассчитать аппроксимацию столбца m с помощью полученной на *шаге 0* модели ортонормированную неотрицательную матричную факторизацию (ONMF-модель): $m_{approx} = (Wm_k \cdot Wm_k^T) \cdot m^i$. После чего сформировать m^{i+1} путем замены пропущенного значения в исходном столбце m соответствующим полученным значением в m_{approx}.
- *Шаг 3*. До тех пор пока значение $\left\|m^i - m^{i+1}\right\| / \left\|m^i\right\|$ меньше заданного порога (как правило, значение порога берут равным 10^{-6}), установить $i = i+1$ и перейти на шаг 2. На практике обычно достаточно 5-6 шагов для сходимости алгоритма.

Заметим, что в приведённом алгоритме вместо ортонормированной неотрицательной матричной факторизации можно использовать и традиционное сингулярное разложение [16].

При прогнозировании многомерных временных рядов можно их прогнозировать как по отдельности, так и с поддержкой перекрестного прогнозирования. Для этого совокупность из m временных рядов нужно объединить в одну матрицу (рис. 5). Тогда получаемые «латентные» признаки (матрица Wm_k) будут содержать наиболее значимую информацию о взаимосвязях между позициями элементов всех временных рядов среди всех подпоследовательностей, таким образом будет учитываться влияние всех анализируемых временных рядов друг на друга.

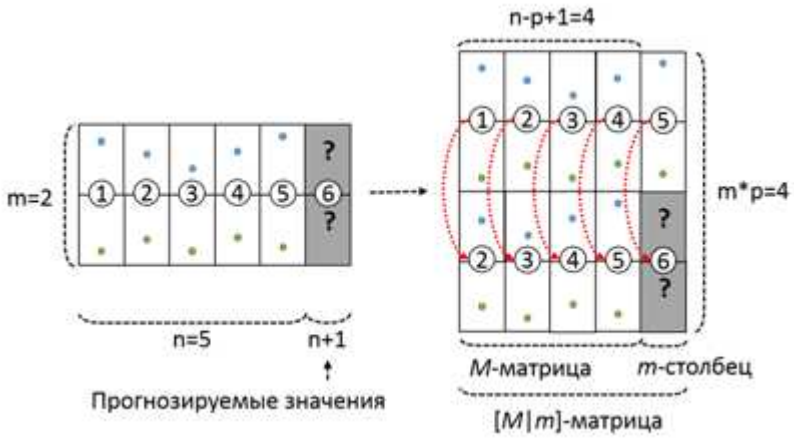

Рис. 5. Формирование матрицы прогнозирования для совокупности временных рядов.

4. Экспериментальные исследования

Первоочередной задачей при проведении экспериментальных исследований предлагаемого подхода анализа работы пользователя с текстовой информацией являлся выбор тестового набора данных. Основными критериями выбора тестового наборы были:

1. контент из корпоративной среды;
2. большое количество текстовых данных;
3. возможность разделения контента по пользователям и времени.

На основе сформулированных выше требований для экспериментального исследования предлагаемого подхода был выбран набор Enron [20]. Набор Enron содержит электронную почту 150 сотрудников (главным образом из высшего руководства) американской энергетической компании, обанкротившейся в конце 2001 года. Кроме того, данный набор широко распространён в работах, посвящённых тематическому анализу текстовых данных [12].

Для экспериментального исследования предлагаемого подхода из набора Enron были выбраны три сотрудника с наибольшим количеством писем: «dasovich-j», «kaminski-v» и «kean-s». В качестве ограничения по времени было выбрано первое полугодие 2001 года, т.к. в это время велась самая активная переписка данных пользователей. Выбранные шесть месяцев были разбиты на пересекающиеся интервалы по пять недель с шагом одна неделя, при этом первые четыре недели каждого интервала использовались в качестве модельного времени, а следующая неделя для прогноза. Таким образом всё рассматриваемое время было разбито на 21 анализируемый временной интервал. В табл. 1 приведена статистика распределения общих писем между

выбранными пользователями. Из представленной в табл. 1 статистике распределения общих писем между выбранными пользователями видно, что пользователь «kean-s» имеет существенное количество (более 25%) общих писем с «dasovich-j», что должно сказаться на качестве классификации.

Табл. 1. Статистика распределения общих писем между пользователями.

Письма пользователя	Общие письма с пользователем		
	dasovich-j	kaminski-v	kean-s
dasovich-j	11480	56	1730
kaminski-v	16	8757	36
kean-s	2628	91	10043

Для демонстрации предлагаемого подхода идентификации работы пользователя рассматривалась следующая задача бинарной классификации: нужно определить временные точки работы с письмами пользователя «dasovich-j» от временных точек работы пользователей «kaminski-v» и «kean-s» на каждой неделе прогноза каждого из 21-го анализируемого временного интервала.

Для каждого из 21-го анализируемого временного интервала производилась следующая процедура, состоящая из 5 шагов:

1. Для каждого пользователя в качестве шага временной точки выбирался не фиксированный шаг времени, а время, за которое пользователь успевал обработать 50 писем. Таким образом, каждая точка временных рядов пользователей представляет конкатенацию из 50 их писем (рис. 6). Текстовые данные в наборе Enron являются англоязычными, поэтому для формирования словаря термов использовались такие методы предварительной обработки текста, как удаление стоп-слов и приведение слов к нормализованной форме на основе семантической сети WordNet [21]. Для вычисления весов термов использовался только локальный логарифмический вес. Векторы временных интервалов нормализовались по евклидовой норме.
2. К сформированной матрице модельных временных интервалов (точек) пользователя «dasovich-j» применялось тематическое моделирование на основе ортонормированной неотрицательной матричной факторизации для получения матрицы «портрета» пользователя (W_k) и матрицы представления временных интервалов в пространстве тематик (H_k). В итоге для пользователя «dasovich-j» получаем k тематических временных рядов для модельного времени (в проводимых экспериментах k=3).
3. Отображения векторов временных точек времени прогноза для всех пользователей осуществлялось с использованием матрицы «портрета» пользователя «dasovich-j» (W_k). Таким образом, получаем реальные тематические данные всех пользователей для временных точек прогноза (рис. 6).
4. С помощью каждого метода прогнозирования (Раздел 3) строились прогнозы тематических временных рядов пользователя «dasovich-j» за модельное время (рис. 6). Далее для обозначения методов прогнозирования используются сокращения: метод линейной авторегрессии — AR, метод на основе авторегрессионной модели дерева решений — MS_ARTXP, предложенный метод прогнозирования на основе ортонормированной неотрицательной матричной факторизации — ONMF.
5. Для каждого метода прогнозирования рассчитывалась оценка отклонения каждой временной точки времени прогноза всех пользователей от спрогнозированных значений. Авторами исследовались два подхода к расчёту оценки отклонения временной точки от прогноза:
   a. *абсолютная оценка*: использование суммарного по всем k тематикам абсолютного отклонения реальных значений весов тематик от спрогнозированных (рис. 3);
   b. *оценка p-значения*: использование p-значения критерия согласия Хи-квадрат [22] для реальных (наблюдаемых) значений весов тематик и прогнозируемых, при этом веса тематик рассматриваются как вероятностное распределение, где вес каждой тематики фактически задает вероятность того, что указанный текст принадлежит данной тематике. Далее спрогнозированные веса рассматривались как ожидаемое распределение, а реальные веса — как наблюдаемое распределение вероятности, что позволяет определить уровень согласия по критерию Хи-квадрат. И чем он ниже, тем более вероятно, что произошла подмена пользователя.

Рис. 6. Временные ряды анализируемых пользователей для одной из тематик.

После проведения вышеописанной процедуры с каждым из 21-го анализируемого временного интервала для каждого метода прогнозирования получаем, что всем прогнозируемым временным точкам (96 точек прогноза для каждого пользователя, всего 288) всех пользователей сопоставлены две оценки отклонения.

Фиксируя значение порога допустимого отклонения от прогноза пользователя «dasovich-j» получаем бинарную классификацию для всех прогнозируемых временных интервалов. Для оценки качества классификации обычно используют ROC-кривые — графическая характеристика качества бинарного классификатора, зависимость доли верных положительных классификаций от доли ложных положительных классификаций при варьировании порога решающего правила (отклонения) [23]. Для сравнения нескольких моделей классификации будем использовать значение AUC (англ. Area Under Curve), которое вычисляется как площадь под ROC-кривой и является агрегированной характеристикой качества классификации, не зависящей от соотношения цен ошибок [23]. Чем больше значение AUC, тем «лучше» модель классификации. Полученные ROC-кривые и значения AUC для рассмотренных методов классификации и подходов расчёта оценки отклонения приведены на рис. 7, рис. 8 и табл. 2.

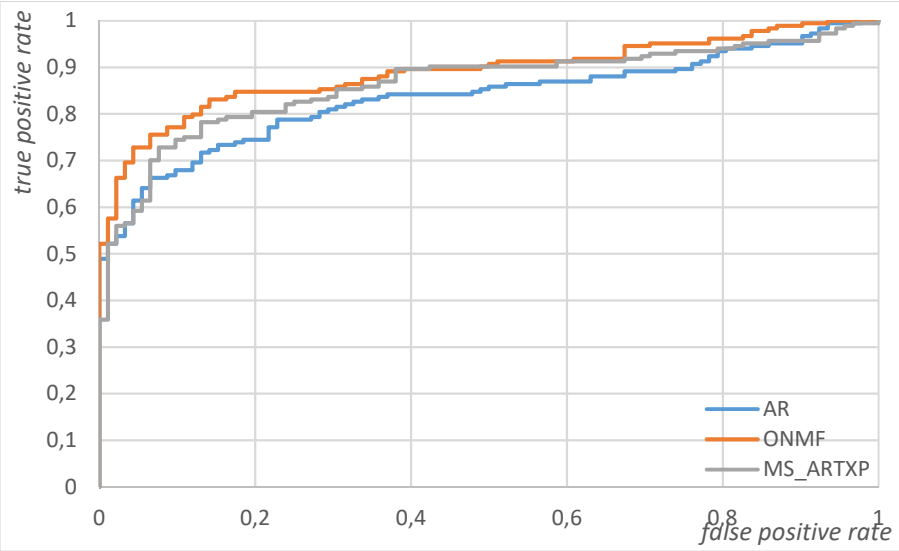

Рис. 7. ROC-кривые методов прогнозирования при абсолютной оценки отклонения.

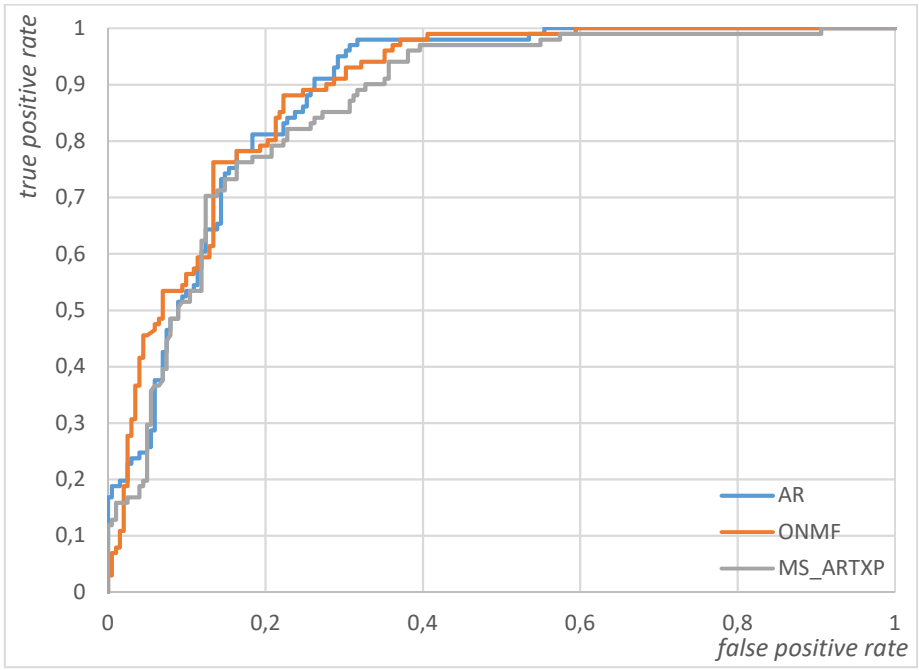

Рис. 8. ROC-кривые методов прогнозирования при оценке p-значения в качестве отклонения.

Табл. 2. Значения AUC для методов прогнозирования и подходов расчёта оценки отклонения.

	AR	MS_ARTXP	ONMF
Абсолютная оценка	0,835539	0,864367	0,889887
Оценка p-значения	0,883884	0,864719	0,889325

Из табл. 2 следует, что использование оценки p-значения показывает более стабильный результат классификации среди набора рассмотренных методов прогнозирования, так значение AUC не опускается ниже 0.86. Поэтому в дополнение приведём табл. 3 с сравнением средних p-значений критерия Хи-квадрат для пользователя «dasovich-j» и рассмотренных методов прогнозирования. При этом для достоверности рассматривались только те эксперименты, в которых число точек прогноз превышало 3, поэтому в табл. 3 приведены результаты 16 экспериментов вместо 21.

Табл. 3 Сравнение средних p-значений для пользователя «dasovich-j».

		Среднее p-значение		
Эксперимент	Число точек прогноза	AR	MS_ARTXP	ONMF
0	5	0.9475	0.8904	**0.9515**
2	7	0.8740	**0.888**	0.869
4	6	0.7773	0.7547	**0.8007**
5	6	0.9129	0.9037	**0.923**
6	6	0.8388	0.8285	**0.8526**
7	7	0.7932	**0.8528**	0.8336
8	4	**0.9462**	0.9038	0.9368
9	7	0.9718	0.9654	**0.9743**
10	7	**0.7456**	**0.7456**	0.6726
11	5	**0.8959**	0.8662	**0.8708**
12	7	0.7696	0.7492	**0.7985**
13	6	0.8609	0.5805	**0.8912**
14	4	**0.9288**	0.8555	0.9055
18	5	**0.8870**	0.5332	0.8639
19	4	0.9766	**0.9958**	0.9899
20	5	0.7365	**0.7409**	0.7267

Данные табл. 2 также согласуются с данными табл. 1 тем, что метод авторегрессии с использованием оценки p-значения продемонстрировал высокое значение AUC. Так в сравнении с остальными методами прогнозирования только метод авторегрессии во всех экспериментах показал средние p-значения большие 0.7.

Из приведённых данных следует, что:

- Предложенный подход идентификации работы пользователя на основе отклонений его тематической направленности от спрогнозированных данных показывает высокое качество идентификации даже при использовании стандартных методов прогнозирования;
- Предложенный в рамках данной статьи алгоритм прогнозирования временных рядов, основанный на ортонормированной неотрицательной матричной факторизации, показал высокое качество прогнозирования и свою применимость в рассмотренном подходе идентификации работы пользователя;

- Оба рассмотренные подхода расчёта оценки отклонения временной точки от прогноза применимы в предложенном подходе идентификации пользователя. Однако использование оценки p-значения показывает более стабильный результат классификации среди набора рассмотренных методов прогнозирования.

5. Заключение

В рамках проведенных исследований изучался вопрос возможности идентификации пользователя на основе анализа отклонений его тематической направленности при работе с текстовой информацией. Для решения указанной задачи был предложен подход, состоящий в тематическом анализе сложившихся в прошлом тенденций работы (поведения) пользователя с текстовым контентом различных (в том числе конфиденциальных) категорий и прогнозировании его дальнейшего поведения. Тематический анализ работы пользователя предполагает определение основных тематик его текстового контента и расчёт соответствующих им весов в заданные интервалы времени. На основе отклонений поведения в работе пользователя с контентом от прогноза осуществляется идентификация данного пользователя.

Для реализации тематического моделирования в предложенном подходе используется ортонормированная неотрицательная матричная факторизация (ОНМФ). В ходе дальнейших исследований был предложен собственный оригинальный метод прогнозирования, который также основан ОНМФ. Важно отметить, что ранее метод ОНМФ не использовался для решения задачи прогнозирования временных рядов.

Экспериментальное исследование предложенного подхода идентификации пользователя проводилось на примере реальной корпоративной переписки пользователей, сформированной из набора данных Enron. В ходе экспериментов было продемонстрировано, что использование разработанного метода прогнозирования на основе ОНМФ, показывает высокое качество классификации тематических характеристик пользователя по сравнению с другими популярны на сегодняшний день методами и свою применимость в рассмотренном подходе идентификации. Кроме того, в работе исследовались два различных подхода оценки отклонений: абсолютная оценка и оценка p-значения. Эксперименты показали, что оба рассмотренные подхода расчёта оценки отклонения временной точки от прогноза применимы в предложенном подходе идентификации пользователя.

В будущем планируется использовать предложенный подход идентификации пользователя при разработке программных средств анализа индивидуальных особенностей поведения пользователей компьютерных систем (поведенческой биометрии) при работе в рамках стандартного человеко-машинного интерфейса.

Список литературы

[1]. R.V. Yampolskiy, V. Govindaraju, Behavioural biometrics: a survey and classification. International Journal of Biometrics (IJBM), Vol. 1, No. 1, 2008.

[2]. Временной ряд (Time Series). March 24 2015. (http://www.machinelearning.ru/wiki/index.php?title=Временной_ряд)

[3]. И.В. Машечкин, М.И. Петровский, Д.В. Царёв. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования. Вычислительные методы и программирование. Том 14, 2013. 91-102.

[4]. I.V. Mashechkin, M.I. Petrovskiy, D.S. Popov, D.V. Tsarev. Automatic text summarization using latent semantic analysis. Programming and Computer Software, 2011, pp. 299-305.

[5]. D.V. Tsarev, M.I. Petrovskiy, I.V. Mashechkin. Using NMF-based text summarization to improve supervised and unsupervised classification. 11th International Conference on Hybrid Intelligent Systems (HIS), 2011. Malacca, MALAYSIA. P. 185-189.

[6]. D.V. Tsarev, M.I. Petrovskiy I.V. Mashechkin. Supervised and Unsupervised Text Classification via Generic Summarization. International Journal of Computer Information Systems and Industrial Management Applications. MIR Labs, Volume 5, 2013, pp. 509-515.

[7]. I.V. Mashechkin, M.I. Petrovskiy, D.S. Popov, D.V. Tsarev. Applying Text Mining Methods for Data Loss Prevention. Programming and Computer Software. January 2015, Volume 41, Issue 1, pp 23-30.

[8]. C.D. Manning, P. Raghavan, H. Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008.

[9]. A. Mirzal. Converged Algorithms for Orthogonal Nonnegative Matrix Factorizations. CoRR abs/1010.5290, 2010.

[10]. Wei Xu, Xin Liu, Yihong Gong. Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, 2003.

[11]. Chris Ding, Tao Li, Wei Peng, Haesun Park. Orthogonal Nonnegative Matrix Tri-Factorizations for Clustering. SIGKDD, 2006.

[12]. M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, R.J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis, pp. 155-173, 2007.

[13]. J. Yoo, S. Choi. Orthogonal Nonnegative Matrix Factorization: Multiplicative Updates on Stiefel Manifolds. Intelligent Data Engineering and Automated Learning – IDEAL 2008, vol. 5326 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, pp. 140–147.

[14]. C. Meek, D.M. Chickering, D. Heckerman. Autoregressive Tree Models for Time-Series Analysis, 2002. (http://go.microsoft.com/fwlink/?LinkId=45966)

[15]. Технический справочник по алгоритму временных рядов (Майкрософт). (http://msdn.microsoft.com/ru-ru/library/bb677216.aspx)

[16]. T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, D. Botstein. Imputing Missing Data for Gene Expression Arrays. Technical report, Stanford Statistics Department 1999.

[17]. O. Troyanskaya. Missing value estimation methods for DNA microarrays. Bioinformatics, , vol. 17, no. 6, 2001. pp. 520-525.

[18]. D.V. Tsarev, R.V. Kurynin, M.I. Petrovskiy, I.V. Mashechkin. Applying non-negative matrix factorization methods to discover user's resource access patterns for computer security tasks. Proceedings of the 2014 International Conference on Hybrid Intelligent Systems (HIS 2014). IEEE Computer Society [New York], United States, 2014. pp. 43–48.

[19]. D. Lee, S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401, 1999. pp. 788-791.

[20]. Enron Email Dataset. March 24 2015. (http://www.cs.cmu.edu/~./enron/)

[21]. Natural Language Toolkit (NLTK). March 24 2015. (http://www.nltk.org)

[22]. М. Кендалл, А. Стьюарт. Статистические выводы и связи. М.: Наука, 1973.

[23]. Кривая ошибок (Receiver Operating Characteristic, ROC curve). March 24 2015. (http://www.machinelearning.ru/wiki/index.php?title=ROC-кривая)

Applying Time Series to The Task of Background User Identification Based on Their Text Data Analysis[2]

D.V. Tsarev <tsarev@cs.msu.su>
M.I. Petrovskiy <michael@cs.msu.su>
I.V. Mashechkin <mash@cs.msu.su>
A.Y. Korchagin <proton.ru@gmail.com>
V.Y. Korolev <bruce27@yandex.ru>
Department of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia.

Abstract. The paper presents the novel approach of user identification based on behavior analytics of user operations with a text information. It is offered to describe user behavior by content of his text documents. The structured representation of the considered behavioral information is carried out based on representation of documents text content in the user topic space, which is created by non-negative matrix factorization. The topic weights in the document characterize the user's topic trend during an operating time with this document. The time variation of the topic weight values creates multidimensional time series that describe the history of user behavior when working with text data. Forecasting of such time series will allow for user identification based on estimated deviation of observed topic trend from the predicted topic weight values. This paper also presents the new time series forecasting method based on orthogonal nonnegative matrix factorization (ONMF) which is

[2] The production of this publication has been made possible through the financial support of the Ministry of Education and Science of the Russian Federation (the subsidy agreement #14.604.21.0056, Unique project identifier RFMEFI60414X0056).

used within proposed user identification approach. It is worth noting that nonnegative matrix factorization methods were not used before for the time series forecasting task. The proposed user identification approach has been experimentally verified on the example of real corporate email correspondence created from the Enron dataset. In addition, experiments with other today popular forecasting methods have shown the superiority of proposed forecasting method in quality of user's topic weights classification. Also we investigated two different approaches to estimates of the deviation of a time series point from the predicted value: absolute deviation and p-value estimation. Experiments have shown that both discussed approaches of deviation estimates are applicable in the proposed user identification approach.

Keywords: computer security; user identification; topic modeling; orthogonal nonnegative matrix factorization; time series forecasting.

DOI: 10.15514/ISPRAS-2015-27(5)-8

For citation: Tsarev D.V., Petrovskiy M.I., Mashechkin I.V., Korchagin A.Y., Korolev V.Y. Applying Time Series to The Task of Background User Identification Based on Their Text Data Analysis. *Trudy ISP RAN/Proc. ISP RAS*, vol. 27, issue 1, 2015, pp. 173-192 (in Russian). DOI: 10.15514/ISPRAS-2015-27(5)-8.

References

[1]. R.V. Yampolskiy, V. Govindaraju, Behavioural biometrics: a survey and classification. International Journal of Biometrics (IJBM), Vol. 1, No. 1, 2008.

[2]. Vremennoi ryad [Time Series]. March 24 2015. (http://www.machinelearning.ru/wiki/index.php?title=Временной_ряд) (in Russian)

[3]. I.V. Mashechkin, M.I. Petrovskiy, D.V.Tsarev. Metody vychislenija relevantnosti fragmentov teksta na osnove tematicheskix modelej v zadache avtomaticheskogo annotirovanija [Methods of text fragment relevance estimation based on the topic model analysis in the text summarization problem]. Vychislitel'nye Metody i Programmirovanie [Numerical Methods and Programming], 2013, vol. 14, pp. 91–102. (in Russian).

[4]. I.V. Mashechkin, M.I. Petrovskiy, D.S. Popov, D.V. Tsarev. Automatic text summarization using latent semantic analysis. Programming and Computer Software, 2011, pp. 299-305.

[5]. D.V. Tsarev, M.I. Petrovskiy, I.V. Mashechkin. Using NMF-based text summarization to improve supervised and unsupervised classification. 11th International Conference on Hybrid Intelligent Systems (HIS), 2011. Malacca, MALAYSIA. P. 185-189.

[6]. D.V. Tsarev, M.I. Petrovskiy I.V. Mashechkin. Supervised and Unsupervised Text Classification via Generic Summarization. International Journal of Computer Information Systems and Industrial Management Applications. MIR Labs, Volume 5, 2013, pp. 509-515.

[7]. I.V. Mashechkin, M.I. Petrovskiy, D.S. Popov, D.V. Tsarev. Applying Text Mining Methods for Data Loss Prevention. Programming and Computer Software. January 2015, Volume 41, Issue 1, pp 23-30.

[8]. C.D. Manning, P. Raghavan, H. Schutze. Introduction to Information Retrieval. Cambridge University Press, 2008.

[9]. A. Mirzal. Converged Algorithms for Orthogonal Nonnegative Matrix Factorizations. CoRR abs/1010.5290, 2010.

[10]. Wei Xu, Xin Liu, Yihong Gong. Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, 2003.

[11]. Chris Ding, Tao Li, Wei Peng, Haesun Park. Orthogonal Nonnegative Matrix Tri-Factorizations for Clustering. SIGKDD, 2006.

[12]. M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, R.J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis, pp. 155-173, 2007.

[13]. J. Yoo, S. Choi. Orthogonal Nonnegative Matrix Factorization: Multiplicative Updates on Stiefel Manifolds. Intelligent Data Engineering and Automated Learning – IDEAL 2008, vol. 5326 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, pp. 140–147.

[14]. C. Meek, D.M. Chickering, D. Heckerman. Autoregressive Tree Models for Time-Series Analysis, 2002. (http://go.microsoft.com/fwlink/?LinkId=45966)

[15]. Tekhnicheskii spravochnik po algoritmu vremennykh ryadov (Microsoft) [Microsoft Time Series Algorithm Technical Reference]. (http://msdn.microsoft.com/ru-ru/library/bb677216.aspx) (in Russian)

[16]. T. Hastie, R. Tibshirani, G. Sherlock, M. Eisen, P. Brown, D. Botstein. Imputing Missing Data for Gene Expression Arrays. Technical report, Stanford Statistics Department 1999.

[17]. O. Troyanskaya. Missing value estimation methods for DNA microarrays. Bioinformatics, , vol. 17, no. 6, 2001. pp. 520-525.

[18]. D.V. Tsarev, R.V. Kurynin, M.I. Petrovskiy, I.V. Mashechkin. Applying non-negative matrix factorization methods to discover user's resource access patterns for computer security tasks. Proceedings of the 2014 International Conference on Hybrid Intelligent Systems (HIS 2014). IEEE Computer Society [New York], United States, 2014. pp. 43–48.

[19]. D. Lee, S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401, 1999. pp. 788-791.

[20]. Enron Email Dataset. March 24 2015. (http://www.cs.cmu.edu/~./enron/)

[21]. Natural Language Toolkit (NLTK). March 24 2015. (http://www.nltk.org)

[22]. M. Kendall, A. Stuart. Statisticheskie vyvody i svyazi [Statistical derivations and associations.]. M.: Nauka, 1973 (In Russian).

[23]. Krivaya oshibok [Receiver Operating Characteristic, ROC curve]. March 24 2015. (http://www.machinelearning.ru/wiki/index.php?title=ROC-кривая) (In Russian)