# АВТОМАТИЗИРОВАННЫЙ АНАЛИЗ ТЕМАТИКИ ТЕКСТОВ НОВОСТЕЙ

**Алла Заболеева-Зотова, Алексей Петровский, Юлия Орлова, Татьяна Шитова**

***Аннотация**: В работе рассматриваются возможные подходы к автоматизированному анализу текста новостей с помощью алгоритмов кластеризации текстов, извлечения ключевых слов и формирования семантической связности блоков текста. Особое внимание уделено выявлению тематики и сюжетов новостей.*

***Ключевые слова**: поток новостей, тематическая кластеризация текстов, извлечение ключевых слов, семантическая связность блоков текста*

***ACM Classification Keywords**: A.0 General Literature - Conference proceedings*

## Введение

Проблема снижения информационной перегрузки людей, и в особенности пользователей Интернета и социальных сетей, становится всё более актуальной. Как сообщает американская исследовательская служба Cyveillance, в начале XXI века количество страниц в Интернете превысило 4 млрд, и с каждым днем увеличивается на 7 млн. Темпы роста аудитории онлайновых новостных ресурсов практически вдвое превышают темпы роста общей численности пользователей Интернета. Большую часть информации, с которой имеют дело пользователи, составляют «сырые» неструктурированные данные. Поэтому велика потребность в эффективных технологиях автоматизированного анализа информации, представленной на естественном языке, выявления групп семантически похожих текстов.

Существует достаточно много программных продуктов, предоставляющих функции анализа текстовых документов. Среди отечественных систем отметим TextAnalyst, Galaktika-ZOOM, из зарубежных – мощный инструмент анализа текстов IBM Text Miner. В TextAnalyst имеются опции создания семантической сети большого текста, подготовки аннотации, автоматической классификации и кластеризации текстов. IBM Text Miner содержит утилиты классификации, кластеризации, поиска ключевых слов и составления аннотации текстов. Однако эти программы не направлены на обработку новостных статей.

Российская система Яндекс Новости позволяет автоматически группировать данные в новостные сюжеты и составлять аннотации статей на основе кластеризации документов. Сервис InfoStream,

обеспечивает доступ к оперативной информации в поисковом режиме с учетом семантической близости документов. Мобильный агрегатор новостей Summly, используемый компанией Yahoo!, также осуществляет группировку новостей по темам. Однако приложение абсолютно неприменимо для обработки текстов на русском языке.

Таким образом, существующие программные системы не решают поставленную проблему полностью. В работе предложена методика комплексного анализа текста новостей, основанная на комбинации алгоритмов тематической кластеризации текстов, статистических алгоритмов извлечения ключевых слов и алгоритмов формирования семантической связности блоков текста. Проведена апробация методики при он-лайн обработке потоков новостных интернет-статей.

## Обработка и анализ текста новости

Анализ новостных текстов включает в себя тематическую кластеризацию текстов и последующую комплексную обработку статей. В основу предлагаемой обобщенной структуры текста новости (рисунок 1) положен принцип «перевернутой пирамиды», который требует размещение основной информации в самом начале материала и последующее ее раскрытие в деталях далее по тексту.

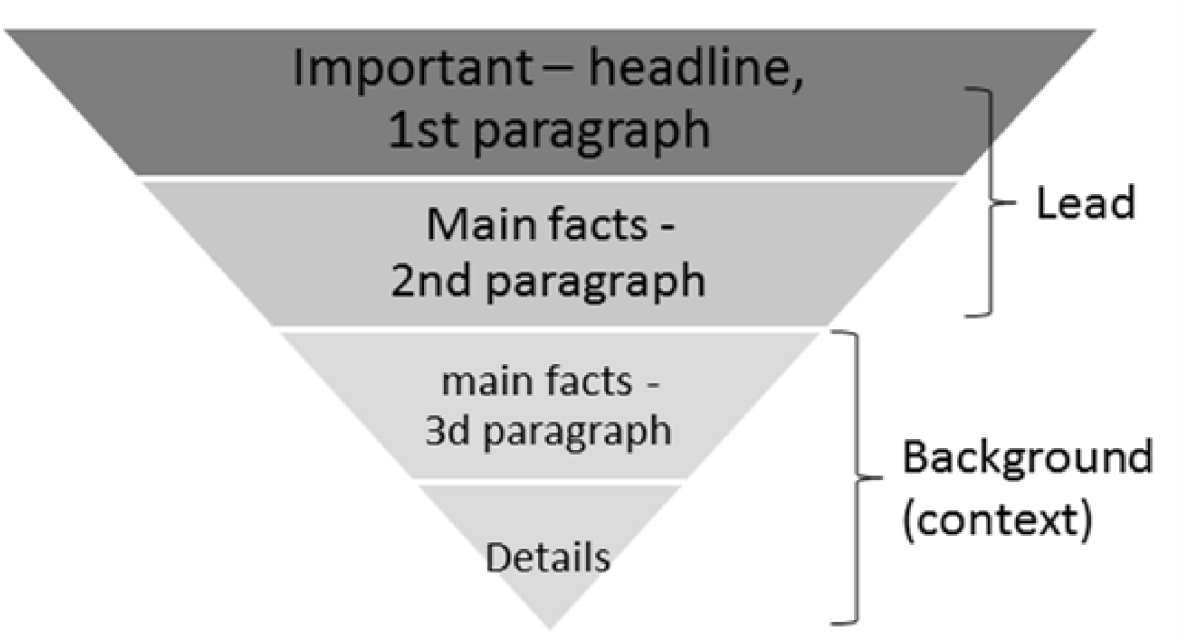


Рисунок 1 – Структура новостного текста.

Заголовок новости отражает ее тему и содержит не более 10 слов (около 80 символов). Так, для примера, в Яндексе отображается в заголовке не более 15-ти слов, Google показывает до 70 слов.

Основные факты, касающиеся события, представляются в 1-2 абзацах, и составляют так называемую вводную часть (lead), которая освещает главную тему новости.

3-й и последующие абзацы составляют контекст (background) новости. Как правило, здесь раскрываются детали происходящего, дается информация, напрямую касающаяся новости.

Таким образом, для содержания новости справедлива формула: (Who? + What? + Where? + Why? + When? + How?) [Добров, 2010]. Это так называемый закон «пяти W и одного H», приписываемый Р.Киплингу. Если бы все новостные сообщения строились по единой структуре, то решение задачи тематического анализа текстов новостей могло бы значительно упроститься.

Процесс обработки и анализа текста новости можно разбить на несколько этапов.

Графематический анализ представляет собой начальный этап обработки текста, в ходе которого вырабатывается информация, необходимая для дальнейшей обработки морфологическим и синтаксическим анализаторами. В задачу графематического анализа входит внутреннее представление структуры новости: T = <P, S, W>, где P – абзацы, S – предложения, W – слова. При этом необходимо корректно выделить заголовок и первое предложение абзаца, содержащее основные факты статьи.

Следующим этапом является морфологический анализ, цель которого – построение морфологической интерпретации слов входного текста. Все методы можно разделить на словарные и вероятностно-статистические (без использования словаря). Недостатками вторых являются большой объем лексиконов, плохая работа на малой выборке, отсутствие точных лингвистических методов. Словарный метод основан на подключении словаря (тезауруса), дает максимально полный анализ словоформы. Для данного блока целесообразно использовать морфологические библиотеки, например, Lemmatizer, FreeLing, NLTK, MCR, tokenizer [Михайлов и др., 2009].

Синтаксический анализ рассматривается как задача построения дерева зависимостей предложения, в ходе которого происходит выделение синтаксических конструкций, определение связности и подчинения фрагментов [Grune, 2012]. Для поиска ключевых слов разработан алгоритм [Soloshenko et al, 2014], сочетающий выделение именованных сущностей из текста

новости (на основе результатов морфоанализа и подключаемого модуля PullEnti), подсчет веса слова с учетом частоты его встречаемости (рисунок 2). Пороговое значение для признания слова ключевым задается значением относительной частоты встречаемости слова-кандидата в ключевые слова, с индексом, равным 0,2×количество сущностей. Такое значение определено экспериментально на выборке из 100 текстов.


Рисунок 2 – Алгоритм поиска ключевых фраз в тексте.

Структуру совокупности знаний S текста новости можно определить следующим образом : S = {M, F}, где M – множество всех понятий данной совокупности знаний, F – отношение «смысловая связь». В качестве формальной модели структуры знаний можно использовать семантическую сеть, определяемую как ориентированный граф G = (E,V), где E – множество вершин, поставленное во взаимно однозначное соответствие с множеством понятий; V – множество ориентированных дуг. Дуга выходит из вершины, соответствующей основному понятию A, и входит в вершину, соответствующую понятию, которое сочетается по смыслу с понятием A [Машечкин и др., 2011; Dmitriev et al, 2013]. Таким образом, содержание новости можно наглядно представить в виде ключевых понятий и связей между ними, либо в виде так называемой интеллектуальной карты (mind map).

Подсчет веса предложения при построении аннотации осуществляется в зависимости от его нахождения в тексте новости и рассчитывается по формуле:

$$Ws = N(kw)\cdot RF(kw)\cdot WP\ C. \quad (1)$$

Здесь Ws – вес предложения; N(kw) – количество вхождений ключевого слова в предложение; RF(kw) – относительная частота ключевого слова; WP – относительный вес параграфа в тексте, равный 0.35 для первого параграфа, 0.2 для второго параграфа, 0.1 для остальных параграфов (контекст); C – коэффициент значимости предложения внутри параграфа, равный 1.0 для первого предложения в абзаце, 0.8 для остальных предложений [Soloshenko et al, 2014]. В итоговую аннотацию в зависимости от заданного коэффициента сжатия включаются предложения с наибольшим весом.

## Методы тематической кластеризации текстов

Напомним основные понятия. Кластеризация – разбиение множества объектов на подмножества (кластеры), число которых может быть произвольным или фиксированным. Основные группы алгоритмов кластеризации: иерархические и неиерархические, четкие и нечеткие. Иерархические алгоритмы строят несколько разбиений исходного множества объектов на непересекающиеся кластеры. Результатом является дерево кластеров, корень которого – все исходные объекты, а листья – итоговые кластеры. Неиерархические алгоритмы строят одно разбиение объектов на заданное число кластеров. Четкие алгоритмы определяют принадлежность каждого исходного

объекта только одному кластеру. Нечеткие алгоритмы ставят в соответствие каждому объекту степень его принадлежности к нескольким кластерам [Bandyopadhyay et al., 2013].

Иерархические алгоритмы разделяются на агломеративные (восходящие) и дивизимные (нисходящие). Первые строят кластеры снизу вверх, начиная с множества кластеров, содержащих по одному одиночному объекту, затем последовательно объединяют пары кластеров, пока не получат один кластер, содержащий все исходные объекты. Вторые разбивают кластеры сверху вниз, начиная с одного кластера, которому принадлежат все исходные объекты, затем этот кластер делится на два и так рекурсивно до тех пор, пока каждый объект не окажется в своём отдельном кластере. Основное их различие заключается в выборе критерия, используемого для принятия решения о том, какие кластеры следует объединить на текущем шаге алгоритма. Большое распространение получили следующие критерии:

- одиночная связь (минимальное расстояние или максимально сходство) – сходство между двумя наиболее похожими объектами/кластерами;

- полная связь (максимальное расстояние или минимальное сходство) – сходство между двумя наиболее непохожими объектами/кластерами;

- групповое усреднение всех показателей сходства – сходство двух кластеров есть среднее сходство всех пар объектов, включая пары объектов из одного кластера, исключая близость объекта самому себе;

- центроидный метод;

- метод Уорда.

Алгоритм k-средних начинает с некоторого начального разбиения объектов на заранее заданное число кластеров и уточняет его, оптимизируя целевую функцию – среднеквадратичную ошибку кластеризации как среднеквадратичное расстояние между объектами и центрами их кластеров:

$$e(D,C) = \sum_{j=1}^{k} \sum_{i:d_i \in c_j} || \vec{d}_i - \vec{\mu}_j ||^2 \qquad (2)$$

где $\mu_j$ – центроид кластера $C_j$. Обычно исходные центры кластеров выбираются случайным образом. Затем каждый объект присваивается тому кластеру, чей центр является наиболее близким документу, и выполняется повторное вычисление центра каждого кластера как центроида, или среднего своих членов. Такое перемещение объектов и повторное вычисление центроидов кластеров продолжается до тех пор, пока не будет достигнуто условие остановки, например: (а) достигнуто пороговое число итераций, (б) центроиды кластеров больше не изменяются, (в) достигнуто пороговое значение ошибки кластеризации.

Нечеткий алгоритм классификации k-средних FCM относит каждый объект к более чем одному кластеру. Как и его чёткий вариант, данный алгоритм, начиная с некоторого начального разбиения данных, итеративно минимизирует целевую функцию, которой является следующее выражение:

$$e_m(D,C) = \sum_{i=1}^{|D|}\sum_{j=1}^{|C|} u_{ij}^m \,||\, \vec{d}_i - \vec{\mu}_j \,||^2 \qquad (3)$$

где $m$ – степень нечеткости, $1<m<\infty$, $u_{ij}$ - степень принадлежности $i$-го объекта $j$-му кластеру.

Алгоритм минимального покрывающего дерева MST сначала строит на графе минимальное покрывающее дерево, а затем последовательно удаляет ребра с наибольшим весом. На рисунке 3 изображено минимальное покрывающее дерево, полученное для девяти объектов. Путём удаления связи, помеченной CD, с длиной равной 6 единицам (ребро с максимальным расстоянием), получаем два кластера: {A, B, C} и {D, E, F, G, H, I}. Второй кластер в дальнейшем может быть разделён ещё на два кластера путём удаления ребра EF, которое имеет длину 4,5 единицы [Pera et al., 2012].

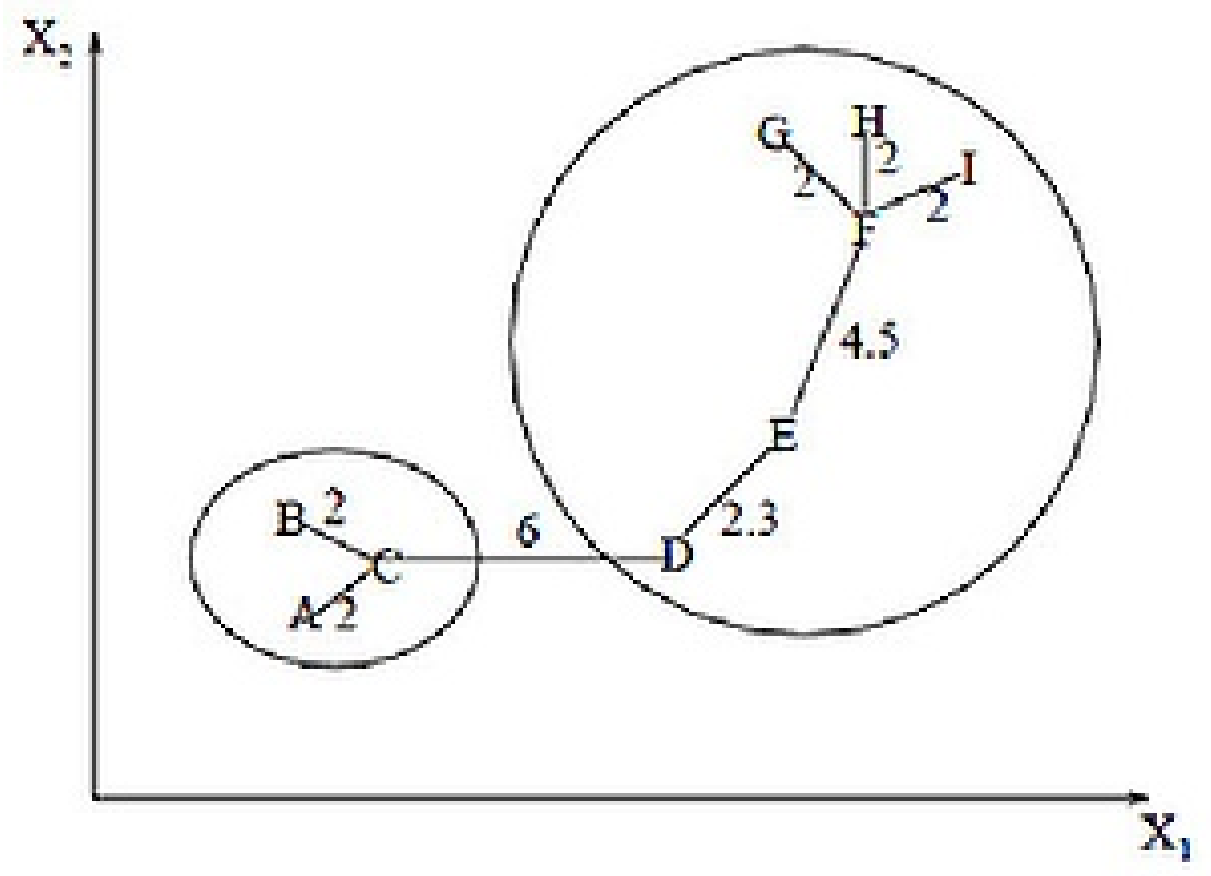


Рисунок 3 – Иллюстрация алгоритма MST.

Алгоритм самоорганизующихся карт (Self Organizing Maps – SOM) был предложены для визуализации и кластеризации данных. Визуализация данных осуществляется путём проецирования многомерного пространства данных в двумерное пространство – карту данных. Такая карта, построенная для массива полнотекстовых документов, может служить как поисковый механизм, альтернативный поиску по запросу, предлагающий обзор/навигацию по коллекции документов [Kiryakov, 2004]. Идея алгоритма заключается в том, чтобы обучить нейронную сеть без учителя. Сеть состоит из некоторого числа нейронов, упорядоченных по узлам двумерной сетки. Каждый нейрон имеет координаты в исходном $|\tau|$-мерном пространстве документов и двумерном пространстве карты. В процессе обучения нейроны упорядочиваются в пространстве документов так, чтобы наилучшим образом описать входной массив документов. Этот процесс является итерационным. На каждой итерации $t$ случайным образом выбирают документ $d_i$ из входного массива $D$; находят нейрон-победитель $m_c$, ближайший к документу $d_i$; корректируют веса соседей нейрона-победителя: $m_i(t+1) = m_i(t) + h_{ci}(t)[d_i - m_i(t)]$.

Алгоритмы групповой иерархической и неиерархической кластеризации объектов разработаны для случаев, когда объекты могут присутствовать в нескольких различающихся версиях [Petrovsky, 2003]. Объекты, описываемые многими количественными и/или качественными признаками, рассматриваются как точки метрического пространства мультимножеств [Петровский, 2003]. Примером может служить одна и та же новость, содержащаяся в разных текстах или опубликованная несколько раз в разное время.

При выборе оптимального алгоритма кластеризации потока текстов новостей необходимо учитывать его следующие особенности: постоянно растущая коллекция документов, одна и та же статья может отражать несколько сюжетов, новости имеют определенную структуру текста, разные части документа должны иметь различный вес при нахождении близости, сюжеты и документы могут иметь перекрестные ссылки друг на друга. Для работы с новостными текстами желательно, чтобы алгоритм был неиерархическим, нечетким, инкрементальным. Поэтому наиболее перспективным для данной задачи видится применение алгоритма FCM или нейронных сетей.

## Апробация методического подхода

Для проверки предложенного подхода к тематической кластеризации текстов новостей была создан программный комплекс, основанный на принципе многокомпонентности программного обеспечения [Zaboleeva-Zotova et al, 2013], а также проведен эксперимент [Солошенко и др., 2014], цель которого – доказать, что за счет автоматизации обработки статей новостного потока снизилось время на обработку и повысилось качество обработки новостных интернет-статей. Были получены следующие результаты.

Время обработки текстов новостей уменьшилось как минимум в два раза. При этом время с помощью программы учитывается не только непосредственно время составления аннотации, но и время, необходимое для окончательной корректировки текстов (рисунок 4). Качество аннотации оценивалось экспертами по следующим критериям: сохранение ключевых фактов, связность новостной статьи, сохранение синтаксической структуры текста после удаления незначащих частей. Каждый из критериев имел шкалу оценок от 0 до 10 баллов. Качество полученной аннотации оценивалось по среднему арифметическому трех показателей для каждого текста. Качество автоматически обработанных текстов новостей осталось на том же уровне, что и при анализе текста человеком.

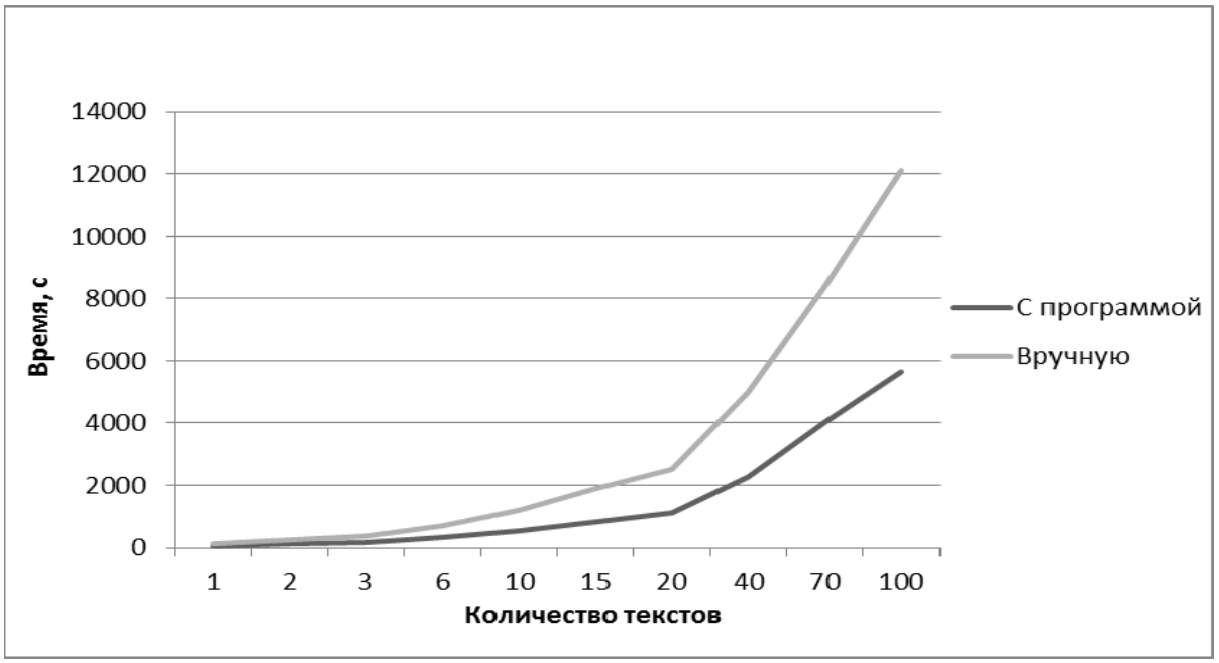


Рисунок 4 – Зависимость времени обработки от количества текстов.

## Заключение

Анализ новостных текстов включает в себя задачу кластеризации и последующую комплексную обработку статей. Проведен анализ потока новостей, выделены особенности документов. Проанализированы методы кластеризации объектов, предложены алгоритмы, наиболее подходящие для анализа текстов новостей. Реализована часть программной системы для он-лайн агрегации новостей из интернет-источников,и проведены исследования эффективности ее работы.

## Благодарности

The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Работа опубликована при частичной поддержке проекта ITHEA XXI общества ITHEA ISS (www.ithea.org) и ADUIS (www.aduis.com.ua).

## Библиография

[Добров, 2010] Добров Б.В. Исследование качества базовых методов кластеризации новостного потока в суточном временном окне // Труды 12-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Казань, 2010. – 287–295.

[Машечкин и др., 2011] Машечкин И.В., Петровский М.И. Латентно-семантический анализ в задаче автоматического аннотирования // Программирование. – 2011. – Т. 37, № 6. – 67-77

[Михайлов и др., 2009] Михайлов Д.В., Емельянов Г.М. Морфология и синтаксис в задаче семантической кластеризации // Математические методы распознавания образов. – Суздаль, 2009. – 1-4.

[Петровский, 2003] Петровский А.Б. Пространства множеств и мультимножеств. – М., Едиториал УРСС, 2003.

[Солошенко и др., 2014] Солошенко А.Н., Розалиев В.Л., Орлова, Ю.А. Автоматизация семантического анализа новостных Интернет-текстов. // Открытые семантические технологии проектирования интеллектуальных систем: матер. IV междунар. науч.-техн. конф. – Минск, БГУИР, 2014. – 435-438.

[Bandyopadhyay et al., 2013] Bandyopadhyay S., Saha S. Unsupervised Classification. – Berlin, Springer, 2013.

[Grune, 2012] Grune D. Tokens to Syntax Tree – Syntax Analysis. – New York, Springer, 2012.

[Dmitiev et al, 2013] Dmitiev A.S., Zaboleeva-Zotova A.V., Orlova Yu.A., Rozaliev V.L. Automatic identification of time and space categories in the natural language text // Applied Computing 2013: Proceedings of the IADIS International Conference. – Fort Worth, 2013. – 187-190.

[Kiryakov, 2004] Kiryakov A. Semantic annotation, indexing, and retrieval // Web Semantics: Science, Services and Agents on the World Wide Web. – 2004. V.2. № 1. – 49-79.

[Petrovsky, 2003] Petrovsky A.B. Cluster analysis in multiset spaces. // Information Systems Technology and its Applications. – Bonn, Gesellschaft für Informatik, 2003. – 109-119.

[Pera et al., 2012] Pera, M.S., Ng, Y.-K.D. Using maximal spanning trees and word similarity to generate hierarchical clusters of non-redundant RSS news articles. // J. Intell. Inf. Syst. – 2012. V.39. – 513-534.

[Soloshenko et al, 2014] Soloshenko A.N., Orlova Yu.A., Rozaliev V.L., Zaboleeva-Zotova A.V. Thematic clustering methods applied to news texts analysis // Knowledge-Based Software Engineering: Proceedings of 11th Joint Conference. – Springer, 2014. – 294-310.

[Zaboleeva-Zotova et al, 2013] Zaboleeva-Zotova A.V., Orlova Yu.A., Rozaliev V.L., Fomenkov S.A., Petrovsky A.B. Formalization of initial stage of designing multi-component software // Multi Conference on Computer Science and Information Systems: Proceedings of the IADIS International Conference – Prague, IADIS, 2013. – 107-111.

**Сведения об авторах**

***Заболеева-Зотова Алла Викторовна*** *– д.т.н., профессор, начальник управления Российского фонда фундаментальных исследований, старший научный сотрудник Института системного анализа ФИЦ «Информатика и управление» РАН, Россия, Москва 119991, Ленинский пр-т, 32А, e-mail::* zabzot@rfbr.ru

***Петровский Алексей Борисович*** *– д.т.н., профессор, заведующий лабораторией Института системного анализа ФИЦ «Информатика и управление» РАН, Россия, Москва 117312, пр-т 60-летия Октября, 9, e-mail::* pab@isa.ru

***Орлова Юлия Александровна*** *– к.т.н., к.п.н., доцент, доцент Волгоградского государственного технического университета, Россия, Волгоград 400005, пр-т Ленина, 28, e-mail:: yulia.orlova@gmail.com*

***Шитова Татьяна Алексеевна*** *– экономист Института системного анализа ФИЦ «Информатика и управление» РАН, Россия, Москва 117312, пр-т 60-летия Октября, 9, e-mail:: tanya-petrovskay@yandex.ru*

## Automated Analysis of Thematic of News Texts

**Alla Zaboleeva-Zotova, Alexey Petrovsky, Yulia Orlova, Tatiana Shitova**

***Abstract****: The paper deals with possible approaches to the automated analysis of news texts using algorithms for clustering texts, keywords extracting and forming the semantic coherence of text blocks. Particular attention is paid to the discovery of themes and subjects of news.*

***Keywords:*** *stream of news, thematic clustering texts, keywords extracting, semantic coherence of text blocks.*