АНАЛИЗ НОВОСТНЫХ ПОТОКОВ НА ОСНОВЕ ИНФОРМАЦИОННОГО ПОИСКА И КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ

ANALYSIS OF NEWS STREAMS BASED ON INFORMATION SEARCH AND COMPUTATIONAL LINGUISTICS METHODS

Казенников Антон Олегович / Anton O. Kazennikov,
аспирант Московского государственного института радиотехники, электроники и автоматики / graduate student of Moscow State Technical University of Radioengineering, Electronics and Automation,
kazennikov@gmail.com

Аннотация

В настоящей статье представлен гибридный алгоритм первичного анализа новостных потоков. Результатом работы алгоритма является кластеризация потока новостей по сюжетам. Представленный алгоритм рассчитан на обработку большого числа новостных лент и предполагает значительное перекрытие сюжетов во входящем новостном потоке. Основная идея представленного подхода состоит в сочетании поверхностных подходов анализа текстов, часто используемых в информационном поиске, с глубоким лингвистическим анализом. В результате гибридного подхода достигается высокая точность и полнота анализа с низким временем анализа одного сообщения.

Abstract

In this article we present a hybrid algorithm for news streams analysis. This algorithm clusters source news stream by topics. The presented algorithm is intended for large scale news feed processing and assumes significant topic intersection in the source news stream. The main idea of the presented method is to integrate surface analysis text processing methods, commonly found in information search field, with deep processing used in field of computational linguistics. The resulting hybrid method reaches high precision and recall with small processing time of a single news message.

Ключевые слова: новостные потоки, информационный поиск, гибридный алгоритм первичного анализа.

Keywords: news streams, information search, hybrid algorithm for analysis.

За прошедшее десятилетие произошел быстрый рост количества электронных лент новостей ведущих информационных агентств. Из-за этого человек практически не в состоянии прочитать и проанализировать все новостные сообщения. Разработка методов для автоматической обработки и агрегации новостных потоков позволяет существенно сократить объем материалов, необходимых для просмотра и анализа человеком. Таким образом, задача автоматической обработки новостных лет является востребованной и актуальной. Кроме того, произошли существенные изменения в составе предлагаемого пользователю материала. В настоящее время новостное сообщение обычно состоит не только из текста новости, но и обладает ссылкой на категорию, к которой она относится, также вероятно наличие краткого списка ссылок по теме, которые поместил редактор.

В настоящее время существует ряд научных работ [1,2,3] и промышленных систем [5,6], которые решают задачу кластеризации новостей. Однако почти все системы ограничиваются поверхностными характеристиками текста. Основным способом представления текста в таких подходах является множество

слов, встреченных в тексте (bag of words). Обычно для такого представления используются заголовки, текстовое содержание новостей. Однако такой способ анализа текста является поверхностным и не охватывает глубинные зависимости и структуру текста.

С другой стороны, за последние годы значительно усовершенствовались методы традиционного глубокого анализа текстов. Такой подход предусматривает последовательный анализ текста на морфологическом, синтаксическом и семантическом уровнях.

Вследствие этого можно предполагать, что подходы к анализу новостных потоков, сочетающие в себе как поверхностные, так и глубинные методы анализа, будут, с одной стороны, сравнимы по скорости работы с поверхностными подходами, а, с другой стороны, позволят получить более высокое качество анализа, нежели чем у только поверхностных методов.

Таким образом, задача анализа потока новостных сообщений является, с одной стороны, как задачей автоматической обработки текстов, так и задачей анализа потока данных.

Центральной задачей для анализа новостных потоков является проблема выявления схожих новостей. Она возникает как самостоятельная задача представления пользователю «сюжетов» - последовательности новостей, которые бы описывали произошедшее событие и его дальнейшее развитие. Также, она входит в состав задачи составления рекомендаций по новостям, которые могли бы заинтересовать пользователя, опираясь на содержание текущей. Кроме того, задача выявления схожих новостей возникает при автоматическом реферировании новостей, когда необходимо выявить первоначальный материал для реферирования.

Постановка задачи

В настоящей статье рассматривается следующая задача. Дан входящий поток новостных сообщений от разных информационных агентств. Необходимо разделить эти сообщения на группы новостей на основе «сюжетов». Под сюжетом в настоящей статье подразумевается некоторое событие, произошедшее в реальном мире, и его контекст. Целью настоящей статьи является разработка метода кластеризации новостных потоков, который бы обладал нижеперечисленными характеристиками.

Онлайновый режим работы. Рамки настоящей задачи предполагают обработку интенсивного потока входящих сообщений. По предварительным оценкам поток новостей от 30 крупнейших информационных агентств и новостных ресурсов превышает 20 тыс. сообщений в сутки. Следовательно, максимальное время обработки одного сообщения не может превышать 8 секунд, а общее число сообщений, которое следует учитывать, составляет около 140 тыс. сообщений в неделю и более 600 тыс. сообщений в рамках одного месяца. Средняя длина новостного сообщения составляет 20 предложений или около 200-300 слов.

Плоская кластеризация. Предполагается кластеризация только новостных сообщений в отличие от иерархической кластеризации, когда кластеры в свою очередь разбиваются на кластеры.

Использование всей доступной информации, которую возможно получить из новостного сообщения. Предварительный анализ показывает, что кроме традиционных характеристик сообщения, таких как время написания, заголовок и текст сообщения, в большинстве случаев доступна краткая аннотация сообщения, а также несколько редакторских ссылок на связанные или похожие события.

Поскольку качественная оценка эффективности кластеризации системы в значительной степени неформальна, то для ее проведения предполагается использование предварительно аннотированного корпуса новостных сообщений, который был бы вручную размечен на сюжетные кластеры.

Ретроспективный обзор проблемы

Проблема обработки новостей не является принципиально новой. Традиционно обработка новостей рассматривается как задача информационного поиска. Основные современные результаты представлены в работах [1,2,3,5], где задача выделения групп новостей, объединенных одним сюжетом, решалась с помощью кластеризации. Однако общая задача кластеризации текстов, и в частности задача кластеризации потока новостей по сюжетным группам имеет ряд особенностей.

Во-первых, в отличие от традиционных задач машинного обучения, задачи обработки текстовой информации обладают высокой размерностью данных. Это связано с тем, что большинство методов кластеризации работает с данными, представленными в виде векторов в пространстве R^n [5]. Представление текстовых данных в таком пространстве обычно осуществляется с помощью процедуры сопоставления каждого признака с функцией-индикатором данного слова. Таким образом, общая размерность задачи определяется общим количеством таких признаков. Из-за этого, общая размерность пространства достигает 100 тыс. - 1 млн. измерений [3,5]. При этом вектор признаков сообщения имеет только малую часть ненулевых значений функций-индикаторов. Кластеризация текстовых данных рассмотрена во множестве источников. Наиболее популярны три вида алгоритмов: алгоритм k-средних, Scatter-Gather [5] и алгоритмы иерархической кластеризации. Однако все эти алгоритмы имеют существенные недостатки для использования в задаче анализа новостей. Все вышеуказанные алгоритмы являются оффлайновыми — они предполагают одновременную обработку всего набора данных. Но для задачи анализа новостей это не совсем верно. Поток новостей не является фиксированным множеством, а, следовательно, для учета поступивших новостей пришлось бы заново вычислять все кластеры, поскольку оффлайновые алгоритмы не подразумевают возможности инкрементальной обработки. Кроме того, алгоритм k-средних предполагает, что количество кластеров заранее известно. Более того, различные оптимизированные версии этого алгоритма предполагают относительно небольшую размерность данных. Алгоритмы же иерархической кластеризации требуют нескольких проходов по каждому экземпляру данных.

Во-вторых, основные алгоритмы кластеризации работают на заранее фиксированном множестве данных. Они не предполагают возможности добавления нового элемента без пересчета всех кластеров.

В-третьих, значительная часть алгоритмов предполагает, что число кластеров, на которые нужно разбить всю совокупность, заранее известно. Для многих задач это верно, но для задачи кластеризации новостного потока это не соблюдается.

В-четвертых, оптимизированные версии основных алгоритмов кластеризации предполагают низкую размерность данных. Их использование на массивах данных высокой размерности приводит к резкому росту времени обработки.

В ряде работ [2,3,6] для кластеризации предлагается следующий алгоритм приблизительной кластеризации. Первоначально множество

кластеров пусто. Для каждого нового сообщения выполняются следующие операции:

1. Оценивается расстояние вектора нового сообщения до центров всех кластеров.
2. Если минимальное расстояние больше некоторого наперед заданного числа, то новое сообщение помещается в отдельный кластер.
3. Если нет, то в один (или несколько ближайших).
4. Пересчитываются центры измененных кластеров.

Особенностью алгоритма является то, что решение о принадлежности какой-либо точки принимается только один раз, в этом смысле алгоритм является линейным по времени.

Кроме того, во всех представленных работах [2,3,4,6] предполагается, что кластеризация для выявления новых информационных сообщений производится только на основе текстового материала. Из-за этого не учитывается довольно большой объем мета-информации: даты публикации материала, наличие взаимных ссылок между статьями, дополнительные ссылки по теме.

Альтернативным подходом является алгоритм иерархической кластеризации [1,4]. Алгоритмически он состоит из двух шагов. Начальный шаг — каждый вектор кластеризуемого множества является отдельным кластером. На каждом следующем шаге находятся два ближайших кластера и сливаются друг с другом. Затем этот шаг повторяется до тех пор, пока не останется один кластер, объединяющий всю обучающую совокупность. Преимуществом этого алгоритма является то, что кластеры организованы иерархически — можно выбирает степень обобщения. Однако сложность этого алгоритма является довольно большой: необходимо выполнить n раз шаг кластеризации, кроме того, каждый поиск ближайших точек для кластеризации занимает как минимум линейное время, а в общем случае — квадратичное. Таким образом, общий алгоритм иерархической кластеризации обладает кубической сложностью.

Наиболее полное описание методов кластеризации новостей для поиска новых информационных сообщений представлено в [3]. В этой работе описывается система Geospace & Media Tool, которая позволяет агрегировать поток новостей. При ее создании приоритетными вопросами была обработка каждой новости только один раз, а также возможность получения результатов кластеризации в любой момент времени, работоспособность без каких-либо предположений о количестве кластеров, динамическая подстройка под материалы новостей.

В указанной работе представлена так же схема ранжирования релевантности слов при кластеризации. При составлении вектора признаков для какого-либо сообщения каждое слово взвешивается относительно метрики TF-IDF [7]. TF-IDF является методом оценки важности слова относительной всей коллекции документов. Он заключается в взвешивании частоты некоторого слова в документе с помощью обратной частоты документа – числа документов в котором это слово присутствует. Таким образом, уменьшается вес малоинформативных слов, которые встречаются в большом количестве документов – например, служебные слова или связки.

Основной алгоритм кластеризации совпадает с алгоритмами, приведенными в работах [1,3]. Однако в самой схеме кластеризации есть существенные изменения. На каждом шаге рассматриваются не все существующие кластеры, а только аппроксимация k ближайших кластеров. Для получения аппроксимации ближайших кластеров используется алгоритм LSH [7]. Он позволяет по-

лучать из фиксированного набора векторов для заданной метрики вероятные наиболее близко расположенные (с наперед заданной верхней границей ошибки) векторы к данному. Эффективные процедуры определены для следующих метрик: коэффициент Жаккарда и расстояние по косинусам. Для увеличения вероятности этот алгоритм используется несколько раз. На основе этого способа составляется таблица наиболее близких точек.

В рамках анализа новостного потока, кроме выделения новых информационных сообщений, важную роль играет разбиение уже существующих новостей по схожим темам. С помощью результатов данного этапа можно выделить как повторяющиеся новостные сообщения, не несущие смысловой нагрузки, так и уточняющие и дополняющие. Информация о том, что данные сообщения относятся к одной теме, используется на этапе частичного семантического анализа для разрешения смысловой омонимии, поскольку уже известно к какой общей теме относятся сообщения.

В отличие от задачи выделения новых информационных сообщений, эта подзадача требует меньшей оперативности решения. В тоже время, эта задача решается так же с помощью кластеризации [1,8].

Другим способом разбиения сообщений по темам является использование вероятностных методов. В частности, на таких подходах основаны алгоритмы латентного семантического индексирования [8]. Алгоритмы данного типа предполагают, что имеется фиксированное число тем. В ходе работы алгоритма составляется матрица сообщений-натермы, над которой проводится сингулярное разложение.

В качестве глубокого лингвистического анализа используются алгоритмы синтаксического анализа. Существует два наиболее используемых подхода к построению систем синтаксического анализа. Более традиционным считаются системы на основе правил [9]. Например, такой системой является лингвистический процессор ЭТАП-3 [9]. Правила для таких систем разрабатываются экспертами-лингвистами. Одним из основных недостатков такого подхода является то, что требуется огромный объем работы лингвистов для построения качественной системы.

Принципиально другим подходом к способу получения лингвистической информации является использование машинного обучения [9]. Тогда структура строится на основе закономерностей, выведенных алгоритмом из большого массива данных. Для этого используется корпус — набор текстов с размеченной синтаксической структурой. У этого подхода есть и слабые стороны: необходим большой корпус, для составления которого требуется много человеческих ресурсов; полученная модель может иметь слабую лингвистическую интерпретацию. С другой стороны, алгоритмы на основе машинного обучения работают очень быстро на этапе анализа.

Эффективными алгоритмами синтаксического анализа на основе машинного обучения является подход на основе максимальных остовных деревьев и подход на основе системы переходов [9].

Перспективными являются гибридные анализаторы, которые сочетают в себе черты систем на основе правил и машинного обучения. В частности в статье [9] представлен такой анализатор. Его основным достоинством является высокая точность построения синтаксических структур, а так же достаточно высокая скорость анализа.

Разработанный подход

В качестве алгоритма кластеризации выбран алгоритм пошаговой

обработки каждого новостного сообщения, который использован, в том числе, в работах [2,3,4]. Его схема приведена ниже.

Основная идея настоящего исследования состоит в использовании ряда дополнительных характеристик, извлекаемых из текста новостного сообщения с помощью глубокого лингвистического анализа текста.

Основными параметром, для процедуры кластеризации является модель признаков, которая используется для сопоставления рассматриваемого новостного сообщения и некоторой точки в пространстве R^n. Другим важным параметром является метрика расстояния между сопоставленными точками в пространстве R^n. Основными метриками являются косинусное расстояние, евклидово расстояние, метрики на основе множеств. В настоящей статье, как и в работах [2,3], используется косинусная метрика:

$$d(x,y) = \frac{x \cdot y}{\|x\| \cdot \|y\|}$$

В качестве базовой модели признаков автор использовал основные поверхностные признаки текстовых сообщений, широко применяемые в области информационного поиска [7]:

– множество слов, присутствующих в данном сообщении;

– их вес, рассчитанный по методике tf-idf для подколлекций документов за сутки, неделю и месяц;

– дата и время составления новостного сообщения;

– множество слов заголовка новости;

– заголовки редакторских ссылок.

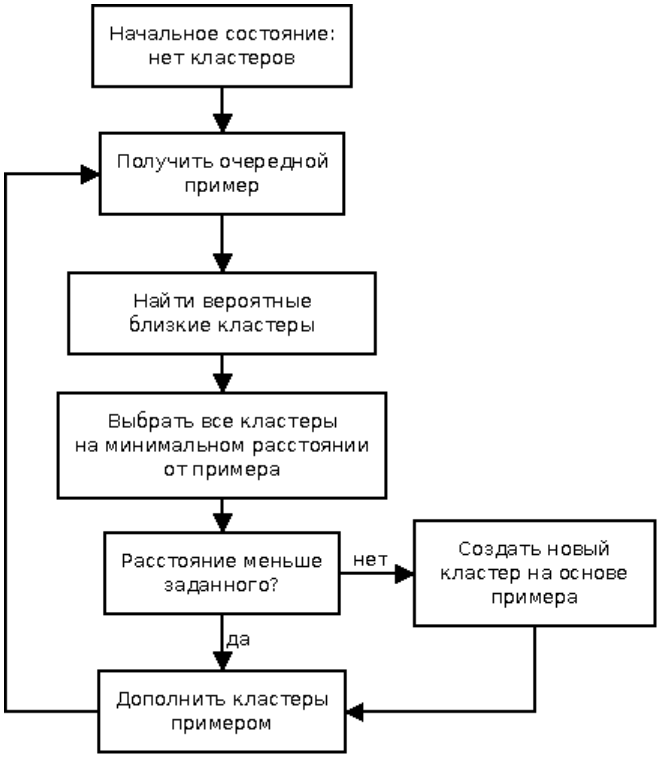

Рис. 1. Алгоритм онлайновой кластеризации

Основное дополнение базовой модели состоит из двух разработанных процедур глубокого анализа текстов на основе синтаксического анализа новостного сообщения. В качестве синтаксического анализатора используется разработанный автором эффективный гибридный алгоритм [9], сочетающий в себе богатые лингвистические знания, представленные в виде правил с высокой скоростью и робастностью подходов на основе машинного обучения. Результатом синтаксического анализа предложения является дерево синтаксических зависимостей слов, входящих в него. Например, синтаксическая структура предложения «Истребитель МиГ-29, разработанный в интересах ВВС Индии, 4 февраля совершил первый полет» представлена на рис 2.

К сожалению, полное представление синтаксического дерева достаточно сложно использовать для алгоритмов кластеризации: дерево синтаксических связей не является сбалансированным. Поэтому для кластеризации используются два представления синтаксической структуры. Первое представление базируется на усеченной синтаксической структуре, включающей в себя только несколько верхних уровней синтаксического дерева, тогда исходное предложение сократиться до следующего: «Истребитель МиГ-29 4 февраля совершил первый полет».

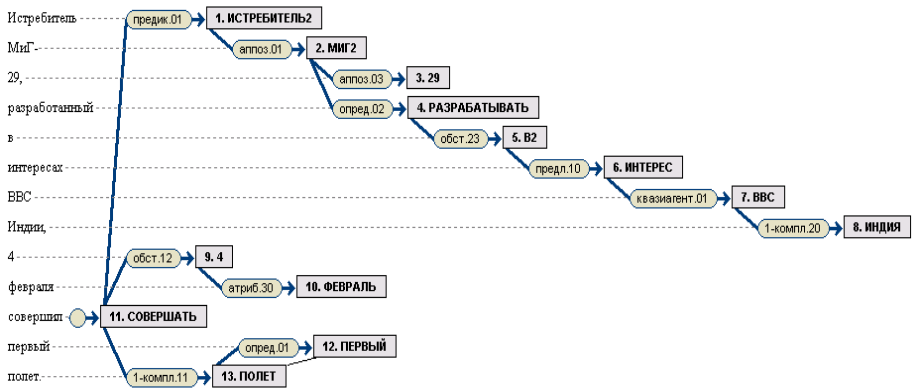

Рис. 2. Синтаксическая структура предложения «Истребитель МиГ-29, разработанный в интересах ВВС Индии, 4 февраля совершил первый полет»

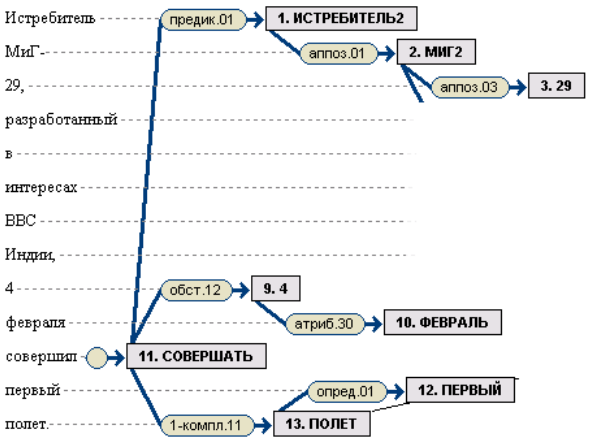

Рис. 3. Базовая синтаксическая структура, используемая для кластеризации новостей

Другим агрегированным представлением синтаксической структуры являются синтаксические группы – последовательности синтаксически связанных слов. Например, из приведенного выше предложения на основе структуры можно выделить следующие группы: «истребитель совершил», «совершил 4 февраля», «совершил первый полет», «истребитель МиГ-29», «истребитель МиГ-29, разработанный в интересах ВВС Индии», «разработанный в интересах ВВС Индии».

Таким образом, кластеризация основывается не только на поверхностных и линейных признаках текста, но так же учитывает и глубинные нелинейные, которые плохо охватываются поверхностной моделью признаков. Для представления синтаксических характеристик используются схема троек «источник-отношение-приемник». В качестве «источника» и «приемника» в тройках используются словоформы, части речи и леммы.

Кроме того, для уменьшения ресурсоемкости основного алгоритма используется процедура хеширования, представленная в работе [3]. Процедура хеширования используется вместо таблицы соответствия признаков и координат, что позволяет с одной стороны не хранить в памяти такую таблицу, а с другой — гибко управлять размерностью пространства признаков.

Эксперименты

Для оценки параметров эффективности предложенного алгоритма была проведена серия экспериментов.

Поскольку целевые параметры кластеризации заданы в основном неформально, то для оценки качественных параметров использовался корпус из 2000 новостей, который был вручную размечен на сюжетные кластеры.

Оценивались следующие параметры эффективности системы:

Среднюю точность кластеризации, определяемую как:

$$p = \frac{1}{k}\sum \frac{n_c}{n_t},$$

где n_c – число новостей с корректно определенным кластером, n_t общее число новостей в кластере.

Среднюю полноту кластеризации, определяемую как:

$$r = \frac{1}{k}\sum \frac{n_c}{n_g},$$

где n_c – число новостей с корректно определенным кластером, n_g - эталонное число новостей в кластере. Так же оценивалась производительность алгоритма – количество обработанных сообщений в секунду.

Была оценена эффективность используемых характеристик: эксперименты проводились на трех разных наборах характеристик:

– Базовый набор. Включает в себя все поверхностные текстовые признаки.

– Расширенный набор (I). Включает базовый набор признаков, а так же дополнительные признаки, полученные на основе именованных сущностей.

– Расширенный набор (II). Включает в себя признаки из расширенного набора, а так же синтаксические признаки.

Результаты экспериментов представлены в таб.1

Из таблицы видно, что использование расширенной модели признаков синтаксическими характеристиками значительно улучшает качество кластеризации, как в оценках точности, так и полноты. Кроме того, использование дополнительных признаков синтаксических групп дополнительно улучшает качество кластеризации.

Таблица 1

Результаты экспериментов

Наборы признаков	Точность, p	Полнота, r	Скорость, сообщений/сек
Базовый	0,78	0,83	4
Расширенный (I)	0,85	0,87	1
Расширенный(II)	0,91	0,89	0,3

В тоже время, производительность всех моделей достаточна для обработки не менее 20 тыс. сообщений в сутки. Даже самый эффективный по качеству, но медленный по скорости алгоритм – на основе расширенной (II) модели признаков, позволяет обрабатывать около 3-х сообщений в секунду, что соответствует более 28 тыс. сообщений в сутки.

Выводы

В настоящей статье представлен алгоритм анализа новостных сообщений, на основе расширенной модели признаков для кластеризации, который сочетает в себе высокую производительность традиционных средств поверхностного анализа текстов с эффективностью глубоких методов анализа.

Экспериментальные исследования показали, что предложенный алгоритм обладает достаточной производительностью для ежедневной обработки более 20 тыс. сообщений.

Алгоритм спроектирован таким образом, что при необходимости система признаков может быть расширена, что позволяет адаптировать представленный алгоритм для решения других задач.

Литература

1. Shen D., Yang Q., Sun J., Chen Z. Thread Detection in Dynamic Text Message Systems, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 06, ACM Press, pp. 35-42.

2. Beringer J., Hullermeier E. Online Clustering of Parallel Data Streams, Data & Knowledge Engineering 58(2006), Elsiever,180-204.

3. Moerchen F., Brinker K., Neubauer C., Any-Time Clustering of High Frequency News Streams, The Thirteenth ACM SIGKDD Int'l. Conference on Knowledge Discovery and Data Mining: Data Mining Case Studies Workshop *(DMCS*), ACM Press, 23-31.

4. McKeown K., Barzilay R., et al. Tracking and Summarizing News on a Daily Basis with Columbia’s Newsblaster, Proceedings of HLT, Morgan-Kaufman, 280-285.

5. Feldman R., Sanger J. The Text Mining Handbook, Cambridge University Press, 2007.

6. Costa G., Mango G., Ortale R., An incremental clustering scheme for data de-duplication, Data Mining and Knowledge Discovery Volume 20 Issue 1, January 2010, Springer, 152-187.

7. Казенников А.О, Трифонов Н.И., Тюрин А.Г., Исследование методов компьютерной лингвистики для задач повышения эффективности информационного поиска. Информатизация образования и науки № 3(7) 2010, 10-20.

8. Yao L., Mimno D., McCallum A. Efficient Methods for Topic Model Inference on Streaming Document Collections, Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, 947-956.

9. Казенников А.О., Куракин Д.В., Трифонов Н.И. Гибридный алгоритм синтаксического разбора для системы анализа новостных потоков, Информатизация образования и науки № 1(13) 2012, 90-97.