OSTIS-2015
(Open Semantic Technologies for Intelligent Systems)

УДК 004.912

МЕТОДЫ ТЕМАТИЧЕСКОЙ КЛАСТЕРИЗАЦИИ ПРИМЕНИТЕЛЬНО К АНАЛИЗУ НОВОСТНЫХ СТАТЕЙ

Солошенко А.Н. [*], Орлова Ю.А. [*], Заболеева-Зотова А.В. [**]

[*] *Волгоградский государственный технический университет, г. Волгоград, Россия*

nastyasolan@gmail.com

yulia.orlova@gmail.com

[**] *Российский фонд фундаментальных исследований, г. Москва, Россия*

zabzot@gmail.com

Данная работа посвящена проблеме выделения сюжетов и тем из потока новостных сообщений. Кратко рассмотрены алгоритмы кластеризации, такие как алгоритмы k-средних, минимальное покрывающее дерево и др. Проанализированы результаты их работы на новостных текстах. В работе представлена методика комплексного анализа новостного текста, основанная на комбинации статистических алгоритмов извлечения ключевых слов и алгоритмов формирования семантической связности блоков текста. Особое внимание уделено особенностям структуры новостного текста.

Ключевые слова: тематическая кластеризация; алгоритмы кластеризации; новостные статьи; представление документов.

Введение

В наши дни заметно усилилась проблема информационной перегрузки. В начале XXI века американская исследовательская служба Cyveillance сообщила о том, что количество страниц в Internet превысило 4 млрд, и с каждым днем увеличивается на 7 млн. В частности, темпы роста аудитории онлайновых новостных ресурсов практически вдвое превышают темпы роста общей численности пользователей интернета, составляя к сегодняшнему дню 43,2% российских интернет-пользователей (исследование сотрудников Nielsen//NetRatings).

«Сырые» неструктурированные данные составляют большую часть информации, с которой имеют дело пользователи, поэтому многие организации и частные лица заинтересованы в эффективных технологиях автоматизированного анализа информации, представленной на естественном языке. При этом автоматическая кластеризация, т.е. выявление групп семантически похожих текстов, является одной из приоритетных задач, решаемых информационными системами.

1. Обзор существующих систем анализа текстов, обеспечивающих возможность кластеризации документов

На международном рынке представлено множество программных продуктов, предоставляющих функцию кластеризации текстовых документов.

Среди отечественных стоит выделить системы TextAnalyst, Galaktika-ZOOM, из зарубежных – мощный инструмент анализа текстов IBM Text Miner. В возможности TextAnalyst входит создание семантической сети большого текста, подготовка аннотации, автоматическая классификация и кластеризация текстов. IBM Text Miner содержит утилиты классификации, кластеризации, поиска ключевых слов и составления аннотации текстов. Однако на обработку новостных статей программы не направлены.

Российская система Яндекс Новости позволяет автоматически группировать данные в новостные сюжеты и составлять аннотации статей на основе кластера новостных документов. Сервис InfoStream, обеспечивает доступ к оперативной информации в поисковом режиме с учетом семантической близости документов. Мобильный агрегатор

новостей Summly, купленный в марте 2013 компанией года Yahoo!, также осуществляет группировку новостных статей по темам. Однако приложение абсолютно неприменимо для обработки текстов на русском языке.

Таким образом, существующие программные системы полностью не решают поставленную проблему. Основная идея заключается в разработке программного продукта для анализа новостных текстов, сочетающего кластеризацию новостных статей с комплексным анализом текста.

2. Особенности структуры новостного текста в решении задачи кластеризации

Проанализировав ряд статей, представленных на новостных сайтах, относящихся к топ-30 новостных порталов Рунета), можно построить обобщенную структуру текста новости (рисунок 1).

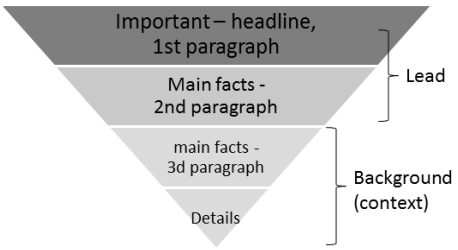

Рисунок 1 – Структура новостного текста

В ее основу заложен принцип «перевернутой пирамиды», который требует размещение основной информации в самом начале материала и последующее ее раскрытие далее по тексту в деталях.

- Заголовок новости отражает ее тему и содержит не более 10 слов (около 80 символов). Так, для примера, в Яндексе отображается не более 15-ти слов в тайтле, Google показывает до 70 слов;
- Основные факты, касающиеся события, отражены в 1-2 абзацах, и составляют так называемый лид текста (освещает главную тему);
- 3-й и последующие абзацы составляют бэкграунд новости (контекст). Как правило, здесь раскрываются детали происходящего, дается информация, напрямую касающаяся новости.

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

3. Кластеризация новостного потока

Прежде чем перейти к описанию алгоритмов кластеризации, определим основные понятия исследуемой области.

Кластеризация - разбиение множества документов на кластеры - подмножества, параметры которых заранее неизвестны. Количество кластеров может быть произвольным или фиксированным. Основные группы алгоритмов кластеризации – это иерархические и плоские, четкие и нечеткие алгоритмы.

Иерархические алгоритмы строят не одно разбиение выборки на непересекающиеся кластеры, а систему вложенных разбиений. То есть на выходе мы получаем дерево кластеров, корнем которого является вся выборка, а листьями — наиболее мелкие кластера. Плоские алгоритмы строят одно разбиение объектов на кластеры.

Четкие (непересекающиеся) алгоритмы каждому объекту выборки ставят в соответствие номер кластера, т.е. каждый объект принадлежит только одному кластеру. Нечеткие (пересекающиеся) алгоритмы каждому объекту ставят в соответствие набор вещественных значений, показывающих степень (вероятность) отношения объекта к кластерам.

3.1. Формализация задачи кластеризации документов

Рассмотрим задачу кластеризации документов более формально.

Пусть X – множество объектов, Y – множество номеров кластеров. Задана функция расстояния между объектами $\rho(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, ..., x_m\} \subset X$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике ρ, а объекты разных кластеров отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_i [Большакова и др., 2011].

Алгоритм кластеризации – это функция $a: X \to Y$, которая любому объекту $x \in X$ ставит в соответствие номер кластера $y \in Y$.

3.2. Методы кластеризации

Далее рассмотрим лишь несколько алгоритмов кластеризации, оптимальных, с нашей точки зрения, для обработки новостного потока.

3.2.1. *Метод агломеративной кластеризации*

Иерархические алгоритмы разделяются на два вида: агломеративные (восходящие) и дивизимные (нисходящие). Первые строят кластеры снизу вверх, начиная с множества кластеров, содержащих по одному одиночному документу коллекции, затем последовательно объединяют пары кластеров, пока не получат один кластер, содержащий все документы коллекции. Вторые разбивают кластеры сверху вниз, начиная с одного кластера, которому принадлежат все документы коллекции, затем этот кластер делится на два и так рекурсивно до тех пор,

пока каждый документ не окажется в своём отдельном кластере.

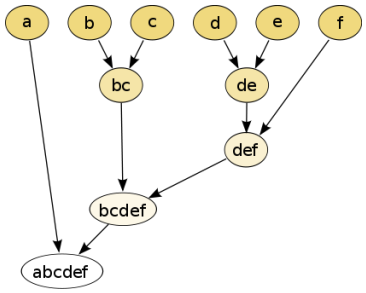

Рисунок 2 – Дендограмма (граф без циклов, построенный по матрице мер близости)

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 двух кластеров - сходство между их наиболее похожими документами;
- полная связь (максимальное расстояние, или минимальное сходство): сходство двух кластеров есть сходство между их наиболее непохожими документами;
- групповое усреднение (усреднение всех показателей сходства): сходство двух кластеров есть среднее сходство всех пар документов, включая пары документов из одного кластера, исключая близость документа самому себе;
- центроидный метод;
- метод Уорда [Bandyopadhyay et al., 2013].

3.2.2. *Алгоритм k-means (k-средних)*

При заранее известном числе кластеров k алгоритм начинает с некоторого начального разбиения документов и уточняет его, оптимизируя целевую функцию – среднеквадратичную ошибку кластеризации как среднеквадратичное расстояние между документами и центрами их кластеров:

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

где μ_j - центроид кластера C_j.

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

3.2.3. *Нечеткие алгоритмы классификации – FCM*

Был предложен как решение проблемы мягкой кластеризации, то есть присвоения каждого документа более чем одному кластеру. Как и его чёткий вариант, k-means, данный алгоритм, начиная с некоторого начального разбиения данных, итеративно минимизирует целевую функцию, которой является следующее выражение:

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

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

3.2.4. *Алгоритм минимального покрывающего дерева*

Алгоритм минимального покрывающего дерева сначала строит на графе минимальное покрывающее дерево, а затем последовательно удаляет ребра с наибольшим весом. На рисунке изображено минимальное покрывающее дерево, полученное для девяти объектов.

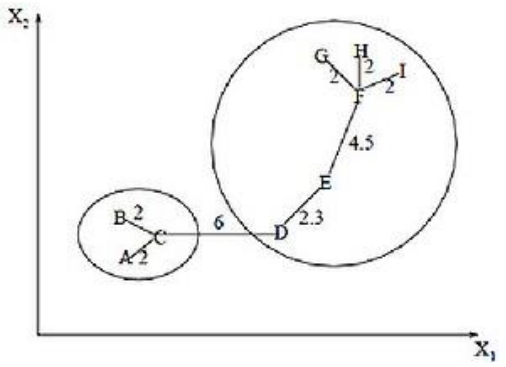

Рисунок 3 – Иллюстрация алгоритма MST

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, равную 4,5 единицам [Pera et al., 2012].

3.2.5. *Применение нейронных сетей*

Алгоритм самоорганизующихся карт (SOM, Self Organizing Maps) был предложены как решение проблемы 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.

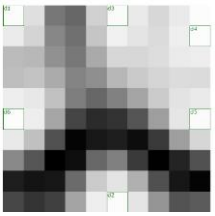

Рисунок 4 – Самоорганизующаяся карта из 6 документов

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 из входного массива $d_i \in D$;
- находят нейрон-победитель $m_c \in M$, ближайший к документу d_i;
- корректируют веса нейрона-победителя, его соседей: $m_i(t+1) = m_i(t) + h_{ci}(t)[d_i - m_i(t)]$.

3.3. Сравнение методов кластеризации применительно к обработке новостных текстов

При выборе оптимального алгоритма кластеризации новостного потока 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.

В соответствии с вышесказанным, проведем сравнение рассмотренных нами алгоритмов кластеризации новостного потока.

- Агломеративный: иерархический, четкий, не требующий задания числа кластеров;
- K-means: плоский, четкий, требующий задания числа кластеров;
- FCM: плоский, нечеткий, требующий задания числа кластеров;
- MST: плоский, четкий, не требующий задания числа кластеров.

Для работы с новостными текстами желательно, чтобы алгоритм был плоским, нечетким, инкрементальным. Поэтому более перспективными методами для данной задачи видится применение алгоритма FCM или нейронных сетей.

Далее рассмотрим аспекты семантического анализа новостного текста.

4. Анализа новостного текста после кластеризации новостного потока

Новостные агрегаторы представляют собой сложные программно-аппаратные комплексы, решающие широкий круг задач: кластеризацию, ранжирование документов внутри кластера, обзорное реферирование, выявление ключевых лиц, тематическую классификацию и поиск по новостям и т.д. Семантический анализ является основной составляющей вышеперечисленных задач, его можно разбить на несколько этапов, рассмотренных далее.

4.1. Предварительная обработка текста новости

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 же метод основан на подключении словаря, тезауруса, дает максимально полный анализ словоформы.

Поэтому для данного блока целесообразно использовать морфологические библиотеки, например, Lemmatizer, FreeLing, NLTK, MCR, tokenizer [Михайлов и др., 2009].

4.2. Задача синтаксического анализа

Синтаксический анализ рассматривается как задача построения дерева зависимостей предложения. В ходе его проведения, происходит выделение синтаксических конструкций, определение связности и подчинения фрагментов [Grune, 2012].

Приведем обзор основных инструментов синтаксического анализа, которые возможно использовать в своем проекте:

Таблица 1 – Модули синтаксического анализа

Название	Методы	Языки
AOT	грамматика HPSG	русский, английский, немецкий
MaltParser	машинное обучение	русский, английский
Link Grammar Parser	грамматика связей	русский, английский
NLTK	машинное обучение	английский
Solarix	правила	русский, английский

4.3. Поиск ключевых слов, построение аннотации

Для новостных текстов уже существуют программные модули для извлечения ключевых сущностей. К таковым относится PullEnti, полностью написанный на C#.NET.

Поэтому был разработан алгоритм поиска ключевых слов [Солошенко и др., 2014a], сочетающий выделение именованных сущностей из текста новости (на основе результатов морфоанализа и подключаемого модуля PullEnti), подсчет веса слова с учетом частоты его встречаемости (рисунок 5).

Рисунок 5 – Алгоритм поиска ключевых фраз

Пороговое значение для признания слова ключевым – это значение относительной частоты встречаемости слова-кандидата в ключевые слова, с индексом, равным (0,2×количество сущностей). Такое значение вычислено экспериментально на выборке из 100 текстов.

В итоге 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 в виде ориентированного графа 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. [Дмитриев и др., 2013; Машечкин и др., 2011]. Таким образом, содержание новости можно представить наглядно в виде ключевых понятий и связей между ними, либо в виде так называемой mind map или интеллектуальной карты (рисунок 6).

Рисунок 6 – Интеллектуальная карта текста новости

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

$$Ws = N(kw) \cdot Rf(kw) \cdot \text{Вес параграфа} \cdot k \qquad (3)$$

Ws здесь – вес предложения; N(kw) – количество вхождений ключевого слова в предложение; Rf(kw) – относительная частота ключевого слова; вес параграфа – относительный вес параграфа в тексте, равен 0.35 для первого параграфа (лид), 0.2 для второго, 0.1 для остальных (контекст); k – коэффициент значимости предложения внутри параграфа. Для первого предложения в абзаце равен 1, для остальных – 0.8 [Солошенко и др., 2014].

В итоговую аннотацию включаются предложения с наибольшим весом, в зависимости от заданного коэффициента сжатия.

5. Основные результаты

Для проверки теоретических рассуждений была реализована система, основанная на принципе многокомпонентности программного обеспечения [Заболеева-Зотова и др., 2013].

Был также проведен эксперимент [Солошенко и др., 2014b], цель которого – доказать, что за счет автоматизации обработки статей новостного потока повысилась эффективность обработки новостных интернет-статей, то есть снизилось время на обработку и повысилось качество получаемого результата.

Были получены следующие результаты:

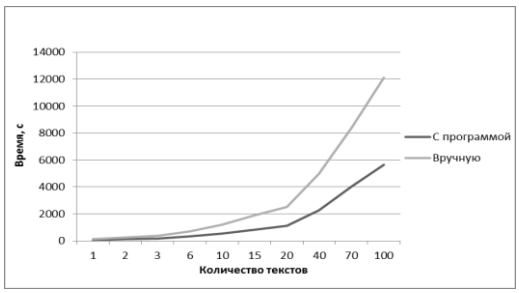

Рисунок 7 – Зависимость времени обработки от количества текстов

Время обработки уменьшилось как минимум в два раза. При этом время с помощью программы учитывает не только непосредственно время

составления аннотации, но и время, необходимое для окончательной корректировки текстов.

Качество аннотации можно оценить по следующим критериям: сохранение ключевых фактов, связность новостной статьи, сохранение синтаксической структуры текста после удаления незначащих частей. Каждый из названных критериев оценивался экспертами по шкале от 0 до 10 баллов, затем для оценки качества полученной аннотации находилось среднее арифметическое трех показателей для каждого текста.

В результате качество обработанных новостей остается на том же уровне, как и при анализе текста человеком.

Заключение

Итак, анализ новостных текстов включает в себя задачу кластеризации и последующую комплексную обработку статей. Был проделан анализ новостного потока, выделены особенности документов. Проанализированы методы кластеризации для новостных статей, предложено наиболее подходящее для нашей задачи решение - алгоритм FCM или нейронные сети. Однако необходимо заметить, что оно не является единственно верным. Реализована часть системы онлайн агрегации новостей из интернет-источников и проведены исследования эффективности ее работы.

Работа частично поддержана Российским фондом фундаментальных исследований (проекты 13-07-00351, 13-07-97042, 14-07-97017, 15-07-07519).

Библиографический список

[Большакова и др., 2011] Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Е.И. Большакова [и др.]. – М.: МИЭМ, 2011. – 272 с.

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

[Дмитриев и др., 2013] Automatic identification of time and space categories in the natural language text / Дмитриев А.С., Заболеева-Зотова А.В., Орлова Ю.А., Розалиев В.Л. // Applied Computing 2013 : proceedings of the IADIS International Conference (Fort Worth, Texas, USA, October 23-25, 2013) / IADIS (International Association for Development of the Information Society), UNT (University of North Texas). – [Fort Worth (Texas, USA)], 2013. – P. 187-190.

[Заболеева-Зотова и др., 2013] Formalization of initial stage of designing multi-component software / Заболеева-Зотова А.В., Орлова Ю.А., Розалиев В.Л., Фоменков С.А., Петровский А.Б. // Multi Conference on Computer Science and Information Systems 2013 (Prague, Czech Republic, July 23-26, 2013) : Proceedings of the IADIS International Conference Intelligent Systems and Agents 2013 / IADIS (International Association for Development of the Information Society). – [Prague], 2013. – P. 107-111.

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

[Михайлов и др., 2009] Михайлов Д. В. Морфология и синтаксис в задаче семантической кластеризации / Д. В. Михайлов, Г. М. Емельянов // Математические методы распознавания образов (ММРО-14), Владимирская область, Суздаль, 21-26 сентября 2009 г. – С. 1-4.

[Солошенко и др., 2014a] Thematic Clustering Methods Applied to News Texts Analysis / Солошенко А.Н., Орлова Ю.А., Розалиев В.Л., Заболеева-Зотова А.В. // Knowledge-Based Software Engineering : Proceedings of 11th Joint Conference, JCKBSE 2014 (Volgograd, Russia, September 17-20, 2014) / ed. by A. Kravets, M. Shcherbakov, M. Kultsova, Tadashi Iijima ; Volgograd State Technical University [etc.]. – [Б/м] : Springer International Publishing, 2014. – P. 294-310. – (Series: Communications in Computer and Information Science ; Vol. 466).

[Солошенко и др., 2014b] Автоматизация семантического анализа новостных Интернет-текстов. / Солошенко А.Н., Розалиев В.Л., Орлова, Ю.А. // Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2014 : матер. IV междунар. науч.-техн. конф. (Минск, 20-22 февр. 2014 г.). Белорус. гос. ун-т информатики и радиоэлектроники, Администрация Парка высоких технологий. Минск, с. 435-438.

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

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

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

[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. In: J. Intell. Inf. Syst, vol 39, pp. 513–534

THEMATIC CLUSTERING METHODS APPLIED TO NEWS ARTICLES ANALYSIS

Soloshenko A.N., Orlova Yu.A.*,

Zaboleeva-Zotova A.V. **

**Volgograd State Technical University, Volgograd, Russia*

nastyasolan@gmail.com

yulia.orlova@gmail.com

*** Russian Foundation for Basic Research, Moscow, Russian Federation*

zabzot@gmail.com

This paper is devoted to a problem of partition documents from the news flow into groups, where each group contains documents that are similar to each other. The existing clustering algorithms such as k-means, minimum spanning tree and etc. are considered and analyzed. It is shown which of these algorithms give the best results working with news texts. This paper also presents a methodic of comprehensive news texts analysis based on a combination of statistical algorithms for keywords extracting and algorithms forming the semantic coherence of text blocks. Particular attention is paid to the structural features of the news texts.

Keywords: thematic clustering, clustering algorithms, news articles, document representation.