# Формирование выпуска новостей на основе автоматического анализа новостных сообщений

Зевайкин А.Н.

Корнеев В.В.

ФГУП НИИ «КВАНТ»
shuraz@yandex.ru

## Аннотация

В настоящее время объем потоков новостных сообщений возрастает, а допустимое для их обработки время сокращается. Это обстоятельство увеличивает интерес к системам, которые позволяют автоматизировать обработку и анализ потоков новостных сообщений.

В работе предлагается технология построения выпусков новостей на основе автоматического анализа новостных сообщений, включающей в себя методы автоматического выделения, ранжирования и реферирования новостных сюжетов.

Разбиение текстовых сообщений на фрагменты позволяет улучшить качество кластерного анализа, качество аннотирования текстовых сообщений и новостных сюжетов.

Вводится единый критерий «актуальность» для ранжирования новостных сюжетов, включающий в себя такие критерии, как время, источник сообщения, количество сообщений в группе, количество чтений пользователями.

Применение предлагаемой технологии позволит улучшить качество обработки потоков новостных сообщений.

# 1. Введение и обзор ключевых работ по исследуемой тематике

В современном мире происходит огромное количество событий, поэтому важная задача СМИ – выбор новостных сюжетов для освещения и определение порядка их появления в новостных выпусках. Выбор новостей в «бумажных» СМИ основывается на практике и опыте редактора. Кроме того, в журналистике существует набор критериев [GR65], по которым отбираются новости, заслуживающие освещения в печати. Чем большему количеству критериев удовлетворяет новость, тем более вероятно, что она будет освещена в СМИ и будет помещена на первое место в выпуске новостей. К таким критериям относятся, например, неожиданность произошедшего события или значимость персоны, о которой идет речь. Стоит отметить, что критерии выбора новостей субъективны, и у каждого СМИ они свои. Значит, конечный результат выбора и ранжирования новостей для каждого СМИ будет разный, то есть каждое СМИ строит свою собственную синтетическую картину мира.

Кроме того, существуют две противоположные тенденции: количество новостных сообщений растет, а допустимое для их обработки время сокращается. Это приводит к тому, что уже невозможно обработать весь поток поступающих сообщений в условиях, когда обработка поступающих сообщений производится вручную. Вследствие этого, создание автоматизированной системы обработки новостей является актуальной проблемой, так как позволяет формализовать и ускорить процесс обработки новостей. Данная система может решать следующие задачи:

- сбор и предварительная обработка новостных сообщений, в частности, преобразование новостных сообщений в формальную модель для того, чтобы их можно было обрабатывать с помощью математических алгоритмов;
- фильтрация новостных сообщений по заданным тематикам, например, селекция новостей о политических событиях;
- поиск по запросам пользователей;
- объединение схожих по содержанию новостных сообщений в группы для упрощения работы системы, так как необязательно обрабатывать десятки и сотни сообщений на одну тему, если можно представить их в виде одной группы;
- ранжирование групп новостных сообщений;
- аннотирование новостных сообщений и групп новостных сообщений.

В настоящее время ведутся исследования по данной тематике в рамках проекта Topic Detection and Tracking [TDT04]. В рамках данного проекта исследуются следующие задачи: выделение новостных сообщений, отслеживание заданного сюжета, разбиение на сюжеты, определение первого сообщения в данном сюжете, обнаружение связанных сюжетов.

В качестве примеров существующих систем автоматического формирования выпусков новостей можно привести следующие системы: Яндекс Новости [Янд03], Google News [Goo03], Новотека [Нов04]. Данные системы имеют общий подход к обработке новостей – минимум человеческого вмешательства. Стоит отметить, что любая автоматизированная система не способна однозначно выделить новостные сюжеты, она может лишь описать их группами сообщений, сами новостные сюжеты складывается в голове у пользователя системы после ознакомления с множеством сообщений.

Рассмотрим подробнее технологию автоматической обработки новостей, для чего сформулируем основные определения.

Новостное сообщение – опубликованное сообщение, обладающее следующими признаками: дата, время опубликования (может отличаться от даты/времени произошедшего события) и источник, название СМИ. Новостной сюжет – совокупность сведений (новостных сообщений) о некоторых сущностях и явлениях (о людях, вещах, отношениях, действиях, процессах, свойствах, и т.д.), а также о связанных сущностях и явлениях. Делается допущение, что группа схожих по содержанию и близких по времени новостных сообщений соответствует новостному сюжету. Таким образом, выделение новостных сюжетов сводится к разбиению сообщений на группы. Выпуск новостей – ранжированный по некоторому признаку список новостных сюжетов.

Процесс формирования выпуска новостей разбивается на следующие части:

- объединение новостных сообщений в новостные сюжеты;
- ранжирование новостных сюжетов для получения выпуска новостей;
- аннотирование новостных сюжетов и новостных сообщений.

## 2. Идея исследования

Основной гипотезой работы является то, что разбиение новостных сообщений на фрагменты позволит улучшить качество обработки новостей.

Подавляющее большинство новостных сообщений и прочих текстовых сообщений разбивается на фрагменты авторами, желающими разделить смысловые блоки текста. Данное разбиение позволяет выделить отдельные мысли в тексте и использовать это для улучшения результатов последующих методов. В качестве фрагментов могут выступать заголовки, абзацы, множество абзацев и прочие части сообщений.

Наиболее часто встречающийся способ фрагментации новостных сообщений – фрагментация на абзацы. Абзац (нем. Absatz) – отрезок письменного или печатного текста от одной красной строки до другой, обычно заключающий в себе сверхфразовое единство или его часть [АЩ99]. Сверхфразовое единство - сложное синтаксическое целое, отрезок речи в форме последовательности двух и более самостоятельных предложений, объединенных общностью темы в смысловые блоки. Может совпадать с абзацем, быть больше или меньше абзаца. Абзац является средством логико-композиционного членения текста.

Типичный пример новостного сообщения имеет вид:

*Самолет Ил-62 МЧС РФ с российскими специалистами вылетел в 14.20 мск из Багдада в Москву, сообщил РИА «Новости» замначальника Управления информации МЧС Виктор Бельцов.*

*Накануне самолет МЧС России Ил-76 забрал из Багдада первую группу сотрудников «Интерэнергосервиса»: 85 россиян, четверых граждан Украины и одного гражданина Белоруссии. Среди прибывших из Ирака был один ребенок. Также в Москву вернулись россияне, получившие ранения в результате нападения. Этим же бортом были доставлены тела двух погибших.*

*Решение об эвакуации было принято руководством «Интерэнергосервиса» после вооруженного нападения в среду, 26 мая, в Багдаде на автобус, перевозивший сотрудников компании. В результате обстрела погибли двое россиян, еще восемь получили ранения.*

В статье отчетливо прослеживается, что автор хотел разбиением на абзацы выделить 2 смысловых блока: описание текущего новостного сюжета - «Эвакуация российских специалистов из Ирака» и более раннего новостного сюжета - «Вооруженное нападение на автобус». Если анализировать данное сообщение без

разбиения на абзацы, то он будет отнесен к одному новостному сюжету, допустим, «влияние обострения конфликта в Ираке на российские компании». Разбиение сообщения на абзацы, позволит разделить данный новостной сюжет на более мелкие новостные сюжеты, в соответствии с тем, как разбил сообщение автор.

Разбиение на тематические фрагменты можно производить двумя способами:

- разделение с помощью структуры текста, например использование HTML тегов <p> и <br>,
- разделение с помощью анализа похожести слов во фрагментах [BG01].

Бывают случаи, когда в сообщении последовательно идут похожие по содержанию фрагменты, тогда нет смысла разделять их и необходимо рассматривать их как один фрагмент. Это можно сделать, например, с помощью кластерного анализа фрагментов. Если два фрагмента отнесены к одному кластеру и находятся в сообщении друг за другом, то их следует рассматривать как один фрагмент.

Таким образом, сообщение можно представить в виде множества своих фрагментов:

$$D_i \overset{def}{=} (f_{i_1},...,f_{i_{k_i}}), \quad i \in N, \tag{1}$$

где $N$ - общее число сообщений, $k_i$ – число фрагментов в i-м сообщении.

В дальнейшем фрагменты сообщений будут рассматриваться как независимые (от сообщений) единицы текста, поэтому для упрощения формул обозначим произвольный фрагмент сообщения как:

$$f_i, \quad i \in \overline{1, N_f}, \quad N_f = \sum_l k_l,$$

где $N_f$ – общее число фрагментов во всех сообщениях.

## 3. Описание методов, алгоритмов и экспериментов

### 3.1. Выделение новостных сюжетов

Выше было сделано допущение, что новостной сюжет соответствует группе схожих по содержанию сообщений. Вследствие этого, выделение новостных сюжетов сводится к выделению однородных групп сообщений. Для этого используется

кластерный анализ. Причем, в качестве объектов кластерного анализа рассматриваются не сообщения, а их фрагменты.

В качестве формальной модели текста предлагается использовать пространственно-векторную модель. Текст фрагмента сообщения рассматривается как неупорядоченное множество независимых информационных признаков (терминов, составных терминов, N-грамм). Согласно данной модели, вектор, представляющий текст фрагмента сообщения, выглядит следующим образом:

$$f_i \overset{def}{=} (w_{i1},...,w_{ij},...,w_{i n_i}), \quad j \in n_i,$$

где $n_i$ – число информационных признаков во фрагменте $f_i$.

Представление фрагментов текста в виде векторов позволяет применять к фрагментам математические методы, в частности кластерный анализ. В результате кластерного анализа множество фрагментов $f_i$ разбивается на $k$ ($k \le N_f$) однородных (в смысле некоторой меры близости) групп. Обозначим полученное множество групп фрагментов как $C = \{C_1,...,C_k\}$. Внутри групп текстовые фрагменты схожи по содержанию. Данные группы соответствуют определенным новостным сюжетам.

Однако пользователю необходимо предоставить новостной сюжет, описанный с помощью множества сообщений, а не фрагментов. Значит нужно перейти от групп фрагментов к группам сообщений. Возвратимся к представлению сообщений с помощью фрагментов, описанному формулой (1). Обозначим полученные группы сообщений $C' = \{C'_1,...,C'_k\}$, которые взаимно однозначно соответствуют группам фрагментов С с теми же номерами.

Сообщение $D_i = (f_{i_1},...,f_{i_{k_i}})$ относится к группе $C'_j$, если в соответствующую группу отнесено число фрагментов данного сообщения, большее некоторого относительного порога $n^c$, в процентах числа фрагментов данного сообщения:

$$\{f_{i_1^c},...,f_{i_k^c}\} \in D_i, \ \{f_{i_1^c},...,f_{i_k^c}\} \in C_j,$$

$$\frac{\left|i_1^c,...,i_k^c\right|}{k_j} * 100\% > n^c \Rightarrow D_i \in C'_j$$

где $i_1^c,\ldots,i_k^c$ - номера фрагментов, принадлежащих группе $C_j$.

Если установить высокий порог, то сообщение будут попадать только в одну группу, в которую попадает большинство его фрагментов. Такой порог имеет смысл, если сообщения, в большинстве своем, на одну тему, тогда уменьшается доля выделенных «шумовых» новостных сюжетов, то есть сюжетов на разные, несвязанные темы. Недостаток данного способа состоит в потере информации, содержащейся в других, не главных темах сообщений.

Если установить нулевой порог, то сообщение будет попадать во все группы, в которые попадают его фрагменты. Такой подход имеет смысл, если сообщения содержат множество тем, например, анонсы новостей. В данном случае все второстепенные темы раскрываются, но появляются шумовые кластеры, за счет фрагментов типа «по материалам агентства ИТАР-ТАСС».

После кластерного анализа фрагментов, каждая группа содержала уникальное множество фрагментов, то есть у групп не было общих элементов: $C_i \cap C_j = \varnothing, \quad \forall i,j \in \overline{1,k}$. После получения групп сообщений, подобное утверждение для групп сообщений стало неверно из-за того, что сообщение может попасть в разные группы. Появилась следующая проблема – появляются группы сообщений с похожим множеством набором сообщений. Решение – объединять группы сообщений, которые содержат некоторый процент одинаковых элементов. Алгоритм объединения выглядит следующим образом:

1. Строим матрицу $R^{C'}$ похожести групп $C'$. Под похожестью групп будем понимать отношение мощности пересечений элементов групп к минимальной мощности групп:

$$R^{C'} = \left(r_{ij}^{C'}\right), \quad r_{ij}^{C'} = \frac{\left|C'_i \cap C'_j\right|}{\min\left(\left|C'_i\right|,\left|C'_j\right|\right)}, \quad i,j \in \overline{1,k}.$$

2. Находим координаты $(i_{\max}, j_{\max})$ и значение $r_{\max}^{C'}$ максимального элемента матрицы $R^{C'}$.

3. Проверяем, больше ли $r_{\max}^{C'}$ некоторого порога $r_{likeness}^{C'}$. Если да, то в матрице нет кандидатов на объединение, конец алгоритма.

4. Объединяем группы сообщений $C'_{i_{\max}}$ и $C'_{j_{\max}}$:

$$C'_{i_{\max}} = C'_{i_{\max}} \cup C'_{j_{\max}}, \quad C'_{j_{\max}} = \varnothing.$$

5. Пересчитываем элементы строк и столбцов матрицы $R^{C'}$ с номерами $i_{\max}, j_{\max}$.

6. Переход к шагу 2.

Недостаток данного алгоритма – при малом пороге остановки $r_{likeness}^{C'}$ группы могут объединиться в одну.

Указанный способ построения и объединения похожих групп сообщений согласуется с определением новостного сюжета в той части, что новостной сюжет может соответствовать нескольким событиям.

## 3.2. Ранжирование новостных сюжетов

Выделенные в результате кластерного анализа группы сообщений (сюжеты) – продвижение по сравнению с исходным множеством сообщений. Но сюжеты остаются между собой неупорядоченными, их необходимо ранжировать для формирования выпуска новостей. Ранжировать сюжеты можно по разным признакам: по времени сообщений сюжета, по количеству сообщений, по важности и другим признакам.

Очевидно, что следует ввести некоторый составной обобщающий признак, по которому будет вестись ранжирование. В англоязычной литературе [Flo05, Wil97] используется понятие «newsworthy», что переводится как «достойный освещения в печати, интересный, важный». Назовем данный признак – актуальностью. Актуальность - важность, значительность чего-либо в настоящее время, современность, злободневность [БСЭ72]. В соответствии с данным определением, будем считать, что новостное сообщение является актуальным, если оно обладает следующими признаками:

1. Новое по времени, то есть появилось недавно.

2. Важное, то есть имеющее большое значение, заслуживающее особого внимания.

По аналогии с актуальностью сообщения, введем определение актуальности новостного сюжета. Новостной сюжет актуален, если он содержит актуальные сообщения, то есть новые по времени и важные.

Понятие актуальности является комплексным показателем, отражающим интерес пользователя к данной предметной области. По данному показателю предлагается вести ранжирование новостных сюжетов.

В данной работе последовательно рассматриваются отдельные признаки актуальности и способы сведения их воедино.

**Ранжирование по времени**

Многие новости со временем стареют, то есть утрачивают практическую полезность для потребителя вследствие различных причин, например, изменения описанного объекта или явления. Степень старения неодинакова для разных новостей. Но в большинстве случаев, зависимость новизны сообщения со временем убывает. Хотя иногда бывают всплески интереса к тому или иному сообщению через неделю, через месяц.

Старение новостей можно оценивать с помощью распределения числа публикаций на определенную тему по времени. Рассматривая диаграммы зависимости числа публикации о новостных событиях в СМИ, можно увидеть, что новизна события убывает с течением времени. Пример такой диаграммы представлен на рис. 1, по оси абсцисс отложено время публикаций о событии «Взрыв электрички в Ессентуках», по оси ординат – число публикаций.

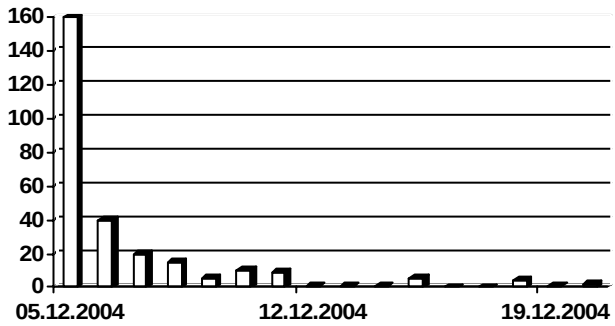


**Рисунок 1. Пример зависимости числа публикаций о событии от времени**

Старение информации в группе сообщений $C_i$ можно оценивать некоторой функцией $s(t_i)$, где $t_i$ – среднее время сообщений в группе $C_i$. $t_i$ определяется по формуле:

$$t_i = \frac{\sum_{D_j \in C_i} t(D_j)}{|C_i|},$$

где $t(D_j)$ – время сообщения $D_j$, $|C_i|$ – число сообщений в группе $C_i$.

Назовем $s(t_i)$ функцией «свежести» группы сообщений. Примерный вид данной функции представлен на рис. 2.

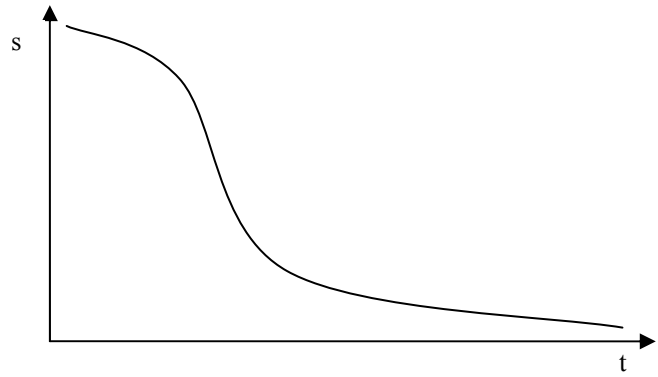


**Рисунок 2. Примерный вид функции "свежести" группы новостных сообщений**

Итак, используя функцию «свежести», можно найти вес группы по времени $R_{time} = s(t_i)$, который можно использовать для ранжирования по времени и итогового ранжирования по актуальности.

**Ранжирование по важности для пользователя**

Важность группы новостных сообщений для пользователя зависит от числа чтений пользователями. Чем больше сообщений прочитано из данной группы, тем более она интересна пользователям. Стоит отметить, что описанный способ хорошо работает лишь при большом количестве обращений пользователей, например на новостном портале в сети Интернет. В случае системы

с малым числом пользователей, например, на рабочем месте аналитика, важность группы пользователь может задавать вручную.

Вес группы $R_{user}$ по важности для пользователя будет равен числу чтений $N_{read}$ или задаваться вручную.

При большом количестве пользователей имеет смысл разделение пользователей на классы по интересам, например, классы «политика», «экономика», «спорт» и подобные. В данном случае, пользователь будет относиться к одному из классов, и ранг групп сообщений по важности для пользователя будет учитывать интересы класса. Например, для пользователя из класса «спорт» на первом месте будут стоять именно свежие спортивные новости.

Для построения данной формальной модели класса пользователей собирается статистика чтения сообщений пользователями. Каждому пользователю $u_i$ можно сопоставить вектор в пространстве групп сообщений:

$$u_i = \left(u_{i,1}, u_{i,2}, ..., u_{i,k}\right),$$

где $u_{i,j}$ – число чтений пользователем $u_i$ сообщений группы $C_j$, $i = \overline{1,n}$, $j = \overline{1,k}$, $n$ – число пользователей, $k$ – число групп сообщений. Полученный вектор зависит от информационных потребностей пользователя.

Стоит обратить внимание, что вектор рассматривается не в пространстве сообщений, а в пространстве групп. Допустим, два пользователя-«спортсмена» прочитали разные сообщения в одной группе. Очевидно, что информационные потребности у них одинаковые, но в пространстве сообщений вектора будут отличаться. В пространстве групп вектора у них будут совпадать.

Класс пользователей определяется решающим правилом – некоторой функцией. На вход функции поступает вектор, на выходе получается класс, к которому принадлежит пользователь.

Построение решающего правила может происходить различными способами: рубрицирование, основанное на примерах, и кластерный анализ [AM01].

В случае рубрицирования, основанного на примерах, необходимо задать примеры сообщений для каждого класса пользователей. Отметим, что данные примеры заранее формируются экспертом, и от качества их подбора во многом зависит качество работы системы. После этого происходит обучение системы рубрицирования, в ходе которого происходит настройка решающего правила.

В случае кластерного анализа, классы пользователей автоматически формируются на основе близости векторов в пространстве. Число классов может быть, как заранее известное, так и нет. Близкие вектора считаются однородными, принадлежащими одному классу.

После определения принадлежности пользователя к классу, нужно определить вес группы новостных сообщений по важности для пользователя. Данный вес по-прежнему будет зависеть от числа чтений сообщений, но с большей степенью нужно учитывать пользователей своего класса, и с меньшей степенью – пользователей других классов. Вес группы новостных сообщений $R_{user}$ по важности для пользователя с учетом классов пользователей будет равен:

$$R_{user} = a_0 N_{read\,0} + (1 - a_0)(N_{read} - N_{read\,0}),$$

где $N_{read\,0}$ – число чтений сообщений пользователями только своего класса, $N_{read}$ – число чтений пользователями всех классов, $a_0$ – коэффициент значимости своего класса, $0 < a_0 < 1$.

Вес группы новостных сообщений будет динамически изменяться в зависимости от класса пользователя, и вес будет выше у тех сообщений, которые больше интересны пользователям «своего» класса. Например, «спортсмену» сначала будут выданы спортивные события, так как другие «спортсмены» больше читали спортивные события, подняв тем самым вес по важности данных событий.

Подобные методы реализованы в алгоритмах контекстной рекламы в таких системах, как www.google.com, www.yandex.ru, www.begun.ru. В алгоритмах контекстной рекламы есть множество рекламных сообщений. Пользователь, вводя запросы в поисковой системе, формирует свои информационные потребности, которые учитываются системами. В зависимости от информационной потребности, пользователю будет выводиться именно та реклама, которая ему наиболее интересна.

**Ранжирование по важности для СМИ**

Количество сообщений, описывающих группу, отображает общий интерес новостных источников к ней. Чем больше пишут об определенной группе сообщений, тем более она интересна СМИ.

СМИ различаются между собой по достоверности. Например, сообщениям «РИА «Новости»» www.rian.ru следует доверять в большей степени, чем сообщениям интернет-агенства «Вокруг новостей» www.vokruginfo.ru. Необходимо введение весов источников. Вес группы сообщений по источникам будет равен:

$\sum_i v_i n_i$, где $i$ – число источников, $v_i$ - вес источника, $n_i$ - количество сообщений группы из данного источника. Этим способом можно отбросить сомнительные новости.

Остается неучтенным вариант, когда один источник, пусть даже с малым весом, будет посылать большое количество сообщений об одной сомнительной новости, в этом случае соответствующая группа сообщений будет иметь большой вес, что неправильно. Значит, следует учитывать и долю источников, пишущих о данной новости. Чем больше источников освещают сообщения данной группы, тем вес группы должен быть больше. Для учета доли источников следует помножить вес по источникам на коэффициент, учитывающий долю источников. Таким коэффициентом может служить $\log \frac{i}{i-k+1}$, где $i$ – общее число источников, $k$ – число источников, имеющих сообщения в данной группе.

Формула веса группы по важности для СМИ будет иметь следующий вид:

$$R_{smi} = \log \frac{i}{i-k+1} \sum_i v_i n_i .$$

Вариацией ранжирования по источникам является явное указание пользователем, какие источники следует обрабатывать. В данном случае будут загружаться и обрабатываться новости только выбранных СМИ, что должно снизить нагрузку на систему и отбросит новости от ненужных СМИ.

**Итоговое ранжирование групп новостных сообщений**

После того, как получены веса групп новостных сообщений по отдельным критериям (по времени $R_{time}$, по важности для пользователей $R_{user}$, по важности для СМИ $R_{smi}$) необходимо перейти к единому весу группы – весу по актуальности. По нему следует производить итоговое ранжирование групп новостных сообщений.

Возможны два варианта перехода к единому весу:

- фиксация критериев,
- учет всех критериев.

**Итоговое ранжирование с фиксацией критериев**

Для любого критерия можно задать ограничение на его значение, например для времени ограничение может выглядеть как «сообщения за последние сутки», для важности по пользователям –

«группа, сообщения которой прочитаны 50% пользователей», для важности по СМИ – «группа, освещаемая 75% источников».

Ограничение $Lim = \{f(r), a, b\}$ на вес $r$ группы новостных сообщений задается с помощью некоторой функции $f(r)$ и граничных значений $a$ и $b$ этой функции. Если $f(r) \in (a, b)$, то группа с весом $r$ удовлетворяет ограничению $Lim$.

Итоговое ранжирование будет выглядеть следующим образом: задаем ограничения Lim на все критерии, кроме одного, и ранжируем группы по весу оставшегося критерия. Итоговый вес будет равен весу одного из критериев:

$$R_{all} = \begin{cases} R_{time}, заданы\ Lim_{user}, Lim_{smi}, \\ R_{user}, заданы\ Lim_{time}, Lim_{smi}, \\ R_{smi}, заданы\ \ Lim_{user}, Lim_{time}. \end{cases}$$

Недостатком данного способа итогового ранжирования является то, что задание ограничений на некоторые критерии может привести к потере некоторых актуальных групп.

**Итоговое ранжирование с помощью нечеткой логики**

Рассмотрим определение актуальности с позиций нечеткой логики [ZL65].

Под универсальным множеством $U$ будем понимать множество всех групп новостных сообщений. На этом универсальном множестве задаются нечеткие множества $A_{time}$ – «свежая группа», $A_{user}$ – «важная группа для пользователя», $A_{smi}$ – «важная группа с точки зрения СМИ», $A_{all}$ – «актуальная группа». Функции принадлежности нечетких множеств будут равны весам групп новостных сообщений:

$$\mu_{time}(C'_j) = R_{time}(C'_j),$$
$$\mu_{user}(C'_j) = R_{user}(C'_j),$$
$$\mu_{smi}(C'_j) = R_{smi}(C'_j),$$
$$\mu_{all}(C'_j) = R_{all}(C'_j),$$

где $C'_j$ – группа новостных сообщений, $j = \overline{1,k}$,

$\mu_{time}(C'_j), \mu_{user}(C'_j), \mu_{smi}(C'_j), \mu_{all}(C'_j)$ – значения функций принадлежности соответствующих нечетких множеств,

$R_{time}(C'_j), R_{user}(C'_j), R_{smi}(C'_j), R_{all}(C'_j)$ – значения соответствующих весов групп (подразумевается, что веса нормированы к отрезку [0,1]).

Из определения актуальности новостного сюжета следует, что сюжет актуален, если он:

- свежий по времени,
- важный для пользователя
- важный, с точки зрения СМИ.

Значит нечеткое множество $A_{all}$ равно пересечению нечетких множеств $A_{time}$, $A_{user}$, $A_{smi}$: $A_{all} = A_{time} \cap A_{user} \cap A_{smi}$.

Существуют различные способы построения функции принадлежности пересечения нечетких множеств. В первом случае [ZL65]:

$$\mu_{all} = \min(\mu_{time}, \mu_{user}, \mu_{smi}).$$

Во втором случае [BZ70]:

$$\mu_{all} = \mu_{time}\mu_{user}\mu_{smi}.$$

Переходя к весам, получаем два варианта итоговой формулы актуальности группы $C'_j$:

$$R_{all}(C'_j) = \min(R_{time}(C'_j), R_{user}(C'_j), R_{smi}(C'_j))$$

Или:

$$R_{all}(C'_j) = R_{time}(C'_j)R_{user}(C'_j)R_{smi}(C'_j)$$

**Итоговое ранжирование с помощью модифицированного метода многокритериального ранжирования**

Для вычисления единого веса групп новостных сообщений можно использовать модификацию метода многокритериального ранжирования. В оригинальном методе [СЛО88] выделяются два этапа: построение функций, вычисляющих веса для каждого критерия, и построение единого веса.

Функции вычисления веса для каждого из критериев у нас уже построены. Способ построения единого веса в методе многокритериального ранжирования использует среднее геометрическое от весов критериев.

$$R_{all} = \sqrt[3]{R_{time} R_{user} R_{smi}}$$.

**Итоговая формула веса актуальности новостного сюжета**

Заметим, что использование нечеткой логики и многокритериального ранжирования дает похожие результаты: для вычисления актуальности можно использовать произведение весов в некоторых степенях.

Запишем общую формулу веса по актуальности :

$R_{all} = R_{time}^{a_{time}} R_{user}^{a_{user}} R_{smi}^{a_{smi}}$, где $a_{time}$, $a_{user}$, $a_{smi}$ – соответствующие коэффициенты весов по времени, важности.

Изменяя коэффициенты, эксперт может добиться желаемого поведения системы. Например, он может установить больший коэффициент для веса по времени, тогда сюжеты со свежими новостными сообщениями будут располагаться выше в выпуске новостей.

## 3.3. Аннотирование новостных сюжетов

Результатом выделения и ранжирования новостных сюжетов будет являться упорядоченный по актуальности список новостных сюжетов.

Сложной задачей является составление качественного описания, чтобы пользователю было понятно, о чем идет речь в данном новостном сюжете, без чтения всех сообщений новостного сюжета. В различных системах кластеризации текстов данная задача решается по-разному.

Наиболее распространенный способ – формирование списка ключевых слов. Данный способ прост в реализации, но обладает существенным недостатком из-за недостаточной информативности. Например, очень часто, в силу особенностей русского языка, ключевыми словами кластеров новостных сообщений являются слова «государство», «человек», «говорить». Данные результаты наблюдаются, если обработать матрицу сообщения-термины с помощью метода главных компонент. В числе главных компонент будут соответствующие вышеуказанным словам.

Другой способ – составление реферата новостного сюжета на основе списка всех сообщений новостного сюжета. Данный способ дает, в зависимости от системы реферирования, довольно понятный реферат новостного сюжета, но сложен алгоритмически.

В данной работе предлагается использование результатов кластерного анализа с разбиением на фрагменты для аннотирования

полученных новостного сюжетов. Стоит отметить, что ниже используется понятие «центр группы», которое имеет смысл лишь в некоторых методах кластерного анализа, в частности в методе «к-средних».

Для наглядности, представим способ составления рефератов в графическом виде. На рисунке ниже представлены 3 группы фрагментов (в виде множеств точек), соответствующие некоторым новостным сюжетам.

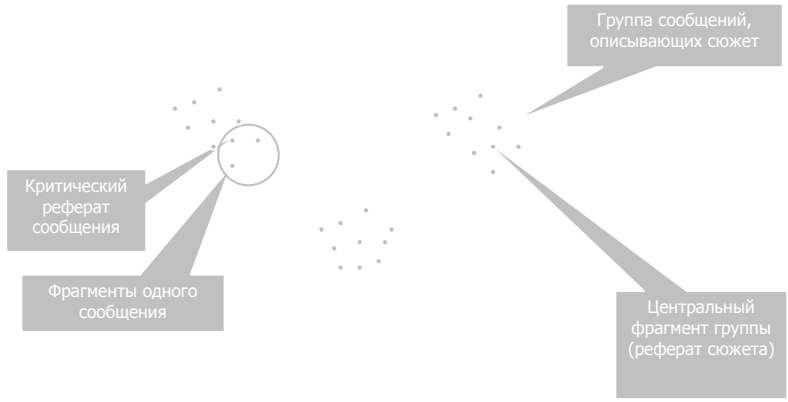


**Рисунок 3. Схематичное представление метода реферирования новостных сюжетов и сообщений**

Выделяются фрагменты, ближайшие к центру группы, содержание каждого такого фрагмента будет наиболее близко к новостному сюжету соответствующей группы. Полученные фрагменты представляют собой законченные смысловые блоки текста, наиболее близкие к данному новостному сюжету, то есть реферат новостного сюжета.

Особый интерес представляет собой получение краткого содержимого сообщения, его реферата. Для каждого сообщения в группе (выделены окружностью), соответствующей некоторому новостному сюжету, можно найти один или несколько фрагментов, которые будет наиболее близки к центру данной группы. Данные фрагменты будет являться выдержкой из текста, которая наиболее близка по содержимому к выбранному новостному сюжету, то есть кратким описанием сообщения как элемента новостного сюжета.

Таким образом, при использовании кластеризации сообщений с разбиением на фрагменты автоматически получается краткое описание новостных сюжетов и сообщений.

## 3.4. Ранжирование сообщений в новостном сюжете

Когда пользователь выбрал новостной сюжет, нужно вывести сообщения из соответствующей группы и ранжировать их. Вес сообщения будет зависеть от времени сообщения и содержимого сообщения.

Ранжирование сообщений по времени использует подобную функцию, как и в случае ранжирования новостных сюжетов. Время сообщения подставляется в функцию «свежести» $s(t)$, и получается вес по времени $R_{message\_time}$. Следует отметить, что функции «свежести» новостных сюжетов и сообщений не обязательно имеют общий вид.

Ранжирование сообщений по содержанию использует степень соответствия сообщения новостному сюжету, то есть точность описания новостного сюжета данным сообщением.

Вес по содержанию $R_{message\_content}$ зависит от расстояния от центра группы в выбранной формальной модели и равен, например, евклидову расстоянию между вектором центра группы и вектором фрагмента сообщения в пространственно-векторной модели. Так как фрагмент в сообщении не один, для рассмотрения берется ближайший к центру группы фрагмент. Чем ближе в данном пространстве ближайший фрагмент к центру группы, тем более точно сообщение близко по содержанию к новостному сюжету, тем выше вес сообщения по содержанию. Таким образом, вес по содержанию для группы $C'_k$ и сообщения $D_i$ будет равен:

$$R_{message\_content} = \min_{f_{ij} \in D_i} \rho(e_k, f_{ij}),$$

где $\rho$ – расстояние в выбранном пространстве, $e_k$ – центр группы $C'_k$, $f_{ij}$ – фрагменты сообщения $D_i$, $j = \overline{1, k_i}$, $k_i$ – число фрагментов в сообщений $D_i$. Стоит отметить, что данный способ годится не для всех методов кластерного анализа.

В случае использования кластеризации с помощью разбиения на фрагменты, вес по содержанию может быть равен доле фрагментов сообщения принадлежащих группе, соответствующей выбранному новостному сюжету. Чем больше фрагментов сообщения принадлежит данной группе, тем выше вес сообщения по

содержанию. В данном случае, вес по содержанию для группы $C'_k$ и сообщения $D_i$ будет равен:

$$R_{message\_content} = \frac{\left| f_{ij} : f_{ij} \in D_i, f_{ij} \in C'_k \right|}{k_i}$$

где $f_{ij}$ - фрагменты сообщения $D_i$, $j = \overline{1, k_i}$, $k_i$ – число фрагментов в сообщений $D_i$.

Итоговый вес сообщения $R_{message}$ в выбранной группе будет функцией от веса по времени $R_{message\_time}$, и веса по содержимому $R_{message\_content}$:

$$R_{message} = F_{message}(R_{message\_time}, R_{message\_content}).$$

Как и в случае с ранжированием новостных сюжетов, в качестве итоговой функции возможно использование либо фиксирования критериев, либо произведения весов с учетом степени значимости критериев.

## 3.5. Эксперименты

**Оценка качества кластерного анализа**

Общие представления о качестве кластерного анализа формулируются в виде некоторого функционала, экстремальные значения которого соответствуют наилучшей классификации. Данный функционал обычно называют оптимизационным. Данное направление пытается задачу кластерного анализа ввести в традиционное математическое русло, четко сформулировать критерий и добиваться его максимизации. При этом, естественно, возникают чисто математические проблемы: определения свойств функционала, путей достижения оптимума, трудоемкости алгоритма. В [Ман88] приводится около 50 наиболее распространенных критериев качества классификации.

Один из показателей качества кластерного анализа основан на сравнении матриц типа "сообщение-сообщение", соответствующих различным разбиениям одних и тех же сообщений [Rand71]. Каждому разбиению $n$ сообщений на кластеры ставится в

соответствие матрица $C_{n*n}$, элементы которой определяются выражением:

$C_{ij} = 1$, если сообщения $d_i$ и $d_j$ принадлежат одному кластеру,

$C_{ij} = 0$, в противном случае,

а в качестве меры качества используются различные меры близости (расстояния) между соответствующими им матрицами. Например, может использоваться следующая мера близости

$$e = 1 - \frac{2}{n(n-1)} \sum_{i>j} \left| C_{ij}^1 - C_{ij}^2 \right|,$$

предложенная в работе [Rand71] и характеризующая процент пар документов, одновременно лежащих либо не лежащих в одном классе в обоих разбиениях, где $C^1$ - матрица, соответствующая первому разбиению документов, а $C^2$ - матрица, соответствующая второму разбиению. Данный показатель меняется от 0 до 1. Чем ближе к 0, тем сильнее совпадают результаты кластерного анализа.

Используя эталонное разбиение экспертов, можно находить похожесть различных разбиений сообщений на кластеры. Стоит отметить, что к данным показателям, основанным на сравнении с эталоном, необходимо подходить критически, так как отсутствие соответствия с эталонной классификацией может быть вызвано тем, что классификация массива документов могла быть проведена по другому основанию. В дальнейшем будем называть меру близости между текущим разбиением и эталонным сокращенно: мера близости.

Используя данный показатель качества, был произведен эксперимент по обработке новостных сообщений на русском языке, скачанных с сайта информационного агентства «РИА «Новости» [Зев05]. Целью данного эксперимента было определение метода кластерного анализа, показывающего лучшее соответствие с экспертной оценкой. Количество сообщений 7039, общий размер 8.6 Мбайт, средний размер 1.3 Кбайт. Сообщения на сайте распределены экспертами по различным тематикам, что позволит впоследствии оценить показатель эффективности кластерного анализа. Эксперименты производились как с разбиением на абзацы, так и без. В результате было установлено, что лучшую меру близости показывает метод «К-средних», причем разбиение текстовых сообщений на фрагменты уменьшает меру близости в 2 раза.

Следующий эксперимент произведен для сравнения с системой «Яндекс Новости» на наборе данных «Новости – Обычная неделя» (из файла). В наборе данных есть соответствие «сообщение-кластер», рассчитанное системой «Яндекс Новости», поэтому возможно применение меры близости для сравнения работы алгоритмов кластерного анализа. В результате получилось, что мера близости с системой «Яндекс Новости» равна 0,02%, то есть матрицы $C_{n*n}$ практические совпадают. Объяснение этому – сильная разреженность этих матриц (количество ненулевых элементов $10^{-6} - 10^{-4}$) для данной подборки. Этот факт еще раз подтверждает, что к показателям качества кластерного анализа, основанным на сравнении с эталоном, нужно подходить очень осторожно.

**Оценка качества аннотирования**

В настоящее время методы оценки рефератов основаны лишь на экспертных оценках, например, оценка «изнутри» (или нормативная оценка). Пользователи судят о качестве реферата, анализируя сам реферат. Они оценивают гладкость текста, делают вывод о том, насколько хорошо реферат отражает основные идеи оригинала, либо сравнивают его с идеальным рефератом, написанным автором исходного текста или другим специалистом.

Поэтому эксперименты по сравнению качества реферирования двух систем можно проводить лишь путем сравнения экспертом результатов реферирования одинаковых новостных сюжетов.

Сравнение производится с системой «Яндекс Новости», одной из функций которой является реферирование новостных сюжетов.

Система «Яндекс Новости» использует алгоритмы представления кластеров, основанные на статистических методах. Данные методы иногда приводят к результатам, из которых невозможно понять, о чем говорится в новостном сюжете. Например, результаты представления новостного сюжета «Выборы президента на Украине» системой «Яндекс Новости»:

| Заголовок сообщения | Аннотация сообщения |
|---|---|
| Милиция не пойдет против народа 11:28 Хартия'97 | Об этом "Трибуне" сообщил высокопоставленный источник в МВД Украины. |
| Украина: взлом сейфа и гонки по вертикали 11:21 Правда.ru | Со всех уголков Украины продолжает поступать информация о нарушениях и ... списков и бюллетеней только в 14 часов в воскресенье, сообщает МВД |

| | |
|---|---|
| | Украины. |
| Оппозиция на улице, в ЦИКе перерыв 11:05 РБК | ЦИК Украины объявил перерыв в подсчете голосов до 15 часов. ... обработки Центральной избирательной комиссией Украины 75,26% протоколов стало ... |
| Уважаемые посетители KM.RU ! 11:03 KM.ru | ... за статью "Битва за Украину-2", размещенную в течение выходных на портале KM.RU ... |
| Олигархи мои. 10:42 From-UA | ... всех кандидатов в Президенты Украины было обещание "разобраться" с олигархами. |

**Таблица 1. Неудачные примеры аннотирования системы "Яндекс Новости"**

Основным недостатком представленных результатов является то, что мысли являются незаконченными, что приводит к неполному пониманию смысла сообщения.

В данной работе предлагается принцип представления результатов с помощью абзацев, то есть законченных мыслей текста, что приводит к более полной картине новостного сюжета.

Например, результаты представления новостного сюжета «Выборы президента на Украине» с помощью предлагаемых алгоритмов:

| Заголовок сообщения | Аннотация сообщения |
|---|---|
| В последний день предвыборной агитации молодежь в Киеве призывает проголосовать спокойно и демократично | На телеканалах известные и уважаемые люди - ученые, писатели, певцы, диджэи и спортсмены призывают голосовать за "своего" кандидата. "За" Януковича на 1-м украинском канале поет в видеоролике Иосиф Кобзон, "за" Ющенко призывают голосовать известные боксеры - братья Кличко - оба в красных фуфайках. |
| По данным параллельного подсчета 67,3%% бюллетеней в штабе Януковича, за премьера проголосовали | Как заявила журналистам представитель штаба Януковича Раиса Богатырева, после обработки 67,3%% бюллетеней центром параллельного подсчета голосов при штабе за Януковича проголосовали |

| | |
|---|---|
| 50,54%. | 50,54%%, за Ющенко - 45,53%%. |
| Наблюдатели от СНГ не зафиксировали серьезных нарушений на выборах президента Украины | В частности, в Одессе, Львове, Киеве наблюдалось несвоевременное открытие избирательных участков, уточнил собеседник агентства. Также, по его словам, во Львове, Херсонской области и Луцке на отдельных избирательных участках в кабины для голосования заходили сразу несколько человек. |
| На избирательном участке в Черкасской области Украины убит милиционер | По предварительным данным, милиционера застрелили из огнестрельного оружия. На месте происшествия работают оперативники. |

**Таблица 2. Примеры аннотирования с помощью предлагаемых алгоритмов**

Из приведенных примеров представления одного и того же новостного сюжета «Выборы на Украине» видно, что разработанная система превосходит «Яндекс Новости».

Сравнение рефератов 5-10 новостных сообщений дает неполное представление о качестве аннотирования, но служит иллюстрацией того, что предлагаемые алгоритмы улучшают качество.

## 4. Выводы и обсуждение результатов

Разбиение текстовых новостных сообщений на фрагменты позволяет без использования сложных методов анализа, таких как синтаксический и семантический, достичь хорошего качества кластеризации, сравнимого с результатами работы экспертов агентства «РИА «Новости». Помимо этого, данный способ позволяет просто решать задачи аннотирования новостных сюжетов и сообщений.

В работе предлагается метод обнаружения групп содержательно близких новостных сообщений и метод выделения актуальных, то есть важных и свежих новостных сюжетов в подборке новостей.

Метод обнаружения групп содержательно близких новостных сообщений основан на использовании кластерного анализа

векторного представления текстов сообщений. Предложено использовать разбиение сообщений на фрагменты, и показано, что это приводит к значительному улучшению качества кластерного анализа.

Введено понятие актуальности новостных сюжетов, оно учитывает время, количество сообщений, количество СМИ, освещающих данный новостной сюжет, количество чтений сообщений пользователями, разбитыми на группы по интересам. Предложен метод расчета ранга новостных сюжетов на основании вышеперечисленных составляющих.

Предложен метод автоматического формирования аннотаций полученных новостных сюжетов и сообщений, учитывающий разбиение сообщений на фрагменты.

Направлениями дальнейших исследований являются:

1. Отслеживание истории новостных сюжетов, то есть описания одного и того же новостного сюжета в разные периоды времени.
2. Принятие во внимание структуры текста, например учет гиперссылок, имен авторов.
3. Более тонкий учет контекста: применение составных терминов, сужение рамок описания контекста до одного предложения.

## 5. Литература

[АЩ99] Азимов Э.Л., Щукин А.И. Словарь методических терминов (теория и практика преподавания языков). Печатное издание СПб.: «Златоуст», 1999.

[АМ01] Айвазян С.А., Мхитарян В.С. Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2т. 2-е изд., испр.- Т.1: Теория вероятностей и прикладная статистика. М.: ЮНИТИ-ДАНА, 2001. 656 с.

[БСЭ72] Большая советская энциклопедия. Гл. ред. А.М. Прохоров, 3-е изд.Т. 10. М.: «Сов. энциклопедия», 1972. 592с.

[ЗЕВ05] Зевайкин А.Н. Об одном подходе к кластеризации текстовых сообщений с разбиением на абзацы. Информационные технологии. 2005. № 5 с. 16-20

[Нов04] Новотека http://www.novoteka.ru/txt/about

[СЛО88] Слотин Ю.С. Ранжирование факторов. Справочник: Надежность и эффективность в технике. Том 5: Проектный анализ надежности. М.: Машиностроение, 1988. с. 310-316.

[Янд03] Яндекс Новости http://news.yandex.ru/about.html

[BG01] Igor A. Bolshakov and Alexander F. Gelbukh. Text segmentation into paragraphs based on local text cohesion. Proc. TSD-2001: Forth International Workshop on Text, Speech and Dialogue, Plzen (Pilsen), Czech Republic, September 10–13, 2001. Lecture Notes in Artificial Intelligence (indexed by SCIE) N 2166, ISSN 0302-9743, ISBN 3-540-4255-7, Springer-Verlag, pp. 158–166.

[BZ70] Bellman R., Zadeh L. Decision-making in fuzzy environment //Management Science. 1970. V. 17. P. 141 - 164.

[Goo03] Google News
http://news.google.com/intl/en_us/about_google_news.html

[FLO05] Chris Flood. Political communication and the media in Britain. Surrey, University of Surrey, 2005. p. E1-E20

[GR65] Galtung J., Ruge M.H. The structure of foreign news. Journal of Peace Research. Vol.2. 1965. P. 64-90.

[TDT04] NIST. The 2004 Topic Detection and Tracking (TDT2004) task definition and evaluation plan.
http://www.nist.gov/speech/tests/tdt/index.htm

[WIL97] Wilson, Elizabeth. Working in journalism: A comprehensive guide to job opportunities in the media. Plymouth: Northcote house, 1997, 189 p.: ill.

[ZL65] Zadeh, Lotfi**.** Fuzzy Sets / Information and Control, 8(3), June 1965, pp.338-353.

News bulletin production, based on automatic analysis of news reports.

A. Zevaikin
V. Korneev

State Research & Development Institute “Quant”
shuraz@yandex.ru

Summary

The news flow keeps increasing nowadays and the time period permissible for the news processing on the contrary becomes more and more limited.
These facts lead the interest to the systems that automate news processing and news analysis to grow.

This work suggests the technology of the news bulletin building, based on the automatic analysis of the news reports. This technology includes methods of the automatic picking out, ranking and abstracting of the news reports.
Splitting of the text messages into the fragments permits to improve the quality of the cluster analysis and the quality of the annotation of the text messages and news reports.
The unified criterion "the topicality" was introduced for the ranking of the news reports. It includes such aspects as time, the source, number of reports in the group, it also reflects how many times the report was read by users.
Application of the suggested methods will help to solve the posed problems (indicated problems).