Формирование исторической справки по корпусу новостей с учетом структуры динамики развития новостного сюжета

© М.М. Тихомиров © Б.В. Добров

Московский государственный университет имени М.В. Ломоносова,
Москва, Россия

tikhomirov.mm@gmail.com dobrov_bv@srcc.msu.ru

Аннотация. Описаны проведенные исследования по тематике формирования исторической справки. Разработаны алгоритмы и реализована программная система, позволяющая автоматически создавать историческую справку по корпусу новостных статей для выбранного новостного документа. Проведено исследование трех новых факторов, учитывающих структуру динамики развития новостного сюжета.

Ключевые слова: обзорное реферирование, историческая справка, информационный поиск.

Using News Corpora for Temporal Summary Formation

© Mikhail Tikhomirov © Boris Dobrov

Lomonosov Moscow State University,
Moscow, Russia

tikhomirov.mm@gmail.com dobrov_bv@srcc.msu.ru

Abstract. The paper describes the research carried out on the subject of the formation of the temporal summary. Algorithms have been developed and a software system has been implemented that allows you to automatically create a timeline summary for the body of news articles for the selected news document. A study of three new factors, taking into account the structure of the dynamics of news story development.

Keywords: timeline summarization, multi-document summarization, information retrieval.

1 Введение

В связи с взрывным ростом количества информации в интернете возникает задача выделения и автоматического обобщения полезной информации в поступающем потоке данных.

Востребованными задачами являются задачи реферирования новостных сюжетов – множества новостных сообщений различных источников, посвященных описанию некоторого события. Такие задачи часто решаются новостными агрегаторами, например, Яндекс.Новости [17], для более полного представления описания произошедшего события. Типичное «время жизни» новостного сюжета (время активного обсуждения произошедшего события) обычно сутки–двое.

Отметим, что некоторые новостные сюжеты имеют «историю» в виде множества предшествующих событий, произошедших в различные моменты времени и в той или иной мере связанных между собой.

Для таких длительных сюжетов, где сама их длительность (повторное возвращение к одной и той же теме) в определенной мере свидетельствует об их значимости, является актуальной задача формирования «исторических справок».

Историческая справка – это тип обзорного реферата (обзорной аннотации), включающего последовательное изложение существенных деталей исследуемого сюжета. Подобная аннотация может содержать в себе основные этапы, события и факты исходного сюжета. Построение подобных аннотаций представляет собой сложную работу, которую выполняют журналисты или аналитики, и, соответственно, автоматизация подобного процесса является востребованной задачей.

В рамках данной работы рассмотрены проблемы и решения при автоматическом построении исторических справок.

Рассматривается ситуация, когда пользователя новостного агрегатора заинтересовала какая-то новость (новостное сообщение), и он хочет получить историческую справку по сюжету, обсуждаемому в данном новостном сообщении, т. е. результатом должен быть упорядоченный по времени перечень описаний произошедших ранее ключевых событий.

Задача рассматривается как задача обзорного реферирования (multi-document summarization) по запросу на представительной коллекции новостных

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

документов. В качестве запроса рассматривается текст новостного сообщения.

На корпусе из 2 миллионов новостных статей на русском языке за первую половину 2015 года была разработана и реализована система, позволяющая автоматизировать процесс построения исторической справки. Проведено исследование трех новых факторов, позволяющих за счет учета структуры новостного корпуса улучшить результаты работы системы. Оценка производилась на 15 новостных сюжетах, из которых для 5 эталонные аннотации были сформированы одним из авторов, а другие 10 взяты с сайта interfax.ru

2 Обзор

2.1 Задача обзорного аннотирования

В настоящее время предложено достаточно большое количество методов автоматического обзорного реферирования [3]. Известны методы как с использованием больших лингвистических онтологий [15], в том числе автоматически пополняемых в процессе анализа [12], так и на основе статистических свойств текстов [16], машинного обучения [13, 17].

Существенными проблемами при составлении аннотации новостного кластера являются [3, 7, 11]:

- обеспечение полноты представления информации, в том числе наиболее свежей информации;
- снижение повторов при представлении информации;
- обеспечение связности и понятности представляемой информации.

Для определения избыточности в порождаемых аннотациях используются различные меры сходства между предложениями. Одним из распространенных подходов является предварительная кластеризация выделение близких по содержанию кластеров предложений [6]. Другим подходом для уменьшения избыточности являются сравнение предложений-кандидатов с предложениями, уже попавшими в аннотацию, и оценка новой (непохожей) информации, например, подход Maximal Marginal Relevance (MMR) [2].

2.2 Историческая справка

Задача построения исторических справок имеет ряд отличий от стандартной задачи обзорного реферирования.

Сначала необходимо определить документы, по которым будет строиться аннотация. Если стандартный новостной сюжет обычно образован близкими документами, посвященными одному событию, которые могут быть получены применением одного из известных методов кластеризации [10, 14].

Для больших коллекций применение методов кластеризации не оправдано. Во-первых, такую задачу придется решать многократно на огромных коллекциях документов. Во-вторых, степень близости между документами, которые описывают далекие по времени, но связанные события, может быть значительно меньше по стандартным мерам сходства.

Требуется выявлять наиболее характерные объекты [1, 9], например, учитывая структурные особенности потока документов [5, 8].

3 Постановка задачи

3.1 Общее описание

Задача построения исторической справки ориентирована на запрос. В самом общем случае пользователь в качестве запроса имеет новостной документ, поэтому данная задача будет рассматриваться как задача автоматического построения аннотации описанного типа по запросу в виде текстового документа. На выходе работы системы должна быть аннотация из n предложений. Связность между предложениями не требуется.

Как пример построенной исторической справки можно рассмотреть аннотацию (таблица 1), построенную по событию, связанному с крушением самолета в Тайване.

Таблица 1. Крушение самолета на Тайване

1	*Самолет ATR 72 авиакомпании TransAsia потерпел крушение 4 февраля на Тайване.*
2	*Операция по поиску жертв крушения самолёта TransAsia Airways завершена, в результате происшествия погибли 35 человек.*
3	*Члены экипажа самолета авиакомпании TransAsia Airways, потерпевшего крушение в феврале на Тайване, отключили работающий двигатель, после того, как второй перестал работать*
…	*…*
n	*Совет по авиационной безопасности Тайваня опубликовал отчет о крушении самолета компании TransAsia Airways в феврале этого года, в результате которого погибли 35 человек.*

В цели работы входит исследование влияния различных факторов на качество построения аннотации, поэтому необходим набор эталонных аннотаций, на которых будет оцениваться качество работы системы.

3.1 Математическая постановка задачи

Описанную выше задачу можно формализовать следующим способом: имеются набор запросов $Q = \{q_1, q_2, \dots, q_m\}$ и ассоциированный с ним набор эталонных аннотаций $D_g = \{D_g^{q_1}, D_g^{q_2}, \dots, D_g^{q_m}\}$. Система в ответ на запросы Q алгоритмом A генерирует набор исторических справок $D_A = \{D_A^{q_1}, D_A^{q_2}, \dots, D_A^{q_m}\}$.

Тогда задача построения исторической справки сводится к задаче максимизации функционала

$$\frac{\sum_{i=1}^{i=|Q|} M(D_A^{q_i}, D_g^{q_i})}{|Q|} \to \max, \quad (1)$$

где M – функция близости между аннотациями. Максимизация происходит по выбору алгоритма A и по всем параметрам выбранного алгоритма.

4 Предлагаемый подход

4.1 Исследуемые факторы

В рамках работы исследовались следующие факторы:

- стратегия расширения запроса;
- учет временного характера новостных сюжетов.
- учет структуры новостной статьи в виде перевернутой пирамиды.

4.2 Стратегия расширения запроса

Информации, которую можно получить из запроса-документа, может быть не достаточно, чтобы эффективно построить историческую справку. Этот факт является следствием того, что большинство новостных статей является не общим описанием события, а обсуждением какого-то частого происшествия или факта. Чтобы избежать подобной проблемы, был разработан алгоритм, использующий кластер близких запросу документов. Алгоритм:

1. Для запроса-документа на основе статистической информации по коллекции (индекс) строится вектор наиболее весомых по tf-idf лемм (нормализованных словоформ) документа.
2. По построенному вектору происходит поиск близких документов в коллекции.
3. По кластеру извлеченных документов происходит анализ важности лемм на основе tf-idf:
    a. Для каждого документа рассматриваются лучшие t лемм.
    b. Происходит ранжирование лемм на основании частоты встречаемости в лучших *t* леммах каждого документа.
    c. Из сортированного списка выбирается *k* наиболее весомых лемм.
4. Повторяются пункты 2–3 (повторное расширение запроса).
5. На выходе имеется вектор из *k* лемм, который отражает семантику документа-запроса.

Как пример работы модуля расширения запроса можно рассмотреть этапы работы алгоритма на новостной статье, посвященной теракту в Париже (порядок в списке обратный по отношению к весу слова):

Олланд назвал нападение на Charlie Herbo терактом
Президент Франции Франсуа Олланд назвал терактом нападение на сотрудников сатирического журнала Charlie Herbo в центре Парижа. По последним данным, в результате стрельбы погибли 11 человек, еще четверо находятся в критическом состоянии. …

Первичный запрос, полученный на этапе 1:

1. *Posten, Jyllands-posten, Jyllands, Herbo, Charlie, Олланд.*

Единожды расширенный запрос, после этапа 3:

2. *Перепечатать, Скандальная, Еженедельник, Карикатура, Олланд, Сатирический, Теракт, Charlie, Herbo.*

Дважды расширенный запрос после этапа 5:

3. *Журнал, Мухаммед, Сатирический, Атака, Пророк, Теракт, Париж, Карикатура, Олланд, Herbo, Charlie.*

Как видно, последний вариант включает в себя наиболее важные элементы.

4.3 Учет структуры новостной статьи в виде перевернутой пирамиды.

Рисунок 1 Перевернутая пирамида «идеального» новостного сообщения

Стратегия написания качественной новостной статьи часто опирается на структуру вида «перевернутая пирамида», Рис. 1.

В дополнительной информации часто встречается описание произошедших ранее событий по теме документа.

Учет данной структуры происходит в 2 аспектах:

1. Построение графа из документов, близких к запросу, где ребром является неявная ссылка между окончанием одной статьи и началом другой статьи, которая была опубликована ранее.
2. Повышение веса предложений, которые располагаются в верхней части новостной статьи и нижней части. Выделение нижней части происходит из-за того, что предложения оттуда часто резюмируют информацию из заголовков других статей.

Алгоритм работы первого способа учета структуры «перевернутая пирамида» выглядит следующим образом:

1. Для набора документов D происходит построение матрицы близости между окончаниями и началами документов.
2. При превышении заданного порога считается, что присутствует ссылка между документами D_i и D_j.
3. На построенном графе происходит ранжирование документов путем использования известного алгоритма LexRank [4]. Веса документов нормируются.

4. Для наиболее весомых документов производится описанная ранее операция построения расширенного запроса.

5. Итого, на выходе имеется ранжированный список документов D и набор из p новых запросов, учет которых будет осуществлен совместно с учетом временной структуры новостного сюжета.

Второй способ учета структуры перевернутой пирамиды реализован в функции ранжирования итоговых предложений, раздел 4.6.

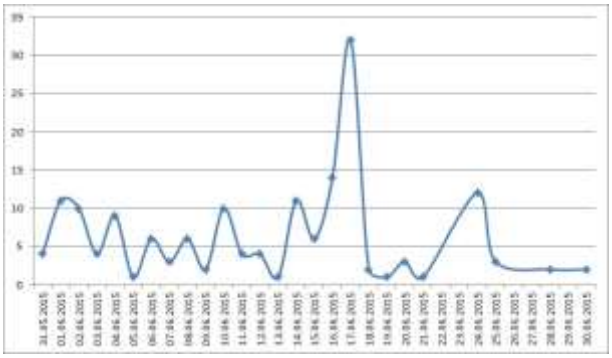

Рисунок 2 Зависимость количества публикаций новостного сюжета от времени

4.4 Учет временного характера новостных сюжетов

Так как любое событие зависит от времени, то публикации и количество публикаций тоже зависят от времени. Как пример, на Рис. 2 изображен график зависимости публикаций по событию «Землетрясение в Непале». Чтобы учесть данный фактор, для набора документов D происходит следующее:

1. Вся временная шкала события разбивается по суткам с метками $T = \{t_1, t_2, \dots, t_n\}$.
2. На основании информации о дате публикации документа каждый документ получает метку из T.
3. Происходит фильтрация дней с малым количеством публикаций. Это происходит за счет анализа количества публикаций для метки t_i к максимальному количеству публикаций в любой день и суммарному количеству публикаций.
4. На выходе имеется сортированный список, где каждый элемент имеет метку t_i из T и набор документов $D_i \in D$.

Помимо прочего, происходит отображение всех ранее построенных расширенных запросов на метки t_i из T, Рис. 3.

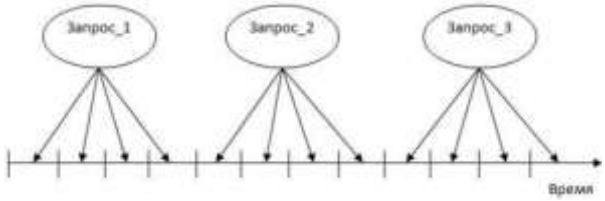

Рисунок 3 Отображение запросов на шкалу времени

4.5 Схема работы программой системы

Описанные в пунктах 4.1 факторы реализуются на различных этапах работы системы. Общая схема работы представлена на Рис. 4.

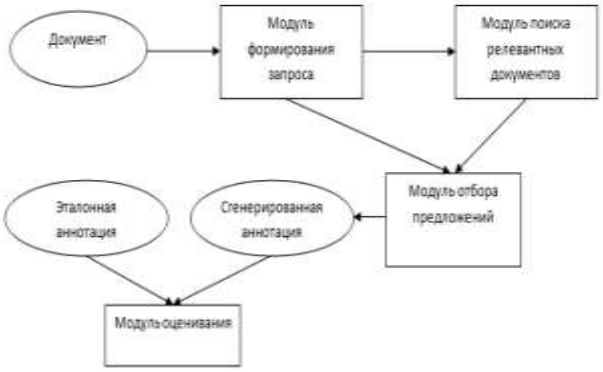

Рисунок 4 Схема работы системы

4.6 Модуль поиска релевантных документов

Поиск релевантных документов происходит путем поиска близких документов для построенного запроса на этапе формирования запроса, описанным в пункте 4.2. Использовалась поисковой машины NearIdx 8, разработанная ООО «Лаборатория информационных исследований».

4.7 Модуль отбора предложений

Данный модуль занимается непосредственно ранжированием предложений из извлеченных документов.

Ранжирование происходит модифицированной версией алгоритма MMR, которая прямо или косвенно учитывает все факторы, описанные в 4.1:

$$MMRT_{s_i^t} = INC_{s_i^t} - DEC_{s_i^t}, \quad (2)$$

где $INC_{s_i^t}$ – член, описывающий положительную составляющую формулы, которая зависит от близости предложения к запросу, веса документа, из которого взято предложение, и позиции предложения в документе;

$$INC_{s_i^t} = (1 + \alpha * I_i) * \gamma * \lambda * Sim(Q^t, S_i^t), \quad (3)$$

$$\gamma = 1 - 0.5 * \sin(\frac{i * \pi}{|D_s|}). \quad (4)$$

Параметры α и λ являются настраиваемыми параметрами алгоритма, I_i – вес документа D_s, в который входит предложение под индексом i, S_i^t – оцениваемое предложение под индексом i и с временной меткой t, Q^t – запрос, отображенный на эту временную метку, γ – слагаемое, понижающее вес предложений из середины документа.

Слагаемое $DEC_{s_i^t}$ – штрафное. Оно зависит от близости к уже извлеченным предложениям:

$$DEC_{s_i^t} = (1 - \lambda) * \max_{S_j \in S} Sim(S_j, S_i^t), \quad (5)$$

S_j – одно из извлеченных предложений, S – множество всех уже извлеченных предложений.

Обработка множества предложений, пришедших из модуля поиска релевантных документов, происходит в хронологическом порядке, на каждом этапе обрабатывается подмножество $D_i \in D$, связанное с меткой $t_i \in T$. Для каждого этапа имеется ограничение на извлечение максимум K предложений за сутки.

4.8 Мера близости

На различных этапах работы программной системы есть ряд моментов, когда вычисляется мера близости между предложениями. В работе использовались два подхода к расчету близости, использующих косинусную меру близости:

$$Sim_{cos}(S_i, S_j) = \frac{(S_i,\ S_j)}{|S_i|*|S_j|}. \quad (6)$$

Для расчета близости на этапе ранжирования предложений для них использовалось стандартное векторное представление, полученное из индекса, где вес элемента – это tf-idf.

Для расчета близости между окончаниями и началами новостных статей (на этапе построения графа) использовались вектора, полученные с помощью word2vec модели, обученной на всей коллекции документов.

5 Оценивание

5.1 Метрики оценивания

Оценивание работы системы происходило на нескольких метриках: ROUGE-1 и ROUGE-2, полноте по предложениям (8) и комбинированной метрики (9):

$$ROUGE - N = \frac{|N_A \cap N_g|}{|N_g|}, \quad (7)$$

где N_A – множество n-грамм словоформ для построенных аннотаций, N_g – для эталонных аннотаций;

$$P^{sent} = \frac{|S_A \equiv S_g|}{|S_g|}, \quad (8)$$

где S_A – множество предложений из построенных аннотаций, S_g – из эталонных аннотаций, а $\equiv$ понимается в том смысле, что в результирующем $S_A \equiv S_g$ остаются только те предложения из S_A, эквивалент которых есть в $S_{g;}$.

$$V^{comb} = 0.8 * R1 + R2 + 2 * P^{sent}, \quad (9)$$

где RN – сумма ROUGE-N и ее F-мера аналога ROUGE-NF.

5.2 Подготовка данных для процедуры оценивания

Так как для процедуры оценки качества работы системы необходим тестовый набор аннотаций, в рамках исследования были вручную подготовлены исторические справки. Процедура формирования такой коллекции происходила следующим образом:

1. На первом этапе происходил отбор ярких событий, которые активно освещались в прессе за период начала 2015 года.
2. Далее для большинства событий на информационном ресурсе interfax осуществлялся поиск соответствующего сюжета. Пример – на Рис. 5.
3. Если соответствующего сюжета на interfax нет, происходили изучение материалов по теме и формирование исторической справки на основе прочитанных документов.
4. Сюжеты просматривались в хронологическом порядке и производился отбор наиболее информативных предложений.
5. На основе отобранных предложений составлялись исторические справки, размер которых, в среднем, около 15 предложений.

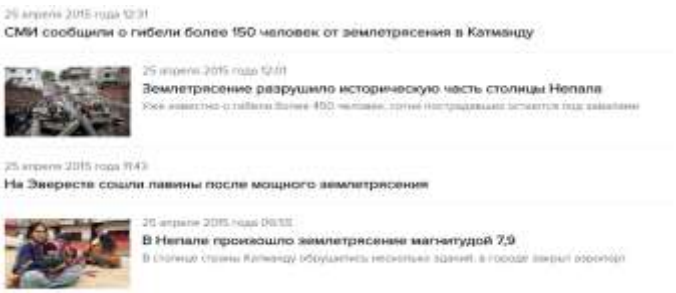

Рисунок 5 Отрывок сюжета с interfax (http://www.interfax.ru/story/151/page_3)

Итого, в результате построенная тестовая коллекция содержит в себе исторические справки по 15 событиям.

6 Результаты

Оценивались 6 конфигураций системы:

1. baseline – простой подход к аннотированию, без учета рассмотренных факторов, с использованием в качестве метода ранжирования обычного MMR;
2. querry-ex – добавление к baseline стратегии расширения запроса, но без повторного расширения запроса;
3. double-ex – querry-ex + двойное расширение запроса;
4. temporal – double-ex + учет временного характера сюжета;
5. importance – temporal + учет структуры перевернутой пирамиды;
6. full – importance + расчет близости на этапе построения графа происходит с помощью word2vec модели.

Каждая конфигурация настраивалась для получения максимального результата по всем внутренним параметрам системы (см. таблицу 2).

Результат измерений качества конфигураций можно увидеть в Таблице 3.

Таблица 2 Параметры системы

Название	Описание
SoftOr	Значение параметра soft_or_coef для поисковой машины.
KeepL	Количество лемм, выбираемых при построении первичного запроса.
KeepT	Количетсво терминов, выбираемых при построении первичного запроса.
DocCount	Значение параметра doccnt при построении расширенного запроса.
QuerrySize	Размер итогового расширенного запроса.
TopLemms	Количество наиболее значимых лемм, извлекаемых в работе

Название	Описание
	алгоритма построения расширенного запроса.
DocCount	Значение параметра doccnt при поиске релевантных документов.
MinSentSize	Минимальный размер предложения.
MaxSentSize	Максимальный размер предложения.
MinLinkScore	Минимальное значение близости окончания и заголовка документа для выявления ссылки.
Power MethodDFactor	Параметр D в алгоритме LexRank.
Power MethodEps	Параметр eps в алгоритме LexRank.
Lambda	Значение параметра λ для $MMRT$.
Alpha	Значение параметра α для $MMRT$.
MaxDaily AnswerSize	Максимальное количество предложений, извлекаемых за сутки.
Doc Boundary	Порог, позволяющий отобрать наиболее важные документы.
Init QuerrySize	Количество лемм, которые используются для повторного расширения запроса.

Таблица 3 Результаты оценивания конфигураций

Конфигурация	R1	R2	P^{sent}	V^{comb}
baseline	0.499	0.136	0.205	1.153
querry-ex	0.529	0.147	0.216	1.276
double-ex	**0.567**	**0.164**	0.260	**1.425**
temporal	0.564	0.162	0.251	1.400
importance	0.548	0.158	**0.261**	1.395
full	**0.566**	**0.162**	**0.262**	**1.433**

Полужирным шрифтом выделены по два лучших результата по каждой метрике.

Из Таблицы 3 можно сделать выводы, что наибольший вклад дало двойное расширение запроса. Факторы временной зависимости событий и структуры новостной статьи показывают неплохие результаты при совместном использовании. Также важную роль играет метрика близости, которая используется на каждом этапе решения.

Таблица 4 Отрывок исторической справки на тему падения самолета на Тайване

11.02.2015	Transasia Airways выплатит родственникам жертв авиакатастрофы на Тайване по 470 тыс.
11.02.2015	Трагедия на Тайване, одна пятая пилотов тайваньской авиакомпании Transasia не прошли тест на профпригодность.
12.02.2015	Спасатели завершили операцию по поиску жертв крушения самолёта авиакомпании Transasia Airways, который потерпел крушение 4 февраля на Тайване.
01.07.2015	Экипаж разбившегося на Тайване самолета Transasia Airways отключил двигатели после потери мощности.
02.07.2015	Самолет Transasia потерпел крушение 4 февраля на Тайване, потому что пилот по ошибке отключил работающий двигатель, когда второй двигатель заглох.

В качестве примера итоговой аннотации можно рассмотреть отрывок аннотации по упомянутому ранее событию падения самолета на Тайване в Таблице 4.

7 Заключение

Проведены исследования по тематике построения исторических справок. Были рассмотрено три фактора, которые могут влиять на качество построения аннотаций. Получены количественные и качественные результаты.

По результатам проведенных исследований оказалось, что выбор стратегии расширения запроса оказывает наибольшее влияние на качество построение аннотации подобного типа. Учет временного характера сюжета совместно с учетом структуры новостной статьи также улучшает результаты по метрикам P^{sent} и V^{comb}, что говорит о том, что данные факторы способны положительно влиять на качество построения исторических справок.

Литература

[1] Binh Tran, G., Alrifai, M., Quoc Nguyen, D.: Predicting Relevant News Events for Timeline Summaries. Proc. of the 22nd Int. Conf. on World Wide Web. ACM. pp. 91-92 (2013)

[2] Carbonell, J., Goldstein, J.: The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. Proc. of the 21st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. ACM. pp. 335-336 (1998)

[3] Dang, H.T.: Overview of DUC 2006. Proc. of the document understanding Workshop. Presented at HLT-NAACL 2006 (2006). http://duc.nist.gov/pubs/2006papers/duc2006.pdf

[4] Erkan, G., Radev, D.R.: Lexrank: Graph-based Lexical Centrality as Salience in Text Summarization. J. of Artificial Intelligence Research, (22), pp. 457-479 (2004)

[5] Hu, P., Huang, M.L., Zhu, X.Y.: Exploring the Interactions of Storylines from Informative News Events. J. of Computer Science and Technology, 29 (3), pp. 502-518 (2014)

[6] Radev, D., Jing, H., Budzikowska, M.: Centroid-based Summarization of Multiple Documents: Sentence Extraction, Utility-Based Evaluation, and User Studies. Proc. of the 2000 NAACL-ANLP

Workshop on Automatic summarization. Seattle. pp. 21-30 (2000)

[7] Radev, D., McKeown, K., Hovy, E.: Introduction to the Special Issue on Summarization. Computational linguistics, 28 (4). pp. 399-408 (2002)

[8] Shahaf, D., Guestrin, C.: Connecting Two (or Less) dots: Discovering Structure in News Articles. ACM Transactions on Knowledge Discovery from Data (TKDD). 5 (4), pp. 24-54 (2012)

[9] Tran, G., Alrifai, M., Herder, E.: Timeline Summarization from Relevant Headlines. Hanbury A., Kazai G., Rauber A., Fuhr N. (eds) Advances in Information Retrieval. ECIR 2015. Lecture Notes in Computer Science, 9022. Springer, Cham. pp. 245-256 (2015). doi: 10.1007/978-3-319-16354-3_26

[10] Yan, R. et al.: Evolutionary Timeline Summarization: a Balanced Optimization Framework via Iterative Substitution. Proc. of the 34th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. Beijing, China. July 24– 28, 2011. ACM. pp. 745-754 (2011). doi: 10.1145/2009916.2010016

[11] Абрамова, Н.Н., Абрамов, В.Е.: Автоматическое составление обзорных рефератов новостных сюжетов. Труды 9-ой Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2007, Переславль-Залесский, Россия. сс. 131-141 (2007)

[12] Алексеев, А.А., Лукашевич, Н.В.: Автоматическое порождение обновления к аннотации новостного кластера. Труды 12й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия. сс. 81-91 (2010)

[13] Браславский П., Густелев, В.: Система автоматического реферирования новостных сообщений на основе машинного обучения. Труды Девятой Всерос. науч. конф. – RCDL'2007, Переславль-Залесский, Россия. Сс. 142-147 (2007)

[14] Добров, Б.В., Павлов, А.М.: Исследование качества базовых методов кластеризации новостного потока в суточном временном окне. Труды 12-й Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия. сс. 287-295 (2010)

[15] Лукашевич, Н.В., Добров, Б.В.: Автоматическое аннотирование новостных кластеров на основе тематического представления. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). М.: РГГУ, Вып. 8 (15), сс. 299-305 (2009)

[16] Тарасов, С.Д.: Исследование и оптимизация параметров алгоритма Manifold Ranking на основе метрики автоматической оценки качества обзорного реферирования ROUGE-RUS. Труды XI Всерос. науч. конф. «Электронные библиотеки. Перспективные методы и технологии, электронные коллекции». Петрозаводск. сс. 86-93 (2009)

[17] Шаграев, А.: Автоматическое аннотирование новостного потока. Семинар: Natural Language Processing (автоматическая обработка естественного языка). Яндекс. 26.11.2011 (2011). https://www.slideshare.net/NataliaOstapuk/ ss-10380447?ref=http://nlpseminar.ru/lecture54/