OSTIS-2013
(Open Semantic Technologies for Intelligent Systems)

УДК 004.912

АВТОМАТИЗАЦИЯ РЕФЕРИРОВАНИЯ НОВОСТНЫХ ИНТЕРНЕТ-ТЕКСТОВ

Солошенко А.Н., Орлова Ю.А., Дмитриев А.С.

Волгоградский государственный технический университет,
г. Волгоград, Россия
nastyasolan@gmail.com
yulia.orlova@gmail.com
dmitriealeksan@yandex.ru

Реферирование текста в последние годы получило значительную актуальность в связи с развитием Internet и каталогов информационных ресурсов. Данная статья посвящена проблеме составления обзорных статей по текстам интернет-новостей. В статье рассматривается построение и структура новостного текста, этапы его анализа, принципы и методика составления рефератов по новостным статьям.
Ключевые слова: автоматизация реферирования, информационные ресурсы, новостные интернет-тексты, методы реферирования.

Введение

Ежедневно в сети Интернет появляются миллионы страниц новостных текстов. Сотни тысяч людей описывают события и явления, берут интервью, создают посты в блогах. Так, например, только число ежедневных сообщений в Twitter приблизилось к отметке 400 миллионов записей в день, или почти 4,5 тысячи сообщений в секунду. Обработка информационного материала вручную требует колоссальных человеческих ресурсов, трудовых и временных затрат, поэтому и возникла задача создания методики для автоматизации реферирования статей. Потребности в средствах автоматического реферирования и аннотирования испытывают: корпоративные системы документооборота, поисковые машины и каталоги ресурсов Internet, автоматизированные информационно-библиотечные системы, каналы вещания, службы рассылки новостей и др.

В России, как и за рубежом, данному направлению исследований придается очень большое значение [Браславский П. и др., 2008; Гридина Е.А., 2011; Заболеева-Зотова А.В. и др., 2008; Орлова, Ю.А. и др., 2011; Тарасов С.Д., 2008]. В настоящее время известно большое количество систем автоматического реферирования текстов. Среди отечественных это TextAnalyst, Content Analyzer, технологии АОТ, RCO, редактор Microsoft Word, МедиаЛингва Аннотатор, система Яндекс Новостей, среди зарубежных - Extractor , QDA Miner (включающий пакеты WordStat и Simstat), системы Inxight Summarizer (компонент поискового механизма AltaVista), Intelligent Text Miner (IBM). Однако многие ведущие системы разработаны на Западе и ориентированы исключительно на обработку западноевропейских языков, что делает их непригодными для анализа текстов на русском языке. Возможности некоторых отечественных систем ограничены выделением и выбором оригинальных фрагментов из исходного документа и соединением их в короткий текст на основе только статистических методов реферирования.

Для упрощения изучения существующих и создания новых интернет-текстов необходимо представление статей в сжатом виде, но с сохранением их смысла. Поэтому для достижения наиболее оптимального результата необходимо совместить несколько алгоритмов реферирования. Сперва рассмотрим специфику новостных статей, их структуру, а затем уже непосредственно методику автореферирования.

1. Структура новостных интернет-текстов

Структура новостного текста может варьироваться в зависимости от задач и определяется авторами индивидуально. Проанализировав ряд статей, представленных на известных новостных сайтах, таких как сайт газеты

Комсомольская правда, Коммерсант.ru, Эксперт и др., построим обобщенную структуру текста новости (рисунок 1).

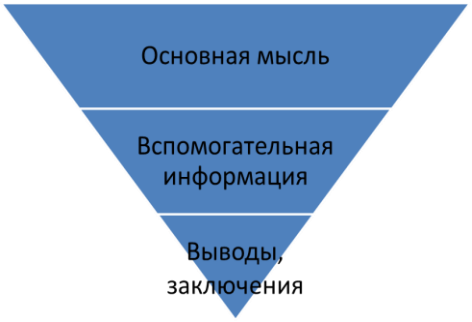

Рисунок 1 - «Перевернутая пирамида» - структура новостных текстов

Как правило, для такого типа текстов характерна пирамидальная структура. 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. [Электронный ресурс 2]

Данная структура соблюдает все каноны построения новостных статей: в тексте отражена только одна главная тема, статья состоит из введения (основная мысль), развития (вспомогательная полезная информация) и заключения. Как правило, заголовок сразу «цепляет» внимание сутью текста, а продолжение вызывает интерес.

Поэтому, зная особенности построения интернет-текстов, можно определить, какие проблемы ставит автор текста, выделить ключевые факты новостной статьи, определить объекта и субъекта новости. Далее в работе будет рассмотрена методика автоматизации реферирования таких текстов.

2. Обзор существующих систем автоматического реферирования

На международном рынке представлено множество программных продуктов, которые позволяют создавать авторефераты для статей. Так как планируется разработать программный продукт для автоматизации реферирования новостных интернет-текстов на русском языке, то наибольший интерес представляют отечественные аналоги. Их можно оценить по следующим практически важным критериям: поддержка русского языка, входные форматы данных, работа со словарями, функция выделения ключевых фраз, составления связного реферата, свободный доступ (freeware лицензия), функция задания коэффициента сжатия текста, удобство интерфейса.

Рассмотрим наиболее известные системы с точки зрения выделенных критериев.

Content Analyzer предназначен для анализа содержания тематических Web-страниц в реальном времени, динамического выделения списков ключевых слов и словосочетаний, построения автореферата текста документа. Определяет и рассчитывает следующие основные характеристики: частота и вес термина/словосочетания в документе, усредненный вес слов словосочетания/предложения. Работа программы основана на использовании алгоритма TF-IDF для выделения ключевых слов.

TextAnalyst позволяет построить дерево терминов, установив связи между ключевыми понятиями, выделяемыми из текста, предоставляет функцию автоматического реферирования текста - формирования его смыслового портрета в терминах наиболее информативных фраз. Основан на использовании нейронных семантических сетей. [Гридина Е.А., 2011]

Такая технология анализа и поиска текстовой информации как RCO (Russian Context Optimizer) позволяет производить морфологический, синтаксический и семантический анализ текста, их анализ и классификацию, автоматическое реферирование. Технология АОТ же (Автоматическая Обработка Текста) состоит из модулей графематического, морфологического, семантического анализа текста. Обе системы работают со словарями и тезаурусами.

Необходимо также упомянуть о пакете MS Word 2003, 2007, в котором реализована функция AutoSummarize. Предусмотрена возможность указать, какую часть от объема документа должно составлять резюме, однако работа осуществляется только с текстами на английском языке. Отдельно стоит сказать о коммерческой системе Яндекс Новости, позволяющей автоматически группировать данные в новостные сюжеты и составлять аннотации статей на основе кластера новостных документов.

Как наименее функциональный инструмент для реферирования текстов можно отметить редактор MS Word. Составление автореферата основано здесь на поиске наиболее часто встречающихся слов, поэтому не гарантируется получения осмысленного или удобочитаемого текста. TextAnalyst как программный продукт, основанный на алгоритмах создания семантических сетей, проявляет наибольшую гибкость при работе с текстом и составлении его смыслового портрета, как, впрочем, и Content Analyzer. Поэтому наибольшие перспективы в области автоматизации реферирования видятся в развитии взаимодействия и совмещения статистических алгоритмов извлечения ключевых слов и алгоритмов формирования семантической связности блоков текста.

3. Методика реферирования интернет-текстов

Существуют статистические, позиционные, логико-семантические методы реферирования. Каждый метод имеет свои достоинства и недостатки, поэтому стоит использовать их комбинированно.

Решение задачи реферирования текста можно разбить на несколько этапов.

Сначала осуществляется предварительная обработка текста, включающая:

- Перевод текста во внутреннее представление - графематический анализ (выделение из исходного текста лексических единиц - определение границ предложений, слов)
- Нормализацию слов с помощью морфологического анализа. Цель состоит в выделении основ слов, то есть словоформ с отсечёнными окончаниями
- Построение частотного словаря словоформ (как раз после объедения грамматических форм одного и того же слова)
- Вычисление весовых коэффициентов слов, предложений в абзацах в зависимости от функциональной структуры. Выбор наиболее весовых терминов, блоков
- Определение темы текста. На основе заголовка формирование для каждой темы списка предложений, которые ее характеризуют (определение веса каждого предложения путем суммирования частот появления в нем слов и словосочетаний, определяющих темы).

Далее осуществляется выделение семантически связных групп предложений, удаление незначащих абзацев/ предложений в соответствии с заданным коэффициентом сжатия. По достижении заданного порога сокращения текста процесс реферирования останавливается, получаем конечный реферат, необходимый пользователю.

Методика реферирования представлена ниже в виде блок-схемы (рисунок 2).

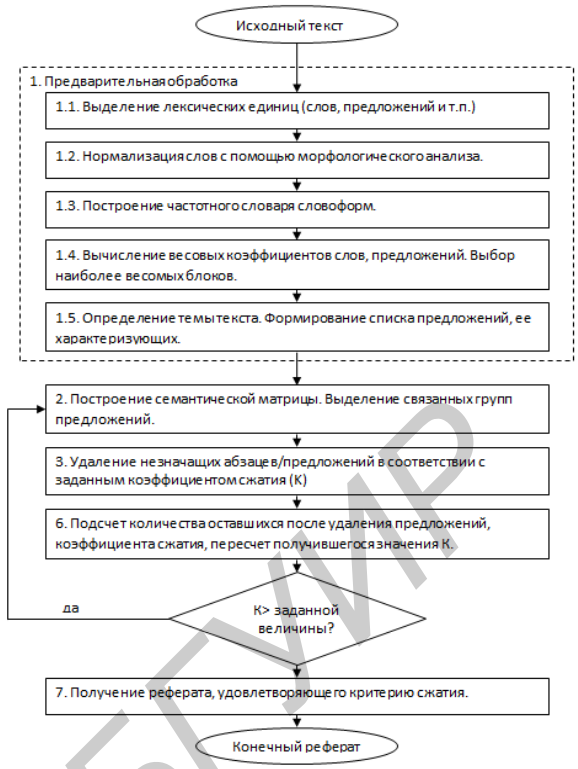

Рисунок 2 - Методика составления реферата

Таким образом, можно выделить следующие необходимые для реферирования текста компоненты:

- модуль графематического анализа;
- модуль морфологического анализа;
- модуль выделения ключевых слов;
- модуль составления связного реферата.

На начальном этапе вырабатывается информация, необходимая для дальнейшей обработки морфологическим и прочими модулями. В задачу графематического анализа входят: разделение входного текста на слова и разделители, выделение предложений из входного текста, абзацев, заголовков и др. Основан на правилах для естественного языка.

Следующим этапом является морфологический анализ, цель которого - построение морфологической интерпретации слов входного текста. Все методы можно поделить на словарные и вероятностно-статистические (без использования словаря). В вероятностно-статистических методах анализ словоформы построен на правилах поиска и сочетания единиц разных лексиконов (лексиконы префиксов, суффиксов, окончаний, основ, баз, корней), приводящий к унификации гипотез. Недостатками являются большой объем лексиконов, плохая работа на малой выборке, отсутствие точных лингвистические методов. Словарный же метод основан на подключении словаря, тезаруса, дает максимально полный анализ словоформы.

Поэтому для данного блока был выбран словарный метод на базе словаря АОТ «Диалинг». Он написан на C++, обладает развитой системой добавления новых слов. Для каждого слова входного текста выдается множество морфологических интерпретаций следующего вида:

– морфологическая часть речи (существительное, глагол, местоимение и т.д.);

– лемма – каноническая форма лексемы (например, существительное в именительном падеже единственного числа, или глагол–инфинитив);

– множество наборов граммем – элементарных описателей, относящих слово-форму к какому–либо морфологическому классу (мр, жр, ср - мужской, женский, средний род; од, но - одушевленность, неодушевленность; ед, мн - единственное, множественное число; им, рд, дт, вн, тв, пр, зв - падежи: именительный, родительный, дательный, винительный, творительный, предложный, звательный).

Структура словаря и алгоритм морфологического анализа представлены на рисунках ниже (рисунки 3 и 4 соответственно).

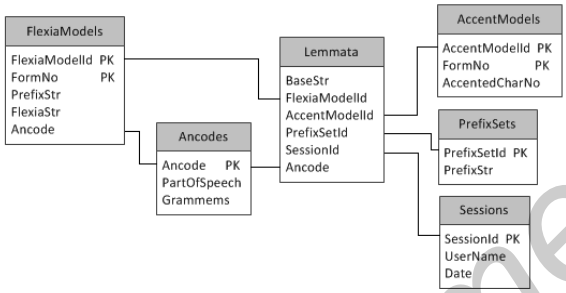

Рисунок 3 - Структура словаря в виде реляционной схемы

Так, таблица Lemmata содержит перечень всех лемм данного словаря, для каждой леммы даны ее свойства: псевдооснова слова (BaseStr), ссылка на набор окончаний (FlexiaModelId), ударений (AccentModelId), приставок (PrefixSetId), на общие граммемы данной леммы (Ancode) (может быть пустым).

Таблица FlexiaModels содержит перечень возможных окончаний всех лемм. Поле PrefixStr содержит префикс данной словоформы, FlexiaStr – окончание словоформы, Ancode – морфологическую интерпретацию данной словоформы (PartOfSpeech содержит часть речи, а поле Grammems набор граммем). [Сокирко А.В., 2004]

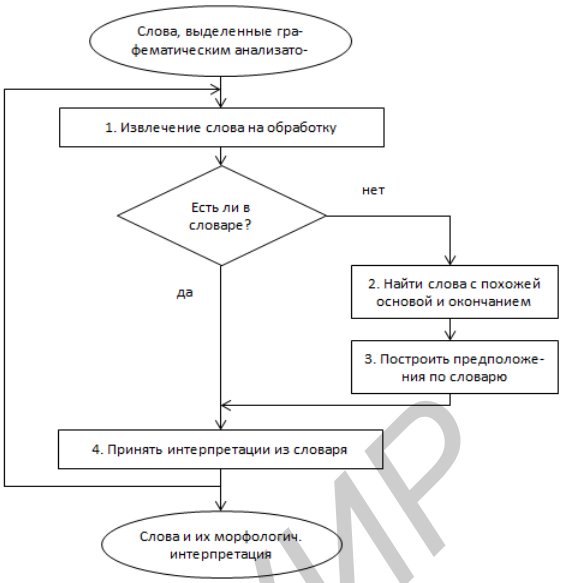

Рисунок 4 - Алгоритмическая модель морфологического анализа

Для выделения ключевых слов в настоящее время разработано множество методов. Будем использовать алгоритмы TF-IDF и C-Value.

TF-IDF (от англ. TF – term frequency, IDF – inverse document frequency) – статистическая мера, используемая для оценки важности слова в контексте документа. Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции. TF – отношение числа вхождения некоторого слова к общему количеству слов документа.

$$TF = \frac{n_i}{\sum_k n_k}. \qquad (1)$$

IDF – инверсия частоты, с которой слово встречается в документе, ее учёт уменьшает вес широкоупотребительных слов.

$$IDF = \log \frac{|D|}{|(d_i \supset t_i)|}, \qquad (2)$$

где $|D|$ — количество документов в корпусе; $|(d_i \supset t_i)|$ – количество документов, в которых встречается t_i (когда $n_i \neq 0$).

Мера TF-IDF является произведением двух сомножителей: TF и IDF.

Метод C-Value позволяет сопоставить каждой извлечённой из текста именной группе значение терминологичности, вычисляемое по формуле:

$$C\text{-}Value(a) = \begin{cases} \log_2 |a| \cdot freq(a), \text{ если не вложен} \\ \log_2 |a| \cdot freq(a) - \frac{1}{P(T_a)} \cdot \sum_{b \in T_a} freq(b) \end{cases}, \qquad (3)$$

где a – кандидат в термины, $|a|$ – длина словосочетания в количестве слов, $\text{freq}(a)$ – частотность a, Ta – множество словосочетаний, содержащих a, $P(Ta)$ – количество словосочетаний, содержащих a.

Путём сортировки списка кандидатов в термины по убыванию значения C-value можно получить список ключевых фраз, наиболее адекватных исходному тексту. [Браславский П. и др., 2008]

Один из самых известных методов реферирования - метод составления выдержек, предполагает акцент на выделение характерных фрагментов (как правило, предложений). Его основу составляет процедура назначения весовых коэффициентов для каждого блока текста в соответствии с такими характеристиками, как расположение этого блока в оригинале, частота появления в тексте, частота использования в ключевых предложениях, а также показатели статистической значимости (ключевые слова). Однако создание итогового документа в данном случае – просто соединение выбранных фрагментов.

Для автореферирования необходимо определять семантическую близость, связность между предложениями текста [Заболеева-Зотова А.В. и др., 2009; Орлова Ю.А., 2007]. Поэтому были предложены следующие методы: выделение фрагментов «объект-действие-субъект» (с использованием POS-таггера АОТ для определения частей речи), реализация которого основана на использовании модификации алгоритма LexRank (более подробно описан в [Dragomir R., etc., 2004]), а также ранжирование связных тематических структур с использованием алгоритма Manifold Ranking, где связная структура текста описывается при помощи матриц [Тарасов С.Д., 2008].

Итак, структурная модель предлагаемого решения, включающая графематический, морфологический модули, модуль выделения ключевых фраз и составления связного реферата приведена на рисунке 5.

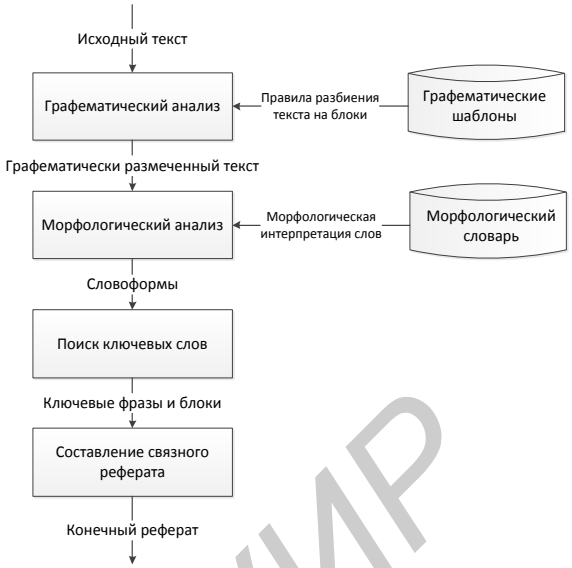

Рисунок 5 - Модель реферирования текста

4. Пример работы алгоритма

Рассмотрим работу алгоритма составления реферата текста на примере небольшого новостного сообщения, взятого из Интернет, заглавием которого является строка *«Медики выяснили, в какое время суток человек чувствует себя счастливым»:*

«Ученые из США потратили два года на то, чтобы узнать - в какое время суток человек чувствует себя наиболее счастливым? Ученые анализировали сообщения Твиттера. Всего изучались 600 млн. сообщений от 2,5 млн пользователей из более 80 государств. В ходе исследования выяснилось, что пользователи оказались более счастливы утром. К вечеру эмоциональное настроение людей постепенно портится.

Кроме того, выяснилось, что самое хорошее настроение у людей бывает в зимние месяцы: с декабря по январь, сообщает epidemiolog.ru. Однако, надо уточнить: счастливый период заканчивается, скорее всего, не в конце января, а 16-го. Это третий понедельник января, который считается самым депрессивным днем в году.

К такому выводу пришел британский психолог Клифф Арнэлл из Университета Кардиффа. Он вывел сложную формулу, которая учитывает отвратительную, как правило, погоду, то, что праздники прошли и надо браться за работу, надо как-то планировать свое будущее. Не радует и то, что до следующих праздников далеко.

Путем деления и умножения специальных коэффициентов, отражающих вышеозначенное состояние человеческой души в середине января, ученый якобы и получил дату, соответствующую третьему понедельнику» [Электронный ресурс 1]

В результате применения алгоритма предполагается получить следующие результаты:

Ключевые фразы: надо, ученые, выяснилось, настроение, людей, человек, Твиттер, узнать, время суток, чувствует, наиболее счастливым, третий понедельник.

Автореферат текста:

«Ученые из США потратили два года на то, чтобы узнать - в какое время суток человек чувствует себя наиболее счастливым? В ходе исследования выяснилось, что пользователи оказались более счастливы утром. Однако, надо уточнить: счастливый период заканчивается, скорее всего, не в конце января, а 16-го. Это третий понедельник января, который считается самым депрессивным днем в году. К такому выводу пришел британский психолог Клифф Арнэлл из Университета Кардиффа. Он вывел сложную формулу, которая учитывает отвратительную, как правило, погоду, то, что праздники прошли и надо браться за работу, надо как-то планировать свое будущее».

Заключение

Таким образом, совмещение нескольких алгоритмов составления реферата позволит улучшить качество реферирования. В условиях нынешнего информационного века, с огромным количеством новостных сообщений в сети интернет, применение таких технологий является необходимостью, так как значительно ускоряет обработку повседневной информации.

Работа частично поддержана Российским фондом фундаментальных исследований (проекты 12-07-00266, 12-07-00270).

Библиографический список

[Электронный ресурс 1] Главные события политической жизни/ KP.RU [Электронный ресурс]. – Режим доступа: http://www.kp.md/daily/25982.5/2915545/.

[Браславский П. и др., 2008] Браславский П., Соколов Е. Сравнение пяти методов извлечения терминов произвольной длины// Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 7 (14).– М.: РГГУ, 2008. - С. 67-74.

[Гридина Е.А., 2011] Гридина Е.А. Анализ алгоритмов автоматического реферирования текстов/ Восточно-Европейский журнал передовых технологий. – 2011.

[Заболеева-Зотова А.В. и др., 2008] Заболеева-Зотова, А.В. Автоматизация семантического анализа текста технического задания / Заболеева-Зотова А.В., Орлова Ю.А. // Системные проблемы надёжности, качества, мат. моделирования, информ. и электронных технологий в инновационных проектах: (Инноватика-2007): матер. междунар. конф. и Рос. науч. школы / Рос. акад. надёжности [и др.]. - М., 2007. - Ч.1. - С. 78-79

[Заболеева-Зотова А.В. и др., 2009] Заболеева-Зотова, А.В. Formalization of text analysis of a technical specification / Заболеева-Зотова А.В., Орлова Ю.А. // Congress on intelligent systems and information technologies (AIS-IT`09), Divnomorskoe, Russia, September, 3-10 : proc. / Южный федеральный ун-т [и др.]. - М., 2009. - Vol. 4. - С. 62.

[Орлова, Ю.А. и др., 2011] Орлова, Ю.А. Обзор современных автоматизированных систем распознавания эмоциональных реакций человека / Орлова Ю.А., Розалиев В.Л. // Изд. ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах". Вып. 10 : межвуз. сб. науч. ст. / ВолгГТУ. - Волгоград, 2011. - № 3. - С. 68-72.

[Орлова Ю.А., 2007] Орлова, Ю.А. Подсистема предварительной обработки текста / Орлова Ю.А., Заболеева-Зотова А.В. // Технологии Microsoft в теории и практике программирования: тр. IV всерос. конф. студ., аспир. и мол. уч., 2-3 апр. 2007: Центральный регион, Москва: [тез. докл.] / Моск. авиац. ин-т (гос. техн. ун-т) [и др.]. - М., 2007. - С. 187-188.

[Сокирко А.В., 2004] Сокирко А.В. Морфологические модули на сайте www.aot.ru// Диалог, 2004 г.

[Тарасов С.Д., 2008] Тарасов С. Д. Алгоритм ранжирования связных структур в задачах автоматического составления обзорных рефератов новостных сюжетов.// RuSSIR'2008, труды Второй Российской конференции молодых ученых по информационному поиску. – Таганрог: Изд-во ТТИ ЮФУ, 2008. – С. 90-100.

[Электронный ресурс 2] Шардаков Д. Структура текста: как ее создать и сделать текст удобным для восприятия/ Shard-Copywriting.Ru [Электронный ресурс]. – Режим доступа: http://shard-copywriting.ru/copywriting-basics/glavnyiy-printsip-kopiraytinga-perevernutaya-piramida.

[Dragomir R. et al., 2004] Dragomir R. Radev, Gunes Erkan. LexRank: Graph-based Lexical Centrality as Salience in Text Summarization// Journal of Artificial Intelligence Research, 2004.

AUTOMATIZATION OF INTERNET NEWS TEXT ABSTRACTING

Soloshenko A.N., Orlova Y.A., Dmitriev A.S.

Volgograd State Technical University, Volgograd, Russia

nastyasolan@gmail.com

yulia.orlova@gmail.com

dmitriealeksan@yandex.ru

Text abstracting received in recent years a considerable urgency in connection with development of Internet and catalogs of information resources. This article is devoted to a problem of an automated abstracting – drawing up reviews in texts of Internet news. Construction and structure of the news text, stages of its analysis, principles and methods of drawing up of author's abstracts under news articles are considered.

Keywords: automated abstracting, information resources, Internet news texts, methods of abstracting.

This work was partially supported by the Russian Foundation for Basic Research (projects 12-07-00266, 12-07-00270).