УДК 004.91

ИЗВЛЕЧЕНИЕ СТРУКТУРИРОВАННОГО НОВОСТНОГО СООБЩЕНИЯ ИЗ ВЕБ-СТРАНИЦ ПРИ ИСПОЛЬЗОВАНИИ ДОПОЛНИТЕЛЬНОЙ ИНФОРМАЦИИ RSS

Казенников А.О., E-mail: kazennikov@gmail.com
Соловьев И.В., д.т.н., профессор, E-mail: soloviev@mirea.ru
МГТУ МИРЭА, Москва, Россия

Аннотация В статье представлен алгоритм автоматического извлечения структурированного новостного сообщения из веб-страницы полного текста новостного сообщения и соответствующего ему элемента RSS-канала новостного потока. Алгоритм основан на выборе целевого узла для заданного поля в DOM-модели (DocumentObjectModel) HTML-документа на основе статистических методов классификации. Основное отличие от существующих алгоритмов состоит, во-первых в использовании признаков, характерных для новостных сообщений, во-вторых в двустадийной схеме работы и значительно расширенной модели признаков для классификации. Результатом работы алгоритма является процедура извлечения содержания, устойчивая к изменению верстки и расположению целевой информации на веб-странице.
Ключевые слова: новостной поток, веб-страница, извлечение содержания, DOM-модель.

EXTRACTION OF THE NEWS MESSAGE STRUCTURE FROM NEWS WEB-PAGES USING ADDITIONAL INFORMATION FROM RSS NEWS FEED

Kazennikov A.O., E-mail: kazennikov@gmail.com
Soloviev I.V., D.ofSci., prof., E-mail: soloviev@mirea.ru
MSTU MIREA, Moscow, Russia

Abstract The paper introduces an algorithm for automatic data extraction from web-pages with support of additional data, extracted from RSS (really simple syndication). The algorithm is based on selection of the target node of the DOM representation of HTML-document using machine learning methods. This algorithm allows to automatically extract data fields based on given pattern. The resulting extractor constructed with this method is robust to layout and position changes of the target information block on this page.
Keywords: news stream, web-page, content extraction, Document Object Model.

Большинство информационных агентств и СМИ в интернете предоставляют доступ к новостям только в виде RSS-каналов с кратким описанием новостных сообщений и HTML-страниц с полным текстом новости. Такой формат представления удобен для отображения у конечного пользователя — протокол RSSпозволяет автоматически получать обновления ленты новостей, а веб-страница новости хорошо оформлена и удобна для чтения. Однако оно не рассчитано на автоматическую обработку сообщений.RSS является машиночитаемым форматом, однако в нем отсутствует полный текст новостного сообщения, а HTMLпредназначен в первую очередь для корректного отображения в брау-

зере пользователя, а не для автоматической обработки. Таким образом, необходимо преобразование из одного типа информации в другой[1, 2].

Для автоматического анализа новостных сообщений необходимо преобразовать исходное представление элемента RSS-ленты и HTML-страницы в структурированный формат данных, содержащий следующие основные элементы новостного сообщения:

- Заголовки новостного сообщения;
- Дата и время создания и обновления новости;
- Текст новости;
- Редакторские ссылки новостного сообщения.

В настоящее время для решения этой задачи применяются составленные вручную правила выделения элементов, основанные на XPath-выражениях[3] и селекторах CSS (CSS – cascadingstylesheet). Этот подход очень трудоемок, поскольку требует составления правил выделения для каждого новостного источника, а также ручного контроля и поддержки работоспособности разработанных правил, поскольку они могут перестать работать при любом изменении структуры или дизайна веб-страницы, даже самом незначительном.

Существуют автоматические средства удаления разметки[3,4,5,6,7,8], основной задачей которых является выделение основной содержательной части документа. Современные веб-страницы имеют сложную структуру, в которой присутствуют навигационные элементы, рекламные блоки, динамически подгружаемая информация. Однако эти средства не предполагают возможности наличия априорной информации, которая может быть использована для улучшения качества извлечения основной части и не поддерживают выделение отдельных полей данных.

В статье представлен алгоритм, который сочетает автоматические методы выделения основного содержания веб-страницы и выделение структурных элементов новостного сообщения. Основное отличие представленного алгоритма от существующих, а также от разработанного авторами в [9], заключается в возможности извлечения отдельных элементов веб-страницы, таких как заголовок, дата публикации, основной текст новости и другие, при этом используя дополнительную информацию, полученную из элемента RSS-ленты. Алгоритм позволяет извлекать а) заголовок новостного сообщения; б) дату и время публикации и обновления; в) полный текст новостного сообщения; г) редакторские ссылки на сообщения схожей тематики. При этом процедура извлечения новостного сообщения должна устойчива к изменениям веб-страницы с полным текстом новостного сообщения.

Существующие решения

Рассмотрим два основных подхода к автоматическому извлечению содержательной части веб-страницы:

- методы для выделения основной части страницы [4,5,6,7,8];
- методы выявления повторяющихся (шаблонных) элементов страницы[6,10,11,12].

Большинство современных методов идентификации и выделения основной части страницы основаны на статистических свойствах блоков HTML-страницы. Принципиальная схема построения этих алгоритмов следующая: страница разбивается на блоки, которые потенциально могли бы быть основным содержимым страницы. Затем оценивается каждый из этих блоков и в итоге выбирается наиболее вероятный блок, который считается основным текстом страницы. Так, в работе[4] предложен метод на основе машинного обучения для автоматического определения блоков, не несущих существенную содержательную информацию.

Метод состоит из следующих шагов: выделяются блоки на основе тегов div, span, table, td. Затем, часть этих блоков размечается экспертами. Полученная выборка используется для обучения классификатора блоков на основе дерева принятия решений на основе алгоритма C4.5[13]. Алгоритм C4.5 последовательно конструирует дерево решений, разделяя текущую обучающую выборку на две части по значению некоторого признака. На каждом шаге используется признак, который для данного множества примеров максимально различает целевые классы. Для обучения использовались признаки геометрических размеров блоков, ссылочная информация, структурных данных. В [5] представлен метод на основе частичной визуализации веб-страницы. Метод преобразует веб-страницу в дерево, узлами которого являются визуально связанные блоки. В этом случае блок – поддерево DOM-представления документа, визуально составляющее единое целое. Для идентификации блока основного содержания страницы оценивается разнородность содержимого данного блока.

В результате ранжирование блоков производится не только по их содержательным характеристикам, но и по пространственным (таким как расположение или размер) элементов. В работе [6] представлен достаточно простой метод выделения содержания на основе плотности информации. За основу взято количество текстовых символов на один тег. Подход основан на предположении, что основная часть страницы содержит много текста и мало html-тегов, что справедливо для новостей и блогов. Обратная ситуация наблюдается в рекламных и навигационных блоках. Для этого в работе [7] предложен усовершенствованный метод, учитывающий статистику ссылок. Для отсечения неинформативных блоков используется пороговый подход – если плотность выше некоторого зна-

чения, то блок считается содержательным. Такой подход эффективен для определения основного содержания страницы. Однако он может вызывать ошибки на содержательной части, в которой имеется высокая плотность тегов. Например, к таким частям относятся списки интересов пользователя в социальных сетях, предыдущие места работы, поля со сложной структурой, в которых каждый элемент оформлен отдельным тегом и др.

Принципиально другим подходом к извлечению информации из веб-страниц являются подходы выделения шаблонных (устойчивых) элементов[4,7,8]. Подход предполагает, что, во-первых, большинство современных сайтов использует автоматическую генерацию своих основных страниц. Это справедливо практически для всех отраслей: новостных сайтов, блогов, форумов, интернет-магазинов и т.п. Во-вторых, метод предполагает, что возможен сбор большого количества страниц с однотипным содержанием с одного ресурса. В-третьих, предполагается, что полученная коллекция документов содержит только один уникальный шаблон для всех документов. Таким образом, кластеризация коллекции по разным шаблонам не входит в постановку задачи определения элементов. Основным принципиальным недостатком подходов этого типа является то, что при достаточной эффективности определения основного содержания страницы, метод не фильтрует изменяющиеся неинформативные блоки, например ссылки на похожие новости на новостном сайте, комментарии к новости и динамически встраиваемую рекламу.

В работе [8] для определения шаблонных элементов используются блоки, которые содержат функциональные элементы страницы одного назначения: навигационный блок, ссылочный блок, меню, основной текст, и т.п. Каждый документ коллекции разбивается на блоки. Для каждого блока оценивается его гомогенность на основе статистических характеристик: количества ссылок в каждом блоке, их плотности на единицу текста и тега и др. Затем блоки кластеризуются. Предполагается, что если большое число блоков разных страниц попали в один кластер, то они являются шаблонными. В работе [10] используется разбиение страницы по тегам, которые часто используются для формирования различных блоков информации. Это теги: table, div, span, td, dd. Затем предлагается выделять блоки не несущие информацию с помощью подсчета статистики слов в каждом блоке по всей коллекции.

Другой подход предложен в работе [11], авторы которой используют метод под названием SiteStyleTree. Метод заключается в построении общего DOM-дерева множества элементов. Если некоторый элемент уже есть в SST, то у него увеличивается счетчик. После построения SST для каждого узла оценивается разнообразие стилей представления и содержания. Меньшее разнообразие указывает на то, что элемент является шаблонным. В результате возможно выявление часто повторяющихся элементов, которые можно считать неинформативными. В работе [12] используют совместную процедуру

идентификации шаблонов и индексирования. Метод заключается в том, что во время индексирования страница разбивается на блоки. Затем происходит кластеризация на основе стиля и позиции блока. Предполагается, что похожие кластеры разных документов являются шаблонной частью, которую можно удалять.

Абсолютное большинство методов извлечения информации базируется на машинном обучении для идентификации основной части страницы или шаблонного элемента. При идентификации основного содержания веб-страницы используются в основном методы классификации, в то время как для определения шаблонных элементов – методы кластеризации.

Кроме того, большинство методов явно использует свойства верстки современных веб-страниц:

- логическое разделение способа отображения веб-страницы и ее структуры;
- выделение каждого информационного элемента отдельным блоком (HTML теги div, span, table, td, li);
- активное использование шаблонов стилей CSS.

Алгоритм извлечения структурированного новостного сообщения из веб-страницы и элемента RSS-ленты

Как было отмечено ранее, основное отличие задачи извлечения структурированной информации состоит в использовании дополнительного источника информации – элемента RSS-ленты, соответствующей новостному сообщению.

На основе элемента RSS-ленты можно извлечь следующую информацию:

- Дата публикации новостного сообщения
- Заголовок новостного сообщения
- Кратное описание новостного сообщения

Однако, использование этой информации имеет ряд ограничений:

- Заголовок RSS-элемента может не соответствовать заголовку полной текстовой версии новостного сообщения
- Краткое описание может отсутствовать
- Может наблюдаться рассинхронизация текстов RSS-элемента и полного текста новости, например, при редакторских правках.

Таким образом, эта информация не является надежной и соответствующие структурные элементы должны извлекаться из полной версии новостного сообщения.

Для конструирования процедуры извлечения каждого структурного элемента (заголовка, полного текста и редакторских ссылок) по аналогии с работой [9] воспользуемся методами машинного обучения.

Задача машинного обучения ставится следующим образом. Пусть дано множество размеченных веб-страниц Xи RSS-элементы Z, в которых отмечены DOM-элементы x, соответствующие i-тым структурным элементам новостного сообщения. Тогда необходимо составить такое множество функций s, что:

$$x = \arg\max_{s} EsExFF$$

Будем искать функции s среди функций вида:

$$s\,ExF = wfExF,$$

Где w – некоторый вектор весов, а $fExF$ –функция преобразования DOM-элемента x и его окружения в вектор в пространстве R.

Таким образом, решение рассматриваемой задачи разбивается на две подзадачи:

1. Процедуру поиска w
2. Определение функции $fExF$

Для решения первой задачи воспользуемся методом SVM[13]:

$$w = \arg\min_{w} \frac{1}{2}w^{2} + C\sum_{i=1}^{n}\max(0, 1 - y_i w x_i),$$

Где C – гиперпараметр, определяющий степень обобщения модели.

Функция извлечения $fExF$ конструируется следующим образом. Каждому признаку g (список которых определен ниже) и его значению v, ставится в соответствие некоторое натуральное число k. Тогда:

$$fExFxiz = \begin{cases} 1, \exists x: g(x) = v_z = i \\ 0, \text{в др. случаях} \end{cases}$$

Для классификации из каждого DOM-элемента извлекаются следующие признаки:

- Признаки текста, содержащегося в данном элементе:
  - Длина текста
  - Число слов
  - Средняя длина слова
  - Количество предложений
  - Средняя длина предложений
- Положение DOM-элемента в DOM-дереве:
  - Число непосредственных потомков текущего элемента
  - Длина пути до вершины DOM-дерева
  - Число родственных элементов (потомков непосредственного родителя данного узла)
  - Наличие вложенных блоков в данный DOM-элемент

- Признаки самого DOM-узла:
  - Название тега
  - Список CSS-классов данного DOM-элемента

Основным отличием предлагаемого алгоритма является его существенное расширение за счет типов признаков:

- биграммы, тригаммы и квадриграммы знаков названий CSS-классов и идентификатора данного элемента;
- *n*-граммы непосредственного родителя данного элемента;
- сочетания классов и идентификатора данного элемента и его родителя;
- средняя доля буквенных, цифровых, пунктуационных и пробельных символов в рассматриваемом элементе.
- Наличие слов из RSS-элемента

Эти признаки позволят выделять не только основное содержимое веб-страницы, но и составные элементы содержимого. Расширение модели необходимо, поскольку решаемая задача более сложная, чем выделение основного содержания страницы. Кроме того, для решения этой задачи можно воспользоваться коллекцией документов, как и в задаче выделения шаблонных элементов, тогда как в постановке задачи выделения основного содержимого такая коллекция документов отсутствует.

Другим важным отличием предлагаемого алгоритма является использование двух стадий для извлечения структурных элементов. На первой стадии выделения очередного структурированного элемента выполняется выбор с помощью обученного классификатора элементов, которые могли бы содержать значение этого структурированного элемента. На второй стадии этот же классификатор используется для удаления из выбранных элементов «лишних» - тех, которые содержат данные, распределение которых сильно отличается от целевого. В результате, улучшается точность извлечения структурных элементов.

Предлагаемый алгоритм содержит две фазы: обучающую и рабочую.

Обучающая фаза состоит из следующих шагов:

1. Разметка достаточно большой коллекции веб-страниц, содержащих информацию, которую необходимо извлекать.

2. Формирование обучающей выборки на основе каждого DOM-узла каждой страницы размеченной коллекции.

3. Конструировании е классификатора DOM-элементов на основе представленной модели признаков.

Рабочая фаза состоит из двух стадий.

На первой стадии из обрабатываемой веб-страницы извлекаются DOM-узлы, потенциально содержащие необходимую информацию. В общем случае извлекаются только DOM-узлы, соответствующие блокам языка HTML: теги div, span, tabletd, li, h1-h6. Затем из этих DOM элементов производится выбор DOM-элемента с наибольшей вероятностью, содержащего необходимую информацию на основе классификатора, построенного в фазе обучения.

На второй стадии производится так называемая «отрицательная» выборка — из оставшихся элементов удаляются те составные части, которые не содержат нужных данных. Блок-схема предложенного алгоритма представлена на Рис. 1.

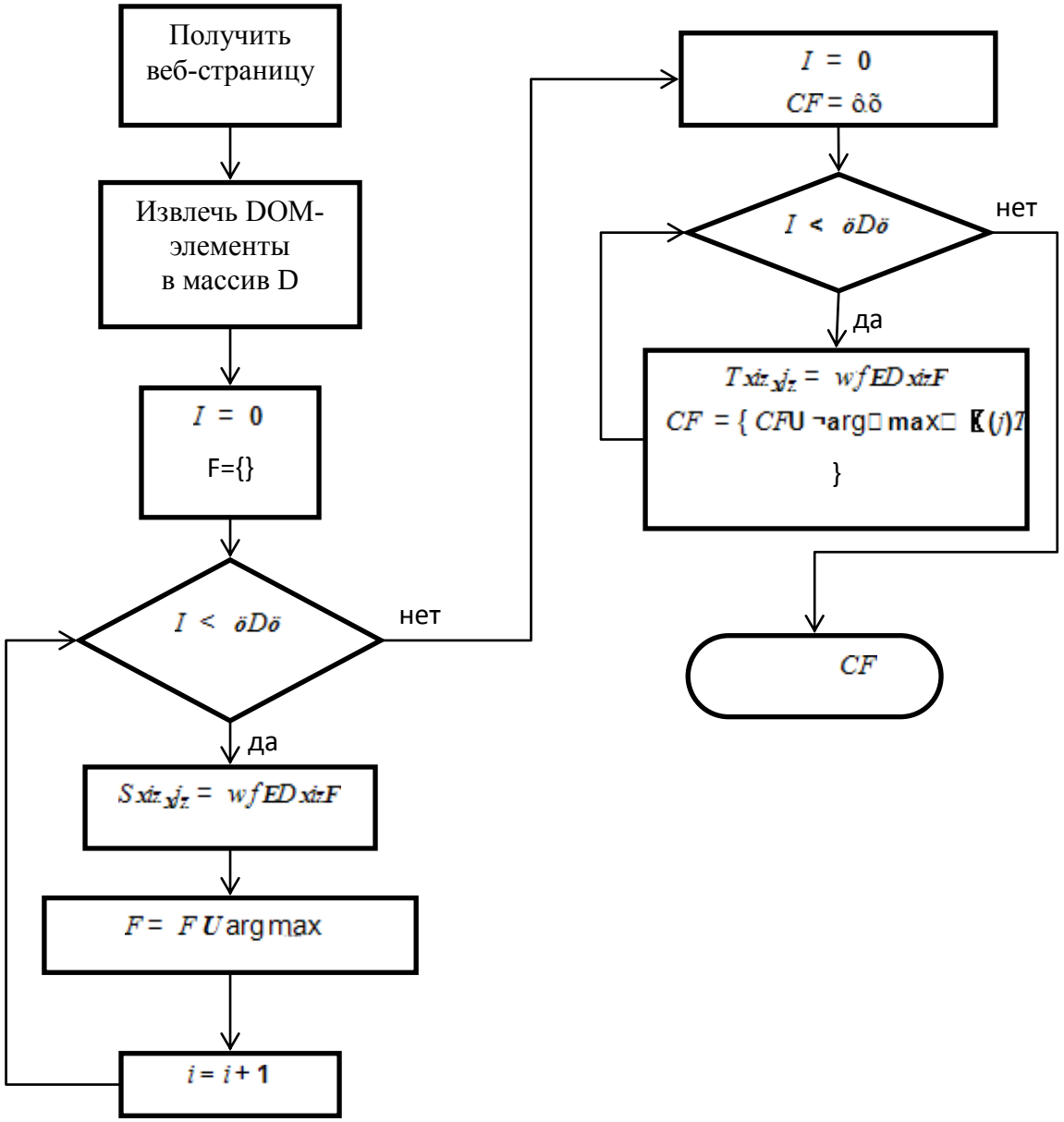

Рис 1. **Блок-схема алгоритма извлечения структурированного новостного сообщения**

Такой алгоритм работы позволяет корректно выделять даже те структурные элементы, в которых в DOM-элементах которых содержится лишняя информация. Так, например, в новостях сайта «Лента.ру» может содержаться врезка на материал по сюжету новости. Одностадийный алгоритм может выделить только новостной текст целиком, вместе с такой врезкой. Разработанный алгоритм может определить такую врезку и удалить ее на второй стадии работы.

Экспериментальная оценка алгоритма

Для экспериментальной оценки эффективности разработанного алгоритма были вручную размечены 2000 новостных сообщений каждого из сайтов «Лента.ру», «NewsRu.COM», «Interfax» и «РИА Новости». В этой выборке размечались:

- Заголовок новости
- Основной текст новости
- Редакторские ссылки и их текст

Проверка эффективности осуществляется следующим образом. На материале сайта «Лента.ру» выполняется обучающая фаза алгоритма. Затем выполняется рабочая фаза на новостных сообщениях сайтов «NewsRu.com», «Interfax» и «РИА Новости» и оцениваются следующие параметры:

- Точность
- Полнота
- F-мера (гармоническое среднее точности и полноты)

Для проверки устойчивости алгоритма обучение дополнительно проводилось на выборке «РИА Новости», а результаты оценивались по всем остальным выборкам.

Проводилось три серии экспериментов. Первая серия была ориентирована на то, чтобы определить влияние дополненной модели признаков на точность выделения составных элементов страницы. Полученный классификатор использовался только для выделения целевых элементов, но не для удаления лишних (одностадийный вариант работы). Вторая серия была направлена на определение эффекта от использования двустадийной схемы выделения составных элементов. И, наконец, третья серия оценивала эффективность предложенного алгоритма выделения основных элементов новостного сообщения для задачи кластеризации новостных потоков. В разработанном автором алгоритме кластеризации новостных потоков [14,15] предполагается, что новостные источники предоставляют поток сообщений в пригодном для обработки формате.

Кроме того, проводилась оценка вклада каждой группы признаков:

1. Текстовые признаки
2. Признаки положения DOM-узла

3. Признаки DOM-узла
4. Предложенное расширение модели
5. Наличие слов из RSS-элемента в оцениваемом узле

Результаты первой серии экспериментов представлены в таблице 1 (жирным обозначены максимальные значения). Как видно результатов, предложенное расширение модели признаков положительно влияет на качество извлечения информации. При этом эффективность предложенной модели признаков не зависела от обучающей коллекции и улучшала качество извлечения в обоих случаях. F-мера для предложенной модели признаков значительно выше значения для базовой модели (группа признаков, обозначенная как «1+2+3»).

Таблица 1. **Результаты экспериментов по оценке влияния модели признаков на качество извлечения информации**

Группы признаков	**Обучающая Коллекция**	**Точность**	**Полнота**	**F-мера**
1	Лента.ру	0.881	0.892	0.885
1 + 2	Лента.ру	0.893	0.891	0.89
1 + 2 + 3	Лента.ру	0.912	0.923	0.915
1 + 2 + 3 + 4	Лента.ру	**0.937**	0.920	0.926
1 + 2 + 3 + 4 + 5	Лента.ру	**0,951**	0,933	**0,941**
1	РИА Новости	0.898	0.901	0.895
1 + 2	РИА Новости	0.895	0.903	0.895
1 + 2 + 3	РИА Новости	0.914	0.921	0.915
1 + 2 + 3 + 4	РИА Новости	0.921	0.942	0.932
1 + 2 + 3 + 4 + 5	РИА Новости	**0,935**	**0,952**	**0,942**

Результаты второй серии экспериментов представлены в таблице 2. При проведении этой серии экспериментов использовалась разработанная модель признаков. Как и в первой серии, оценивалось влияние двустадийной схемы при разных обучающих коллекциях. Из результатов экспериментов видно, что двустадийная схема фильтрации дополнительно улучшает качество извлечения и также не зависит от обучающей коллекции. Улучшение результата достигается за счет тонкой фильтрации.

Таблица 2. **Результаты экспериментов по оценки влияния двустадийной схемы на качество извлечения информации**

Тип схемы	**Обучающая Коллекция**	**Точность**	**Полнота**	**F-мера**
Одностадийная	Лента.ру	**0,951**	0,933	**0,941**
Двустадийная	Лента.ру	**0.962**	**0.933**	**0.946**
Одностадийная	РИА Новости	**0,935**	**0,952**	**0,942**

Двустадийная	РИА Новости	**0.941**	**0.952**	**0.944**

Сводный график результатов экспериментов представлен на Рис. 2. Разработанная модель и алгоритм позволяют значительно улучшить качество выделения информации из веб-страниц.

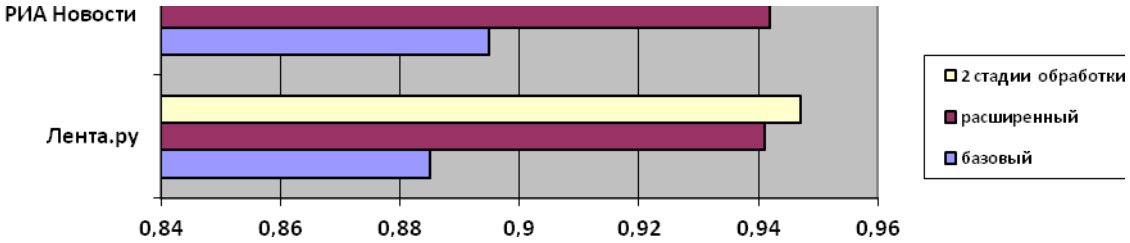

Рис 2. **Сводная диаграмма результатов экспериментов по коллекциям данных (F-мера)**

Третья серия экспериментов была проведена для оценки влияния разработанного алгоритма извлечения структурных элементов на задачу кластеризации новостных потоков. В этой серии экспериментов для корректности сравнения результатов с работой [12] использовалась коллекция «РИА Новости», поскольку в [14] для оценки качества кластеризации использовался материал сайта «Лента.ру». Результаты этой серии представлены в таблице 3.Предложенный алгоритм показал результаты, сравнимые с составлением правил для извлечения информации вручную. Разница результатов поF-мере составляет 0.006, что является достаточно малой величиной, которой на практике при массовом добавлении новостных источников можно пренебречь.

Таблица 3. **Результаты применения разработанного алгоритма для задачи кластеризации новостных потоков.**

Тип метода	**Точность**	**Полнота**	**F-мера**
Ручной метод извлечения новостного сообщения	**0,912**	**0,891**	**0.901**
Автоматический метод извлечения новостного сообщения	0.909	0.890	0.899

Выводы

В статье представлен эффективный алгоритм извлечения структурного новостного сообщения из веб-страницы с полным текстом и соответствующим ему RSS-элементом. Он позволяет извлекать не только основную содержательную часть документа, но и его отдельные структурные элементы.

При проведении экспериментов проверена переносимость результирующей процедуры извлечения, обученной на одном источнике на другие. Алгоритм продемонстрировал хорошую адаптивность в рамках сходного характера материала для извлечения информации.

Произведена экспериментальная оценка разработанного алгоритма. Оценено влияние расширенной модели признаков и двустадийной схемы для извлечения структурных элементов. Предложенный алгоритм показал существенное улучшение качества извлечения.

Кроме того, произведено экспериментальное сравнение ручного и автоматического подходов извлечения структурных элементов для задачи кластеризации новостного потока. Разработанный алгоритм показал результаты, сравнимые с ручным методом.

Разработанный алгоритм может использоваться для подключения источников новостных сообщений к системам анализа новостных сообщений.

Список литературы

1. Соловьев И.В., Цветков В.Я. О содержании и взаимосвязях категорий «информация», «информационные ресурсы», «знания» // Дистанционное и виртуальное обучение. – 2011. - №6 (48) - с.11-21

2. Иванников А.Д., Тихонов А.Н., Соловьев И.В., Цветков В.Я. Инфосфера и инфология. – М: ТОРУС ПРЕСС, 2013. -176с.

3. Sun F., Song D., Liao L. DOM Based Content Extraction via Text Density. Inproc. OfSIGIR’2011, Beijing, China, 2011

4. Kohlschutter C., Fankhauser P., Nejdl W. Boilerplate detection using shallow text features. In Proc. Of WSDM’10 pp 441-450, 2010

5. Cai D., Yu S., Wen J., Ma W. Extracting content structure for web pages based on visual representation. In Proceedings of APWeb’03 pp 406-417, 2003

6. Gottron D. Content code blurring: A new approach to content extraction. In Proc of DEXA’08, pp 29-33, 2008

7. Vieira K., da Silva A., Pinto N. et al. A Fast and Robust Method for Web Page Template Detection and Removal. In Proc. Of CIKM’06, 2006

8. Chen L., Ye S., Li. X. Template Detection for large scale search engines. In Proc. Of SAC'06, pp 1094-1098, NY USA, 2006

9. Казенников А.О., Трифонов Н.И. Алгоритм автоматического извлечения информации из сообщений новостного потока. Информатизация образования и науки. - 2014. № 1 (21). С. 111-119.

10. Lin S., Ho J. Discovering informative content blocks from web documents. In Proc. Of SIGKDD'02, pp 588-593, NY USA, 2002

11. Weninger T., Hsu W.H., Ma W. Learning block importance models form web pages. In Proc. Of WWW'04, pp 971-980, NY USA, 2004.

12. Bar-Yossef Z., Rajagopalan S. Template Detection via data mining and its applications. In proc. Of WWW'2002, pp 580-591, 2002

13. Wu X., Kumar V., Quinlan J. Top 10 algorithms in data mining. Knowledge Information Systems Vol. 14, pp 1-37, 2008.

14. Казенников А.О. Анализ новостных потоков на основе информационного поиска и компьютерной лингвистики. Информатизация образования и науки №4(16) 2012, стр.155-164

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