Интернет-журнал «Науковедение» ISSN 2223-5167 http://naukovedenie.ru/

Том 8, №6 (2016) http://naukovedenie.ru/vol8-6.php

URL статьи: http://naukovedenie.ru/PDF/73TVN616.pdf

Статья опубликована 02.12.2016

Ссылка для цитирования этой статьи:
Редькин О.К. Подходы к представлению текста для определения типа источника информационного сообщения // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 8, №6 (2016) http://naukovedenie.ru/PDF/73TVN616.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ.

УДК 004

Редькин Олег Константинович[1]
ФГБОУ ВО «Московский технологический университет», Россия, Москва
Аспирант
E-mail: o.k.redkin@gmail.com

Подходы к представлению текста для определения типа источника информационного сообщения

Аннотация. В работе рассматривается задача выделения из множества интернет страниц, составляющих новостные и информационные веб-ресурсы, наиболее информационно насыщенных страниц. Для решения поставленной задачи предлагается условное разделение интернет страниц по выполняемым ими функциям на информационные и навигационные. Рассматриваются особенности формирования страниц обоих типов на примере новостных ресурсов. Для решения задачи классификации страниц предлагается два подхода к представлению размещённого на них текстового содержимого. В рамках первого подхода для выявления разделяющего свойства был осуществлён анализ содержимого страниц обоих типов на наличие у них основных свойств текста (в лингвистическом понимании), в результате которого было выделено свойство глобальной связанности текста, характерное лишь для страниц одного типа. В рамках второго предлагаемого подхода учитывается особенность использования наиболее информативных частей речи при формировании навигационных страниц и в качестве разделяющего признака предлагается использовать частоту появления различных частей речи в текстах страниц обоих типов. Для обоих предложенных подходов разработаны соответствующие модели представления текстового содержимого интернет страниц и осуществлена их проверка на применимость для решения поставленной задачи средствами математической статистики.

Ключевые слова: компьютерная лингвистика; обработка текста; извлечение данных; интернет страница; классификация интернет страниц; текстовое содержимое интернет страниц

Растущее с каждым годом количество данных, генерируемых человеком и различными устройствами, а также увеличивающееся доля содержащейся в них потенциально полезной информации привело к созданию широкого класса автоматизированных систем обработки информации и поддержки принятия решений. Подобные системы позволяют автоматизировать процесс поиска, сбора, хранения, передачи, обработки и защиты

[1] 111675, г. Москва, ул. Лухмановская, д. 27, кв. 317

информации, повышая тем самым эффективность работы человека с большими объёмами данных.

Одной из разновидностью описываемых систем являются информационные системы (ИС), основной задачей которых является своевременное обеспечение надлежащих людей надлежащей информацией (1). ИС разрабатываются для решения конкретных задач в рамках заданной предметной области и их эффективность во многом зависит от характеристик данных, поступающих на вход и используемых системой, а также применяемых методов их обработки и преобразования.

Первичной задачей, решаемой при разработке ИС, является задача сбора целевых данных. Тип и форма входных данных зависит от области применения системы и используемых источников: данные могут быть структурированными, слабоструктурированными или неструктурированными и быть представлены в виде чисел, текста, документов, изображений, видео, аудио и т.д. При этом большинство методов анализа данных предполагают работу только со структурированными данными: данные, не соответствующие этому требованию, должны быть преобразованы к структурированной форме. Текстовые данные, являющиеся наиболее используемыми, информативными и удобными для восприятия человеком, относятся к классу неструктурированных данных.

Сбор и анализ данных в сфере информационного взаимодействия является одной из областей применения ИС: полученная в результате их работы информация используется для принятия решений специалистами конкретных областей. Одной из возможных областей применения такого рода систем является выявление информационных сообщений, содержащих явные или скрытые угрозы безопасности общества и государства, а также источники и каналы их распространения. Под информационным сообщением понимается сообщение, содержащее сведения, информацию о каких-то фактах или событиях.

Задача сбора и анализа данных в среде информационного взаимодействия предполагает осуществление постоянного мониторинга информационного пространства с целью выявления сообщений, которые могут содержать информацию, потенциально полезную для специалистов. Эффективность анализа поступающих данных во многом зависит от используемых для их получения целевых источников: их количества, разнообразия, а также формы представления информационных сообщений.

В зависимости от целевого назначения ИС, используемые источники могут содержать как структурированные, так и неструктурированные данные. Зачастую в качестве целевого источника используются общедоступные интернет ресурсы, среди которых особое внимание уделяется веб сайтам. Методы извлечения данных из подобных источников могут быть условно разделены на две группы: использующие предоставляемые ресурсом инструменты и основанные на анализе html-разметки источника (2), (3). К первой группе относятся интерфейсы прикладного программирования (API), позволяющие осуществлять доступ к данным и предоставляемые разработчиками ресурсов, а также средства, представляющие данные источника в форматах, удобных для автоматической обработки: RSS, Atom, XML и т.д. В том случае, если целевой источник не предоставляет подобных средств, для извлечения данных используются методы веб-скрейпинга (web scraping) - технологии получения данных из веб страниц (4). Одним из наиболее используемых методов веб-скрейпинга является анализ структуры страниц исходного ресурса, определяемой html-разметкой, в результате которого специалистом формируется набор правил, позволяющих осуществлять доступ к целевым данным, размещенным в определённых узлах полученной структуры веб-документа.

Основным достоинством описанного метода является высокое качество и скорость извлечения данных, размещённых на веб-страницах. Однако этот подход обладает рядом

недостатков. Для каждого добавляемого в систему целевого источника специалист должен составлять собственный набор правил для разбора содержимого составляющих его страниц, а при изменении структуры страниц веб-ресурса соответствующие поправки должны быть внесены и в существующие наборы правил.

Перечисленные недостатки описанного метода связаны с его сильной зависимостью от структуры разметки интернет страниц обрабатываемых источников и они могут быть устранены за счёт разработки и использования более универсального метода, учитывающего особенности формирования структуры самой страницы, а не её разметки. В рамках данной работы в качестве общедоступных источников информационных сообщений рассматривались русскоязычные сайты интернет СМИ, поскольку именно новостные сайты обычно выступают первичным источником и распространителем новостных информационных сообщений.

Проведённый анализ структуры интернет-страниц, составляющих различные веб-ресурсы, показал, что они могут быть условно разделены в зависимости от выполняемых ими функций на навигационные и информационные страницы. Задачей навигационных страниц является предоставление пользователям удобных средств навигации по ресурсу для доступа к интересующей его информации. Страницы этого типа преимущественно состоят из блоков, включающих в себя ссылку перехода на связанную страницу и краткую информацию о данных, содержащихся на ней – такие блоки в дальнейшем мы будем называть навигационными блоками. Страницы, содержащую основную и наиболее полную информацию, задачей которых является её представление пользователям, мы будем называть информационными страницами. Под информационным блоком в дальнейшем мы будем понимать часть интернет страницы, содержащей информацию по заданной теме, основной задачей которого является конечное представление данных пользователю. Как правило, информационные страницы содержат не только информационный блок, но и набор навигационных блоков, позволяющих пользователю осуществлять переходы на страницы ресурса, связанные с текущей. Таким образом, навигационные страницы используются для организации пользователям удобного доступа к информационным страницам, которые содержат данные по конкретной теме.

Описанное разделение страниц, составляющих интернет ресурсы, характерно для большинства сайтов, однако наиболее явно оно прослеживается на новостных (СМИ) или информационных (энциклопедии, каталоги документов, информационные порталы) ресурсах. Так, на сайтах интернет СМИ навигационными страницами, как правило, являются главные страницы сайтов, а также начальные страницы разделов новостей: они содержат краткое описание и ссылки на страницы с новостями, относящимся к текущей категории. Эти страницы позволяют пользователю, опираясь на заголовки и краткое описание, выбрать интересующие его новости и перейти на страницу, содержащую полный текст заинтересовавшей его новости (информационную страницу). При решении задач, требующих анализа текстовых данные, получаемых из общедоступных источников, зачастую данные содержащиеся в навигационных блоках навигационных страниц представляют меньший интерес, чем данные связанных с ними информационных страниц, поскольку они являются менее содержательными и валидными (информационно насыщенными). Таким образом, одной из задач при анализе данных, получаемых из общедоступных источников, является определение типа источника информационного сообщения: использование информационных страниц в качестве источника информационных сообщений позволит повысить информационную насыщенность исходных данных, а исключение из обработки навигационных страниц позволит уменьшить нагрузку на систему обработки.

Следует отметить, что говоря о текстовом содержимом интернет страниц, помимо значимых с точки зрения получения новой информации частей (содержательные блоки), как

правило, страницы обоих типов содержат незначимые для анализа части: блоки рекламы, заголовки ссылок перехода на другие страницы, служебную и навигационную информацию и т.д. Такие блоки далее мы будем называть незначимыми блоками. В зависимости от используемого способа представления текстового содержимого и методов его обработки, наличие незначимых блоков может повлиять на результаты определения типа страницы.

Для исключения из обработки подобных частей могут быть использованы различные подходы, одним из которых является разработка шаблонов разбора html-разметки страницы, которые содержали бы универсальные для всех видов источников правила для выделения ключевых блоков на странице. Однако, учитывая то, что в рамках одного ресурса подобные малоинформативные блоки имеют одинаковую структуру, схожее содержимое и используются на страницах обоих типов, возможны ситуации, когда их наличие не будет оказывать значительного влияния при определении типа страницы и могут рассматриваться в качестве информационного шума. В зависимости от результатов, полученных в процессе тестирования разрабатываемых в рамках данной работы подходов, текстовое содержимое незначимых блоков может быть исключено из текста страницы на этапе предобработки или же не исключаться вовсе. В дальнейшем, говоря о текстовом содержимом страницы, мы будем иметь в виду текст, размещаемый в содержательных блоках интернет страницы.

Как было отмечено ранее, навигационные страницы состоят из множества навигационных блоков, позволяющих осуществлять пользователям навигацию по страницам интернет ресурса. Информационные страницы помимо информационного блока, содержащего целевые данные по заданной теме, также включают и навигационные блоки, используемые для перехода на связанные страницы текущего или сторонних ресурсов. Таким образом, необходимо определить разделяющие признаки содержимого интернет страниц, позволяющие определить тип конкретной страницы.

Анализ содержимого информационных страниц новостных веб-ресурсов показал, что текст информационной страницы может быть условно разделён на три части: заголовок, лид и основное содержимое. Заголовок выполняет информативную и контактную функции: он должен содержать в себе информацию о предмете сообщения и побудить читателя к его прочтению. Лидом (lead) называется первый абзац новости, содержащий основную информацию из новостного сообщения; обычно он состоит из нескольких предложений, в которых формулируются проблема и вывод. Лид позволяет передать смысл сообщения в сжатом виде и его задачей является привлечение к прочтению основного содержимого информационного сообщения, если освещаемая тема интересна читателю. Заголовок и лид формируют заголовочный комплекс, и на них приходится порядка 70% от общего смысла новости (5). Основное содержимое является наиболее объёмной частью информационного сообщения и содержит детали об описываемом событии.

При формировании содержимого навигационных блоков обычно используются заголовки со связанных с ними информационных страниц. Помимо заголовков многие новостные ресурсы включают в навигационный блок лид (или его часть) соответствующей новости.

В общем виде, текстовое содержимое интернет страницы может быть представлено в следующем виде:

$$T = \{w_1, w_2, ..., w_i, ..., w_n\}, \quad (1)$$

где: T – текстовое содержимое, w_i – слово, занимающее i-ую позицию в тексте.

Содержимое информационной страницы может быть представлено в следующем виде:

$$T_{\mathrm{i}}^{\mathrm{inf}} = \{\mathrm{H}_i, \mathrm{L}_i, \mathrm{B}_i\}, \quad (2)$$

где: $T_{\mathrm{i}}^{\mathrm{inf}}$ – текст i-ой информационной страницы, $\mathrm{H_i}$ – заголовок i-ой страницы, $\mathrm{L_i}$ – лид i-ой страницы, $\mathrm{B_i}$ – основное содержимое новости.

Текст навигационной страницы включает заголовки связанных с ней информационных страниц, иногда сопровождаемыми лидами:

$$T^{nav} = \bigcup_{i=1}^{k} H_i[\mathrm{L}_i], \quad (3)$$

где: T^{nav} – текст навигационной страницы, H_i – заголовок i-ой новости, $[\mathrm{L}_i]$ – лид i-ой новости, который может присутствовать или нет, в зависимости от конкретного ресурса, k – число информационных страниц, связанных с навигационной. Обычно, при формировании навигационных страниц новостных ресурсов, лиды либо дополняют заголовки всех новостей, либо не используются вовсе.

Учитывая формулы (2) и (3), текст навигационной страницы может быть представлен в виде объединения текстов информационных страниц, связанных с ней, за исключением основного содержимого:

$$T^{nav} = \bigcup_{i=1}^{k} (T^{inf} / T^{body}), \quad (4)$$

где T^{body} – часть текста, удаляемая из исходного текста информационной страницы для формирования его представления на навигационной странице.

Для решения задачи классификации интернет-страниц и выделения информационных страниц, основываясь на их текстовом содержимом, нами должна быть разработана модель представления текста, формализующая особенности формирования содержимого страниц обоих типов и позволяющая различать их.

Подход, учитывающий различия в структуре текстового содержимого страниц разных типов

С лингвистической точки зрения, для того, чтобы текстовое содержимое интернет страницы могло считаться единым текстом, оно должно обладать рядом определённых свойств – свойств текстуальности. Существует несколько подходов к определению свойств текста: они различаются как по числу выделяемых признаков, так и отношением к обязательности наличия и значимости конкретных признаков (6), (7). В результате проведённого анализа наиболее используемых подходов, нами были выбраны свойства текста, выделяемые в рамках различных подходов, как наиболее значимые: глобальная связанность, локальная связанность, информативность (8), членимость и прагматичность (9).

Проведённый анализ ряда русскоязычных интернет СМИ на наличие выбранных признаков у текстов, размещённых на информационных и навигационных страницах этих ресурсов, позволил определить глобальную связанность как свойство, характерное только для информационных страниц. Под глобальной связанностью (далее – связанность), или цельностью, понимается связанность содержания текста на всём его протяжении (8). Отсутствие данного свойства в текстах навигационных страниц связана с особенностями их формирования: содержимое множества навигационных блоков связано внутри себя, но если рассматривать их как единое целое, то части получившегося текста не будут связаны между

собой. Таким образом, определения наличия у текстового содержимого интернет страницы свойства связанности позволит отнести её к классу информационных страниц.

Связанность текста предполагает наличие между частями текста смысловых, коммуникативных и структурных связей; в тексте она достигается за счёт использования трёх различных видов связей: грамматических, логико-семантических и стилистических. Для организации каждого вида связи используются определённые методы и средства связи, которые могут быть выявлены на разных уровнях анализа текста: при этом признаки, характерные для конкретного вида связи, как правило, выявляются на нескольких уровнях анализа (таблица 1).

Таблица 1

Признаки наличия текстообразующих связей (разработано автором)

Тип связи	Признак наличия связи	Уровень анализа
Грамматическая связанность	Согласование словоформ и синтаксических конструкций	Синтаксический, морфологический
	Деепричастные обороты	Синтаксический, морфологический
	Параллельные синтаксические конструкции	Синтаксический
	Неполные синтаксические конструкции	Синтаксический
Логико-семантическая связанность	Полный тождественный повтор	Морфологический
	Частичный лексико-семантический повтор	Морфологический
	Тематический повтор	Семантический
	Синонимический повтор	Семантический
	Антонимический повтор	Семантический
	Дейктический повтор	Морфологический
	Логико-смысловые отношения	Морфологический
Стилистическая связанность	Повторяющиеся стилистические приёмы	Синтаксический, семантический

Для формализации свойства связанности текста нами были выбраны средства связи, определяемые на морфологическом уровне анализа текста. Данный выбор обусловлен сравнительной простотой морфологического анализа (по сравнению с синтаксическим и семантическим) текста и существованием готовых автоматизированных решений для его проведения.

Наличие признаков связи, определяемых на морфологическом уровне представления текста, проявляется в виде использования слов с определёнными морфологическими характеристиками, при этом частота появления конкретного формального признака связанности в тексте может быть выражена через частоту использования соответствующих ему слов. На выбранном уровне представления могут быть выявлены следующие признаки: использование деепричастий, полных тождественных (повторение словоформ, имеющих одинаковый корень), частичных лексико-семантических (использование разных словоформ с одним корнем) и дейктических повторов (использование дейктических слов: местоименных существительных, наречий, числительных), а также наличие логико-смысловых отношений (использование союзов) (10). Наличие в тексте описанных признаков не позволяет однозначно определить, является ли текст связанным: они указывают на наличие определённых связей между частями текста, которые могут не сохраняться на всём протяжении текста. Однако,

частое использование в тексте средств связи значительно увеличивает вероятность того, что текст является связанным, поэтому в дальнейшем мы будем называть описанные признаки формальными признаками связанности текста. В рамках данной работы рассматриваются тексты, состоящие из предложений, соответствующих принятым в языке грамматическим нормам, которые могут быть прочитаны и понятны пользователем. Для таких текстов наблюдается следующая зависимость: чем реже в тексте используются различные средства связи, тем менее связаны его части, т.е. текст в целом является менее связанным.

В рамках решаемой задачи связанность текста будет рассматриваться нами как наличие у него определённого набора формальных признаков:

$$E_T = \{s_1, ..., s_m\},\ s_i \in S_E, \tag{5}$$

где: E_T – связанность текста T, s_i – i-ый формальный признак связанности, S_E – множество формальных признаков связанности, m – число учитываемых формальных признаков связанности.

Введём двухместную операцию сравнения двух текстов, для определения более связанного из них:

$$F_{comp}(T_1, T_2) = F_{comp}(E_{T_1}, E_{T_2}) = \sum_{i=1}^{m} comp(As_i^{T_1}, As_i^{T_2});$$

$$comp(As_i^{T_1}, As_i^{T_2}) = \begin{cases} 1, As^{T_1} > As^{T_2} \\ 0, As^{T_1} = As^{T_2} \\ -1, As^{T_1} < As^{T_2} \end{cases};\quad As^T = \frac{Ns^T}{n^T}, \tag{6}$$

где: $F_{comp}(T_1, T_2)$ – операция сравнения степени связанности текстов T_1 и T_2, E_{T_j} – степень связанности текста T_j, As_i^{Tj} – отношение числа вхождений i-ого формального признака связанности (Ns^T) в текст T_j к общему числу слов в нём (n^T), $comp(As^{T_1}{}_i, As^{T_2}{}_i)$ – операция сравнения частоты вхождения i-ого формального признака связанности в тексты T_1 и T_2 соответственно.

Результатом операции сравнения $F_{comp}(T_1, T_2)$ является число - сумма результатов поэлементного сравнения соответствующих вхождений формальных признаков в текстах T_1 и T_2. Полученный результат может быть интерпретирован следующим образом:

$$\begin{cases} E_{T_1} > E_{T_2}, & F_{comp}(T_1, T_2) > 0 \\ E_{T_1} \approx E_{T_2}, & F_{comp}(T_1, T_2) = 0 \\ E_{T_1} < E_{T_2}, & F_{comp}(T_1, T_2) < 0 \end{cases}, \tag{7}$$

При сравнении текстов навигационной и информационной страниц с учётом формул (2) и (3) операция сравнения примет вид:

$$F_{comp}(T^{inf}, T^{nav}) = F_{comp}(\{H^{inf}, L^{inf}, B^{inf}\}, \sum_{i=1}^{n}\{H_i^{inf}, L_i^{inf}\}), \tag{8}$$

где: $H_i^{\inf}$ – заголовок i-ой, а $L_i^{\inf}$ – лид i-ой информационной страницы, используемый в тексте навигационной страницы T^{nav}.

Введённая операция сравнения позволяет определить более связанный текст из двух текстов, однако для решения поставленной задачи классификации должна быть введена числовая мера связанности текста, применимая для разграничения информационных и навигационных страниц. Ранее мы определили, что чем чаще в тексте встречаются средства связи, тем больше вероятность того, что текст в целом является связанным – используем данный подход для количественного выражения степени связанности текста:

$$E_T = \sum_{i=1}^{m} Ns_i^T; \quad E_{T\inf} > E_{Tnav}, \tag{9}$$

где E_T – числовое значение степени связанности текста.

Выделенные ранее формальные признаки связанности текста могут быть разделены на две группы по способу их определения: на признаки, выражаемые при помощи определённых частей речи и признаки, выражаемые при помощи использования схожих словоформ.

Признаками первой группы является использование в тексте деепричастий, местоименных существительных, наречий, числительных и союзов. Числовым представлением частоты использования является отношение числа появления соответствующей части речи к общему числу слов в тексте:

$$p_i^s = \frac{Ns_i}{n}; \quad i = \{1,...,\mathrm{k}\}, \tag{10}$$

где: p_i^s – числовое представление s_i – ого признака связанности, определяемого по количеству использований частей речи, Ns_i – число использований в тексте частей речи, соответствующих s_i-ому признаку, n – общее число слов в тексте, k – число признаков связанности, относящихся к первой группе.

Признаками второй группы является использование словоформ имеющих одинаковый корень, а также разных словоформ с одним корнем. Числовое представление этих формальных признаков может быть получено через отношения числа повторяющихся словоформ к общему числу слов:

$$r_j^s = \frac{Nwd}{n}, \quad Nwd = (\mathrm{n} - n_u), \quad \mathrm{j} = \{1,2\}, \tag{11}$$

где: r_j^s – числовое представление s_j-ого признака связанности, определяемого по количеству повторяющихся словоформ, Nwd – число повторяющихся словоформ, соответствующих s_j-ому признаку связанности, n – общее число слов в тексте, n_u – число уникальных слов в тексте.

Учитывая формулы (5), (10) и (11), для определения степени связанности текста, исходный текст страницы должен быть представлен в следующем виде:

$$T = \{\mathrm{p}_1^s,...,\mathrm{p}_k^s,\mathrm{r}_1^s,\mathrm{r}_2^s\} = [\mathrm{p}_1^s,...,\mathrm{p}_k^s,\mathrm{r}_1^s,\mathrm{r}_2^s] \tag{12}$$

Таким образом, предлагаемая модель представления текста учитывает 8 признаков связанности, выявляемых на морфологическом уровне анализа: частоту использования деепричастий, местоимённых существительных, местоименных прилагательных, общего числа местоимений, числительных, союзов, а также полных дейктических и лексико-семантических повторов.

Подход, учитывающий различие в используемых частей речи при формировании текстового содержимого страниц разных типов

Ранее мы описывали различия в принципах формирования содержимого навигационных и информационных страниц, связанные в первую очередь с задачами, выполняемыми страницами каждого типа. Так, основной задачей содержимого навигационных страниц является передача информации об описываемых событиях в наиболее сжатом виде. Проведённый анализ содержимого навигационных страниц ряда интернет СМИ показал, что при формировании текстов навигационных блоков предпочтение отдаётся определённым частям речи, позволяющим передать суть сообщения в нескольких словах (например, [кто и что сделал] или [когда, что и где произошло]), т.е. используются наиболее информативные части речи. При формировании информационного блока ограничение, накладываемые на объём тексте, не является столь жёстким: описания события является достаточно подробным и зачастую сопровождается его оценкой или описанием предшествующих ему событий: т.е. менее информативные части речи используются наравне с более информативными. Таким образом, текстовое содержимое навигационных страниц включает больше информативных и меньше менее информативных частей речи, чем текстовое содержимое информационных страниц, т.е. частота использования определённых частей речи на страницах разных типов должна различаться.

Для русского языка нами было выделено 25 морфологических параметров, позволяющие различать как части речи, так и их формы: при формировании этого множества мы опирались на возможности автоматического морфоанализатора русского языка (11).

Рисунок 1. *Учитываемые моделью представления части речи и их формы (разработано автором)*

Соответствия между используемыми в тексте словами и морфологическими характеристиками из соответствующих множеств задаются с помощью функции отображения:

$$f_{morph}(\mathrm{w}): f_{morph}(\mathrm{w}_i) = p_j; \mathrm{w}_i \in W; p_j \in P, \quad (13)$$

где: p_j – j-ая часть речи из множества всех частей речи P; w_i – i-ое слово из множества всех слов W; $f_{morph}(w_i)$ – функция отображения i-ого слова на j-ую часть речи.

Приведём исходное представление текста в виде последовательности слов к морфологическому представлению, которое может быть выражено в виде объединения:

$$T = \bigcup_{j=1}^{l} Bp_j = \bigcup_{j=1}^{l} \{\mathrm{w}_1^j, ..., \mathrm{w}_i^j, ..., \mathrm{w}_n^j\}; \; w_i^j \in P = \{\mathrm{p}_1, ..., \mathrm{p}_l\} \quad (14)$$

где: Bp_j – множество, состоящее из слов, относящихся к p_j-ой части речи, w_i^j – i-ое слово исходного текста, относящееся к p_j-ой части речи P – множество частей речи языка, l – количество различаемых частей речи,

Для нормализации числа использований конкретной части речи относительно длины текста мы также будем использовать их отношение. Введём функцию определения частоты вхождений каждой части речи в текст:

$$F_{count}(T): F_{count}(T) = (f_{count}^1(T), ..., f_{count}^j(T), ..., f_{count}^l(T)) = [\mathrm{c}_1, ..., \mathrm{c}_j, ..., \mathrm{c}_l], \quad (15)$$

где: $f_{count}^j(T)$ – функция определения частоты использования j-ой части речи в тексте T; $[\mathrm{c}_1, ..., \mathrm{c}_j, ..., \mathrm{c}_l]$ – результирующий вектор, в котором c_j – частота вхождения j-ой части речи в исходный текст.

Применим данную функцию к тексту навигационной страницы, представленной в виде (4):

$$F_{count}(T_{nav}) = F(\bigcup_{i=1}^{k}(\mathrm{T}_i^{\inf} / T_i^{body})) = \bigcup_{i=1}^{k} F(\mathrm{T}_i^{\inf} / \mathrm{T}_i^{body}) = \bigcup_{i=1}^{k}(F(T_i^{\inf}) - F(T_i^{body})) =$$
$$= \sum_{i=1}^{k}([c_{i1}^{\inf}, ..., c_{ik}^{\inf}] - [c_{i1}^{body}, ..., c_{ik}^{body}]) = \sum_{i=1}^{k}[c_{i1}^{\inf} - c_{i1}^{body}, ..., c_{ik}^{\inf} - c_{ik}^{body}], \quad (16)$$

где: $c_{ij}^{\inf}$ – частота использования j-ой части речи в i-ом тексте информационной страницы, c_{ij}^{body} – частота использования j-ой части речи в содержимом тела новости на i-ой странице, k – число информационных страниц связанных с заданной навигационной.

Полученная формула показывает, что при формировании текстового содержимого навигационной страницы на основе связанных с ней информационных страниц количество вхождений различных частей речи в текст будет изменяться. Таким образом, число вхождений различных частей речи в текст может выступать в качестве параметров модели, использующейся для классификации интернет страниц на навигационные и информационные.

Стоить отметить, что описанная модель представления текста позволяет учитывать все выделенные морфологические характеристики используемых слов, часть из которых могут быть избыточны для решения поставленной задачи. Наиболее значимые параметры могут

быть определены с помощью методов сокращения пространства и ранжирования исходного множества учитываемых параметров.

Одним из возможных недостатков, связанных с использованием описанного подхода, является его зависимость от способа формирования текстового содержимого навигационных страниц конкретного источника: на некоторых ресурсах навигационные блоки состоят только из заголовков, на других они включают краткое описание, которое может включать несколько слов или предложений. Таким образом, добавление новых источников данных может привести к снижению точности работы классификатора и необходимости его переобучения.

Описанный недостаток возникает из-за сильной связи между рассматриваемыми параметрами и особенностями формирования содержимого навигационных блоков.

Сравнение описанных подходов к представлению текста

Для проверки применимости предложенных подходов к представлению текстового содержимого интернет страниц для определения её типа (информационная или навигационная) нами была сформирована тестовая выборка. В качестве источников информационных сообщений было выбрано несколько сайтов русскоязычных интернет СМИ, полученная выборка содержит по 500 страниц каждого типа. В ней представлены данные о текстовом содержимом страниц интернет ресурсов: частоте использования различных частей речи, уникальных и схожих слов и словоформ. При формировании данных под текстовым содержимым понималось содержимое как значимых блоков (информационные и навигационные), так и незначимых (рекламные блоки, заголовки переходов на связанные таблицы). Затем, для проверки применимости описанных подходов, текстовое содержимое страниц было представлено в виде, предлагаемом в рамках конкретного подхода и к полученным представлениям были применены методы статистического анализа.

Для представления, учитывающего различия в структуре содержимого страниц обоих типов и формализующего свойство связанности текста, кластерный анализ также показал наличие двух больших и одного малого кластера. В один из больших кластеров входят преимущественно информационные страницы, во второй большой и малый кластеры входят преимущественно навигационные страницы.

Исходное пространство параметров, используемых в рамках данного представления, при помощи факторного анализа было преобразовано в пространство трёх главных компонент. Полученная трёхмерная диаграмма рассеяния для наблюдений представлена на рисунке 2. На полученной диаграмме видно, что информационные страницы преимущественно входят в один кластер, а навигационные страницы образуют два других кластера. При данном подходе к представлению текстового содержимого страниц наблюдается наличие выбросов, характерных преимущественно для информационных страниц, а также попадание некоторых страниц в несоответствующие их типу кластеры.

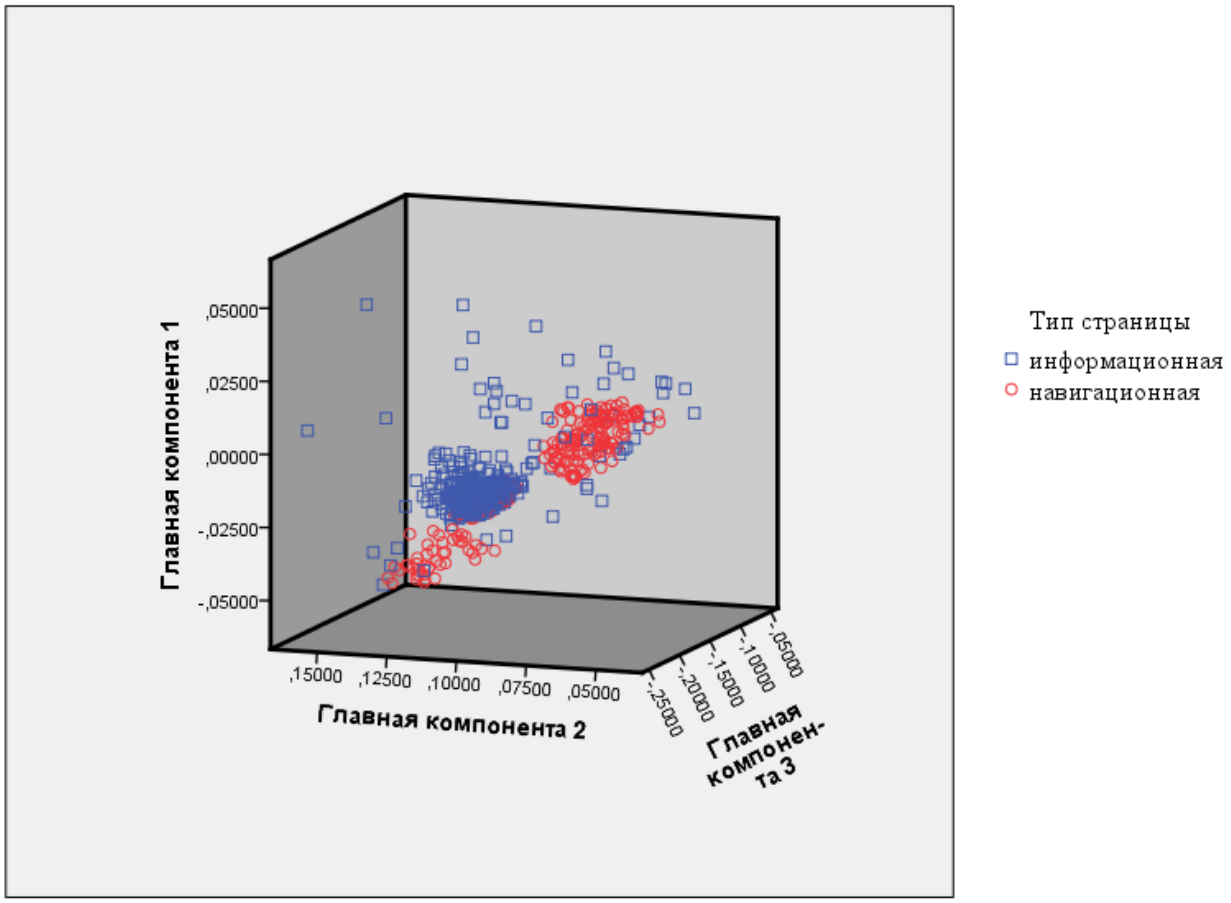

Рисунок 2. *Разделение страниц по их содержимому на основе анализа связанности текстового содержимого (разработано автором)*

Для представления, учитывающего различия в используемых частях речи при формировании текстового содержимого интернет страниц, кластерный анализ показал наличие двух кластеров: в один из них входят преимущественно информационные страницы, а во второй – навигационные.

Для наглядного представления полученных результатов был проведён факторный анализ полученной выборки, позволивший преобразовать исходное пространства параметров в пространство трёх главных компонент. Полученная трёхмерная диаграмма рассеяния наблюдений представлена на рисунке 3. Близость обоих кластеров может быть связана с особенностями формирования исходной выборки, учитывающей текстовое содержимое как значимых, так и незначимых блоков.

При представлении текстового содержимого интернет страниц в виде частоты использования различных частей речи наблюдается разделение страниц по их типам, при этом полученные результаты говорят о наличии выбросов (часть страниц не могут быть отнесены к конкретному кластеру, относится преимущественно к информационным страницам), а также о попадании некоторых страниц в несоответствующий их типу кластеру, что может привести к ошибке определения типа страницы при классификации. При этом различие между кластерами, включающими страницы разных типов, является менее явным, чем при первом рассмотренном подходе. Одной из возможных причин данного различия является избыточность параметров, учитываемых в рамках рассматриваемого подхода.

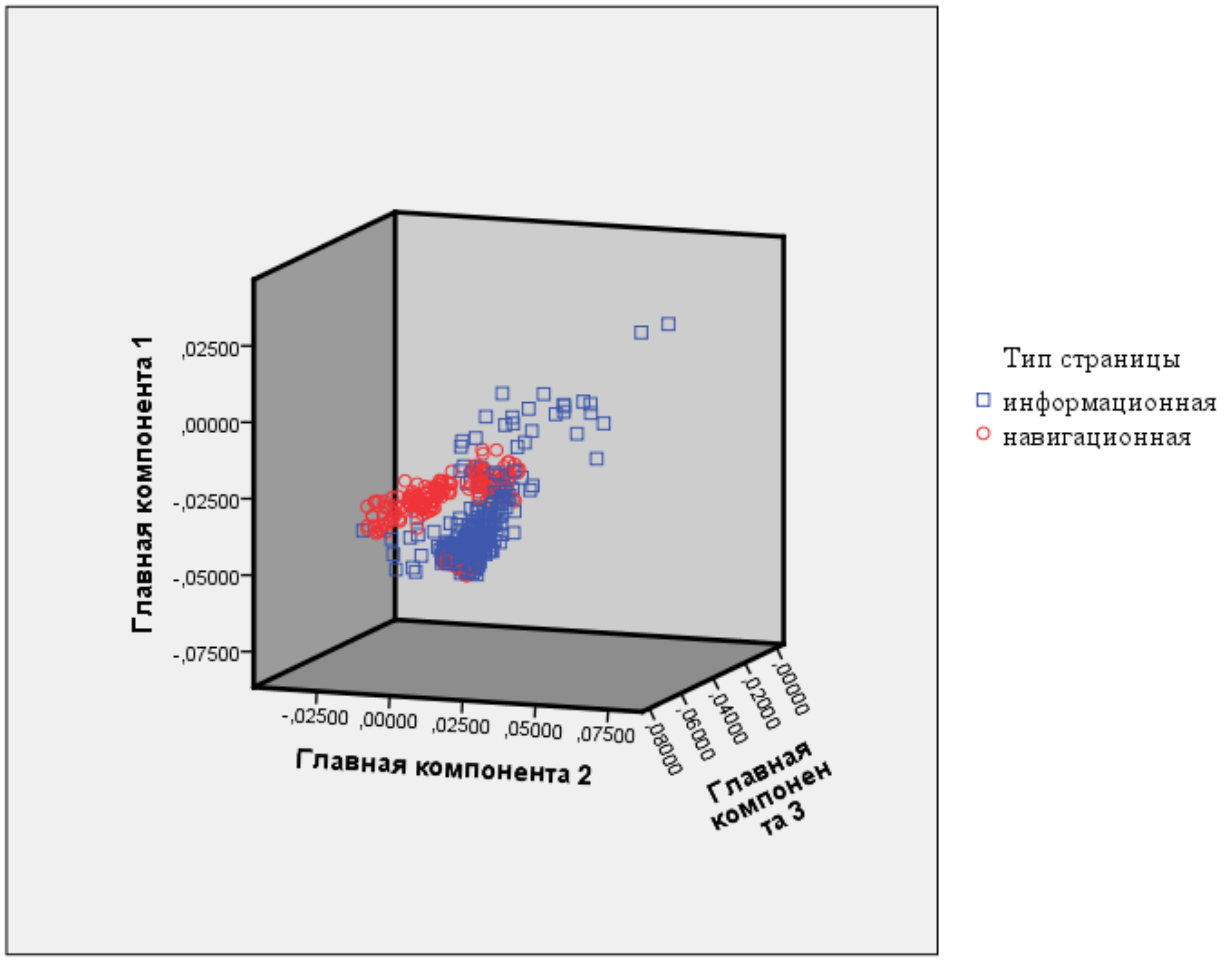

Рисунок 3. *Разделение страниц по их содержимому на основе анализа используемых частей речи (разработано автором)*

Оба описанных подхода к представлению текстового содержимого интернет страниц в качестве значимых параметров используют частоту вхождения слов с определёнными характеристиками, определяемыми на морфологическом уровне анализа. Основное различие между ними заключается в том, что при использовании представления, формализующего свойство связанности текста, в качестве учитываемых параметров выступают определённые части речи, соответствующие формальным признакам связанности текста, выделяемым лингвистами. В рамках подхода, рассматривающего отличия в частоте использования различных частей речи при формировании информационных и навигационных страниц, в качестве параметров используется информация о частоте появления всех частей речи (используется 25 параметров). Избыточность данной модели может быть одной из причин близости формируемых наблюдениями кластеров. При этом необходимо отметить, что при использовании этого подхода к представлению текста, в отличие от представления, формализующего свойство связанности, страницы разных типов образуют два, а не три кластера.

Опираясь на результаты, полученные в ходе анализа двух предложенных подходов к представлению текстового содержимого интернет страниц, можно сделать вывод о необходимости разработки обобщённой модели представления. В качестве базового представления предлагается использовать пространство параметров модели, формализующей свойство связанности текста, которое должно быть расширено путём добавления наиболее значимых параметров из другой предложенной модели представления. Это позволит учесть

наиболее значимые параметры обоих подходов, исключив из пространства рассматриваемых параметров избыточные данные. Помимо этого, для повышения точности анализа текстового содержимого интернет страниц должен быть разработан алгоритм, позволяющий извлекать содержимое только значимых блоков интернет страницы.

Заключение

В данной работе была сформулирована задача выделения из множества интернет страниц, составляющих открытые интернет ресурсы (в первую очередь, новостных и информационных) страницы, содержащие наиболее содержательные и валидные данные. Для решения задачи классификации страниц были предложены и рассмотрены два подхода к представлению их текстового содержимого: подход, формализующий свойство связанности частей текста и подход, учитывающий разницу в частоте использования различных частей речи при формировании содержимого страниц разных типов. В рамках каждого подхода была разработана соответствующая ему модель представления текста для решения задачи классификации. Методами математической статистики была осуществлена проверка предложенных моделей на их применимость для решения поставленной задачи классификации.

Дальнейшие работы должны быть направлены на разработку обобщенной модели представления текстового содержимого интернет страниц, учитывающей наиболее значимые параметры обеих предложенных моделей. Также должен быть разработан инструментарий для извлечения из обрабатываемых интернет страниц содержимого только значимых блоков, позволяющих повысить качество классификации.

ЛИТЕРАТУРА

1. William S. Davis, David C. Yen. The Information System Consultant's Handbook: Systems Analysis and Design. б.м.: CRC Press, 1998. стр. 800.
2. Nitin Agarwal, Huan Liu. Modeling and Data Mining in Blogosphere. б.м.: Morgan & Claypool Publishers, 2009.
3. Jonathan Grey, Lucy Chambers, Liliana Boungeru. The Data Journalism Hadnbook. б.м.: O'Reilly, 2012. стр. 242.
4. Jones, M. Tim. Extract information from the web with Ruby. [В Интернете] 17 Декабрь 2013 г. http://www.ibm.com/developerworks/opensource/library/os-extractruby/os-extractruby-pdf.pdf.
5. Колесниченко А.В. Прикладная журналистика. Москва: Издательство Московского университета, 2008.
6. Понятие текста и критерии текстуальности. Москвин, В.П. Москва: Наука, 2012 г., Русская речь: Научно-популярный журнал, стр. 37-48.
7. Кронгауз М.А. Семантика: Учебник для студ. лингв. фак. высш. учеб. заведений. 2-е. Москва: Издательский центр "Академия", 2005. 5-7695-2016-7.
8. Валгина Н.С. Теория текста. Москва: Логос, 2003.
9. Гальперин И.Р. Текст как объект лингвистического исследования. 4-е. Москва : КомКнига, 2006. 978-5-484-00618-2.
10. Казарин Ю.В., Бабенко Л.Г. Теория лингвистического анализа художественного текста. Учебник. Москва: Флинта, Наука, 2009. стр. 182-193.
11. Яндекс технологии - Mystem. [В Интернете] Яндекс. https://tech.yandex.ru/mystem.

Redkin Oleg Konstantinovich
Moscow technological university, Russia, Moscow
E-mail: o.k.redkin@gmail.com

Approaches to the text representation to define type of information message source

Abstract. This paper considers the allocation problem of web pages constituting single web-resources the most information-packed pages. To solve the problem the conditional-section of the web pages to their functions for information and navigation is offered. The features of the formation of both types of pages on the example of news resources. To solve the problem of classification of pages two approaches is offered to the presentation of the posted text content. In the first approach to identify classification properties it was carried out the analysis of the contents of the pages of both types having-the basic properties of the text (in the linguistic sense). The analysis has been allocated global connectivity property of the text which is characteristic for the pages of the same type only. The second feature of the proposed approach takes into account the use of the most informative parts of speech in the formation of the navigation pages. As a separating sign using the frequency of the different parts of speech in the texts of both types of pages is proposed. For both the proposed approaches the models representations of web pages text content were developed and they were tested for applicability to solve the problem by means of mathematical statistics.

Keywords: computational linguistics; word processing; data mining; web page; web pages classification; text web page content