УДК 004.75

Обобщение динамического контента в веб-коллекциях

Д.Л. Зайцев, А.Н. Зеленина, А.Д. Степченков

Воронежский институт высоких технологий, Воронеж, Россия

Статья посвящена новому исследовательскому проекту, направленному на обобщение динамического контента веб-страниц. Значительная часть информации в Интернете теряется из-за временной природы веб-документов. Таким образом, адаптация методов обобщения для веб-документов является актуальной задачей. Исследование нацелено на разработку методов обобщения изменяющегося контента, который извлечен из коллекции тематических веб-страниц, которые связаны между собой, за определенные промежутки времени. В данной статье рассматриваются наиболее известные тематики и концепции, обнаруженные в веб-коллекциях, которые являются ретроспективными. По причине разнообразия содержимого, а также изменений веб-контента, связанных со временем, необходимо применить методы, отличающиеся от базовых, которые применяются для документов, являющихся статическими.

Ключевые слова: динамический контент, веб-коллекции, веб-страницы.

Summarization of Dynamic Content in Web Collections

D.L. Zaitsev, A.N. Zelenina, A.D. Stepchenkov

Voronezh Institute of High Technologies, Voronezh, Russia

The paper is devoted to a new research project aimed at generalizing the dynamic content of web pages. Much of the information on the web is lost due to the temporal nature of web documents. Thus, adapting generalization methods for web documents is a relevant task. The research aims to develop methods for summarizing the changing content, which is extracted from a collection of topic web pages that are related to each other, over certain time intervals. This paper focuses on the most prominent topics and concepts found in web collections that are retrospective. Due to the diversity of content, as well as changes in web content over time, it is necessary to apply methods that differ from the basic ones that are applied to documents that are static.

Keywords: dynamic content, web collections, web pages.

Введение

Раньше исследование обобщения документов было основано статьях, напечатанных в газетах, или других документах, которые являлись статическими, но в век интернета, веб-технологий и телекоммуникационных услуг существует потребность инвестировать больше внимания и ресурсов обобщению веб-страниц. Есть техники, которые разработаны исключительно для суммирования контента в веб-страницах [1, 2]. Интернет представляется динамичным и разнообразным местом. Данные характеристики вызывают трудности при использовании традиционного анализа текста в веб-пространстве. Одной из основных отличительных черт веб-сайтов от всех других вариантов документов является способность этих самых веб-сайтов менять со временем как структуру, так и содержимое. Большое количество известных веб-страниц очень часто подвергается обновлению и выдает обновленную информацию. Как итог, веб-документ, который размещен на веб-ресурсе или сервере и который управляется той или иной операционной системой, правильнее всего будет оценивать как динамический объект или какое-либо место, которое связано с URL-адресом. Данный взгляд дает

возможность учесть контент, который изменен, а также который подвергается добавлению или удалению из цифровых документов для создания синопсиса. Данная проблема реформирования разнится от привычного общего вывода тем, что нацелена на подвижную информацию онлайн-страниц. Различие между традиционным и новым подходами к обобщению изменений заключается в следующем: новый метод включает сравнение временных версий двух и более документов для выявления их изменений, которые затем суммируются (рис. 1).

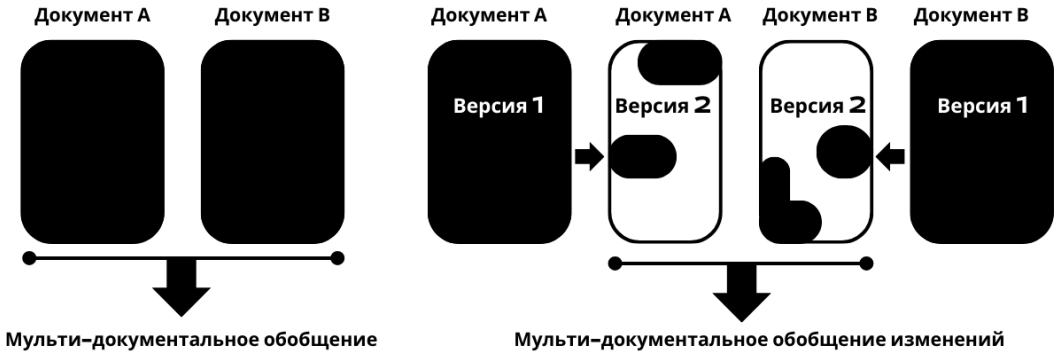

Рисунок 1. Различие между традиционным и новым подходами к обобщению изменений

Суммирование модификаций в онлайн-документах имеет возможность стать достаточно ценным в тех или иных обстоятельствах. Например, потребитель может быть заинтересован изучить информацию о самых модных фактах в его избранной веб-коллекции за конкретный промежуток. Реализовывать данный процесс мануально, изучая все онлайн-документы с целью найти малейшие изменения, слишком трудоемко. Создание веб-коллекции по определённой теме с тщательно отобранными источниками информации может служить решением этой задачи. Такая коллекция может рассматриваться как целостный и комплексный источник сведений о сфере интересов пользователя. Основные события и популярные изменения, относящиеся к заданной теме, могут быть выявлены в зависимости от качества и характеристик исходной коллекции [3]. По данным Cisco до конца 2019 больше половины населения планеты будут иметь доступ к Интернету, при этом количество устройств с выходом в сеть превысит численность населения в три раза, а до 2021 года видео составит около 80% мирового интернет-трафика. Наиболее массовыми сервисами, просматриваемыми в Интернете, можно считать youtube-ролики, онлайн игры, IPTV и применения IT в учебном процессе, но не несмотря на востребованность прием-передача значительных объемов информации в Интернете остается сложной задачей.

Решение задач приема-передачи контента (ППК) в значительной степени осложняется гетерогенностью информационных сред, используемых в системах приема-передачи контента (СППК). одновременно используются очень разные аппаратные средства, характеристики которых (процессор, память, сетевое оборудования) лежат в широком диапазоне. Программное обеспечение (операционные системы, браузеры, средства сжатия информации, сетевые протоколы различных уровней) также очень разное. Все это может варьироваться в разных сеансах даже для одного пользователя.

Для передачи одного и того же полезного контента могут использоваться различные виды информации (текст, графика, мультимедиа), типы, форматы и метод сжатия (JPEG, AVI, MP4, 3GP, MPEG и т. д.).

Несоответствие типов, форматов и объемов информации характеристикам СППК может либо сделать получение информации невозможным (разъединение текущего

сеанса связи из-за временных задержек), либо значительно ухудшить качество воспроизведения: задержка изображения, его качество, отсутствие синхронизации видео и звука. Общим недостатком современных исследований и разработок в области адаптации контента к характеристикам программно-аппаратных средств СППК являются значительные затраты на их реализацию. Более того, многие решения оказываются специализированными и накладывают жесткие ограничения на использование определенных программно-аппаратных конфигураций.

Кроме того, обзор наиболее распространенных информационных технологий и систем, которые используются в учебном процессе (Moodle, Blackboard, SAKAI, Lotus Workplace Collaborative Learning т. п.), показывает, что ни одна из них не анализирует характеристики конкретной СППК и не обеспечивает динамическое формирование контента, который должен быть передан пользователю.

Таким образом, исследование, направленное на разработку моделей, метода и информационной технологии динамического формирования контента является актуальным.

Для решения задачи повышения качества процесса приема/передачи контента необходимо решить следующие задачи:

– исследовать программно-аппаратные конфигурации СППК;

– разработать модель приемной части СППК, включающую характеристики канала связи и программно-аппаратных средств пользователя;

– создать базовую модель передаваемого контента и провести её апробацию на примере учебного курса, преподаваемого с использованием информационных технологий;

– разработать модель СППК, содержащую минимально необходимый набор компонентов для обеспечения приема и передачи контента;

– разработать метод динамического формирования контента (тип, формат), который учитывает характеристики всех компонентов СППК и обеспечивает определение необходимого контента для передачи в реальном времени в сети Интернет;

– разработать информационную технологию и информационную систему динамического формирования контента в реальном времени в сети Интернет, обеспечивающие соответствующее качество процесса ППК, а также выполнить их апробацию.

Метод обобщения динамического контента на веб-страницах, представленный в этой статье, может применяться к любым типам веб-сайтов. Этот метод ориентирован на сохранность семантической и композиционной последовательности между отличающимися вариациями одной онлайн-страницы. В контексте настоящей гипотезы фокус направлен на различные цифровые страницы, и ссылки или страницы, которые находятся рядом, не принимаются во внимание. Несмотря на это есть возможность сконфигурировать формулу для взаимодействия с набором веб-документов или совокупностью взаимосвязанных онлайн-страниц. В подобных ситуациях существует возможность указать конкретную степень исследования для мониторинга данных цифровых документов. Например, существует возможность рассмотреть все документы, к которым обращается веб-сайт компании: карточки ассортимента, штата, открытых возможностей и т. п. Выход находится в слиянии всех указанных онлайн-страниц в единственный документ, который демонстрирует нужную часть сайта. Для реализации данного процесса наполнение каждой электронной страницы может быть изучено в зависимости от числа перемещений от первоначального веб-документа. В результате, каждая электронная страницы или сегмент онлайн-документов в сборнике может быть оценена как независимый веб-контент, и схема взвешивания будет применяться к содержимому каждой страницы.

Основная часть

Существуют два основных подхода к получению тематических коллекций веб-страниц. Первый способ предполагает использование существующих веб-каталогов, таких как ODP. Однако эти каталоги часто имеют ограниченное число тематических категорий, которые могут быть устаревшими. Это ограничивает пользователя фиксированной иерархией доменов, не позволяя выбирать произвольные темы. Второй способ заключается в применении поисковых систем, где можно задать любое количество критериев, что обеспечивает большую гибкость в выборе. Однако набор полученных веб-страниц не всегда соответствует интересам пользователя, что требует дополнительного анализа результатов поиска. Также критически важно сегментировать накопленные электронные файлы для исключения копий онлайн-записей по причине того, что они могут существенно испортить валидность окончательного итога.

На дальнейшей стадии вариации электронных страниц машинально вносятся с заданной регулярностью. Период t между загрузкой новых версий веб-страниц следует выбирать, учитывая временные характеристики коллекции. Чем больше интервал T, тем меньше вероятность зафиксировать изменения из-за кратковременного контента, что характерно для новостных лент или популярных страниц. Некоторые части веб-страниц могут изменяться несколько раз в течение интервала T, что увеличивает риск потери информации. Напротив, высокая частота выборки страниц может увеличить количество обнаруженных изменений, но также приведет к более интенсивному использованию сети. Пусть $C_a = \{C_1, C_2 \dots C_n\}$ представляет собой набор всех изменений, которые происходят на одной веб-странице в течение заданного интервала, и $F_a = \{F_1, F_2 \dots F_n\}$ набор обнаруженных изменений. Если сделать предположение, что страница меняется с фиксированной частотой t_a , то отклик изменений возможно аппроксимировать следующим образом:

$$R_a = \frac{|F_a|}{|C_a|} \approx \frac{t_a}{t}, \textit{если } t_a \le t,$$

$$R_a = \frac{|F_a|}{|C_a|} = 1, \textit{если } t_a > t.$$

Обозначим весь период времени, за который будет составлена сводка, как T. Если принять малый интервал T, включающий лишь несколько таких промежутков, это приведет к небольшому числу веб-страниц с изменениями. В этом случае вклад этих страниц в итоговый результат будет относительно большим. Следовательно, качество окончательного результата может быть ниже в отношении фактических изменений по теме сбора, поскольку он определяется всего несколькими веб-страницами. Если же выбрать длинный интервал T, включающий множество таких промежутков, можно ожидать выявление большего числа изменений, что уменьшит влияние одной веб-страницы на итоговый результат [6].

Время реакции для нескольких, например пары, электронных страниц без каких-либо отличий на ситуацию, которая случилась в том или ином временном промежутке, имеет шанс не быть идентичным. Пусть данные онлайн-документы постоянно затрагивают только исключительно приоритетные и важнейшие действия, которые указывают на предпочтения потребителя. Обычно новостные веб-сайты опубликовывают информацию в определенное время, тогда как для других типов сайтов это может занять больше часов. Эта разница называется «временным разнообразием» веб-страниц, отличая его от «содержательного разнообразия». Выбор чересчур короткого интервала T может привести к менее точному итоговому результату, так как реакция на определенное событие может растягиваться во времени на различных страницах. И наоборот, другой и наиболее длинный интервал T может увеличить вероятность учета многих несвязанных и не по теме изменений, что повлечет за собой

снижение качества материала. Для извлечения изменяющегося контента создаются две последовательные версии всех веб-страниц, чтобы сравнить их между собой. Сравнение производится в формате предложений. Предложения из самых близких версий документа анализируются с целью определения вставок и поиска удалений. Мы фокусируемся только на контенте в формате текста и не используем изображения и другие мультимедийные элементы. На странице возможно наличие двух типов текстовых изменений: добавление и удаление. Если определенное предложение появляется лишь в более новой версии страницы, это считается вставкой. В случае, если предложение существует исключительно в старой версии, это рассматривается как удаление.

Далее происходит реализация элементарных последовательностей начального анализа текстовых элементов: стемминг, а также исключение слов, состоящих в стоп-списке. Модифицированные текстовые единицы и биграммы учитываются как перечень функционала. Все определения взвешиваются, базируясь на своих распределениях в движущихся фрагментах архива на протяжении отрезка *t*. Подобная техника базируется на гипотезе, что с нашумевшей терминологией есть возможность познакомиться в практически идентичных модификациях огромной вариации электронных документов, которые расположены рядом. По этой причине определения, появляющиеся в частях большого количества документов, которые были изменены, могут получить наивысшие оценки в сравнении со встречающимися только в частях некоторых измененных веб-страниц. Также, если определение много раз может быть встречено в большом количестве версий страниц, которые изменили, то рейтинг этого определения может быть увеличен. Концепция «популярности» определяет, что частота отображения термина в различных документах намного приоритетнее частоты его использования в единичном документе. По этой причине часть уравнения, которая касается частоты документов, может содержать экспоненциальный характер. Частота документов (*DF*) подразумевает под собой количество нескольких версий документа, содержащих данный термин. Регулярность формулировки (TF_j) имеет в виду регулярность той или иной формулировки в пределах движущегося фрагмента одной из нескольких вариацией страницы j. В модели (1) регулярность разбивается по целой сумме всех возможных частей в рамках всех возможных модификаций, то есть размеру изменения S_j, а следом сводится к среднему по экземплярам электронных документов $N \cdot n$: N – сумма различных онлайн-страниц. В общем стандартная схема оценки похожа на популярную систему взвешивания TFIDF [7].

$$S_{term} = \frac{\sum_{j=1}^{N \cdot n} \frac{TF_j}{S_j}}{N \cdot n} \cdot exp\,(DF). \quad (1)$$

Как и упоминалось ранее, изменения делятся на два типа: вставка и удаление. Последнее на интуитивном уровне можно оценивать как удаление старого контента, не имеющего более какого-либо значения, которое будет заменено другим текстом. Однако, если одновременно в нескольких веб-документах удаляется схожий контент, это может указывать на то, что связанное с этим содержанием важное событие подошло к концу.

В этом случае термины, встречающиеся в таких вставках, должны иметь высокое значение присвоенной важности. С другой стороны, термины, которые встречаются во многих ближайших вставках, также будут иметь высокие баллы. Наконец, общая оценка за семестр будет равна комбинация обеих частичных оценок, рассчитанных по удаленному и вставленному текстовому содержимому [8].

Пусть d_x – удаление, а i_x, допустим, – вставление одиночного электронного документа, в котором x – нумерация вариации онлайн-страницы (рис. 2), а именно, d_x и

i_x обозначают содержимое, которое было ликвидировано из вариации x-1. Данные, прибавленные в вариацию x, будут являться цифровой страницей (2).

Суммарный размер удалений d в единичном онлайн-документе за отрезок времени T, которое может быть определено как «отрицательная модификация» онлайн-документа, подразумевает внутри себя полный текстовый набор, который был изъят из всего количества вариаций данного цифрового элемента за обозначенный отрезок времени:

$$D = \bigcup_{x=1}^{x=n} d_x \,. \tag{2}$$

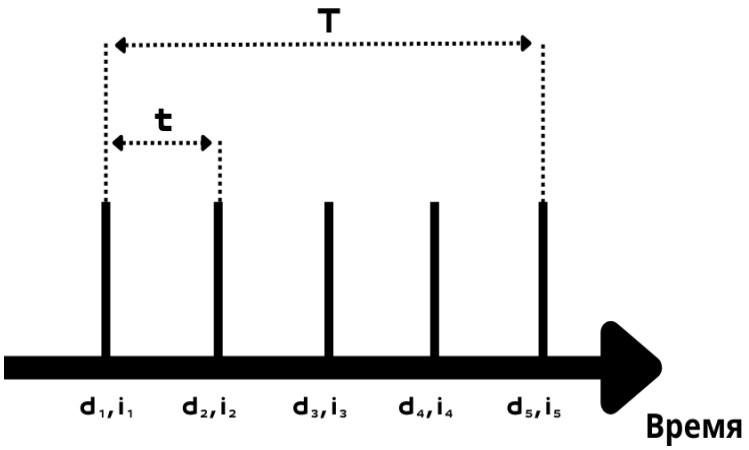

Рисунок 2. Графическое представление временных изменений на веб-странице

С другой стороны, весь пул вставок, выраженный I, может быть назван «положительным изменением» для документа (3).

$$I = \bigcup_{x=1}^{x=n} i_x \,. \tag{3}$$

Необходимо присвоить самый большой вес определениям, вставляющимся или удаляющимся из огромного списка документов в самые короткие сроки. Чтобы это реализовать используется временное сходство определений, похожее на осмотр наборов данных типа событий от TDT (тематическое обнаружение и отслеживание). Пользователь определяет диапазон сходства по движущемуся отрезку, которое уравнивается по отрезку или длине L. Такой отрезок имеет возможность направляться в сторону градиентно отрегулированного архива. По этой причине равномерно одинаково имеют возможность оцениваться исключительно вариации L/T полностью всех цифровых документов (рис. 3). Определения имеют возможность находится не только в негативных, но и в положительных вариациях модификаций, но в рамках всех отрезков. Темные области символизируют вставки и удаления в версиях веб-страниц (рис. 3). Термины могут оцениваться, равняясь на схему взвешивания (4). С другой стороны, в таком случае видно, что разница между частотой документов и определениями в двух типах изменений равномерна.

Оценка термина обозначается S_{term}^{win} и выражается формулой 4 [9]:

$$S_{term}^{win} = \frac{\Sigma_{j=1}^{N \cdot n} \left| \frac{TF_J^I}{S_J^I} - \frac{TF_J^D}{S_J^D} \right| \cdot exp\left| DF^I - DF^D \right|}{N \cdot n \cdot L}. \tag{4}$$

В уравнении 4 метки I и D, которые сфокусированы над линией деления, показывают фокусные вариации модификаций в области единичного расположения отрезка. Выходит, регулярность понятий и страниц каждого без исключения понятия имеет возможность быть посчитанной для площади, которая останавливается

исключительно отрезком, и только. Суммарное число понятия (5) демонстрирует усредненная дистрибуция всех модификаций понятия в рамках всех расположений отрезка *Nw*.

$$S_{term}^{overrall} = \frac{\sum_{win=1}^{N \cdot w} S_{term}^{win}}{N \cdot w}. \quad (5)$$

В контексте регулярного столкновения понятия в единичной вариации модификаций в рамках обширного количества пунктов отрезка совокупное число имеет возможность оказаться выше среднего, но понятия, которые демонстрируют почти зеркальную дистрибуцию не только негативных, но и положительных модификаций в преимущественной сумме пунктов отрезка имеют возможность быть с показателями ниже среднего. Другими словами, первостепенно важны понятия, которые находятся в виде нескольких удалений или добавлений в преимущественном диапазоне пунктов конкретного отрезка. Данный процесс может быть достигнут с помощью учёта значений различий, которые абсолютны, в частоте документов и терминов двух видов изменений. Сама длина окна определяется пользователем на основании факта, является ли это окно долгосрочным или, наоборот, предназначено для целей обнаружения [10].

На последнем этапе происходит извлечение предложений, которые содержат заданные понятия за период *T*. Для выбора таких предложений необходимо вычислить средний бал термина, для каждого заданного понятия и извлечь только те, которые имеют наивысший бал. Длина результата задается пользователем. Мы также задаем вводимый потребителем предел на сумму вариантов, которые есть возможность получить из определенной версии электронной страницы. Подобный предел дает шанс исключить события, как, например, когда единичная страница или группа страниц преобладает в финальной выдаче. Для повышения ясности результатов мы добавляем предшествующие и последующие предложения вокруг тех, которые имеют наибольшее количество баллов.

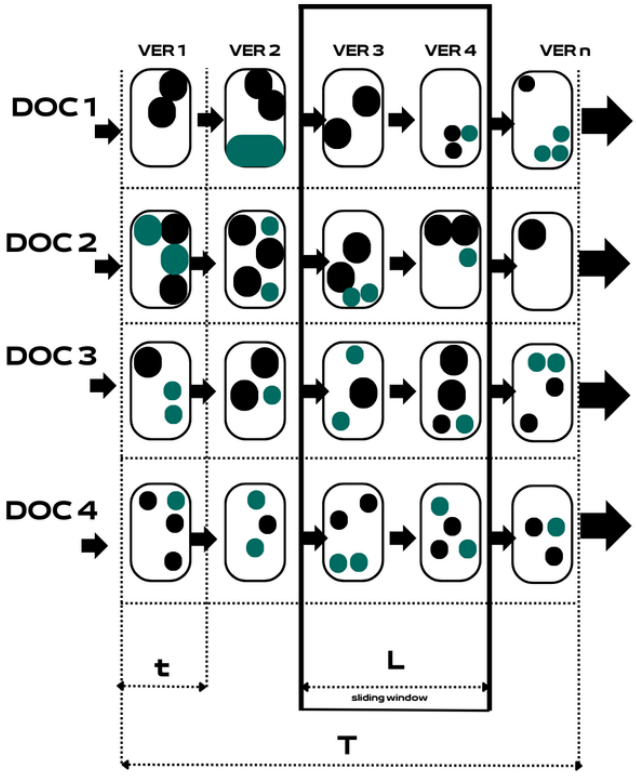

Рисунок 3. Скользящее окно в веб-коллекции из N документов с n версиями

Выводы

Был предложен метод по обобщению динамического контента в ретроспективных веб-коллекциях, который использует скользящее окно с типами вставки и удаления изменений. Эта техника концентрируется на временных условиях упоминания в цифровых страницах. Рекомендуется оценивать исключения как компонент модификаций движущегося потока данных цифровых страниц и применять компонованную технику, чтобы проанализировать все вариации модификаций. Были изучены положительные стороны и сложности суммирования в движущихся цифровых архивах.

Предложенный метод может найти применения при формировании динамического контента систем дистанционного образования в научно-исследовательской работе «Модели и метод динамического формирования контента в зависимости от характеристик системы приема-передачи».

СПИСОК ИСТОЧНИКОВ

1. Allan J. Temporal Summaries of News Topics / J. Allan, R. Gupta, V. Khandelwal // SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. – New York, 2001. – P. 10-18.

2. Ларичев О.И. Системы поддержки принятия решений: современное состояние и перспективы развития / О.И. Ларичев, А.Б. Петровский // Итоги науки и техники. Теория вероятностей. Математическая статистика. Теоретическая кибернетика. – 1987. – Т. 21. – С. 131-164.

3. Большаков А.А. Разработка стенда для оценки применимости транспортных протоколов в задачах обработки потоковой информации для создания адаптивной системы преобразования данных / А.А. Большаков, И.В. Егоров, В.В. Лобанов, Д.В. Лачугин // Вестник Тамбовского государственного технического университета. – 2014. – Т. 20. – № 3. – С. 440-451.

4. Jatowt A. Web Page Summarization Using Dynamic Content / A. Jatowt, M. Ishizuka // WWW Alt. '04: Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. – New York, 2004. – P. 344-345.

5. Mani I. Advances in Automatic Text Summarization / I. Mani, M.T. Maybury. – Cambridge: MIT Press, 1999. – 434 p.

6. McKeown K.R. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster / K.R. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J.L. Klavans, A. Nenkova, C. Sable, B. Schiffman, S. Sigelman // HLT '02: Proceedings of the Second International Conference on Human Language Technology Research. – San Francisco, 2002. – P. 280-285.

7. Radev D.R. NewsInEssence: A System for Domain-Independent, Real-Time News Clustering and Multi-Document Summarization / D.R. Radev, S. Blair-Goldensohn, Z. Zhang, R.S. Raghavan // HLT '01: Proceedings of the First International Conference on Human Language Technology Research. – Stroudsburg, 2001. – P. 1-4.

8. Salton G. Term-Weighting Approaches in Automatic Text Retrieval / G. Salton, C. Buckley // Information Processing & Management. – 1988. – Vol. 24. – No. 5. – P. 513-523.

9. Зайцев Д.Л. Формальное представление деятельности пользователей с выявлением информационно-значимых объектов / Д.Л. Зайцев, А.Н. Зеленина // Вестник Воронежского института высоких технологий. – 2021. – Т. 15. – № 2 (37). – С. 45-56.

10. Зайцев Д.Л. Классификация интерактивных взаимодействий пользователя с программным обеспечением / Д.Л. Зайцев, А.Н. Зеленина // Вестник Воронежского института высоких технологий. – 2022. – Т. 16. – № 3 (42). – С. 43-48.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Зайцев Даниил Леонидович, аспирант, Воронежский институт высоких технологий, Воронеж, Россия.
e-mail: turbo_char@mail.ru

Зеленина Анна Николаевна, кандидат технических наук, доцент, ведущий специалист проектного отдела, Воронежский институт высоких технологий, Воронеж, Россия.

Степченков Андрей Дмитриевич, аспирант, Воронежский институт высоких технологий, Воронеж, Россия.
e-mail: stepchenkov136@gmail.com