КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА

УДК 004.942

МЕТОД СЕМАНТИЧЕСКОЙ КЛАСТЕРИЗАЦИИ ГИПЕРТЕКСТОВОЙ СТРУКТУРЫ С ПРИМЕНЕНИЕМ СТАТИСТИКИ ПЕРЕХОДОВ ПОЛЬЗОВАТЕЛЕЙ ПО ВНУТРЕННИМ ССЫЛКАМ

В. С. Салин, С. В. Папшев, А. А. Сытник

Саратовский государственный технический университет им. Гагарина Ю.А.

Поступила в редакцию 10.04.2015 г.

Аннотация. Кластеризация веб-документов по различным признакам, включая их семантическую близость, лежит в основе различных современных подходов к поиску и извлечению информации в Интернете. В данной статье предложен метод построения семантических кластеров в гипертекстовой структуре веб-сайта на основе статистики переходов пользователей между узлами. Кластеризация документов по отслеживаемым маршрутам пользователей применяется к графовой модели гипертекстовой структуры веб-сайта. Приводятся результаты апробации предложенного подхода на реальных веб-сайтах.
Ключевые слова: гипертекстовая структура, семантическая кластеризация, статистика посещений, графовая модель.
Annotation. Clustering of web documents by various attributes, including their semantic similarity, underlies the modern approaches to information retrieval and data extraction. In this paper, we present a method for constructing semantic clusters in hypertext structure of the website, which is based on website statistics of user visits and their transitions by hyperlinks. We apply documents clustering by user routes to graph model of the hypertext structure. We also present evaluation results of the method, tested on the real-world websites.
Keywords: hypertext structure, semantic clustering, website statistics, graph model.

ВВЕДЕНИЕ

Объемы неструктурированной информации, представленной в Интернете, постоянно возрастают, равно как и общее количество веб-сайтов и объемы их содержимого. Задачи поиска релевантной информации и анализа данных, размещенных на веб-страницах, являются неотъемлемой частью при работе со значительными объемами информации. Важным этапом в таких задачах является кластеризация обрабатываемых документов, которая позволяет выявлять группы семантически похожих документов.

Кластеризация веб-страниц по тематике актуальна в различных приложениях поиска и интеллектуального анализа данных, таких как распознавание шаблонов, извлечение ключевых слов [1]. Группировка семантически связанных страниц лежит в основе рекомендательных алгоритмов и адаптивных интерфейсов, которые могут «подсказывать» пользователю наиболее релевантные для него страницы. Также, определение кластеров близких по теме страниц используется в веб-аналитике и поисковой оптимизации, веб-разработке и реинжиниринге. Подробный обзор по данной проблематике представлен в работе [2].

© Салин В. С., Папшев С. В., Сытник А. А., 2015

Обычно кластеризация документов построена на методах статистического анализа текста или медиаконтента. В таких случаях, она разбивается на два этапа – индексация, с выявлением термов для дальнейшего анализа, и собственно кластеризация на основании общих термов между страницами.

Реализация таких методов имеет определенные сложности: необходимость предварительной индексации, высокое время выполнения при больших объемах текстов, с другой стороны – необходимость в наличии минимального объема текста с релевантными ключевыми словами.

В данной работе для определения кластеров семантически близких страниц предлагается использовать не анализ содержимого страниц, а собранную статистику о переходах пользователей сайта между страницами. В основе подхода лежит гипотеза, что при навигации по веб-сайту, его пользователи придерживаются определенных интересов и выбирают страницы схожей тематики, соответствующие их информационным запросам. Повторяющиеся маршруты пользователей формируют группы семантически близких страниц, объединение которых покрывает веб-сайт.

Для моделирования переходов между страницами гипертекстовой структуры веб-сайта, ей в соответствие ставится ориентированный граф, который расширяется данными статистики посещений, собранными внешней системой. На полученной модели мы определили метод кластеризации, группирующий узлы графа на основании статистики переходов. В результате работы алгоритма мы получаем кластеры веб-страниц и на примере задачи статистического анализа текста данных веб-страниц показываем, что страницы в данных кластерах семантически связаны и имеют больше общих ключевых слов.

В первой части данной статьи рассматриваются известные результаты исследований в областях кластеризации и извлечения информации, а также работы со статикой посещений. Во второй части, приводится построение графовой модели гипертекста, ее расширение статистикой посещений и определение метода кластеризации на основе данной статистики. Далее, показано различие между статистическим анализом текстов одиночных документов и кластеров документов и то, как кластеризация может влиять на эффективность поиска ключевых слов. Приведенные в завершении работы результаты экспериментальной апробации метода демонстрируют эффективность кластеризации гипертекста для задачи информационного поиска на основе информации о статистике переходов пользователей между страницами на примере трех различных веб-сайтов.

ИССЛЕДОВАНИЯ В ОБЛАСТИ СЕМАНТИЧЕСКОЙ КЛАСТЕРИЗАЦИИ И ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ

Задача семантической кластеризации, как задача извлечения информации (InformationExtraction), относится к области исследования WebDataExtraction (WDE). Обзор существующих систем WDE и их прикладного применения в различных сферах дан в [3]. В частности, авторы выделяют два вида веб-приложений – социальные и для предприятий. Данное исследование позволяет понять в целом техники, применяемые в области WDE, а также демонстрирует примеры возможных вариантов применения. Однако в данной работе авторы не углубляются во внутренние технические особенности работы WDE систем, поэтому информации о применяемых подходах к кластеризации и их использования для задач WDE в исследовании не представлено.

Другой интересной работой в области WDE является [4], сконцентрированная в основном на обзоре техник, применяемых в данной области. Здесь авторы различают два основных типа техник – первая основана на определении шаблона (wrapper/templateinduction), вторая – на автоматическом извлечении данных. В задаче определения шаблона, методы кластеризации используются для подтверждения, что полученные кластеры не пересекаются (т.е. одна страница не может входить сразу в два различных кластера одновременно). Что касается анализа русскоязычных текстов, под-

робный обзорсистем и инструментов представлен авторами в [5].

В различных системах, кластеризация веб-страниц построена на различных их свойствах. Например, в [6] авторы используют кластеризацию на основе путей в древовидной DOM-структуре веб-страниц, при этом пути в дереве определяются на основе анализа веб-сайта и журнала его посещений. В [7] представлен подход извлечения информации EXALG, который проводит кластеризацию веб-страниц по классам эквивалентности, определяемых наборами маркеров (токенов). Подход EXALG является практическим примером применения подхода определения шаблона без учителя. Основное различие между подходами к кластеризации, упомянутыми выше, и данной работой в том, что предлагаемый подход к кластеризации веб-страниц основан на данных о гипертекстовой структуре веб-сайта и о связях между страницами.

Применение средств веб-аналитики, использование статистики посещений, отслеживание поведения пользователейактуально, прежде всего, для задач маркетинга и продвижения веб-сайтов [8]. В области поиска и анализа данных, механизмы отслеживания переходов пользователей между страницами применяются в [9]. Автор применяет метод формальных концептов для подтверждения гипотезы, что маршрут представителя определенной группы интересов примерно совпадает с маршрутами всей группы. Подход к реинжинирингу и возможностям семантической оптимизации структуры образовательных веб-сайтов на основе статистики посещений представлен в [10].

В целом, рассмотренные методы решают частные и смежные задачи в области семантической кластеризации, однако поиск эффективных методов семантической кластеризации остается актуальным.

КЛАСТЕРИЗАЦИЯ ГРАФОВОЙ МОДЕЛИ ГИПЕРТЕКСТА НА ОСНОВЕ ДАННЫХ О СТАТИСТИКЕ ПЕРЕХОДОВ

В данном разделе мы рассматриваем графовую модель гипертекстовых структур как основу для построения семантически связанных кластеров, а также подход ее кластеризации на базе информации о статистике переходов между узлами.

Пусть дана гипертекстовая структура $H = \{P, L\}$, состоящая из множества гипертекстовых страниц $P = \{p_1, \ldots, p_n\}$ и множества гиперссылок между ними $L = \{l \mid \exists p_1, p_2 \in P : p_1, p_2 \in l(p_1, p_2)\}$. Так как гиперссылки имеют направленный характер, то $\forall p_1, p_2 \in P, l(p_1, p_2) \neq l(p_2, p_1)$.

Поставим данной структуре в соответствие орграф $G = \{V, E\}$, в котором V множество вершин графа, E – множество его ребер. Заметим, что полученный граф является невзвешенным. Чтобы применить к этому графу алгоритмы для взвешенных графов, допустим, что вес каждого ребра равен единице. Тогда получим следующий граф: $G = \{V, E, w\}$, где для любых $x, y \in V$ вес ребра $w(e(x, y)) = 1$.

КЛАСТЕРИЗАЦИЯ НА ОСНОВЕ СТАТИСТИКИ ПЕРЕХОДОВ

Рассмотрим систему отслеживания статистики посещений пользователями гипертекстовых страниц и переходов между ними. Пусть данная система внедрена на $P' \subset P$ гипертекстовых страницах.

Обозначим маршрут пользователя, переходящего между страницами по гиперссылкам, как упорядоченный список посещенных им страниц:

$$r = \{p'_1, \ldots, p'_n\}, p'_i \in P', i \in [1..n]. \quad (1)$$

Пусть на заданном, интересующем нас интервале времени, гипертекстовую структуру посетили m пользователей. Запишем их, согласно выражению (1):

$$\begin{aligned} r_1 &= \{p'_{11}, \ldots, p'_{1n}\}, p'_{1i} \in P', i \in [1..n] \\ r_2 &= \{p'_{21}, \ldots, p'_{2k}\}, p'_{2i} \in P', i \in [1..k] \\ &\ldots \\ r_m &= \{p'_{m1}, \ldots, p'_{ml}\}, p'_{mi} \in P', i \in [1..l]. \end{aligned} \quad (2)$$

Далее, сгруппируем маршруты (2), первые элементы которых совпадают и получим набор множеств маршрутов $R_1, \ldots, R_t$, $t \leq m$ общей стартовой страницей p'_{i1}.

Поместим все уникальные страницы, содержащиеся в маршрутах каждого множества R_i, $i \in [1..t]$, в $P_i = \{p \in P \mid \exists r \in R_i : p \in r\}$.

Для каждого полученного множества R_i, составим квадратную матрицу M^i размерности $N \times N$, где $N = |P_i|$ и заполним её значения следующим образом:

$$M_{j,k}{}^{i} = \left| \left\{ \begin{array}{c} r \in R_i \mid \exists p_j, p_k \in P_i, \\ p'_x, p'_{x+1} \in r, x \in [1..|r|] : \\ p_j = p'_x, p_k = p'_{x+1} \end{array} \right\} \right|. \quad (3)$$

То есть, каждый элемент матрицы $M^i_{j,k}$ (3) представляет количество путей c_{jk}, в которых страницы p_j, p_k были посещены последовательно:

$$\begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & \ddots & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}, 0 \le c_{jk} \le |R_i|, c_{jk} \in \mathbb{N}. \quad (4)$$

Данная матрица (4) будет являться матрицей смежности ориентированного взвешенного графа $G_i = \{V_i, E_i, W_i\}$, в котором вершины множества $V_i = P_i$ соответствуют посещенным страницам, ребра $E_i = \{e \mid \exists p_j, p_k \in P_i : c_{jk} \neq 0 \to p_j, \quad p_k \in e(p_j, p_k)\}$ соответствуют хотя бы единожды совершенным переходам, а веса $W_i = \{w \mid \forall e_{jk} \in E_i \wedge c_{jk} \neq 0 : w_{jk} = c_{jk}\}$ показывают количество таких переходов.

Сопоставим полученный граф G_i с исходным графом гипертекстовой структуры G. Так как множество страниц с отслеживанием статистики посещения пользователей не превосходит общее число страниц, а каждая страница соответствует ровно одной вершине графа, то $V_i \subset V$.

Что касается ребер, то мы допустили, что пользователь совершает переходы между страницами по гиперссылкам, каждой из которых соответствует ребро в исходном графе. Поэтому очевидно, что E_i не может превосходить E. Следовательно, граф G_i можно считать подграфом G.

Чтобы исключить случайные переходы с одной страницы на другую, вычтем из множества вершин V_i вершины, которые являются конечными для ребер с весами, равными 1. Получим множество X_i:

$$X_i = \{v \in V_i \mid \forall u \in V_i : w(e(v,u)) \rangle 1, \; w \in W_i, e \in E_i\} \quad (5)$$

Повторим данную процедуру для каждого из множеств маршрутов $R_1, \ldots, R_t$, $t \le m$. В результате получим набор множеств вершин $X_1, \ldots, X_t$, $t \le m$, причем их объединение $X = \bigcap_1^t X_i$ будет являться подмножеством множества вершин V исходного графа G.

Таким образом, получим кластеризацию графа гипертекста на базе статистики посещений страниц: $X_{st} = \{X_1 \ldots X_t\}$, $X_i \subset V$, $i \in [1..t]$.

Рассмотренный выше метод мы применяем к графу, построенному на основе гипертекстовой структуры сайта. Полученные кластеры документов могут быть использованы для повышения эффективности извлечения ключевых слов в задаче информационного поиска. При оптимальном разбиении на кластеры, частотные характеристики слов, специфичных конкретному домену, повышаются по сравнению с частотами вхождения общеупотребимых слов.

Покажем различие в выделении ключевых слов из одиночных документов и из кластеров документов. Рассмотрим документ d. Опустив порядок слов в нем, мы можем представить его в виде двух множеств: $T_d = \{t_1, \ldots, t_n\}$ – множества уникальных слов (лексем), встречающихся в документе и $C_d = \{c_1, \ldots, c_n\}$ – множества количеств вхождения каждого слова в документ.

Определим функцию $tf(t,d)$, которая определяет частоту вхождения каждого слова t в документ d:

$$tf(t,d) = \frac{c_t}{\sum_1^n c_i}. \quad (6)$$

Образом данной функции будет вес каждого слова в документе:

$$W_d = \{\forall t_i \in T_d : w_i = tf(t_i, d)\}. \quad (7)$$

Теперь мы можем определить функцию $kw(d)$, которая бы определяла множество ключевых слов для документа d, наиболее статистически достоверно описывающих его содержание:

$$kw(d) = \{t \in T_d \mid \forall t_i \in T_d : w_i > w_0 \to t = t_i\}. \quad (8)$$

В данном случае, w_0 является пороговым весом, который позволяет отбросить редко встречающиеся слова в документе. Значение

данной переменной подбирается эмпирическим путем на основе результатов анализа частот вхождения термов в документах, с целью исключения ключевых слов с низкой частотой вхождения.

Результаты отображения функции (8) формируют множество T_{kw}:

$$T_{kw} = kw(d), T_{kw} \subset T_d. \quad (9)$$

Пусть для каждого документа определен список релевантных ключевых слов T_{rel}, полученный при помощи экспертной оценки, который является эталоном.

Для каждого документа, будем сопоставлять множества найденных ключевых слов с эталонными множествами. Определим три функции:

Точность:

$$Precision(T_{kw}, T_{rel}) = \frac{|T_{kw} \cap T_{rel}|}{|T_{kw}|}. \quad (10)$$

Полнота:

$$Recall(T_{kw}, T_{rel}) = \frac{|T_{kw} \cap T_{rel}|}{|T_{rel}|}. \quad (11)$$

Сбалансированная F-мера:

$$F(T_{kw}, T_{rel}) = \frac{2Recall(T_{kw}, T_{rel}) * Precision(T_{kw}, T_{rel})}{Recall(T_{kw}, T_{rel}) + Precision(T_{kw}, T_{rel})}. \quad (12)$$

Функция F-меры (12) показывает, насколько множество найденных ключевых слов T_{kw} (9) соответствует эталонному множеству релевантных слов. Максимальное значение функции равно единице и достигается при $T_{kw} = T_{rel}$.

Отсюда следует ряд требований к исходному документу, необходимых для достижения максимальной релевантности указанным методом поиска ключевых слов:

1) $T_{rel} \subset T_d$, т. е. исходный документ должен содержать в себе все релевантные ключевые слова;

2) $\forall t_r \in T_{rel}, t_{nr} \in T_d \setminus T_{rel} : tf(t_r, d) > tf(t_{nr}, d)$, т. е. релевантные ключевые слова должны встречаться чаще любых нерелевантных слов.

Очевидно, что не каждый документ может соответствовать указанным требованиям, поскольку в числе релевантных слов могут присутствовать синонимы тех, что встречаются в тексте документа, частота релевантных слов в документе может быть ниже частоты широко распространенных и общеупотребимых слов языка, а также длина документа может быть слишком короткой и не репрезентативной для проведения анализа.

В частном случае, проблемы могут быть решены внесением изменений в сам документ (что иногда невозможно ввиду их большого количества или запрета на их изменение), либо усложнением методики анализа.

Покажем, как данные проблемы могут быть решены без изменения функций статистического анализа (10–12), а также без внесения изменений в исходные документы, при условии, что мы имеем коллекцию документов D.

Пусть задан оператор $f(D) = \{D_1, \ldots, D_k\}$, который некоторым образом разбивает коллекцию документов на k подмножеств. Каждое из этих подмножеств состоит в общем случае из нескольких документов: $D_i = \{d_1, \ldots, d_m\}$, $i \in [1..k]$.

Объединим все документы подмножества D_i в один путем последовательной конкатенации текстов каждого из документов. В результате получим документ d', который представим в виде множеств C_d' и T_d' через соответствующие множества C_d и T_d каждого из исходных документов:

$$T_d' = \bigcup T_{d_j}, T_{d_j} = \{t_1, \ldots, t_n\}, j \in [1..m] \quad (13)$$

$$C_d' = \{(c_{11} + c_{21} + \ldots + c_{m1}), \ldots, (c_{1n} + c_{2n} + \ldots + c_{mn})\} = \{c_1', \ldots, c_n'\}. \quad (14)$$

Выбор той или иной функции группировки страниц, по-своему формирующей коллекцию документов, очевидно, будет приводить к различным значениями T_d' (13) и C_d' (14). Определим критерии эффективности такой функции, исходя из требований к документам (13) и (14) выше, необходимых для максимизации F-меры:

1) T_d' должно включать в себя все релевантные ключевые слова.

2) Частота вхождения релевантных слов tf (6), рассчитываемая по C_d', должна быть выше соответствующей частоты нерелевантных слов.

Иными словами, функция группировки должна выбирать документы таким образом, чтобы конкатенация их содержимого: а) содержала как можно больше релевантных ключевых слов, б) частота релевантных слов после конкатенации документов оставалась бы выше частоты нерелевантных.

Рассмотренный выше метод кластеризации графовой модели гипертекста на основе данных о переходах между узлами вполне может быть применен для получения разбиений документов, соответствующим гипертекстовым страницам, на кластеры.

МЕТОДИКА ЭКСПЕРИМЕНТА

Покажем на практике эффективность предлагаемого метода кластеризации с точки зрения выявления групп семантически близких документов. Производить численную оценку эффективности будем при помощи функций Recall, Precision и F-measure (10–12), рассмотренных выше.

Для оценки эффективности подхода, использовались следующие средства:

1) Определение семантической близости групп веб-страниц путем ручной оценки вошедших в кластеры страниц.

2) Извлечение и сопоставление с эталоном ключевых слов как из одиночных веб-документов, так и из кластеров веб-страниц.

Первый метод оценки предполагает значительный объем ручной работы по просмотру каждого кластера и каждой веб-страницы из него. Однако он дает качественные результаты, поскольку эксперт самостоятельно оценивает содержимое веб-страницы и сам оценивает схожесть с другими страницами.

Другой подход основан на извлечении ключевых слов и выполняется автоматически для каждой веб-страницы. Для оценки качества кластеризации с точки зрения семантической близости, используется статистический анализ – веб-страницы с похожим по смыслу содержанием будут включать в себя большее количество терминов, соответствующих данной области. В то же время, анализ одиночных страниц покажет усредненные показатели как для терминов, так и для общеупотребимых слов.

ЭТАЛОН ДЛЯ СРАВНЕНИЯ РЕЗУЛЬТАТОВ КЛАСТЕРИЗАЦИИ И АНАЛИЗА КЛЮЧЕВЫХ СЛОВ

Оба метода оценки эффективности рассматриваемого метода требуют эталонного множества веб-страниц, соответствующих кластеру, и релевантных ключевых слов, соответствующих веб-страницам.

Для оценки работы кластеризации, вручную были сформированы наборы возможных кластеров веб-страниц, где каждый кластер соответствует определенной тематике или разделу веб-сайта. Чтобы оценить эффективность, вычисленные кластеры сопоставлялись с подходящим эталонным кластером. Среднее количество совпадений эталонных кластеров и вычисленных кластеров для веб-сайта показало точность подхода, с одной стороны.

С другой стороны, для оценки эффективности извлечения ключевых слов, экспертами был составлен набор наиболее релевантных ключевых слов для репрезентативной выборки страниц веб-сайта. И чтобы оценить эффективность статистического анализа кластеров страниц, подсчитывался уровень соответствия между релевантными словами из эталона и автоматически найденными ключевыми словами.

ИСТОЧНИКИ ДАННЫХ

Для оценки эффективности предлагаемого подхода, были использованы три веб-сайта в качестве источников данных. Первый ресурс – веб-сайт Саратовского Государственного Технического Университета им. Гагарина Ю. А.[1] Всего сайт насчитывает около 12 000 веб-страниц.

Второй источник данных – веб-сайт Информационно-образовательного центра «Виртуальный филиал Русского музея»[2]. Веб-сайт насчитывает более 1500 страниц.

Третий веб-сайт, Исследовательской группы Agile Knowledge and Semantic Web (AKSW

[1]http://www.sstu.ru/

[2]http://museum.seun.ru/

Research Group)[3]. На основном домене веб-сайта обнаружено около 500 веб-страниц.

Все три веб-сайта подключены к системе сбора статистики посещений Google Analytics[4], счетчик посещений которой установлен на всех страницах веб-сайтов, за исключением служебных. В выборку попали данные о переходах пользователей между страницами за последний месяц.

РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ МЕТОДА

Процессы обхода веб-сайтов, кластеризации и анализа ключевых слов обрабатывались на аппаратном обеспечении с четырехъядерном процессором Intel(R) Xeon(R) E5-2670 частотой 2.50ГГц и с 64Гб общей оперативной памяти.

На первом этапе эксперимента был выполнен процесс обхода веб-сайтов – источников данных, и построение графовой модели данных веб-сайтов. Результаты построения графовой модели – количество вершин и ребер, средняя степень связанности – представлены в табл. 1.

Для разбиения полученных графов веб-сайтов на кластеры, через программный интерфейс GoogleAnalytics были выгружены данные о переходах между страницами веб-сайтов за последний месяц. После чего, на основании маршрутов пользователей, были построены кластеры семантически близких документов, информация о которых представлена в табл. 2.

К полученным кластерам был применен первый способ оценки: в ручном режиме, найденные кластеры сопоставлялись с эталоном. Как видно из результатов сопоставления на рис. 1, 90 и более процентов эталонных кластеров включены в автоматически найденные кластеры. То есть, найденные кластеры также объединяются по принципу схожего по теме содержания.

Далее, были взяты отдельные полученные кластеры документов и к ним применен метод статистического анализа текста к конкатенации содержимого каждого из документов (таким образом, получим один макро-документ, включающий содержимое всех веб-страниц). Для индексации и выделения ключевых слов из текста, использовалась библиотека Apache Lucene[5].

В результате, получено множество входящих в данный макро-документ ключевых

Таблица 1

Метрики построенной графовой модели для веб-сайтов

	sstu.ru	**aksw.org**	**museum.seun.ru**
Количество вершин	11987	563	1780
Количество ребер	562996	12103	47193
Средний показатель связанности	46.96	21.49	26.51
Время построения, сек.	20 573	934	3 190

Таблица 2

Результаты, полученные при кластеризации трех веб-сайтов на основе статистики посещений

	sstu.ru	**aksw.org**	**museum.seun.ru**
Число кластеров	44	36	5
Средний размер кластера	32.5	27.4	33.4
Максимальный размер кластера	152	150	112
Всего страниц покрыто кластерами	573 (4.7%)	195 (34.6%)	254 (14.2%)
Посещений за последний месяц	13721	4654	314
Длительность кластеризации, сек.	43	32	29

[3]http://aksw.org/

[4]http://www.google.com/analytics/

[5]http://lucene.apache.org/core/

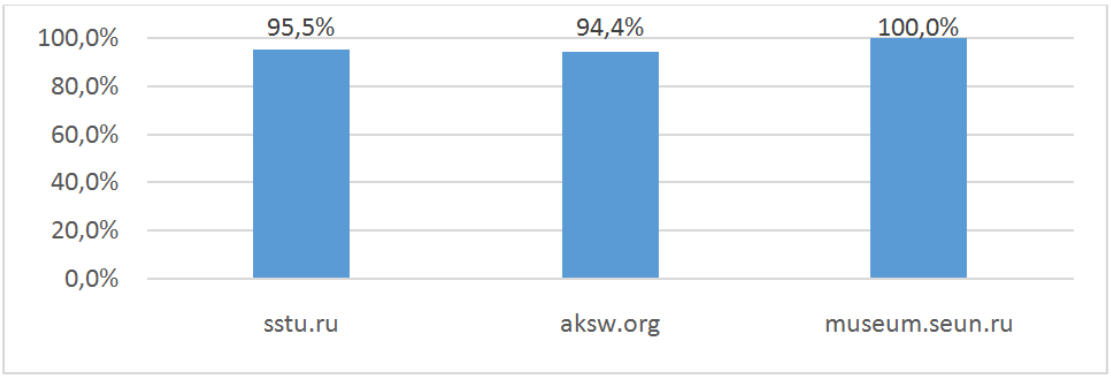

Рис. 1. Процент совпадения автоматически найденных кластеров с эталонными кластерами при кластеризации веб-сайтов sstu.ru, aksw.org и museum.seun.ru

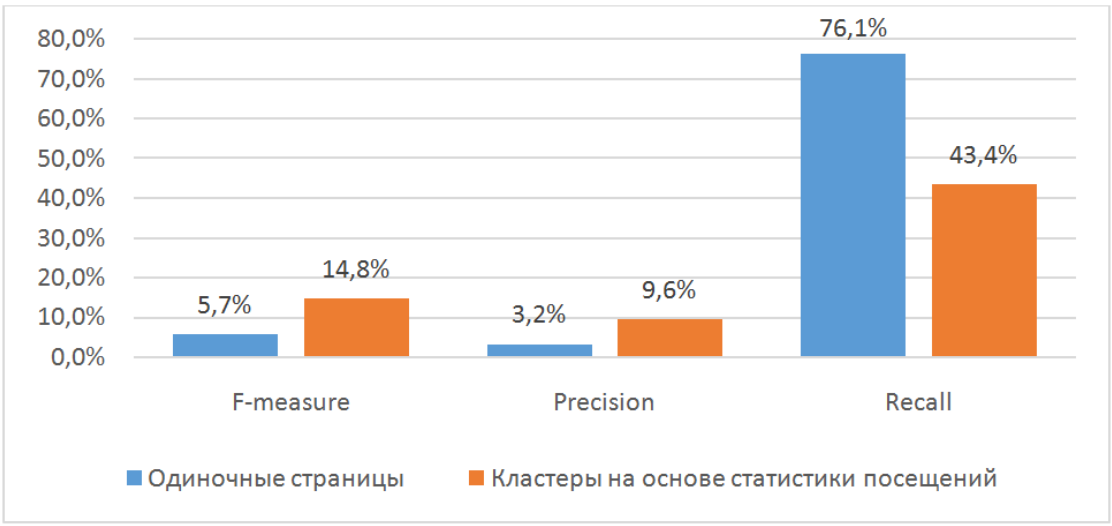

Рис. 2. Средние значения показателей F-measure, Precision и Recall по трем сайтам

слов с распределением их частот. После отсева слабо значимых ключевых слов (с частотой ниже 0.05), оставшееся множество ключевых слов было сопоставлено с эталоном и посчитаны значения трех функций: Recall, Precision и F-measure (10–12).

Подобная процедура также была выполнена для статистического анализа отдельных веб-документов. Пороговое значение для данного случая было установлено ниже, равное 0.01, так как частоты вхождения слов в случае одиночных страниц оказались ниже. Результаты представлены в табл. 3.

Для наглядности полученных результатов, ниже на рис. 2 представлены средние значения показателей для всех трех веб-сайтов.

Таблица 3

Эффективность статистического анализа текста на кластере и на одиночных страницах веб-сайтов

	sstu.ru	aksw.org	museum.seun.ru
Одиночные страницы			
F-measure	6.1%	6.2%	4.8%
Precision	3.5%	3.3%	2.8%
Recall	69.4%	91.1%	67.9%
Кластеры на основе статистики посещений			
F-measure	19.2%	11.8%	13.3%
Precision	13.9%	7.1%	7.7%
Recall	46.8%	33.3%	50.1%

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Как видно из табл. 3, значения ключевой функции F-measure для всех трех веб-сайтов выше в случае анализа кластеров страниц, полученных методом кластеризации на основе статистики переходов между страницами, чем в случае анализа одиночных страниц, что говорит о более точном попадании в эталонный набор ключевых слов. В то же время, в результатах анализа одиночных страниц, значительную величину имеет Recall. Это легко объясняется тем, что при низком пороговом значении было получено больше ключевых слов, которые включили больше слов из эталона. Однако низкое значение Precision показывает, что при этом в конечный список ключевых слов попало множество лишних, общеупотребимых слов.

В целом, по полученным результатам можно судить об эффективности метода в плане получения кластеров семантически связанных страниц. Значения показателей F-measure и Precision при кластеризации выше, чем при анализе одиночных страниц – это означает, что попавшие в кластеры страницы содержат больше совпадений терминов, специфичных для конкретных страниц.

Существенные отличия между значениями Precision и Recall можно объяснить особенностями методики эксперимента. Так как эталоны ключевых слов составлялись вручную экспертами, размеры множеств эталонных слов для каждой страницы оказались невелики (в среднем, 3–5 слов), поэтому точность попадания в эталон (Precision) оказалась в целом невысокой, в то время как полнота (Recall) оказалась существенно выше за счет слов со средней и низкой частотой вхождения. Кроме того, при сопоставлении ключевых слов, полученных при помощи описанного метода, с эталонными не учитывались синонимы, что также сказалось на результатах. Влияние данных факторов на результаты может быть ослаблено за счет 1) расширения эталонных множеств ключевых слов, 2) усовершенствования техники сопоставления найденных ключевых слов с эталоном.

Сам метод также может быть улучшен и доработан. На данный момент, результаты напрямую зависят от качества исходных статистических данных, а именно количества посещений и процента веб-страниц, покрытых счетчиками. Одним из вариантов доработки является подход к учету не только маршрутов пользователей, но и их исходных поисковых запросов, по которым пользователи попали на анализируемый ресурс. Подобные идеи предлагались также в [9]. Другим вариантом является совместное использование предлагаемого метода с альтернативными способами кластеризации гипертекстовой структуры.

ЗАКЛЮЧЕНИЕ

В статье рассмотрен метод кластеризации гипертекстовых структур на основе статистики переходов между узлами, а также применение данного метода при кластеризации гипертекстовой структуры по принципу семантической близости документов. При кластеризации, данный метод не анализирует содержимое веб-документов, а опирается на графовую модель структуры гипертекста и заранее собранной статистики посещений. Это позволяет избавиться от предварительного индексирования всех веб-страниц и дополнительно дает возможность анализа документов без текстовой информации.

На примере трех веб-сайтов, показана эффективность подхода к кластеризации при помощи предлагаемого метода. Результаты работы метода оценивались по двум принципам: 1) сопоставление полученных кластеров с эталонными, на основании семантической близости документов в них; 2) применение статистического анализа к кластерам и сопоставление автоматически найденных ключевых слов с эталонными. Также приведены результаты статистического анализа одиночных страниц и сопоставление автоматически найденных в них ключевых слов с тем же эталоном. Результаты показывают более высокую эффективность кластеризации на основе предлагаемого в данной работе метода, по сравнению с анализом одиночных страниц.

Рассмотренный подход к кластеризации может быть применен на практике в различных областях, связанных с документами в Вебе. В веб-разработке, подобная кластеризация применяется для построения адаптивных интерфейсов, подсказывающих пользователю релевантные страницы. В веб-поиске, применяется разбиение результатов выдачи на группы по тематикам. В веб-администрировании и реинжиниринге, результаты кластеризации предоставляют данные о структуре веб-сайта и возможных направлениях его реструктуризации. Предложенный метод также может быть использован в веб-аналитике и поисковой оптимизации, с целью анализа кластеров веб-сайта с семантически близкими документами.

Отметим некоторые ограничения рассмотренного в данной работе подхода. Во-первых, применяемая графовая модель соответствует статической гипертекстовой структуре. Если структура веб-сайта будет динамической (к примеру, содержать асинхронные AJAX-запросы, добавляющие в процессе навигации в пользователя новые элементы структуры), то предлагаемая графовая модель будет не в полной мере отражать реальную структуру веб-сайта. Использование более сложной модели позволит снять это ограничение.

Другой важной особенностью является необходимость в наличии статистики посещений страниц веб-сайта пользователями. Современные системы, такие как GoogleAnalytics или Яндекс Метрика, предлагают удобные средства сбора статистики и понятные документированные программные интерфейсы. Однако некоторые веб-сайты могут не использовать счетчики данных инструментов. В таком случае, информация о переходах между страницами может быть получена из журнала доступа к веб-ресурсу, который чаще всего ведется контейнером веб-приложения.

СПИСОК ЛИТЕРАТУРЫ

1. *Сегалович И. В.* Как работают поисковые системы // Мир Internet. – 2002 г. – 10. – С. 24–32.

2. *Hollink V., Someren M. and Wielinga B. J.* Navigation behavior models for link structure optimization // User Model User–Adapted Interaction. – [s.l.] : Springer, 2007. – 4 : Vol. 17. – pp. 339–377.

3. *Ferrara E.* [et al.] Web data extraction, applications and techniques: A survey [Article] // CoRR. – 2012.

4. *Devika K. and Surendran S.* An overview of web data extraction techniques // International Journal of Scientific Engineering and Technology. – 2013. – 2(4).

5. *Артемов М. А., Владимиров А. Н. и Селезнев К. Е.* Обзор систем анализа естественного текста на русском языке // Вестник Воронежского ун-та. Серия: Системный анализ и информационные технологии. – 2013. – № 2. – С. 189–194.

6. *Chakrabarti D. and Mehta R.* The paths more taken: matching dom trees to search logs for accurate webpage clustering // Proceedings of the 19th international conference on World wide web. – [s.l.] : ACM, 2010. – pp. 211–220.

7. *Arasu A. and Garcia–Molina H.* Extracting structured data from web pages // Proceedings of the 2003 ACM SIGMOD international conference on Management. – [s.l.] : ACM, 2003. – pp. 337–348.

8. *Кошик А.* Веб-аналитика 2.0 на практике. – [б.м.] : Диалектика, 2011.

9. *Beydoun G.* Formal concept analysis for an e-learning semantic web // Expert Systems with Applications. – [s.l.] : Elsevier, 2009. – 36. – pp. 10952–10961.

10. *Салин В. С., Папшев С. В. и Сытник А. А.* About a Method of Educational Web Resources Optimization // International Conference on Engineering Education and Research (iCEER 2013). – Маракеш : [б.н.], 2013. – С. 829–835.

Салин Владимир Сергеевич – аспирант Саратовского государственного технического университета (СГТУ) им. Гагарина Ю. А.
E-mail: salinvs@gmail.com

Папшев Сергей Владимирович – кандидат физико-математических наук, доцент кафедры информационных систем и технологий СГТУ им. Гагарина Ю. А.
E-mail: spapshev@list.ru

Сытник Александр Александрович – доктор технических наук, Лауреат премии Президента РФ, Заслуженный деятель науки Российской Федерации, профессор, заведующий кафедрой информационных систем и технологий СГТУ им. Гагарина Ю. А.
E-mail: as@sstu.ru

Salin Vladimir S. – post-graduate student of the department of information systems and technologies, Yuri Gagarin State Technical University of Saratov (SSTU).
E-mail: salinvs@gmail.com

Papshev Sergey V. – PhD, associate professor of the department of information systems and technologies, SSTU.
E-mail: spapshev@list.ru

Sytnik Alexander A. – Doctor of technical sciences, Laureate of the President prize of the Russian Federation, Honored Worker of Science of Russian Federation, professor of the department of information systems and technologies, SSTU.
E-mail: as@sstu.ru