УДК 004:378.1

РАСПРЕДЕЛЕННАЯ СИСТЕМА СБОРА И АНАЛИЗА ЦИФРОВОГО СЛЕДА ОБУЧАЮЩЕГОСЯ ВУЗА

Михайлова С.С., Данилова С.Д., Веселов А.В.

Восточно-Сибирский государственный университет технологий и управления, Улан-Удэ, e-mail: ssmihailova@mail.ru, dan-soelma@yandex.ru, xayk2803@gmail.com

В статье проведены исследования и разработка распределенной системы сбора и анализа цифрового следа обучающегося вуза. Проведено краткое описание модели системы, проведен сравнительный анализ алгоритмов решения задач сбора и анализа цифрового следа, разработаны рекомендации по построению оптимальной системы по сбору цифрового следа студента. Представлено описание процесса проектирования модели электронной траектории образования обучающихся, включающей в себя модель распределенной системы сбора и анализа цифрового следа обучающегося вуза. Порядок движения информации в системе состоит из получения цифрового следа с компьютера, передачи его на сервер данных, загрузки данных с сервера данных на сервер обработки с целью последующего анализа посещенных сайтов путем сравнения их с исходным документом, вывода итогового коэффициента занятости студентов на занятиях в виде процента посещенных сайтов, относящихся к процессу обучения. Для решения задачи анализа текста будет использоваться готовое решения на основе BigARTM. Обоснован выбор методов и средств реализации распределенной системы. Разработаны технические требования и алгоритмы работы системы по сбору и анализу цифровых следов.

Ключевые слова: цифровой след, распределенная система сбора и анализа, траектория образовательного процесса, тематическое моделирование, парсинг

DISTRIBUTED COLLECTION SYSTEM AND ANALYSIS OF THE DIGITAL FOOTPRINT OF A STUDENT UNIVERSITY

Mikhaylova S.S., Danilova S.D., Veselov A.V.

East Siberia State University of Technology and Management, Ulan-Ude, e-mail: ssmihailova@mail.ru, dan-soelma@yandex.ru, xayk2803@gmail.com

In this article, research and development of a distributed system for collecting and analyzing the trace of a student university was carried out. A brief description of the system model was carried out, a comparative analysis of the algorithms for solving the problems of collecting and analyzing a trace was carried out, recommendations were developed for building a unique system for collecting a student's previous trace. A description of the process of developing models of electronic trajectories of education of students is presented, including a model for the distribution of a system for collecting and analyzing traces of a student at a university. The mechanism of information movement in a system based on reading data from calculations, transferring it to a data server, downloading data from a data server to a data processing server, taking into account the analysis of visited sites of visited sites, to the learning process . The solution to the text analysis problem will be ready based on BigARTM. The choice of methods and means of implementing a distributed system is substantiated. Technical requirements and algorithms for the operation of the system for collecting and analyzing digital traces have been developed.

Keywords: digital footprint, distributed collection and analysis system, educational process trajectory, topic modeling, parsing

В современном мире обучающийся не привязан больше ни к учителю, ни к своей среде обитания. Цифровые коммуникационные технологии дают ему возможность выбирать, где и чему учиться, в какой среде развиваться, в какую деятельность включаться.

Успех в этой новой, все более цифровой системе образования определяет, насколько обучение адаптирует человека к текущему социально-экономическому укладу. Его развитие все больше зависит от способности постоянно адаптироваться, изменяться, эффективно осваивать новую деятельность и приобретать новые профессиональные качества [1].

Это предъявляет новые, принципиально другие требования к системе образования. В мире, где студент имеет возможность выбирать, где, как, когда и чему учиться, задача системы не в том, чтобы обеспечить качественно высокий уровень каждого конкретного преподавателя, обучающего конкретному предмету, а в том, чтобы:

− обеспечить студента инструментами для осознанного выбора;

− технологиями навигации в пространстве образовательных возможностей;

− надежными средствами оценки эффективности того или иного образовательного процесса.

Цифровой след – это постоянно пополняемый набор данных, включающий значения показателей, созданных самими студентами [1]. В их числе рефераты и обзоры литературы, курсовые, отчеты по практикам, описания проектов, выпускные квалификационные работы, эссе и мотивационные письма на конкурсы. Именно

из этих текстов современные методы анализа данных позволяют извлечь объективную информацию для диагностики профессиональной компетентности выпускника и выявить факторы, которые повлияли на ее формирование [2].

Целью исследования является разработка системы траектории электронного обучения на основе цифрового следа с использованием распределенной системы сбора и анализа цифрового следа обучающегося вуза.

Материалы и методы исследования

К основным методам исследования относятся анализ литературы, моделирование, вычислительный эксперимент. Анализ литературы (интернет-статьи о методах и способах организации систем сбора данных), связанной с обработкой цифрового следа, использован при анализе предметной области. Метод моделирования использован при создании модели распределенной системы. Экспериментальный метод исследования использован для проверки работоспособности созданной системы. Разработка системы сбора и анализа цифрового следа студента включает совокупность выбранных методов и технологий. К ним относятся работа с базами данных на основе SQL запросов, стемминг текста, парсинг сайтов, работа с базой данных PostgresSQL, тематическое моделирование текстов при помощи BigARTM.

Результаты исследования и их обсуждение

Электронное образование – это система обучения, осуществляемая при помощи информационных и электронных технологий. На данный момент развитие электронной информационно-образовательной среды (ЭИОС) является одним из приоритетных направлений всех образовательных учреждений, так как это позволяет всем получать своевременный, удобный и равноценный доступ к материалам для обучения, а сам процесс обучения становится более прозрачным и понятным как студентам, так и людям вне образовательных учреждений. Основное положение, регламентирующее, из чего должна состоять электронная образовательная среда, включает в себя организационно-методические средства, совокупность технических и программных средств хранения, обработки, передачи информации, обеспечивающая оперативный доступ к информации и осуществляющая образовательные научные коммуникации [3].

Траектория образовательного процесса – путь или движение по образовательной среде обучающегося индивида. Основным отличием новой системы станет возможность ее анализа, а значит, и появления способа повлиять на нее.

Отличительной особенностью модели траектории электронного обучения на основе цифрового следа станет включение в нее распределенной системы сбора и анализа цифрового следа обучающегося [4]. Обычно цифровой след собирается по результатам достижений обучающихся, что не дает представления о путях к этим достижениям. Поэтому было решено изменить место его сбора в процессе получения знаний по конкретной дисциплине.

Основной задачей исследования является отслеживание и нахождение цифрового следа пользователей, его обработки и извлечения из него полезной информации. В работе в качестве основного источника информации цифровых следов рассмотрен компьютер дисплейного класса университета. Сбор и анализ цифрового следа обучающегося включает в себя сбор данных с компьютера дисплейного класса; передача полученных данных на сервер; сохранение данных в хранилище. После сбора данных будет происходить их первоначальный анализ для определения наиболее частых употребляемых слов или фраз на посещенных ими сайтах при помощи вычисления тематики сайтов, а также частотности их посещения [5]. Для этого надо решить следующие задачи, связанные с разработкой и реализацией способов сбора данных с компьютера, хранения и анализа, просмотра и вывода полученных данных, а также более детального рассмотрения каждой записи данных в системе.

При разработке ПО для анализа цифрового следа в университете будут использованы следующие модули:

− приложения C# для сбора истории посещенных сайтов из браузера Google Chrome;

− PostgreSQL для хранения всех полученных данных;

− алгоритм классификации данных при помощи алгоритмов через получение тематики текстов (собранных при помощи алгоритмов, описанных ранее), реализованные на языке Python.

Выбор алгоритма анализа данных проведен по следующим критериям: сложность разработки и настройки, скорость работы приложения, системные требования к аппаратному обеспечению.

Сама же обработка данных будет из двух этапов:

1) формирование выборки текста сайта при помощи парсера. Данный алгоритм будет рассмотрен в следующей главе;

2) построение модели классификации сайта по тематикам на основе полученных текстов, определение их приближенности к эталонному исходному документу. В качестве исходного документа можно использовать рабочие программы дисциплин (РПД).

Для решения поставленных задач необходимо реализовать следующие задачи:

1) изучить тематику дисциплины на основе РПД. Необходимо будет ввести в систему данный документ для возможности сравнения посещенных сайтов с выявленной тематикой и определения схожих элементов;

2) составить анализ тематик на основе тематического моделирования всех посещенных во время занятий сайтов в виде весов ключевых слов текста (токенов). Текст предварительно выбелить путем стемминга текста и удаления стоп-слов или словосочетаний, то есть частей текста, не несущих смысловой нагрузки;

3) построить модель и сравнить все построенные ранее тематические модели документов на предмет совпадений.

Преподаватель занимает место тьютора в процессе образования, что поможет разгрузить его время от монотонных действий, а студент может получить доступ к предлагаемой информации в любое удобное время. Основным плюсом является возможность быстрой эффективной оценки не только конкретных достижений в процессе обучения, но даже экспертной оценки самого процесса обучения, что и является основной новизной данной модели и работы в целом (рис. 1).

Модель работы системы по сбору и анализу цифрового следа обучающегося, включая все процессы от начала сборки данных до получения выходных данных, позволяет понять механизм распределения обязанностей внутри системы цифрового следа (рис. 2).

Система состоит из следующих компонентов: N клиентов для сбора информации (приложение на C# для сбора данных по истории посещения сайтов с браузера); сервера хранения данных (СУБД – PostgreSQL для хранения всех данных); сервера анализа и обработки данных (приложения на Python для классификации сайтов).

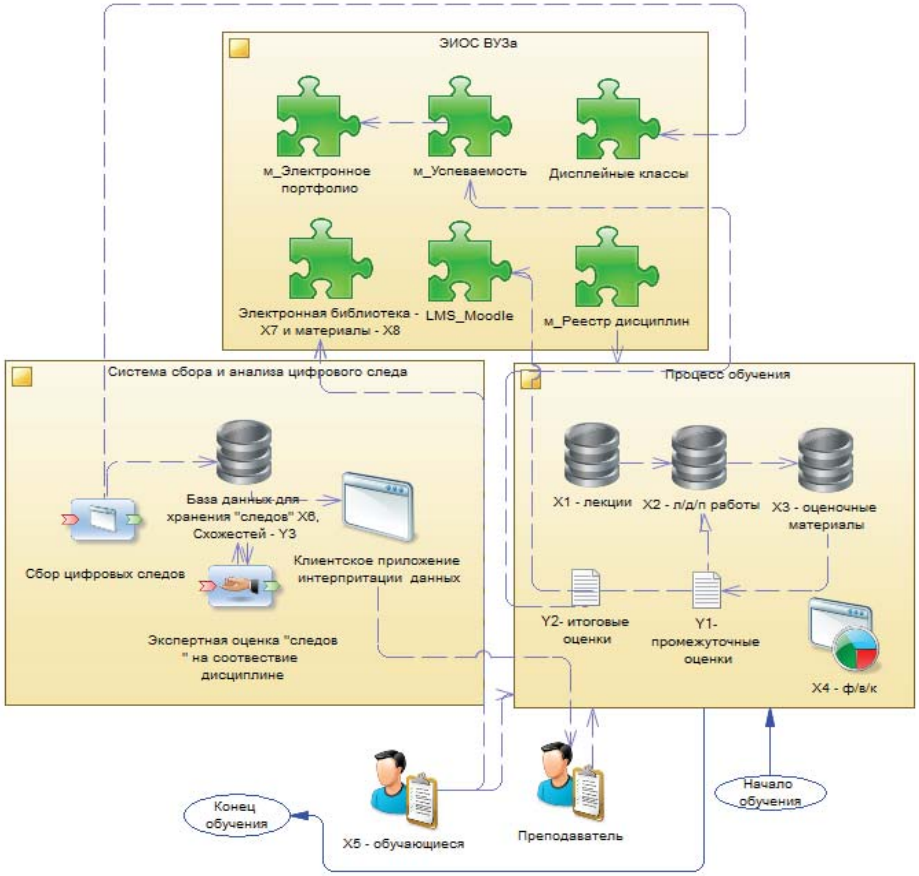

Рис. 1. Архитектура системы «Электронное образование»

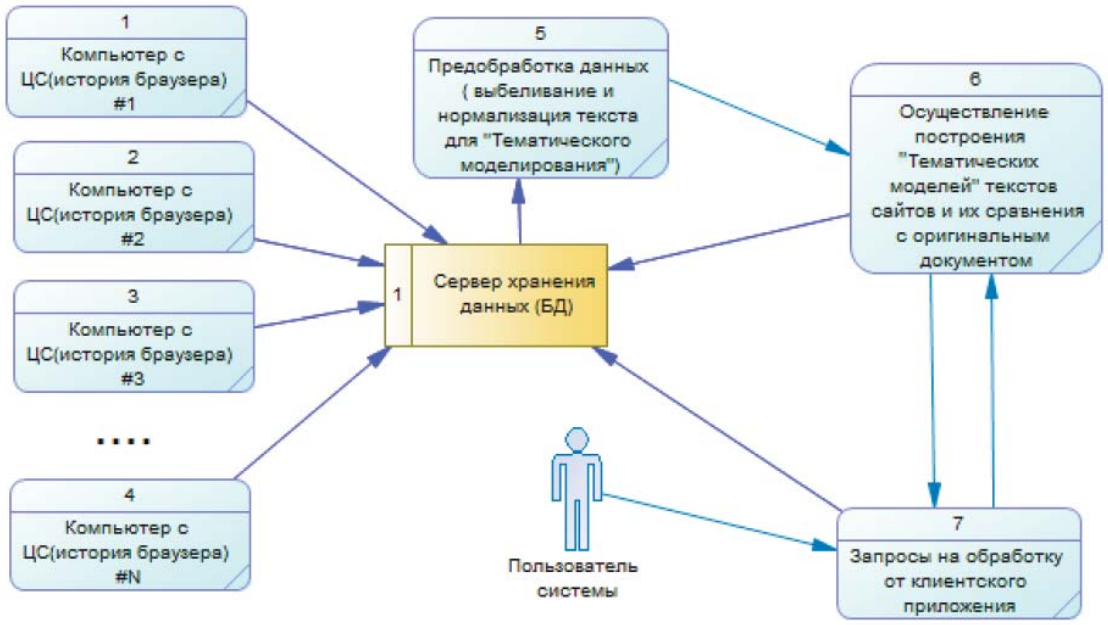

Рис. 2. Схема работы системы по получению и обработке цифрового следа

Основным отличием представленной модели является наличие клиента для работы с системой. Запросы на получение выходных данных, а также ввод исходного документа будет осуществляться при помощи клиентского приложения, но система будет полностью автономной и может работать без него. Клиентское приложение должно решать задачу создания удобного интерфейса работы с распределенной системой сбора и анализа цифрового следа.

Сервер системы должен поддерживать работу с несколькими потоками для осуществления своевременной обработки данных, используя параллельную работу, что обеспечит ускорение обработки данных.

Анализ данных будет осуществляться при помощи алгоритма тематического моделирования методом вероятностного латентно-семантического анализа. Формула разделения текстов на тематики, используемая в технологии bigARTM, имеет вид

$$p(d, w) = \sum p(t)p(w|t)p(d|t),\ t \in T$$

где t – тема;

$p(t)$ – неизвестное априорное распределение тем во всей коллекции;

$p(d)$ – априорное распределение на множестве документов: $p(d) = nd / n$ – длина документов;

$p(w)$ – априорное распределение на множестве слов: $p(w) = nw / n$, где nw – число вхождений слова w во все документы.

Пример разделения сайтов по тематическим моделям можно представить в виде схем тематического анализа двух документов из статей «Ботаника» и «Астрология» (рис. 3).

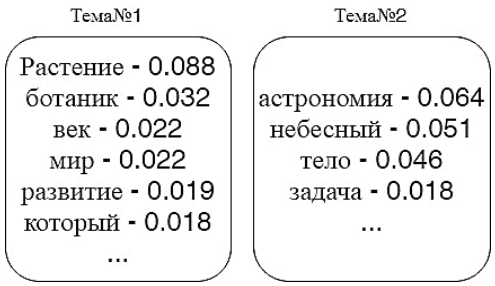

Рис. 3. Пример тематического моделирования текста

После получения списка вероятностей совпадения тем необходимо получить модель поведения студентов. Сравнение исходного документа со всеми сайтами, посещенными студентами во время занятий, отсеянных при помощи фильтра по времени посещения, использование описанного ранее алгоритма даст представление о качестве получаемой студентами информации из сети Интернет.

Система берет на входе исходный документ, описывающий изучаемую дисциплину, сравнивая со всеми известными сайтами, подходящими по условию. В итоге получаем коэффициент полезности времени, проведенного студентами на сайтах.

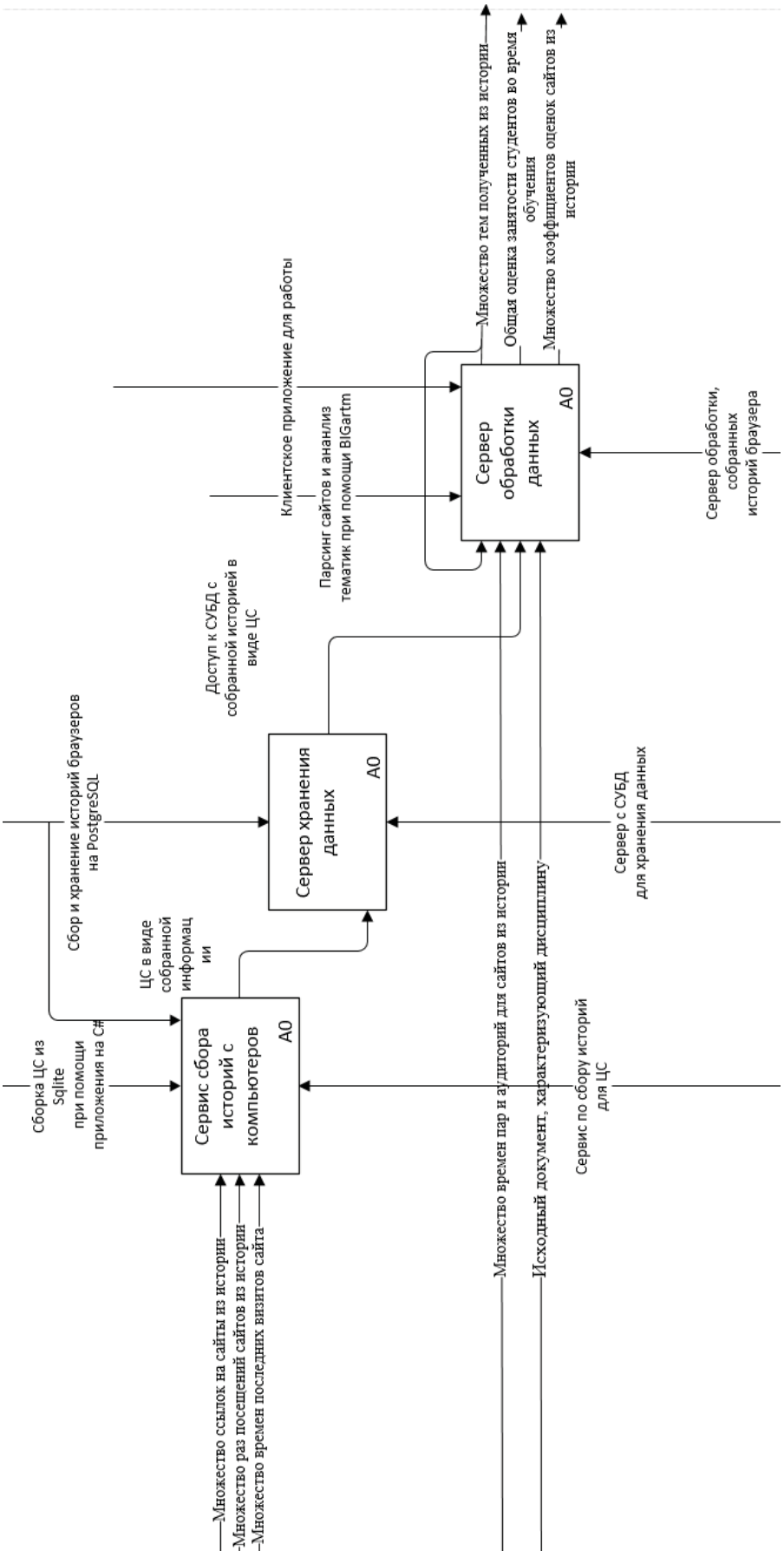

Рис. 4. Полная модель данных

Выборки данных для распознавания полезности сайта

Тестовая выборка	**Ожидаемый ответ**	**Тестовая выборка**	**Ожидаемый ответ**
Текст с сайта «youtube.com»	30% полезности	Текст с сайта «esstu.ru»	50% полезности
Текст с сайта «intuit.ru»	70% полезности	Текст с сайта «esstu.ru»	0% полезности
Текст с сайта «tiktok.com»	0% полезности	Текст с сайта «esstu.ru»	80% полезности

В таблице приведен пример ожидаемого ответа от программы. То есть система будет говорить, был ли полезен каждый сайт по отдельности, а потом интегрировать все полученные данные и давать итоговый вердикт.

При решении данной комплексной задачи будет использоваться готовое решение на основе библиотеки BigARTM, функциональная модель которой состоит из трех блоков (рис. 4).

Запрос такого анализа при помощи BigARTM будет осуществляться через удобный пользователю графический интерфейс, который получает на вход исходный документ, характеризующий дисциплину.

На основе полученных выходных данных система может дать рекомендации по сайтам, которые обучающиеся посещали чаще остальных. Это должны быть сайты, подходящие под тематику заданной дисциплины, так как они помогут обучающимся следующих курсов лучше освоить преподаваемую дисциплину.

Заключение

Предложенный в работе «цифровой переход» позволяет улучшить взаимодействие между участниками образовательного процесса и повысить качество освоения образовательной программы. Разработка системы обработки цифрового следа в образовательном учреждении является одним из ключевых элементов в цифровизации основного процесса и позволяет выстраивать индивидуальную образовательную траекторию обучающегося.

Список литературы

1. Цифровой след: новые задачи системы образования в эпоху данных [Электронный ресурс]. URL: https://habr.com/ru/post/513616/ (дата обращения: 10.03.2022).

2. Цифровой след показал профессиональную компетентность студентов [Электронный ресурс]. URL: https://openscience.news/posts/2401-tsifrovoy-sled-pokazal-professionalnuyu-kompetentnost-studentov / (дата обращения: 10.03.2022).

3. Абакумова Н.Н., Алексеев А.А. Информационная среда как ресурс для развития образовательного учреждения. Открытое и дистанционное образование. Томск, 2008. С. 35–41.

3. Положение «Об электронной информационно-образовательной среде в университете» [Электронный ресурс]. URL: https://esstu.ru/uportal/document/download.htm?documentId=18044 (дата обращения: 10.03.2022).

4. Шамсутдинова Т.М. Когнитивная модель траектории электронного обучения на основе цифрового следа // Журнал «Open education». 2020. V. 24. № 2. Секция Математическое обеспечение – 2020. URL: https://openedu.rea.ru/jour/article/view/726 (дата обращения: 10.03.2022).

5. «Цифровой след» как инструмент повышения качества исходных данных скоринг-моделей потенциального заёмщика [Электронный ресурс]. URL: http://earchive.tpu.ru/bitstream/11683/52597/1/conference_tpu-2018-C04_p389-390.pdf (дата обращения: 10.03.2022).