УДК 681.5

Жуманбаева С.К. [1], Пащенко Г.Н [1]

[1] Международный университет информационных технологий, Алматы, Казахстан

ПРОЕКТИРОВАНИЕ И РАЗРАБОТКА ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ ОБРАБОТКИ НАУЧНЫХ ТРУДОВ

Аннотация: *Рост объемов научных трудов вызывает необходимость широкого использования информационных технологий для этого процесса. В последнее время возрастает потребность в разработках информационных систем различного характера для работы с научными трудами. Следовательно, разработка и исследование таких систем является актуальной задачей. В данной статье рассматриваются существующие информационные системы и программы для обработки научных трудов, и выделяются особенности данных систем. Приводится анализ процессов и разбор алгоритма системы. В результате анализа сформирована оптимальная модель процессов. Приводится описание разработанной информационной системы для обработки научных трудов.*

Ключевые слова: *обработка научных трудов, информационная система, научные труды, плагиат, научная работа, проверка грамматики.*

Введение

Научная работа, в широком смысле слова, это исследование, включающее в себя любой формальный сбор данных, информации и фактов для развития знаний. Для некоторых людей написание научных статьей является частью бытовой жизни или работой, а для кого-то, это настолько чуждо, что они даже не читали ни одной научной статьи в своей жизни. Большинство хотя бы раз в жизни сталкивались с необходимостью написать научную статью, диплом, диссертацию. Для каждого написания научной статьи в первый раз было самым сложным и непонятным из всех. Возникают трудности, начиная от формулирования названия научной работы, заканчивая ее оформлением, не говоря уже о том, что сам процесс исследования и написания научной работы занимает не малое время. Поэтому возникает необходимость оптимизировать и повысить эффективность исследователей. В настоящее время, возрастает потребность в разработках информационных систем различного характера для работы с научными трудами. В связи с этим, разработка и исследование таких систем является актуальной задачей.

Самый лучший способ достижения оптимизации и эффективности это автоматизация процессов создания научной работы. Среди всех этапов написания статьи, проверка на плагиат, на грамматические ошибки и на соответствие с требованиями является самой муторной, но немаловажной частью работы. Именно эти вышеперечисленные этапы проверки можно автоматизировать и тем самым дать исследователям сосредоточиться на самом анализе и исследовании научной области.

Методы исследования

В статье используется метод сравнительного анализа, применяемый к существующим информационным системам и программам для обработки научных трудов. Сравнение информационных систем проводится по нескольким критериям. Полученные данные анализируются и выделяются особенности данных систем. В результате анализа сформирован список необходимых функциональных требований к разрабатываемой информационной системе, создана модель процессов и продуман алгоритм.

Результаты исследования

AS-IS - модель "как есть", модель существующего состояния системы. Данная модель позволяет систематизировать протекающие в данный момент процессы, а также использу-

мые информационные объекты [1]. На основе этого выявляются уязвимые места в организации и взаимодействии бизнес-процессов, определяется необходимость тех или иных изменений в существующей структуре.

После детального исследования процессов работы аналогов созданной системы создана модель процессов, чтобы оптимизировать и улучшить систему. На рисунке 1 показана модель процессов AS-IS, она состоит из двух пулов и показывает основной процесс исправления документа пользователя на сайте, начиная с его входа на сайт, заканчивая сохранением и скачиванием документа.

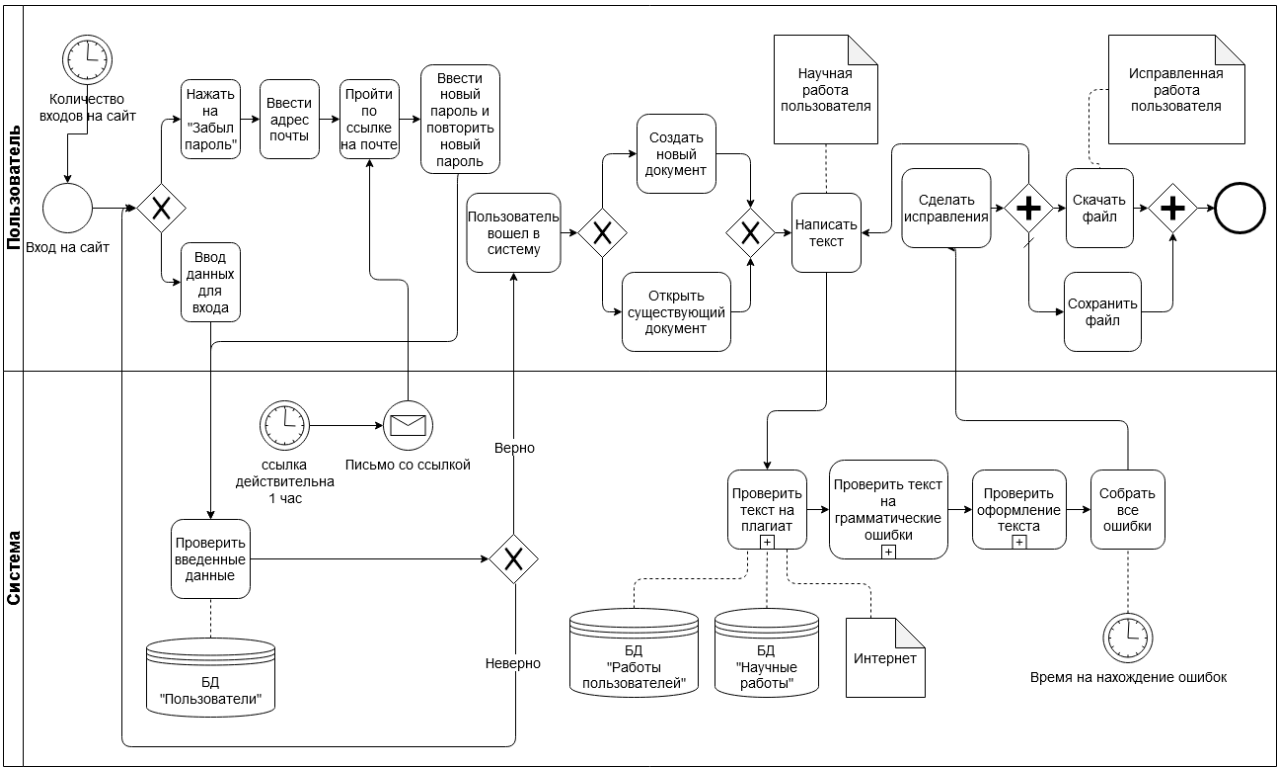

Рисунок 1 - Модель процессов AS-IS

TO-BE - модель "как должно быть". Как правило, данная модель создается на основе AS IS, с устранением недостатков в существующей организации бизнес-процессов, а также с их совершенствованием и оптимизацией. Это достигается за счет устранения выявленных на базе анализа AS-IS уязвимых мест. На рисунке 2 изображена модель процессов TO-BE.

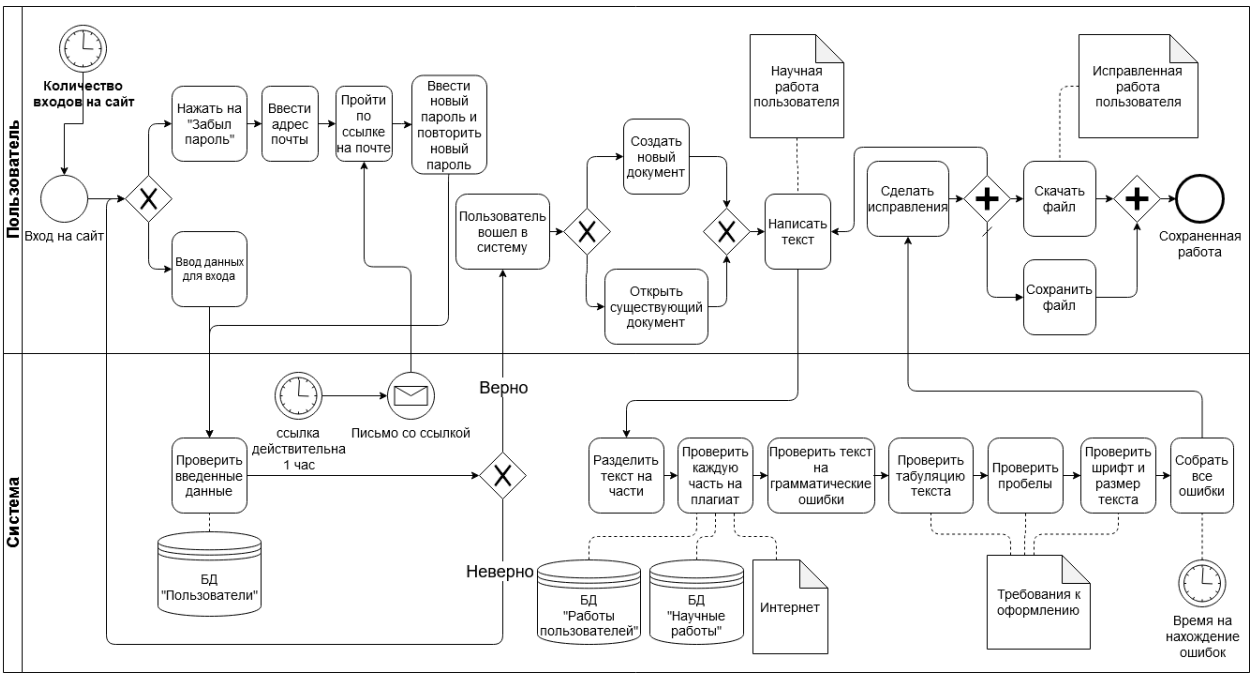

Рисунок 2 - Модель процессов TO-BE

Эта модель является улучшенной версией модели на рисунке 2. Были сделаны следующие изменения:

– после восстановления пароля, пользователя отправят на страницу входа, где он будет подтверждать свою личность;

– пользователь может добавить готовый документ с устройства;

– для того чтобы проверка шла быстрее, три процесса проверки будут идти параллельно и независимо от друг друга;

– добавлены требования к оформлению;

– в базу данных «Работа пользователей» будет автосохранением добавляться текущая работа пользователя.

На данный момент имеются информационные системы для проверки на плагиат и на грамматические ошибки. Одним из больших представителей инструментов для проверки на грамматические ошибки, является Grammarly. Grammarly - это инструмент для письма, который поможет вам проверить несколько типов ошибок.

Grammarly используют разнообразные инновационные подходы - в том числе передовое машинное обучение и глубокое обучение. На данный момент они постепенно открывают новые возможности в исследованиях по обработке естественного языка (NLP) [2]. Чтобы правильно обрабатывать тексты на естественном языке, они должны были понять, как функционирует язык, как его изучают и как он развивается. Синтаксические и семантические парсеры как инструменты компьютерной лингвистики позволяют им извлекать структурированную информацию из миллионов фрагментов необработанного текста. Синтаксический анализ является ключевой частью конвейера обработки текста, а языковая структура, которую он создает, позволяет Grammarly обеспечивать обратную связь при написании в реальном времени.

Вторая система, которая является одним из лучших помощников для проверки на плагиат, это Whitesmoke. Whitesmoke - одно из самых надежных и точных программ для проверки плагиата. Помимо проверки грамматики и корректора, программному обеспечению Whitesmoke для борьбы с плагиатом доверяют многие научные работники. Whitesmoke сканирует миллиарды веб-страниц и ресурсов в Интернете, чтобы проверить неоригинальное

или скопированное содержимое в вашем документе, и отображает их. Интеграция с браузером, MS Word и Outlook поможет улучшить научную работу.

Преимущества:

– сканирует и сопоставляет научную работу с миллиардами веб-страниц, чтобы обнаружить сходство в обрабатываемой работе;

– этот продукт является кроссплатформенным, а также доступен онлайн, поэтому его легко использовать;

– лучшая проверка плагиата для научных работ.

Недостатки:

– не имеет такой большой базы данных, как некоторые аналоги.

Теперь рассмотрим алгоритм нахождения плагиата. Система создана с помощью языка Python, и для того, чтобы сделать алгоритм для проверки текста на плагиат нужно добавить библиотеку для машинного обучения. При работе была использована библиотека scikit-learn.

Компьютеры могут понимать только нули и единицы и для выполнения некоторых вычислений с текстовыми данными нужен способ преобразования текста в числа [3]. Процесс преобразования текстовых данных в массив чисел обычно известен как word embedding, для этого использованы встроенные функции библиотеки scikit-learn.

Для обнаружения сходства в документах используется базовая концепция вектора, скалярного произведения, вычислив значение Cosine similarity между векторными представлениями текстов [4].

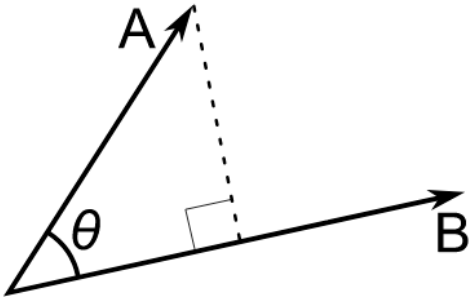

Рисунок 3 - Cosine similarity

Cosine similarity - это показатель, используемый для определения того, насколько похожи документы независимо от их размера [5]. Его формула имеет следующую форму:

$$similarity = \cos(\theta) = \frac{A * B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Математически он измеряет косинус угла между двумя векторами, проецируемыми в многомерное пространство. В этом контексте два вектора представляют собой массивы, содержащие количество слов в двух документах.

При нанесении на многомерное пространство, где каждое измерение соответствует слову в документе, cosine similarity отражает ориентацию (угол) документов, а не величину.

Преимуществом cosine similarity в том, что даже если два похожих документа находятся далеко друг от друга по евклидовому расстоянию, но угол между документами может быть маленьким. Чем меньше угол, тем больше сходство. В итоге с помощью cosine similarity слова представляются в виде позиции в пространстве.

Далее рассмотрим сам алгоритм. Используется OS Module для загрузки путей текстовых файлов, а затем TfidfVectorizer для встраивания слов в наши текстовые данные и Cosine similarity для вычисления плагиата.

После создается две лямбда-функции, одна для преобразования текста в массивы чисел, а другая для вычисления сходства между ними.

Обсуждение результатов

Разработанная система для обработки научных работ учитывает все вышеперечисленные особенности иностранных систем. Планируется создать информационную систему, которая будет проверять на плагиат, на грамматические ошибки и на соответствие с требования ми.

1. Плагиат. Информационная система будет обнаруживать сходства или копии работы. Все найденные результаты будут выделяться от основного текста и показывать на источник, где этот текст встречается.

2. Грамматические ошибки. Разработанная система обнаруживает грамматические ошибки и мгновенно рекомендует все возможные варианты исправления.

3. Проверка оформления текста. После добавления текста можно будет сразу же настроить проверку по требованиям к оформлению, такие как стиль и размер шрифта, поля, межстрочный интервал, выравнивание, абзацный отступ и так далее.

На данный момент не существует информационной системы для проверки статьи на соответствие требованиям. Разработанная информационная система имеет интуитивный и минималистичный дизайн.

Заключение

В данной работе проведен анализ модели процессов, который был оптимизирован в процессе. Также были детально рассмотрены алгоритмы для нахождения плагиата. Разработанная информационная система будет полезна для научных работников, студентов, магистрантов, докторантов и преподавателей, которые пишут статьи, диссертации и другие научные работы. Система имеет преимущества над приведёнными выше аналогами и привносит новую функцию проверки оформления научной работы. Разработанная информационная система поможет многим специалистам сэкономить время и сконцентрироваться на исследовании.

ЛИТЕРАТУРА

1. Guven, A. How good is an AS-IS model really? / A. Guven, A. R. Hajo, R. H. Roy // Lecture notes in business information processing. – 2012. – № 132. – С. 89-100.
2. Dreher H., Automatic Conceptual Analysis for Plagiarism Detection / Dreher H. // Informing Science and Information Technology – 2007. – №4. – С. 601-614.
3. Sorin, A. NLP applications in external plagiarism detection / A. Sorin, C. Dan, B. Theodor // The scientific bulletin. – 2014. – №76(3). – С. 29-36.
4. Xia, P., Learning similarity with cosine similarity ensemble / P. Xia, L. Zhang, F. Li // Information Sciences. – 2015. – №307. – С. 39-52.
5. Muflikhah, L. Document clustering using concept space and cosine similarity measurement / L. Muflikhah, B. Baharudin // International Conference on Computer Technology and Development. – 2009. – С. 58-62.

REFERENCES

1. Guven, A. How good is an AS-IS model really? / A. Guven, A. R. Hajo, R. H. Roy // Lecture notes in business information processing. – 2012. – № 132. – С. 89-100.
2. Dreher H., Automatic Conceptual Analysis for Plagiarism Detection / Dreher H. // Informing Science and Information Technology – 2007. – №4. – С. 601-614.
3. Sorin, A. NLP applications in external plagiarism detection / A. Sorin, C. Dan, B. Theodor // The scientific bulletin. – 2014. – №76(3). – С. 29-36.

4. Xia, P., Learning similarity with cosine similarity ensemble / P. Xia, L. Zhang, F. Li // Information Sciences. – 2015. – №307. – С. 39-52.
5. Muflikhah, L. Document clustering using concept space and cosine similarity measurement / L. Muflikhah, B. Baharudin // International Conference on Computer Technology and Development. – 2009. – С. 58-62.

Жуманбаева С.К. [1], Пащенко Г.Н. [1]

Ғылыми еңбектерді өңдеуге арналған ақпараттық жүйені жобалау және зерттеу

Аңдатпа. Ғылыми еңбектердің көлемінің өсуі ақпараттық технлогияларды кеңінен қолдануды керек етеді. Қазіргі таңда, ғылыми жобаны жазу үшін әртүрлі ақпараттық жүйелерді енгізудің қажеттілігі артуда. Демек, мұндай жүйелерді әзірлеу және зерттеу өзекті мәселе болып табылады. Бұл мақалада ғылыми еңбектерді өңдеуге арналған қазіргі кезде қолданыстағы ақпараттық жүйелер мен бағдарламалар қарастырылады және осы жүйелердің ерекшеліктері айқындалады. Салыстырмалы талдау әзірленеді және әрбір ақпараттық жүйенің артықшылықтары көрсетіледі. Процестердің талдау нәтижелері және ақпараттық жүйенің алгоритмі талданады. Талдау нәтижесінде ақпараттық жүйенің тиімді процестер моделі құрастырылады. Ғылыми еңбектерді өңдеуге арналған ақпараттық жүйенің сипаттамасы келтіріледі.

Түйінді сөздер: ғылыми еңбектерді өңдеу, ақпараттық жүйе, ғылыми еңбек, плагиат, грамматиканы тексеру.

Zhumanbaeva S.K. [1], Pachshenko G.N. [1]

Desining and development of information system for the processing scientific works

Abstract: An increase in the volume of scientific works necessitates the widespread use of information technology in its processing. Recently, there is an increasing need for the development of information systems of various nature for working with scientific papers. Therefore, the development and research of such systems is an urgent task. This article discusses the existing information systems and programs for processing scientific works, and highlights their features. It presents the analysis of such processes and of the system algorithm, an optimal process model built as a result of the analysis, and describes the developed information system for the processing of scientific papers.

Key words: processing of scientific works, information system, scientific works, plagiarism, scientific paper.

Сведения об авторах:

Пащенко Галина Николаевна, к.т.н., ассоциированный профессор кафедры «Информационные системы» Международного университета информационных технологий.

Жуманбаева Сымбат Кажмухаметкызы, магистрант Международного университета информационных технологий.

Авторлар туралы мәлімет:

Пащенко Галина Николаевна, т.ғ.к., Халықаралық ақпараттық технологиялар университеті «Ақпараттық жүйелер» кафедрасының қауымдастырылған профессоры.

Жуманбаева Сымбат Кажмухаметкызы, магистрант, Халықаралық ақпараттық технологиялар университеті.

About authors:

Galina N. Pachshenko, Cand. Sc. (Technology), Associate Professor of the Department of «Information Systems» of the International Information Technology University.

Symbat K. Zhumanbayeva, master student, International Information Technology University.