НАДЕЖНОЕ ОБНАРУЖЕНИЕ ПЛАГИАТА МАЛЫМ ЧИСЛОМ ПОИСКОВЫХ ЗАПРОСОВ

В.В. Дягилев,
аспирант кафедры высшей математики и математического моделирования Алтайского государственного технического университета им. И.И. Ползунова

А.А. Цхай,
доктор технических наук, профессор, заведующий кафедрой математики и прикладной информатики в экономике Алтайской академии экономики и права

С.В. Бутаков,
кандидат технических наук, доцент кафедры математики и прикладной информатики в экономике Алтайской академии экономики и права

E-mail: dyagilev@mail.ru, taa1956@mail.ru, sergey.butakov@gmail.com
Адрес: г. Барнаул, проспект Ленина, д. 46

Представлены усовершенствованная архитектура системы обнаружения плагиата и эксперимент, показывающий эффективность предложенного подхода. В предложенном варианте поисковый Интернет-сервис выделен в отдельный модуль, размещаемый на стороне организации-контролера. В отличие от традиционных архитектур предполагается, что контролер не получает проверяемый документ целиком. Проведенный эксперимент демонстрирует преимущества предложенного подхода.

Ключевые слова: определение плагиата, архитектура сервиса.

1. Введение

В данной статье плагиат определен как использование чужих работ без прямого указания источника. Повсеместное развитие сети Интернет способствует распространению плагиата во многих сферах деятельности. Плагиат является острой проблемой, как в образовании, так и в других отраслях. Так в результатах исследований, проведенных в Национальном исследовательском университете «Высшая школа экономики», утверждается, что около 50% студентов российских вузов используют рефераты и курсовые работы из сети Интернет [1]. Росфиннадзор отметил проблему плагиата при аудите использования бюджетных средств выделенных на НИОКР в 2009 году [2]. Не менее актуально эта проблема стоит при рассмотрении конкурсных проектов в государственные инновационные программы и фонды, а также при регистрации заявок в Роспатенте. Примитивный плагиат заключается в прямом копировании мате-

риалов источника, в то время как более сложные формы стараются скрыть связь с последним, например, при помощи перефразирования. Само понятие «сходство» может быть также формализовано различными способами [3], включая, например, сходство индексов [4] или семантическое сходство. Само понятие «Сходство документов» не является предметом данной работы и поэтому в данном случае мы ограничимся рассмотрением плагиата как прямого копирования текста из ранее опубликованных документов с минимальными вариациями.

При определении возможных источников плагиата необходимо определиться с границами поиска, а также с исполнителем, чьими силами будет осуществляться этот поиск. Если поиск необходимо производить по коллекции документов какой-нибудь библиотеки, то это один случай. Например, сотрудник университета сможет выполнить такой поиск, используя локальную информационную систему. Другой вариант – это поиск схожих документов, опубликованных на web-страницах в сети Интернет. Для данного вида поиска, чтобы сравнить миллиарды документов, необходимо иметь собственную поисковую машину, периодически сканирующую всю сеть Интернет, либо воспользоваться услугами сторонней организации.

При передаче текстов в стороннюю организацию на проверку неизбежно возникают вопросы, связанные с защитой интеллектуальной собственности. Как поступить в случаях, когда необходимо осуществить проверку текста на наличие плагиата, но при этом желательно ограничить доступ третьей стороны к информации, содержащейся в документе? Подобные ситуации могут возникнуть, например, в случаях обработки заявок на изобретения, полезные модели, промышленные образцы, проверки отчетов по научно-исследовательской работе или при рассмотрении конкурсных проектов в государственные инновационные программы и фонды и т.п. Ограничение при помощи лицензирования, используемое в настоящее время большинством подобных сервисов, должно быть дополнено применением методологии создания средств ограничения доступа.

Использование подобных средств, таких как, например, хэширование и передача хэшей вместо оригинального текста, с одной стороны, гарантирует невосстановимость документа, но, с другой стороны, накладывает существенные ограничения на практическое применение данного подхода. Ограничения возникнут в силу того, что проверяющая организация будет вынуждена поддерживать хэш поискового индекса на своей стороне. При этом основная проблема защиты содержания проверяемого текста не решится, так как поисковый индекс будет в любом случае содержать открытый текст из потенциального источника плагиата. Кроме того, применение хэширования сделает невозможным использовать для целей поиска обычные поисковые системы, такие как Яндекс, Google или Bing.

Возможным вариантом решения проблемы является существенное изменение проверяемого текста таким образом, чтобы его восстановление потребовало бы значительных усилий. Кроме того, такие действия по восстановлению могут быть квалифицированы как неправомерный доступ к интеллектуальной собственности.

Далее предлагается оригинальный способ построения системы обнаружения плагиата, при котором требуемое качество поиска сохраняется, но существенно затрудняется возможность полного восстановления проверяемого текста из переданных данных. Предлагаемая архитектура системы основана на использовании общедоступных поисковых машин Интернета и базового постулата о том, что обнаружение местонахождения потенциальных источников плагиата в сети Интернет обеспечивается использованием ограниченного объема информации из проверяемого документа.

Оставшаяся часть статьи организована следующим образом. Во втором разделе характеризуются типовые схемы системы обнаружения плагиата. В третьем разделе детально описывается предлагаемая клиент-серверная архитектура системы обнаружения плагиата. В заключении приводятся результаты эксперимента, показывающие зависимость качества обнаружения источников плагиата от объема переданной информации и числа запросов, а также делаются выводы о применимости предложенного подхода.

2.Существующие решения

Архитектура сервиса обнаружения плагиата (СОП), приведенная далее, типична для систем, применяемых в таких областях как журналистика, научные исследования и образование. Именно на последнюю область мы будем опираться в качестве примера, но выводы, предложенные в работе, могут быть расширены на другие смежные сферы деятельности. Анализ декларируемых принципов работы и детальное рассмотрение открытых СОП

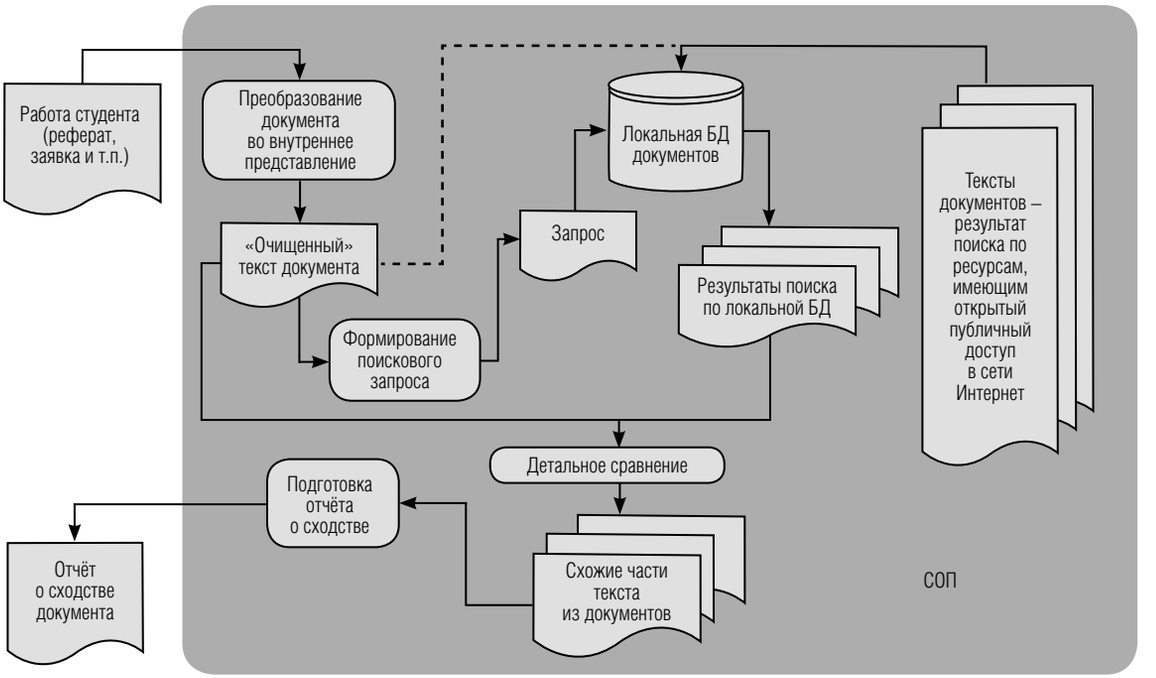


Рис. 1. Типовая архитектура СОП

позволяют выявить их типовую структуру. Последняя может быть охарактеризована описанием продуктов – лидеров рынка, таких как системы «Антиплагиат» (лидер русскоязычного рынка) и «Turnitin» (лидер рынка в англоговорящих странах). Упомянутая структура СОП отображена на схеме (*рис. 1*).

Пользователь (студент или преподаватель) передает на проверку документ в СОП через информационную систему своего университета, либо напрямую через web-интерфейс СОП. Затем содержимое документа преобразуется системой, с целью выделения «чистого» текста, т.е. избавления от форматирования документа, присущего текстовым процессорам. На основе полученного текста строится запрос к базе данных документов СОП, результатом которого является набор документов, потенциально возможных источников плагиата.

Детальное сравнение текстов документов позволяет определить подобные части и сформировать отчет о найденных совпадениях. В результате, пользователь получает отчет о проведенной проверке, с указанием частей текста и источников «заимствования», если таковое обнаружено СОП. При этом база данных документов СОП может содержать индексы открытых сегментов сети Интернет (как в случае с продуктом «Turnitin»), так и доступ к некоторым библиотекам с ограниченным доступом. Например, продукт «Антиплагиат» имеет доступ к базе данных диссертаций ВАК РФ.

Рассматривая процесс обнаружения плагиата, следует отметить ключевой момент – содержимое проверяемого документа передается в СОП, то есть третьей стороне. Ограничение при помощи лицензирования [5] не исключает технической возможности неправомерного использования передаваемой информации.

Если исключить применение СОП из процесса, то один из возможных способов поиска плагиата – это использование общедоступных поисковых машин, таких как, например, Яндекс или Google (т.е. использования ограниченного объема информации из текста документа, для того, чтобы получить потенциальные источники плагиата из сети Интернет). Общепринятый способ – это, когда часть текста помещается в кавычки и осуществляется глобальных поиск схожих документов, т.е. производится попытка найти точные совпадения текста в документах, опубликованных в открытом доступе в сети Интернет. Исследования показали, что такая техника поиска, с использованием публичных поисковых машин по критерию точного совпадения фраз, может быть эффективной [6], но такой поиск – медленный, так как осуществляется вручную. При этом остается открытым вопрос определения ключевых фраз в тексте, по которым следует вести поиск.

Для данной цели возможно использование открытого программного обеспечения, например, системы «CROT» [7]. Данная система имеет типовую архитектуру СОП, за тем исключением, что локальная база данных документов состоит только из внутренних ресурсов организации пользователя, а нахождение внешних потенциальных источников плагиата, выполняется путем отправки запросов к поисковой машине Интернета.

Система «CROT» выполняет исчерпывающий поиск, посылая запросы, сформированные «скользящим окном». Алгоритм «скользящего окна» осуществляет прямой перебор фраз. Например, для шекспировской фразы «to be, or not to be: that is the question» при длине окна X=4 алгоритм сформирует 7 следующих запросов: «to be or not», «be or not to», «or not to be», «not to be that», «to be that is », «be that is the», «that is the question». Авторы системы «CROT» указывают, что, если значительная часть текста документа была взята из какого-либо источника в Интернете, то нет необходимости посылать все возможные запросы, а достаточно только 10% от этих числа, чтобы определить местонахождение этого источника [7].

3. Предложенная архитектура

На схеме (*рис. 2*) отображена основная концепция предлагаемой архитектуры СОП. Сам сервис разделен на внутреннюю (клиентскую) часть, работающую на инфраструктуре пользователя, и на внешнюю (серверную) часть, использующую инфраструктуру сторонней организации. Внутренняя часть выполняет функции сервера для обращений со стороны пользователей и одновременно с этим является клиентом, выполняющим запросы к внешней части сервиса.

Предполагается, что разделенная структура сервиса будет выполнять работы в следующем порядке.

1. Клиентская часть системы, получив документ, переданный пользователем для проверки, преобразует его в «чистый» текст.

2. Клиентская часть создает специальный запрос, путем случайного выбора определенного количества запросов, сформированных методом «скользящего» окна.

3. Клиентская часть отправляет специальный запрос в серверную часть СОП.

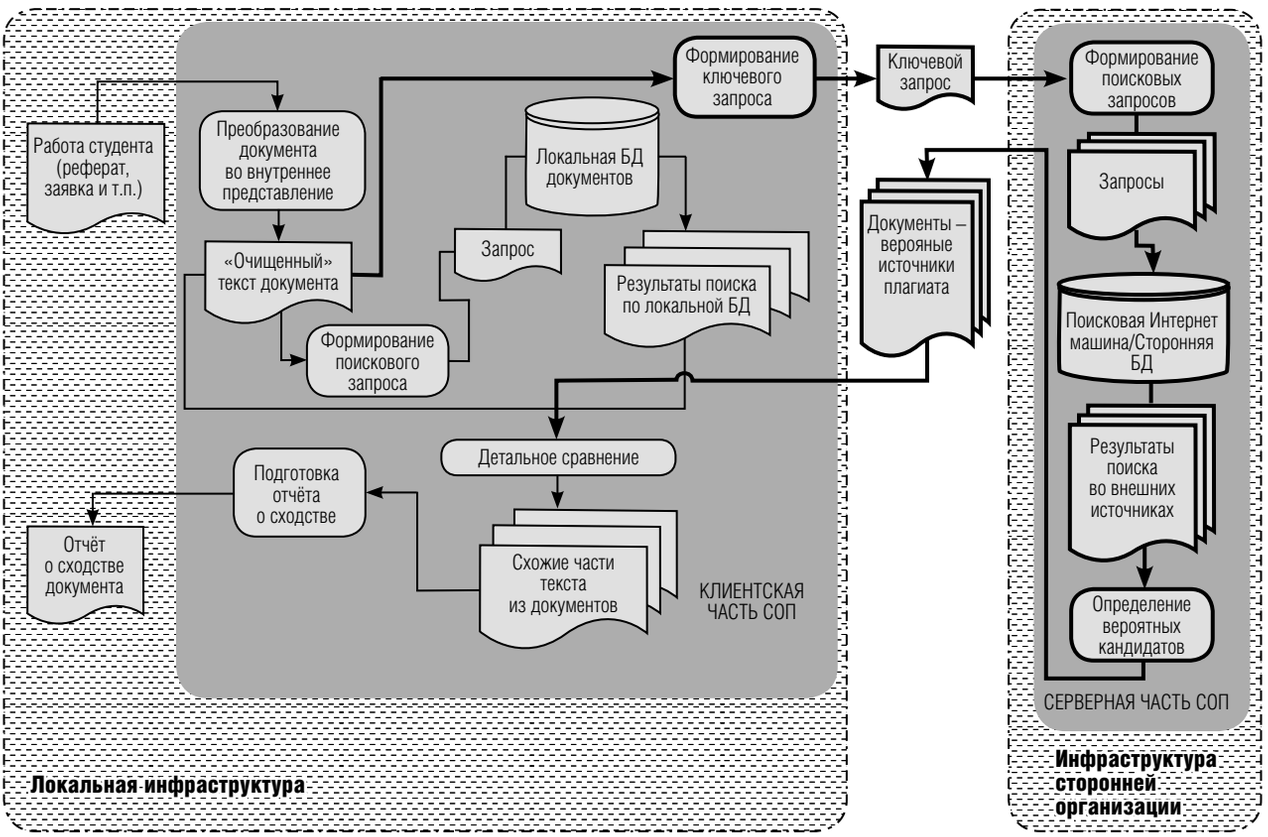


Рис. 2. Схема предлагаемой архитектуры СОП

4. Серверная часть, получив специальный запрос, формирует к поисковой машине Интернета запросы на нахождение документов, опубликованных в открытом доступе в сети Интернет.

5. Серверная часть, как результат выполнения запросов, получает множество ссылок на документы и выбирает те ссылки, которые встречаются чаще всего.

6. Серверная часть по выбранным ссылкам загружает найденные документы, и отправляет последние в клиентскую часть.

7. Клиентская часть производит детальное сравнение исходного текста и полученных документов.

8. Клиентская часть определяет схожие части текста и направляет отчет о сходстве исходного документа пользователю.

Большая часть работы сервиса определения плагиата происходит на оборудовании организации-пользователя. Сторонняя же организация выполняет только глобальный поиск возможных источников плагиата из сети Интернет.

Для составления специального запроса в клиентской части СОП можно использовать алгоритм «скользящего окна», реализованный в системе «CROT». Данный алгоритм осуществляет полный перебор всех фраз длины X, доступных в проверяемом документе. С учетом того, что большинство поисковых машин допускает в обрабатываемых запросах не более 10 слов, то можно ограничить $X <= 10$.

Очевидно, что при длине текста Y слов, общее количество запросов N, определяется формулой $N = Y - X + 1$, где $Q = \{q_1, q_2, ..., q_n\}$ – массив запросов. С учетом того, что Y существенно больше X, можно утверждать, что подобный алгоритм сформирует число запросов, близкое к числу слов в документе: $N \approx |Q|$.

Планируя выполнение большого числа запросов к поисковой машине, необходимо рассмотреть возможность появления следующих затруднений:

- ✦ увеличения времени поиска, что обуславливается повышенными требованиями к Интернет – каналу в случае распараллеливания выполнения данных запросов;
- ✦ возможности восстановления документа из фраз, переданных поисковой машине.

Исследования показывают, что порядок передачи поисковых фраз Q не влияет на результаты поиска [7]. Иными словами, если фразы массива Q будут перемешаны в случайном порядке, то результат поиска совпадет с результатом, полученным при последовательной передаче. Случайное перемешивание не решает проблемы возможного восстановления документа – случайно перемешанная мозаика может быть восстановлена простым перебором, так как в полном массиве соседние элементы q_i и q_{i+1} содержат $X - 1$ совпадающих слов, что делает восстановление тривиальной задачей сбора мозаики фраз по пересечениям из $X - 1$ слов.

Чтобы определить местонахождение источника «заимствования», нет необходимости посылать все возможные запросы Q. Достаточно использовать только небольшой процент случайно выбранных элементов данного массива [7]. Насколько возможно использование этих свойств в случае передачи текста сторонней организации?

Пусть Q_1 – массив случайно выбранных элементов из полного массива запросов $Q : |Q_1| < |Q|$. Общее число слов в запросах, передаваемых в поисковую машину, будет равно $Y_s = |Q_1| * X$. Таким образом, если $|Q_1|$ удовлетворяет неравенству $Y_s < Y$, где Y – общее число слов в документе, то можно гарантировать, что полное восстановление исходного текста на стороне поисковой машины из запросов, переданных ей, становится невозможным.

То есть, если

$$|Q_1| < \frac{Y}{X},$$

то исходный документ – гарантированно невосстановим в первоначальном виде без использования специальных методов лингвистического анализа, активно развиваемых в последнее десятилетие (см., например, [8]). Это ограничение рекомендуется для определения доли запросов при формировании специального запроса, направляемого в серверную часть СОП.

Очевидно, что приведенная архитектура увеличивает требования к мощности вычислительных ресурсов, используемых на стороне клиентской части СОП. Данное требование относится как к повышению вычислительной мощности (требуются вычислительные затраты для детального сравнения документов), так и увеличению дискового пространства для хранения данных документов.

В части дискового пространства в случае использования алгоритмов хеширования, схожих с алгоритмом Винновинг [9], для надежного обнаружения плагиата требуется хранить около 5% хешей, в расчете от числа символов в документе. При исполь-

зовании 128- или 256-битных хешей и однобайтной кодировки текста можно говорить о том, что объем хешей будет примерно равен объему чистого текста в документе. Данное увеличение дискового пространства не представляется сколько-нибудь значимым с учетом постоянно снижающейся стоимости дисковой памяти.

В части вычислительных ресурсов хеширование одного документа не требует существенных вычислительных затрат при использовании локальных алгоритмов, подобных Винновинг [9], так как вычислительные затраты на сравнение документов линейны по отношению к длине текста. Повышенные требования могут предъявляться к СУБД на клиентской части СОП. Фактически именно затраты на приобретение лицензии и обслуживание СУБД будут увеличивать стоимость клиентской части СОП в предложенной архитектуре. Однако большинство организаций уже имеют и поддерживают СУБД, поэтому стоимость дополнительного лицензирования может быть несущественной.

Потенциальным недостатком предложенной архитектуры является то, что она не сможет найти «заимствованный» документ, если он не публиковался в сети Интернет, и поисковые машины Интернета не получали к нему доступ. Однако поисковые машины, такие как Яндекс, Bing или Google, постоянно расширяют свои возможности, получая доступ к научным библиотекам и журналам. Также следует отметить, что по данным специального исследования – ресурсы сети Интернет являются главный источником плагиата [10]. Поэтому авторы склонны считать, что указанный недостаток не является критичным по сравнению с преимуществами, возникающими при использовании предложенной архитектуры СОП.

4. Эксперимент

Главная цель эксперимента была направлена на проверку возможности применения и оценку практического значения предложенного подхода. Основные цели эксперимента:

✦ исследовать связь качества поиска с объемом используемой информации в различных случаях;

✦ определить объем информации, необходимой для гарантируемого обнаружения плагиата.

Для эксперимента был построен набор документов, имитирующий плагиат. В качестве основы набора был использован оригинальный текст, длиной около 3000 слов, который никогда не был опубликован в Интернете. Источником плагиата служили тексты 50 документов, сопоставимых по длине и опубликованных в составе общедоступного ресурса Wikipedia.org. Часть текста оригинального документа заменялась текстом из публичного источника. Место вставки «публичного текста» выбиралось случайным образом. Выборка имитировала десять различных уровней плагиата. Размер вставляемых блоков текста изменялся от 5% до 50% общего объема документа. Таким образом, были сгенерированы 500 документов с различной степенью содержания плагиата в них. Далее, для каждого документа были сформированы поисковые запросы методом «скользящего окна» длиной 6 слов. По случайно отобранным запросам осуществлялся поиск в сети Интернет. Объем запросов составлял 5%, 10%, ..., 50% от общего их числа.

На *рис. 3* представлена поверхность, отображающая процент обнаружения источников плагиата в сети Интернет в зависимости от количества заимствованного текста и числа запросов, отправленных поисковой машине Интернета.

Светлая часть графика указывает на лучшие результаты. Как видно из параметров графика надежное обнаружение плагиата может быть достигнуто для низкого числа запросов, если существенная часть текста была заимствована из источников сети Интернет. А именно, если больше 50% текста в документе заимствовано, то для того, чтобы достоверно обнаружить плагиат, до-

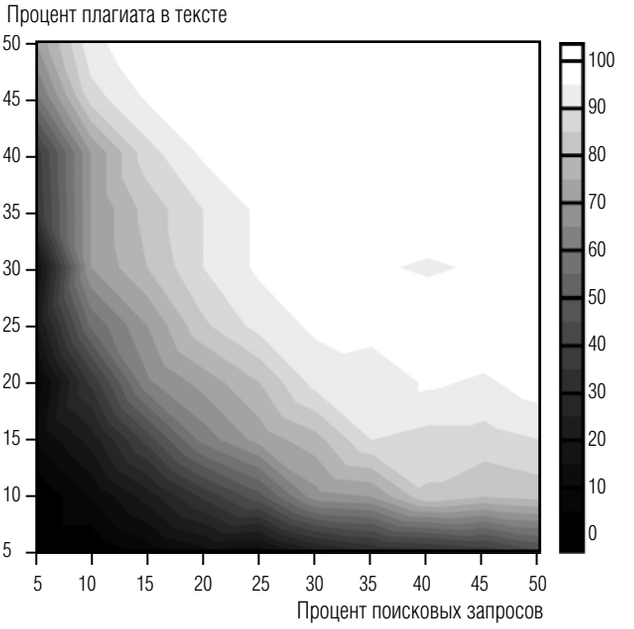


Рис. 3. Результаты обнаружения источников плагиата

статочно будет отправить только 15% запросов, сформированных «скользящим окном». При узко-специфической тематике документа, при отправке тех же 15% запросов достаточно будет уже 25% заимствованного текста для стопроцентного обнаружения плагиата. Это может объясняться тем, что запросы с более специфическими ключевыми словами в большинстве случаев получают более релевантные ссылки на возможный источник плагиата.

Худшие результаты поиска для документов с низким уровнем плагиата могут быть объяснены тем, что фрагменты текста с плагиатом не попали в случайно выбранное число запросов. Можно предположить, что для увеличения качества поиска следует предварительно проанализировать документ, с целью определения сегментов текста, которые наверняка должны участвовать в поисковых запросах. Один из возможных способов реализации – это использование стилеметрии (определения изменений стиля текста) [11]. Использование стилеметрии планируется авторами в качестве одного из направлений развития исследований в данной области.

5. Направления дальнейших исследований

Одно из возможных направлений исследований – это включение стилеметрии [11] (определения стиля текста) во внутреннюю часть СОП. Определение изменений стиля текста позволит улучшить процесс формирования поисковых запросов, что в свою очередь должно положительно отразиться на качестве поиска документов в сети Интернет при малом объеме информации, передаваемой в стороннюю организацию.

Другим возможным направлением продолжения исследований является использование языковой семантики при проведении поиска плагиата. Известно, что избыточность естественноязыковых конструкций позволяет восстанавливать тексты по их фрагментам. Известен ряд проектов (например, [12]), использующих семантику для поиска плагиата. Для затруднения восстановления текстов необходимо рассмотреть возможность исключения семантических индикаторов из поисковых запросов. С этой целью будут использованы модели и методы математической лингвистики и математической информатики, в частности, излагаемые в монографии [8]. Такое исключение должно существенно снизить вероятность восстановления исходного текста документа из содержания запросов. Однако нужно исследовать, как это повлияет на качество поиска документов в сети Интернет.

6. Заключение

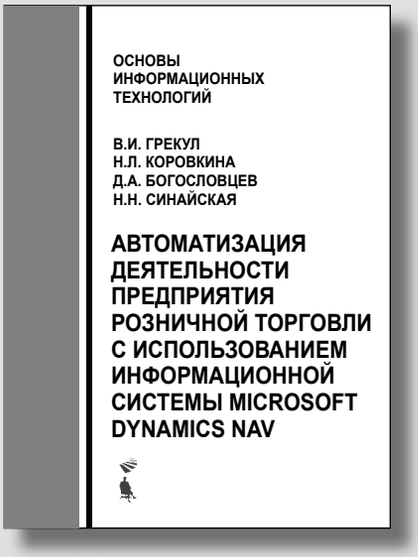
Предложена новая архитектура СОП, позволяющая определять в проверяемом документе наличие присвоенного материала из текстов, опубликованных в сети Интернет. При этом сторонняя организация, осуществляющая поиск похожих документов, получает не содержимое документа целиком, а только набор поисковых запросов. Качество поиска документов при таком подходе остается на прежнем, высоком уровне.

Проанализированы результаты экспериментального поиска по различным вариантам смоделированного плагиата. Наилучшие результаты получены для поиска в случае узкоспецифических текстов. В дальнейшем планируется совершенствование программной реализации предложенной архитектуры СОП. ■

Литература

1. Ивойлова И. Украденные мысли: Половина студенческих рефератов и курсовых скачивается из Интернета // Российская газета. Федеральный выпуск. – № 4830. – 20 января 2009 г. (http://www.rg.ru/2009/01/20/referaty.html).
2. Информационная справка о результатах проверки использования министерствами, ведомствами, внебюджетными фондами и подведомственными им организациями бюджетных средств, выделенных в 2009 году на научно-исследовательские и опытно-конструкторские работы. – Росфиннадзор, 2010. (http://www.rosfinnadzor.ru).
3. Федотов А.М., Барахнин В.Б. К вопросу о поиске документов «по аналогии» // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2009. – Т. 7. – Вып. 4. – С. 5–7.
4. Козлов А.В., Мальцева С.В. Методы повышения эффективности автоматического индексирования документов // Автоматизация и современные технологии, 2004, № 6. – а С. 22-27.

5. Жарова А.К. Правовая защита интеллектуальной собственности / Отв. ред. С.В.Мальцева. – М.: Юрайт, 2011.
6. Culwin F., Child M. Optimizing and Automating the Choice of Search Strings when Investigating Possible Plagiarism // Proceedings of 4th International Plagiarism Conference, Newcastle, June 2010. – 2010.
7. Butakov S., Shcherbinin V. On the Number of Search Queries Required for Internet Plagiarism Detection // Proceedings of the 2009 Ninth IEEE international Conference on Advanced Learning Technologies (July 15 - 17, 2009). – Washington, DC: ICALT. IEEE Computer Society, 2009. – P. 482-483.
8. Fomichov V.A. Semantics-Oriented Natural Language Processing: Mathematical Models and Algorithms / Series: IFSR International Series on Systems Science and Engineering, Vol. 27. – New York, Dordrecht, Heidelberg, London: Springer, 2010.
9. Schleimer S., Wilkerson D., Aiken A. Winnowing: Local Algorithms for Document Fingerprinting // Proceedings of the ACM SIGMOD International Conference on Management of Data, June 2003. – 2003. – P. 76-85.
10. Jones S., Johnson-Yale C., Millermaier S., Pérez F.S. Academic work, the Internet and U.S. college students // The Internet and Higher Education. – 2008. – Vol. 11. – Issues 3-4. – P. 165-177.
11. Stamatatos E. A survey of modern authorship attribution methods // Journal of the American Society for Information Science. – 2009. – 60. – P. 538–556.
12. Lin W.Y., Peng N., Yen C.C, Lin S.D. Online plagiarism detection through exploiting lexical, syntactic, and semantic information // Proceedings of the ACL 2012 System Demonstrations (ACL '12). – Stroudsburg, PA: Association for Computational Linguistics, 2012. – P. 145-150.



АВТОМАТИЗАЦИЯ ДЕЯТЕЛЬНОСТИ ПРЕДПРИЯТИЯ РОЗНИЧНОЙ ТОРГОВЛИ С ИСПОЛЬЗОВАНИЕМ ИНФОРМАЦИОННОЙ СИСТЕМЫ MICROSOFT DYNAMICS NAV

В.И. Грекул, Н.Л. Коровкина, Д.А. Богословцев, Н.Н. Синайская

Москва: Бином. Лаборатория знаний, 2009.

В книге рассмотрены процедуры настройки информационной системы Microsoft Dynamics Navision 4.0 и работа в системе при ее использовании в качестве инструмента автоматизации управления деятельностью торговой компании. Книга предназначена для подготовки пользователей системы, а также может использоваться в качестве пособия при проектировании информационных систем автоматизации торговли. В отличие от традиционных инструкций по применению программных продуктов, материал излагается в контексте выполнения задач производственной деятельности, который задается моделями типичных для отрасли бизнес-процессов.