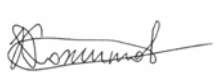
На правах рукописи



Кожитов Сергей Львович

СРЕДСТВА ИНТЕГРАЦИИ, УЛУЧШЕНИЯ КАЧЕСТВА И КООРДИНАЦИИ ДАННЫХ В ИНФОРМАЦИОННЫХ ПОТОКАХ МЕТАЛЛУРГИЧЕСКОГО ХОЛДИНГА

Специальность 05.13.01. — «Системный анализ, управление и обработка информации (в производственной сфере)»

Автореферат

диссертации на соискание ученой степени

кандидата технических наук

Москва 2009

Работа выполнена на кафедре АСУ в Государственном технологическом университете «Московский институт стали и сплавов»

Научный руководитель: кандидат технических наук, доцент, Громов Сергей Владимирович

Официальные оппоненты:
доктор технических наук, профессор, Прошин Иван Александрович
кандидат технических наук , Сергеев Леонид Георгиевич

Ведущая организация: ФГУП “ЦНИИчермет им И.П. Бардина”

Защита состоится " 04 " марта 2009 г. в 16-00 часов на заседании диссертационного совета Д.212.132.07 в Государственном технологическом университете «Московский институт стали и сплавов» по адресу: 119049, г. Москва, Крымский вал дом 3

Аудитория № К-325

С диссертацией можно ознакомиться в библиотеке ГТУ «МИСиС»

Автореферат разослан " 04 " февраля 2009 г.

Учёный секретарь Диссертационного Совета — к.т.н., профессор, Калашников Е.А

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность работы обуславливается необходимостью создания методики и инструментальных средств, обеспечивающих интеграцию и взаимную согласованность данных в информационных потоках управления металлургического холдинга (МХ).

В последнее время в черной металлургии стали заметны организационные преобразования. Если раньше металлургические комбинаты являлись унитарными хозяйственными субъектами, то за период 1999 по 2002 годы они активно стали преобразовываться в холдинги.

Рост холдингов резко нарушил устоявшийся на уровне унитарных предприятий баланс между объемами основных учетно-аналитических операций, совершаемых в приложениях ERP-контура с их жестко контролируемой методологией обработки данных.

Проблему низкой эффективности управления в холдингах пытаются решать не всегда верными способами. Полная автоматизация на базе решений ERP на уровне холдинга будет стоить намного дороже внедрения ERP-системы на отдельном предприятии.

Основные трудности, возникающие при создании информационных систем на крупном промышленном предприятии, связанны с наличием неоднородной среды, включающей различные аппаратные платформы, операционные системы, СУБД и средства разработки приложений.

Одним из перспективных направлений является применение интеграционных технологий для построения гибких, легко адаптируемых информационных систем, а также разработка методов и моделей интеграции, улучшения качества и координации данных в неоднородных системах.

Цель работы. Разработка методики, моделей и алгоритмов интеграции, улучшения качества и координации данных в условиях КИС (корпоративной информационной системы) металлургического холдинга. Разработка технологии, позволяющей быстро и с минимальными затратами устранить дублирование записей в информационных массивах разнородных баз данных.

В соответствии с поставленной целью были решены следующие задачи исследования:

- исследование особенностей построения КИС в условиях металлургического холдинга (КИС МХ);
- выявление проблематики, связанной с интеграцией, улучшением качества и координацией данных при создании КИС МХ;
- разработка методов поиска текстовой информации на основе анализа близости текстовых строк на основе методов с использованием генетических алгоритмов;
- разработка методов автоматической классификации электронных документов и оценка их качества.

Методы исследования. В работе использовались методы определения редакционного расстояния, метод N-грамм, генетические алгоритмы, метод динамического программирования и методы классификации с использованием мер близости, оценки качества по мере F1.

Научная новизна диссертации заключается в следующих положениях:

- решена задача унификации справочника контрагентов КИС МХ.

- разработана модель и алгоритм расчета сходства текстовых реквизитов разнородных баз данных;
- разработана модификация генетического алгоритма для решения задачи поиска оптимальных параметров модели расчета сходства текстовых реквизитов;
- разработана модификация модели и алгоритма классификации Rubryx, основанная на подборе оптимальных коэффициентов учета вклада различных словосочетаний;
- методом перебора решена задача выбора оптимальных коэффициентов в модели Rubryx.

Практическая ценность состоит в том, что разработанная методика нашла свое применение в КИС ПК “Брэдфорд” на ОАО "ГМК "Норильский никель", ОАО «ММК», Северсталь, Евраз-Холдинг, Мечел.

Апробация работы. Основные положения и результаты работы обсуждались на Российско-Японских семинарах “Перспективные технологии и оборудование для материаловедения, микро и наноэлектроники” в 2003, 2004, 2005, 2006, 2007 годах.

Публикации. По теме диссертации опубликовано 12 работ, включая одну работу в издании, рекомендованном ВАК РФ.

Структура и объем диссертационной работы. Диссертация состоит из введения, трёх глав, заключения и списка литературы, включающего 158 наименований. Основной объём работы занимает 186 страниц, в том числе 52 рисунка и 18 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе рассматривается особенности обработки информационных потоков и документооборота в металлургическом холдинге.

Переход к рыночной экономике в начале 90-хх годов резко изменил систему управления металлургическими предприятиями в России. На отраслевом пространстве появились новые субъекты управления, не похожие ни на унитарные предприятия, ни на министерства – появились, и стали стремительно расти многочисленные холдинговые структуры.

Согласно классическому определению "холдинг" - это предприятие, являющееся держателем акций одной или нескольких организаций с целью контроля и управления их деятельностью. Хотя российское законодательство и не предусматривает такой организационной формы предприятия как холдинг, в последние годы все мы являемся свидетелями становления и бурного развития множества корпораций, являющихся, по сути, холдинговыми структурами. Преимущества такой формы предприятия очевидны: во-первых, увеличение доли присутствия на рынке и, соответственно, контроль значительной его части; во-вторых, возможность привлечения инвесторов за счет слияния капиталов подконтрольных фирм; в- третьих, целый ряд преимуществ при осуществлении непосредственной деятельности холдинга, например, перераспределение средств для поддержания менее прибыльных, но стратегически необходимых его элементов или оптимизация налогообложения за счет оффшорных зон.

При возникновении металлургических холдингов в виде сильно диверсифицированных структур остро возникла проблема в их управляемости.

Одновременно произошло осознание того факта, что в области ИТ-технологий наметился сдвиг в идеологии построения КИС – Applications (приложения,

функциональные модули) могут морально устаревать и замещаться на более современные, тогда как данные и структура их отношений являются более постоянной категорией и в большей степени отражают сущность конкретного бизнеса. Осознание такого рода объективных закономерностей привело к новой волне интереса к интеграционным технологиям.

Мониторинг проблем, с которыми сегодня сталкиваются многие предприятия при построении своих информационных систем, позволяет сделать некоторые выводы:

1. Проблемы автоматизации в значительной степени лежат в плоскости методологии эффективного использования имеющихся информационных ресурсов для обеспечения общей управляемости хозяйствующими субъектами, чем в выборе тех или иных программных средств для автоматизации повседневно исполняемых операций. На рисунке 1. представлены основные факторы, влияющие на уровень управляемости холдинга
2. Обеспечение управляемости холдинга неразрывно связано с решением теоретических проблем экономической модели функционирования современного предприятия, холдинга в условиях крайней нестабильности (как внешней, так и внутренней).
3. В условиях наметившегося сокращения инвестиций важно правильно сделать выбор в пользу проведения политики построения гибких, легко адаптируемых к изменяющейся действительности, информационных систем.
4. При наличии эффективной экономической модели предприятия можно вполне рассчитывать на успешное построение гибких информационных систем управления с помощью массового использования наиболее перспективных сегодня и при этом широко доступных для рядовых пользователей ETL- технологий обработки данных (Технологии извлечения, ьрансформации и загрузки).
5. ИТ-Результатом массового использования ETL- технологий является построение распределенных интеллектуальных информационных систем, в которых рядовые пользователи смогут самостоятельно инкапсулировать свои повседневные рутинные операции обработки данных в разного рода сетевые сервисы и программные интеллектуальные агенты.

Одним из приоритетных направлений исследований в рамках КИС МХ является качество первичной нормативно-справочной информации (НСИ). Учитывая, что эта проблема существенно обострилась в связи с процессом образования холдингов на основе предложенных решений на базе системы ПК “Брэдфорд” в АО “Промсталь” в течение ряда лет выполнялся комплекс работ по выверке массивов нормативно-справочной информации (НСИ):

- Справочников контрагентов, материалов, продукции, основных фондов, бюджетных статей и показателей, других источников, нормативов и классификаторов, входящих в состав информационного обеспечения автоматизированных систем заказчика.
- Интеграция данных — необходимый этап работ при внедрении корпоративной информационной системы (КИС) в условиях крупной компании. На основе существующей, но неоднородной информации, формируется хранилище данных, создается единое информационное поле, позволяющее с легкостью оперировать данными из различных источников (программных сред).

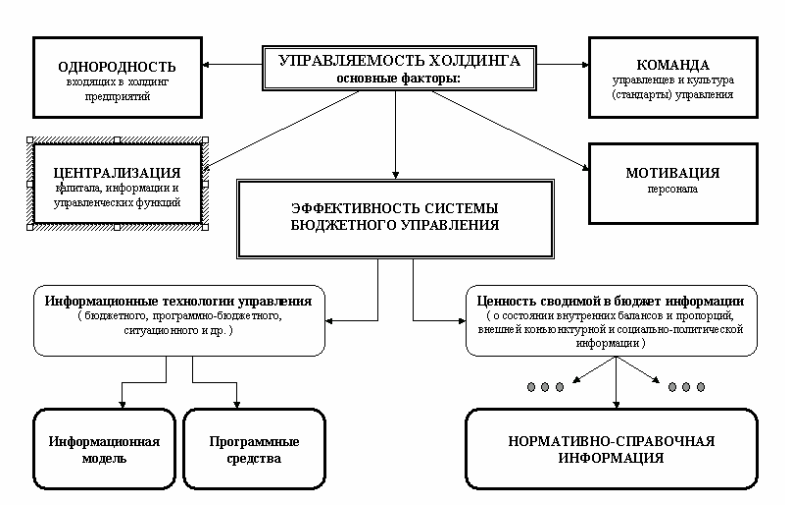


Рисунок 1. Основные факторы, влияющие на уровень управляемости холдинга.

- Интеграция данных из различных корпоративных информационных источников, как в головном офисе, так и в филиалах, прежде всего позволяет осуществлять работу по построению и обработке запросов в едином формате (SQL, OLAP-средства).

Уже на стадии создания массива однородных данных закладываются основы для реализации механизмов доступа. Выигрыш в скорости очевиден: грамотно выполненная интеграция данных обеспечивает проведение транзакций в реальном времени. После охвата всех информационных источников, в том числе сильно рассредоточенных географически, и консолидации их содержимого в единое информационное пространство могут быть выявлены и сгруппированы "родственные" данные. Интеграция данных позволяет упростить процесс сбора первичных данных бухучета из автоматизированных систем убрать "внутренние перегородки" между различными информационными источниками, сделать корпоративный информационный "ландшафт" более "прозрачным" и безопасным.

Чтобы успешно конкурировать на рынке, предприятию необходимо использовать информационные технологии на всех направлениях его деятельности.

Для гарантии успеха в бизнесе информационные технологии должны предоставлять пользователям целостный и непротиворечивый доступ ко всем данным предприятия, быстро реагировать на запросы данных независимо от увеличения объема данных, а также предоставлять разработчикам инструменты, сокращающие время разработки. Для достижения этих целей предприятие должно объединить данные с помощью мощной платформы интеграции данных.

Неудачи интеграции на крупных предприятиях объясняются не только технологическими причинами, но и рядом других факторов.

В последнее время все более актуальным становится использование различных программ, осуществляющих поиск документов различных форматов, информации в СУБД и информационных системах, сообщений электронной почты и прочих данных, содержащихся как на жестком диске персонального компьютера или в локальной сети предприятия, так и в других источниках знаний.

Настоящая работа касается проблемы интеграции, улучшения качества и координации данных в информационных потоках металлургического холдинга. Эти проблемы обусловлены неоднородностью информационных систем, которая возникает, с одной стороны в процессе создания и развития приложений в рамках КИС каждого предприятия, и, с другой стороны, при интеграции нескольких предприятий в холдинг. Зарубежный и отечественный опыт создания КИС позволяет предложить целый ряд технологических и научных решений по преодолению проблемы неоднородности. Наиболее перспективными среди них являются ETL-технологии (Extract Transform Load технологий – технологии извлечения, трансформации и загрузки их в хранилища данных) и MDM-системы (Master Data Management). Эти подходы позволяют унифицировать информационное пространство холдинга с минимальными затратами.

В процессе интеграции данных из различных информационных источников возникают проблемы классификации текстов.
Задача автоматической классификации текстов считается традиционной и довольно хорошо изученной областью компьютерной лингвистики.

Большинство методов автоматической классификации используется в технологии машинного обучения и требует для эффективной работы большой выборки текстов, размеченной вручную.

В различных системах для поиска информации в базах данных, системах орфокоррекции решается задача построения расстояния между текстовыми строками. Задачу построения расстояний и меры близости между текстовыми строками и реквизитами можно рассматривать с разных позиций.

В процессе поиска дубликатов в нормативно-справочной базе данных (материалы, контрагенты) используются методы и алгоритмы анализа строк. За последние 50 лет накоплен большой положительный опыт решения задачи анализа строк. Наилучшими решениями для задачи сопоставления строк является алгоритм Бойера-Мура, для задачи определения всех вхождений образца в текст - метод Бойера-Мура-Хорспула; для расчета расстояний между строками используют преимущественно метрики, основанные на расстоянии Хэмминга, расстоянии Левенштайна и расстоянии редактирования. Для решения задачи о наибольшей общей подпоследовательности в большинстве случаев наилучшие результаты показывает метод Хиршберга. Выбор оптимального метода, метрики и алгоритма во многом зависит от характера решаемой задачи интеграции, улучшения качества и координации данных. Метод генетических алгоритмов позволяет также эффективно решить ряд задач сопоставления данных и выявления наилучшего результата за приемлемый отрезок времени.

В качестве платформы для проверки новых методов и алгоритмов классификации текстов и поиска дубликатов в НСИ очень хорошо подходит система «Бредфорд», которая прошла апробацию на целом ряде металлургических холдингов.

Во <u>второй главе</u> рассматривается конкретные задачи интеграции, улучшения качества и координации данных в информационных потоках металлургического холдинга, а также задачи унификации и выверки справочников контрагентов.
В современных крупных холдингах, использующих для работы с различными рода партнерами, клиентами и поставщиками нецентрализованные базы данных, очень часто возникает ситуация, когда один и тот же контрагент в территориально различных отделениях холдинга отвечает многим записям в справочниках «на местах».

Централизация управления холдингом предусматривает необходимость создания единой системы кодирования используемых справочников и классификаторов и, в том числе, - справочника контрагентов.

В разных подразделениях холдинга и даже в разных автоматизированных системах (АС) у одного и того же подразделения, как правило, используются собственные справочники контрагентов, коды и наименования которых могут не совпадать с реквизитами таких же по сути контрагентов в справочниках других АС и подразделений. Помимо этого дублирование контрагентов нередко встречается и внутри одного и того же локального справочника.

Все это имеет целый ряд негативных последствий как в виде прямых экономических потерь, так и в виде снижения эффективности системы управления. В частности, это приводит к тому, что собрать достоверный баланс дебиторов-кредиторов в рамках холдинга становится невозможным. И естественно, что первым требованием при создании централизованной информационной системы становится задача унификации и выверки справочника контрагентов.

С другой стороны, одномоментно перевести все АС подразделений холдинга на единый справочник невозможно: какое-то время региональные подразделения должны одновременно использовать и старые коды для работы унаследованных приложений, и новые коды – для обеспечения правильного свода и консолидации данных в рамках холдинга.

В качестве объекта исследований данной работы выбран бизнес-процесс выверки и согласования разных версий справочника контрагентов: эталонного, поддерживаемого на уровне управляющей компании холдинга и рабочего справочника контрагентов, ведущегося на одном из подразделений холдинга.

Актуальность выбранного объекта исследований, подтверждается серьезностью экономических потерь, который может нести холдинг из-за рассогласованности справочников контрагентов и, в частности, из-за наличия дублированных записей на одного и того же контрагента.

Разработка алгоритма однопараметрического поиска аналогов не является оптимальной. Запись в справочнике контрагентов идентифицируется целым набором полей: «наименование», «ИНН», «КПП», «адрес юридический», «адрес почтовый» и т.д. Осуществлять в этом случае поиск аналогов только по одному наименованию не правильно, т.к. существуют примеры организаций с одинаковыми наименованиями, но разными адресами или одинаковыми ИНН, но разными наименования и т.п.

Таким образом, как минимум при поиске двойников в справочнике контрагентов необходимо помимо наименования учитывать адрес и ИНН контрагента. Если вычислять суммарный коэффициент релевантности, как среднее арифметическое коэффициентов по каждому отдельному параметру, то результат получается неадекватным, так как влияние на результат поиска отдельных коэффициентов неодинаково, поэтому предложена следующая формула расчета суммарного коэффициент релевантности:

$$K2_{общий} = \frac{KV_1 * K2_1 + KV_2 * K2_2 + \ldots + KV_N * K2_N}{KV_1 + KV_2 + \ldots + KV_N} \quad (1)$$

где $K2_{общийи}$ - суммарный коэффициент релевантности при многопараметрическом поиске;

KV_i – коэффициент влияния параметра на суммарный коэффициент релевантности;

$K2_i$ – коэффициент релевантности параметра.

Таким образом, задача состоит не только в определении коэффициентов релевантности для каждого из параметров описания объектов при расчете суммарного коэффициента релевантности при многопараметрическом поиске, но и в определении их коэффициентов влияния. Для исчисления оптимальных величин коэффициентов влияния каждого из параметров на суммарный коэффициент релевантности необходимо разработать алгоритм математического анализа эталонной выборки данных, в которой экспертным путем уже были выявлены случаи отнесения позиций к явным двойникам при наличии в параметрах их описания существенных различий, и наоборот – позиции, не признанные экспертами явными двойниками, несмотря на полное совпадение значений тех или иных параметров описания. В дальнейшем, на базе разработанного алгоритма необходимо реализовать программные решения, позволяющие на практике реализовывать модель многопараметрического поиска аналогов, в частности, в сводных справочниках контрагентов металлургических холдингов.

Практическая проверка правильности выявленного алгоритма расчета суммарного коэффициента релевантности при многопараметрическом поиске аналогов рассмотрена в третьей главе на примере работы подсистемы «Контроль контрагентов» Системы ведения нормативно-справочной информации ОАО «ГМК «Норильский никель».

В данной работе рассмотрены формальные методы решения задачи нечеткого сравнения строк и нахождения коэффициента релевантности $K2_i$ из (1).
Определим коэффициент сходства двух строк как число от 0 до 1. Пусть имеется функция $f(S_1, S_2)$ от двух строк, которая осуществляет отображение меры сходства двух строк на отрезок [0;1], где 1 соответствует полному смысловому совпадению строк.

Пусть существует множество пар строк, которые считаются одинаковыми в смысловом плане даже при их формальном несовпадении, то есть похожими, и множество пар строк, которые считаются разными, то есть непохожими. В этом случае задача состоит в том, чтобы определить функцию коэффициента сходства таким образом, чтобы на одинаковых парах строк функция принимала бы максимальные значения, на непохожих парах строк функция принимала бы минимальные значения. Ослабив условие строгого максимума и минимума, можно сказать, что функция сходства на множестве одинаковых пар строк не должна принимать значения меньше порогового коэффициента, на множестве непохожих пар строк функция сходства не должна принимать значения больше порогового коэффициента

Итак, если множество похожих пар строк обозначить как S^0, множество непохожих пар строк как $\bar{S}^0$, а функцию от двух строк – $f(S_1, S_2)$, то формальная постановка задачи записывается так:

$f(S_1, S_2) \to \max$ для всех пар строк $<S_1, S_2> \in S^0$

при условии $f(S_1, S_2) \geq K_n$,

$f(S_1, S_2) \to \max$ для всех пар строк $<S_1, S_2> \in \bar{S}^0$

при условии $f(S_1, S_2) \leq K_n$,

где K_n – пороговый коэффициент сходства. K_n может принимать значения из интервала (0,1).

При ослабленном условии необходимо всего лишь найти такую функцию *f*, значение которой на множестве похожих пар строк будет больше заданного порогового коэффициента K_n, на множестве непохожих пар строк меньше заданного порогового коэффициента K_n

Основная работа по нахождению функции *f* в данной задаче находится тремя методами: методом вычисления редакционного расстояния (расстояния Левенштайна) и двумя методами, использующими N-граммы. В качестве дополнительных условий сравнения строк используется частичный формальный синтаксический анализ строк.

Задача нахождения редакционного расстояния (иначе, оптимального выравнивания) представляет собой задачу динамического программирования. Редакционное расстояние представляет собой последовательность операций вставки, удаления и замены символа. Редакционное расстояние будет минимальным, если будет выполнено

$$d(S_1,S_2) \rightarrow min \tag{2}$$

$$d(S_1,S_2) = k_1 I + k_2 R + k_3 D + k_4 M, \tag{3}$$

где S_1, S_2 – сравниваемые строки,
I – количество операций вставки символа,
R – количество операций замены символа другим символом,
D – количество операций удаления символа,
M – количество «пустых» операций, когда на шаге не делается ничего.

Коэффициенты k_i в зависимости от конкретного применения выбираются различными. В данной задаче $k_1=k_2=k_3=1$, а $k_4=0$. Программная реализация системы позволяет выбирать k_2 в зависимости от того, какой именно символ каким заменяется (алфавитно-взвешенное редакционное расстояние).
Соответственно, искомая функция сходства строк выглядит так:

$$f(S_1,S_2) = 1 - \frac{d(S_1,S_2)}{\max(|S_1|,|S_2|)}, \tag{4}$$

где $d(S_1,S_2)$ – редакционное расстояние между строками S_1 и S_2,
$|S_1|$, $|S_2|$ – длины строк S_1 и S_2.

В качестве других методов вычисления сходства строк в работе применяются два метода, основанные на использовании N-грамм.

Пусть задан некоторый конечный алфавит $V_T = \{w^i\}$, где w^i – отдельный символ. Множество цепочек (строк) конечной длины, состоящих из символов алфавита V_T, называется языком на алфавите V_T и обозначается $L(V_T)$. Отдельную цепочку из языка $L(V_T)$ будем называть высказыванием на этом языке. N-граммой на алфавите V_T называется цепочка длиной N. N-грамма может совпадать с каким-то высказыванием, быть его подстрокой или вообще не входить в $L(V_T)$.

При использовании методов, основанных на N-граммах, математически задача сводится к разбиению обеих сравниваемых строк на множество N-грамм, нахождение среди них повторяющихся элементов и вычисление пересечения и объединения множеств N-грамм первой и второй строки .

Алфавитом для строк в данном случае являются буквы русского алфавита, цифры и специальные символы (например, кавычки). Допустим, слово $A = a_1a_2a_3a_4a_2a_3$ и слово $B = b_1b_2a_4a_2a_3$, а длина N-граммы равна двум. Задача заключается в том, чтобы найти все кортежи вида $<a_1, a_2>$, $<a_2, a_3>$,…, $<b_1, b_2>$,…. Также предполагается дополнение слов начальным и конечным пробелом, т.е. в данном примере необходимо учесть кортежи вида $<\_, a_1>$, $<a_3, \_>$,….

После разбиения строк на множество кортежей необходимо вычислить коэффициент Пфайфера, определяющий меру сходства строк:

$$\delta = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|}, \qquad (5)$$

где α и β – множества N-грамм слов *A* и *B*.

Ниже приведено подробное описание каждого из этапов алгоритма определения коэффициентов сходства строк.

1. На начальном этапе у пользователя системы имеется некоторое количество пар строк, предназначенных для сравнения. Заранее неизвестно, какие именно строки похожи, какие нет. Пользователь может обладать лишь общей смысловой информацией о характере представляемых строк.

2. После поступления на вход набора данных производится вычисление меры сходства для каждой пары строк. Изначально все параметры модели получены при обучении системы на тестовых выборках, и в первой итерации построения модели используются именно они.

3. Полученные для всех пар строк результаты приводятся к единой шкале, удобной для просмотра и оценки экспертом.

4. Эксперт производит визуальную оценку качества построения модели. Возможность такой оценки продиктована не только ее неформальным характером, но и плотностью распределения ложных срабатываний относительно выбранного порогового коэффициента. Как правило, подобные модели строятся так, что при любых значениях параметров модели и зафиксированном пороговом коэффициенте на единой шкале количество ложных срабатываний 1-го и 2-го рода уменьшается при увеличении расстояния от местоположения порогового коэффициента на шкале до положения порогового коэффициента. Соответственно, пользователю достаточно просмотреть малую часть пар строк со значениями меры сходства, близкими к пороговому.

5. На этапе обнаружения ложных срабатываний пользователь принимает решение о характере изменения параметров модели. Количественно адекватность модели нужно оценить двумя способами: простым подсчетом числа ложных срабатываний 2-го рода при данных значениях коэффициентов, и построением функции распределения количества ошибок от значения разницы между рассчитанным на модели коэффициентом и пороговым его значением на паре строк, который является ошибкой модели. Причины ошибки не поддаются формальному определению в рамках данной системы. Именно поэтому пользователь является экспертом в рамках данной системы.

6. Пользователь может изменить все доступные ему параметры модели в зависимости от того, какой характер ошибок определен на предыдущем этапе. Задача эксперта состоит в том, чтобы при известных синтаксических факторах, влияющих на адекватность построенной модели, выделить наличие того или иного фактора в тех парах строк, при сравнении которых были допущены ошибки. В подавляющем большинстве случаев пользователю достаточно менять только пороговый коэффициент на единой шкале, как и принято во всех подобных системах.
Блок-схема алгоритма представлена на рис. 2.

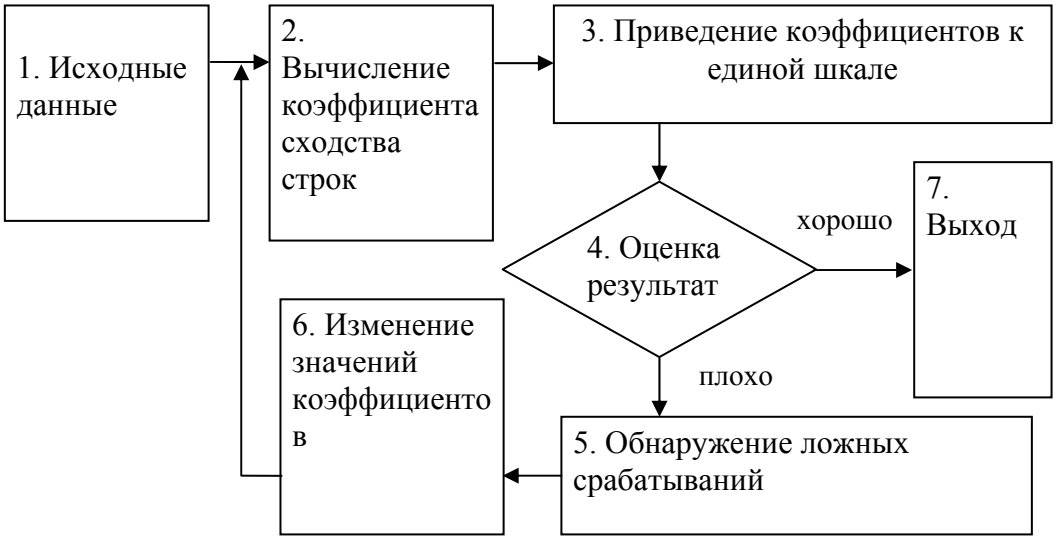


Рисунок 2 – Алгоритм определения коэффициентов сходства строк

Главными критериями оценки качества построенной модели являются отсутствие ложных срабатываний 1-го рода, минимальное число ложных срабатываний 2-го рода и характер распределения количества ошибок в зависимости от значения разницы подсчитанного и порогового коэффициентов.

Графически плотность распределения ошибок можно представить в виде гистограммы, изображенной на рис. 3. На представленном графике Δ – это разница между пороговым значением коэффициента и значением коэффициента сходства для всех пар строк с ложными срабатываниями 2-го рода, N – это количество ложных срабатываний 2-го рода, для которых значение Δ находится в соответствующем интервале.

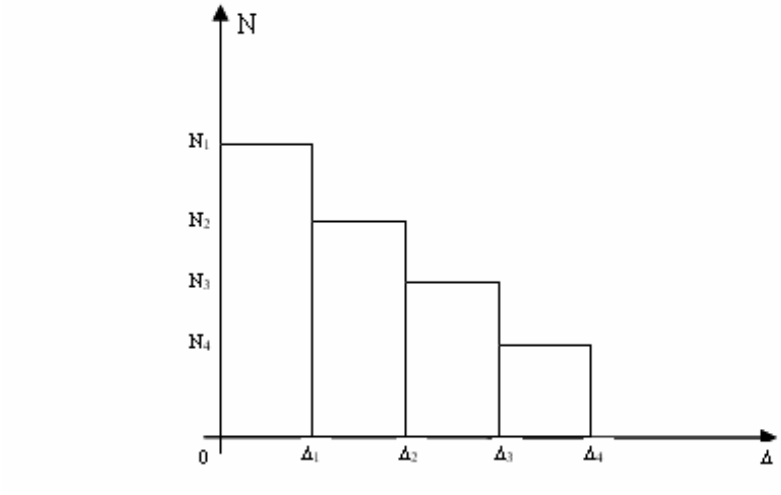


Рисунок 3. Плотность распределения ошибок

Первым очевидным критерием качества модели является интегральная оценка количества ложных срабатываний 2-го рода. Согласно рисунку 3, модель считается хорошей, если значение $\int_0^{\Delta_4} N(\Delta)$ минимально. В данном случае:

$$N(\Delta)=\begin{cases} N_1, 0\le\Delta<\Delta_1 \\ N_2, \Delta_1\le\Delta<\Delta_2 \\ N_3, \Delta_2\le\Delta<\Delta_3 \\ N_4, \Delta_3\le\Delta<\Delta_4 \\ 0, \Delta>\Delta_4 \end{cases}. \qquad (6)$$

Второй критерий качества предназначен для оценки зависимости распределения количества ложных срабатываний от разницы между пороговым коэффициентом и мерой сходства, рассчитанной для каждой пары строк.

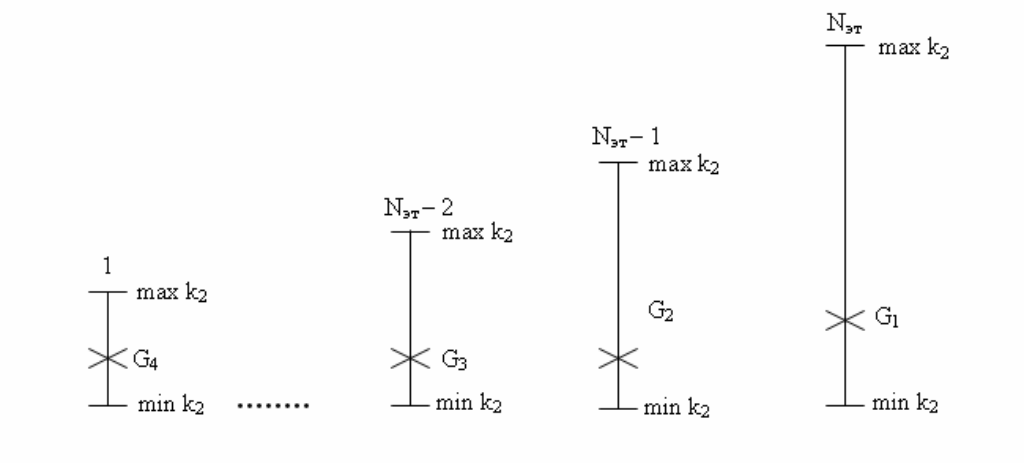


Рисунок 4. Отсев значений на первоначальных шкалах

Согласно рисунку 4, функция распределения будет считаться хорошей, если в каждом последующем интервале количество ложных срабатываний будет много меньше количества ложных срабатываний в предыдущем интервале. Формально подобную оценку можно записать так:

$$F=\sum_{i=1}^{L}\frac{N_i}{N_{i-1}}, \qquad (7)$$

где L – общее количество интервалов, на которых производится сравнение,
N_i – количество ложных срабатываний 2-го рода на i-м интервале.

Если $N_i = 0$, то N_{i+1}/N_i принимается равным нулю. N_0 определяется как $\sum_i N_i$. Значение F будет стремиться к нулю с улучшением качества модели.

Для получения адекватных оценок интервалы должны быть одинаковой длины. Меньшее значение интервала считается более предпочтительным, т.к. дает более точную оценку.

Рассматривая методы основанные на вычислении редакционного расстояния и N-граммах, получилось, что на коротких и достаточно похожих между собой строках алгоритм вычисления редакционного расстояния даёт более адекватные результаты. Тем не менее, различия результатов работы алгоритмов стремятся к 0 при увеличении длины строк. Затем была использована проверка адекватности построенной модели. В результате эксперимента была рассчитана оценка качества модели, которая показала эффективность, данных алгоритмов.

Задача классификации.

В рамках системы электронного документооборота (СЭД) КИС холдинга задача автоматической классификации текстовых документов имеет особую актуальность. Документооборот крупных металлургических холдингов достигает до 10000 документов в день.

Нормативно-правовая база, как правило, представлена в виде отдельной специализированной системы, для остальных документов используется одна из распространенных систем автоматизации документооборота, либо единая система отсутствует вовсе, и документы хранятся только в виде отдельных файлов на компьютерах породивших их пользователей.

Как правило, при внедрении СЭД[1], новые документы классифицируются в соответствии с имеющимся рубрикатором автоматически, однако при этом лицо, принимающее решение (ЛПР) осуществляет дополнительный контроль за правильностью рубрикации. Сложнее обстоит дело с архивом документов, который может насчитывать сотни тысяч и даже миллионы файлов. В этом случае применение полностью автоматической системы классификации является не только экономически оправданным, но часто и единственно возможным решением проблемы.

Задача автоматической классификации текстов в общем виде формулируется как задача нахождения значения величины из {0,1} для каждого входа a_{ij} матрицы решения в соответствии с рисунком 5.

	d_1	...	...	d_j	...	...	d_n
c_1	a_{11}	...	...	a_{1j}	...	...	a_{1n}
...	...	...	...	...	...	...	...
c_i	a_{i1}	...	...	a_{ij}	...	...	a_{in}
...	...	...	...	...	...	...	...
c_m	a_{m1}	...	...	a_{mj}	...	...	a_{mn}

Рисунок 5. Матрица решения

На рисунке 5. $C = \{c_1, \ldots, c_m\}$ является набором определенных категорий, а $D = \{d_1, \ldots, d_n\}$ – набор документов, которые нужно классифицировать. Величина равная 1 для a_{ij} указывает, что документ d_j относится к категории c_i, тогда как величина равная 0 указывает, что документ d_j не относится к категории c_i.

Технология классификации Rubryx включает следующие элементы.

1) Тематический (макро-) словарь специального формата.
2) Набор рубрик.
3) Набор текстов-образцов (по 2-5) для каждой рубрики.
4) Микро-словари специального формата (по одному для каждой рубрики).
5) Пороговые значения для каждой рубрики.
6) Формула для расчета коэффициента близости рубрики и документа.
7) Процедуры обучения классификатора и решения задачи классификации.

[1] Таких систем, как Дело, Dokumentum, Евфрат, Инталев-Документооборот и т. п.

Формула для расчета коэффициента близости K (без дополнительных эвристик) выглядит следующим образом.

$$K = \frac{w_1 * K_1 + w_2 * K_2 + w_3 * K_3}{3} \cdot 100\% \quad (8)$$

Тогда условие Pi вхождения в класс (рубрику) документа di выглядит как:

$$Pi = \begin{cases} 1, \text{ если } Ki \geq K^* \\ 0, \text{ если } Ki < K^* \end{cases}$$

где , K*-пороговое значение K

В формуле (8) K_1, K_2, K_3 - промежуточные коэффициенты по терминам из 1-го, 2-х и 3-х слов, отражающие степень вхождения терминов в классифицируемые документы di ($N_1 i$, $N_2 i$, $N_3 i$) относительно общего количества терминов в микрословаре ($|M_1|$, $|M_2|$, $|M_3|$).

$$\left.\begin{aligned} K_1 &= (N_1 i / |M_1|) \cdot 100\% \\ K_2 &= (N_2 i / |M_2|) \cdot 100\% \\ K_3 &= (N_3 i / |M_3|) \cdot 100\% \end{aligned}\right\} \quad для\ i = 1 \ldots n$$

w_1, w_2, w_3 - весовые коэффициенты, отражающие вклад однословных, двухсловных и трехсловных терминов в общее значение коэффициента близости K. Сумма коэффициентов подчиняется условию:

$$w_1 + w_2 + w_3 = 3 \quad (9)$$

Рассмотрим формальную постановка задачи определения значений коэффициентов w_1, w_2, w_3 в формуле (8) для расчета меры близости документа к рубрике.
Имеется случайная выборка объемом 3299 документов из тестовой коллекции Reuters-21578 по множеству рубрик c_j. Для каждого документа d_i рассчитывается коэффициент K_{ij}, определяющий принадлежность документа i к рубрике j. Будем менять веса w_1, w_2, w_3. Для каждой комбинации весов w_1, w_2, w_3 оценивается качество классификации по мере F1. Найти такую комбинацию w_1, w_2, w_3, которая обеспечит выполнение следующих критериев

$$\text{Max} \left(\sum_{k=1}^{m} F1_k\right) \quad (10)$$

где $F1_k$ – значение меры F1 для k-ой рубрики, m – число рубрик.

$$w_1 \leq w_2 \leq w_3 \quad (11)$$

В целях сокращения пространства поиска в качестве шага дискретизации было выбрано значение 0,2. Также на значения весов w_1, w_2 и w_3 были наложены дополнительные ограничения:

$$(0{,}2 \leq w_1 \leq 1{,}4) \wedge (0{,}2 \leq w_2 \leq 1{,}8) \wedge (0{,}2 \leq w_3. \leq 2{,}6) \quad (12)$$

Таким образом, с учетом ограничений (9), (11) и (12) пространство поиска сократилось до 25 комбинаций.

Для проведения расчетов, были выбраны пятнадцать наиболее популярных рубрик коллекции Reuters-21578.

В диссертации представлены результаты расчетов каждого из 25 вариантов комбинаций весовых коэффициентов по рубрикам.

В таблице 1 приведены сводные данные по максимальным значениям показателя качества классификации F1.

Таблица 1. Сводные данные по максимальным значениям показателей качества

Топик	**Старое значение F1**	**Максимальное значение F1**	**Номер комбинации весовых коэффициентов, обеспечивающей максимум F1**
Ship	0,82	0,882	3,5,9,10, 13-15, 18-25
Acq	0,85	0,886	10, 14, 15, 19
Corn grain wheat	0,89	0,898	6, 9, 10, 13, 18, 22
Crude	0,9	0,921	3, 5, 8, 12, 17, 21, 24
Earn interest	0,85	0,861	6, 9, 13, 14, 19
Dlr	0,85	0,800	1,2,4,7,11,16,20,23, 25
Gnp	0,84	0,921	14, 19,
Money-fx	0,88	0,880	2,3,15
Money-supply	0,81	0,836	3,5,8,12,17,21,24
Oilseed	0,73	0,800	10,14,15,19
Trade	0,88	0,880	3,5,6,8,9,12,13,17,18,21,22,24
Sugar	0,75	0,750	10,14,15,19

Можно видеть, что разброс по комбинациям весовых коэффициентов весьма велик. Для установления номера (или номеров) комбинаций, удовлетворяющих условию (10) необходимо просуммировать значение меры F1 по всем рубрикам и таким образом выявить максимальное значение суммы (10). Эти данные представлены в таблице 2.

Таблица 2. Значения суммы F1 по каждому варианту комбинации весовых коэффициентов

Номер комбинации	**Сумма показателей качества F1 по всем рубрикам (3')**	**Число максимумов в рубриках**	**Оценка значения**
Исходный вариант	10,050		Исходное значение
1	9,803	1	Хуже исходного
2	9,844	1	Хуже исходного
3	10,115	4	Лучше исходного
4	9,847	1	Хуже исходного
5	10,100	4	Лучше исходного
6	9,953	3	Хуже исходного
7	9,877	1	Хуже исходного
8	10,066	3	Лучше исходного
9	10,017	4	Хуже исходного
10	10,110	5	Лучше исходного
11	9,854	1	Хуже исходного
12	10,046	3	Хуже исходного
13	10,017	4	Хуже исходного
14	**10,189**	**6**	**Max**
15	10,150	5	Лучше исходного

16	9,827	1	Хуже исходного
17	10,048	3	Хуже исходного
18	10,045	3	Хуже исходного
19	**10,189**	**6**	**Max**
20	9,843	2	Хуже исходного
21	10,112	4	Лучше исходного
22	10,015	3	Хуже исходного
23	9,855	2	Хуже исходного
24	10,130	4	Лучше исходного
25	9,873	2	Хуже исходного

Как можно видеть из таблицы 2 лишь 9 из 25 комбинаций весовых коэффициентов имеют значения F1 лучше исходной комбинации. Это говорит о том, что поставленная задача не тривиальная и имеет решение. Максимальное значение суммы F1 (10,189) имеют две комбинации с номерами 14 и 19. Для этих комбинаций наблюдается и максимальное число локальных максимумов F1 в отдельных рубриках. Для каждой из комбинации это число равно 6. Весовые коэффициенты, соответствующие максимальным значениям целевого условия, приведены в таблице 3.

Таблица 3. Результаты численного эксперимента

Номер эксперимента	**W1**	**W2**	**W3**	**Сумма F1 по рубрикам**
Исходные значения	0,2	1,3	1,5	10,050
14	0,8	1	1,2	10,189
19	0,8	1,2	1*	10,189

Наличие максимума в контрольной комбинации (с номером 19 в табл. 3), который не соответствует условию (11), можно проинтерпретировать таким образом, что весовые коэффициенты двухсловных и трехсловных термов не должны отличаться. Таким образом, условие (11) необходимо переформулировать следующим образом:

$$w_1 \leq w_2 = w_3 \qquad (13)$$

Для решения задачи определения оптимальных значений весовых коэффициентов слов, обеспечивающих максимальное выявление двойников, была разработана модификация генетического алгоритма, в которой для получения асимптотической сходимости к глобальному минимуму, использована стратегия элитного отбора. Для этого после формирования следующего поколения, если лучшая хромосома не была в него скопирована, то они копируются вместо худшей хромосомы следующего поколения. В целях предотвращения преждевременной остановки алгоритма в области локального минимума, принудительное копирование одной и той же лучшей хромосомы выполняется не более пяти поколений подряд.

В <u>третьей главе</u> рассмотрена практика работы поискового механизма АРМа «Контроль контрагентов» ОАО «ГМК «Норильский никель» в вариантах монопараметрического поиска и в варианте многопараметрического поиска с использованием алгоритмов расчета весовых коэффициентов и логарифмического коэффициента уровня качества поиска, разработанных в рамках настоящей диссертации.

Автоматизированное рабочее место эксперта (АРМ) «Контроль контрагентов» разработано как специализированная конфигурация программного конструктора «BRADFORD», используемого как средство интеграции, улучшения качества и координации, данных металлургического холдинга. Такой выбор для практической реализации решений настоящей диссертации обусловлен не только тем, что этот софт с

1995 года активно используется на основных металлургических холдингах в России, но и тем, что по данным Агентства CNews данный программный комплекс входит в число восьми ведущих российских разработок в классе Master Data Management[2] (управления нормативно-справочной информацией).

Предназначение программного комплекса:

1. Для выполнения работ по объединению и выверке массивов НСИ (в том числе – перед загрузкой в ERP-системы), создания на их основе централизованных корпоративных справочников и классификаторов с таблицами переходных ключей.
2. Для обеспечения обязательной перепроверки на дублирование всех вновь вводимых записей в справочники и классификаторы системы НСИ
3. Для построения и поддержки распределенных систем ведения НСИ, автоматизации процессов упорядочивания, классификации, кодирования и верификации нормативно-справочной информации, обеспечения качества и взаимосогласованности информации в различных системах НСИ, а также в Хранилищах данных и контента.

В главе представлено описание действующей поисковой системы АРМа «Контроль контрагентов» с вариантами расчета коэффициента релевантности на основе сравнения контрагентов по одному параметру (последовательно) и настройка многопараметрического поиска по разработанной в настоящей диссертации модели. АРМ «Контроль контрагентов» предназначен для автоматизации регулярных процессов выявления двойников в корпоративном справочнике контрагентов ОАО «ГМК «Норильский никель».

Основные функции АРМ «Контроль контрагентов»:

1. Формирование таблицы новых поступлений контрагентов в КУДС (MDM-система, играющая роль хранилища сводного корпоративного справочника контрагентов) и мониторинга результатов ее обработки.
2. Поиск возможных двойников контрагентов для каждой позиции таблицы новых поступлений контрагентов по специально настроенному и согласованному с Заказчиком Сценарию с использованием методов точного поиска, приближенного (LIKE) поиска и поиска с механизмом нечеткой логики (NOM).
3. Поиск возможных двойников для любого произвольно заданного пользователем контрагента (возможность генерации пользователем произвольных форм ввода данных).
4. Оптимизации результатов работы поисковой системы за счет изменения и подстройки пользователем основного Сценария поиска , дополнения и изменения содержимого вспомогательных справочников и адаптации параметров настройки механизма NOM-поиска.
5. Уточнение отдельных реквизитов контрагентов по резервным базам данных с предварительно настроенными сценариями параллельного поиска.
6. Использование вспомогательных справочников - аббревиатур и сокращений, транскрипции латинских названий инофирм, исключаемых слов, слов с минимальным влиянием на расчет коэффициента релевантности – для повышения эффективности процессов поиска двойников.
7. Экспорт выявленных двойников для формирования переходных ключей и последующего выполнения операций замещения.

В результате доработки поисковой системы по предложенной модели многопараметрического поиска двойников в настройке параметров НОМ-поиска реализована возможность настройки расчета коэффициентов релевантности по дополнительным полям со степенью их влияния на суммарный коэффициент релевантности по набору параметров.

[2] Иногда для MDM используется альтернативное название — управление справочными данными (Reference Data Management, RDM

Примеры двойников в справочнике контрагентов с расчетом суммарного коэффициента релевантности по полям «наименование» и «адрес» см. в таблице 4.

Таблица 4. Двойники в справочнике контрагентов с суммарного расчетом суммарного коэффициента релевантности

№	Наименование1	Адрес1	K2 общий	K2 наим.	K2 адр.	Наименование2	Адрес2
66	Архангельская контора филиал ОАО ГМК горно металлургический металлургическая Норильский никель	Россия, 660059, Красноярский край, г. КРАСНОЯРСК, ул. КОММУНАЛЬНАЯ, д.2-А	100	100	100	Архангельская контора - филиал ОАО "ГМК "Норильский никель"	Россия, 660059, Красноярский край, г. КРАСНОЯРСК, ул. КОММУНАЛЬНАЯ, д.2-А
66	Архангельская контора филиал ОАО ГМК горно металлургический металлургическая Норильский никель	Россия, 660059, Красноярский край, г. КРАСНОЯРСК, ул. КОММУНАЛЬНАЯ, д.2-А	79	100	69,1	Архангельская контора - филиал ОАО "ГМК "Норильский никель"	Россия, 354008, Краснодарский край, г. СОЧИ, ул. ПИРОГОВА 10,
28	Заполярный филиал ОАО ГМК горно металлургический металлургическая Норильский никель	Россия, 354008, Краснодарский край, г. СОЧИ, ул. ПИРОГОВА 10,	100	100	100	Заполярный филиал ОАО "ГМК"Норильский никель"	Россия, 354008, Краснодарский край, г. СОЧИ, ул. ПИРОГОВА 10,
28	Заполярный филиал ОАО ГМК горно металлургический металлургическая Норильский никель	Россия, 354008, Краснодарский край, г. СОЧИ, ул. ПИРОГОВА 10,	72	49,56	100	Главный офис ОАО "ГМК "Норильский никель"	Россия, 354008, Краснодарский край, г. СОЧИ, ул. ПИРОГОВА 10,
781	ОАО "Кольчугинский завод по обработке цветных металлов им С.Орджоникидзе"	Россия, 600017, Владимирская область, г. ВЛАДИМИР, ул. ГОРОХОВАЯ, д.15	100	100	100	ОАО "Кольчугинский завод по обработке цветных металлов им. С. Орджоникидзе"	Россия, 600017, Владимирская область, г. ВЛАДИМИР, ул. ГОРОХОВАЯ, д.15

Использование результатов исследования по многопараметрическому методу поиска двойников в подсистеме «Контроль контрагентов» Системы ведения НСИ ОАО «ГМК «Норильский никель» позволило повысить общие показатели эффективности работы информационно-поисковой системы и обеспечить требуемый уровень качества сводного корпоративного справочника контрагентов.

Основные результаты и выводы:

В настоящей работе произведено исследование методов повышения качества и координации данных в информационных потоках, а также средств интеграции этих данных и создание конкретных решений на безе этих методов.

В ходе выполнения работы выяснилось, что в современных условиях металлургическим холдингам необходимы новые модели построения корпоративного управленческого учёта с использованием мощных математических и программных аппаратов, способных выявлять дублирования записей. А также максимально автоматизировать операции с данными и разработка технологии автоматической классификации архива документов перед запуском системы электронного документооборота. Большинство опубликованных исследований ориентированы на развитие традиционных подходов к созданию КИС, в то же время исследования, ориентированных на использование методик разрешения дублирования и выверки НСИ, или построения поисковых систем практически отсутствуют.

Данная работа, в свою очередь, была направлена на исследование различных методик основанных на методе редакционного расстояния, N-грамм, формальном синтаксическом анализе, генетических алгоритмах, методе оценки качества мер близости.

Построенные в результате работы модели позволили убедиться в правильности сделанных предположений о применимости данных методик для повышения качества поиска двойников в системах НСИ металлургических предприятий, построении автоматической классификации архива документов. Созданные приложения продемонстрировали применимость данных моделей для выверки НСИ и удаления дублирования записей в информационных потоках разнородных баз данных металлургического холдинга.

Проделанная работа привела к следующим результатам и выводам:

Была создана поисковая система, которая выявила совпадения строк между собой и вывела результаты в виде коэффициентов релевантности.

После математического описания взятых методов была решена задача принятия решений и обучение алгоритма, а затем и проверка адекватности построенной модели. Рассматривая методы основанные на вычислении редакционного расстояния и N-граммах, получилось, что на коротких и достаточно похожих между собой строках алгоритм вычисления редакционного расстояния даёт более адекватные результаты.

В результате эксперимента была рассчитана оценка качества модели, которая показала эффективность данных алгоритмов.

Бала разработана модификация генетического алгоритма для оптимизации весовых коэффициентов слов поискового запроса, после чего был рассчитан суммарный коэффициент релевантности.

Была решена задача поиска весовых коэффициентов при классификации документов по методу Rubryx, с использованием общепринятой для тестирования автоматических классификаторов коллекции текстов Reuters-21578.

Рассмотрена система “Брэдфорд” в которой были воплощены ранее освещённые алгоритмы и методы поиска и идентификации контрагентов с учётом случайных ошибок и расхождений, вызванных разными обычаями написания наименований, адресов и других реквизитов разными операторами, а также для выверки справочников и реестров контрагентов, для их объединения в эталонный сводный массив, для разработки таблиц перекодировки.

Была описана система с автоматизированным рабочим местом для сотрудников службы ведения НСИ.

Результаты исследований применяются во множестве систем документооборота металлургических холдингов России, таких как ОАО "ГМК "Норильский никель" Мечел, ММК, Северсталь, Евраз-Холдинг.

Основные результаты диссертации опубликованы в следующих работах

1.Калашников Е.А., Дубравина Т.В., Кожитов С.Л. Гибридный генетический алгоритм для решения транспортных задач // Материалы 4-го Российского-Японского семинара “Перспективные технологии и оборудование для материаловедения, микро и наноэлектроники”: 22-23 мая 2006 Астрахань. Астраханский государственный университет, 2006 год, с.453-456.

2.Бодров Д.А., Поляков В.Н., С.Л. Кожитов Автоматизация текстового оборота на металлургическом предприятии и новые поисковые технологии // Материалы 4-го Российско-Японского семинара “Перспективные технологии и оборудование для материаловедения, микро и наноэлектроники”: 22-23 мая 2006 Астрахань, Астраханский государственный университет, 2006 год, с.487-494.

3. Громов С.В., Кожитов С.Л. Подход к созданию инструментов поддержки принятия решений при разработке технологических процессов. // Материалы Российско-Японского семинара “Материаловедение и металлургия. Перспективные технологии и оборудование” 25 марта 2003 года Москва. Московский государственный институт стали и сплавов, 2003 год, с.361-366.

4.Yu.N. Pronin, S.L. Kozhitov. Wide use of integration tools as the best means of optimization of quality and coordination of information data flows in holding structures/ of 2nd Russian-Japanese Seminar ”Perspective Technologies, Materials and Equipments of Solid-State Electronic Components ”: April 6, 2004-Moscow, Moscow State Institute of Steel and Alloys (Technological University), 2004, p. 417-426.

5. S.V. Gromov, S.L. Kozhitov. Development of Tools of an in-line Processing of the Data and Decision Making for the Companies with a Complex Organization Structure on the Basis of Technologies of Web-Services // Proceedings of 2nd Russian-Japanese Seminar ”Perspective Technologies, Materials and Equipments of Solid-State Electronic Components ”: April 6,2004-Moscow, Moscow State Institute of Steel and Alloys (Technological University), 2004, p. 428-435.

6. Е.А.Калашников, Т.В. Дубравина, С.Л. Кожитов. Применение генетического алгоритма для решения модифицированных специальных задач линейного программирования с множеством квазиоптимальных решений // Металл оборудование инструмент, май-август 2005. Издательский дом “ИКАР”, Московский институт стали сплавов с.57-59.

7. S.V. Gromov, S.L. Kozhitov. Development and research of components for the distributed data processing and decision-making for the companies with complex organizational structure/ Труды III Российско-Японского семинара “Оборудование и технологии для производства компонентов твердотельной электроники и наноматериалов”, 12 апреля Москва 2005. Московский институт стали и сплавов, 2005 год с.163-167.

8. Ю.Н. Пронин, А.М. Перепёлкина, С.Л. Кожитов. О построении корпоративных информационных систем // Образование, наука и производство, Межвузовый сборник научных трудов. Т. II. Экономика и Менеджмент. Московский государственный институт стали сплавов 2001 г. с. 134-137//

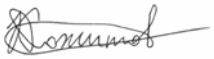
9. Ю.Н. Пронин, С.Л. Кожитов. Возможности ETL-технологий для построения гибких информационных систем управления холдингами на примере построения системы управления нормативно-справочной информацией //Научно-практический семинар “Научно-техническое обеспечение деятельности предприятий, институтов и фирм” Москва 1 июля 2003г., Москлвский Государственный Институт Стали и Сплавов с. 208-218//

10. Бодров Д.А., Кожитов С.Л., Поляков В.Н. Задачи интерактивной обработки поисковых запросов в теоретико-множественной постановке. //Известия Саратовского

унив. Новая серия. Серия «Математика. Механика. Информатика» - Саратов, 2007, т.7. Вып.1, стр. 78-83.//

11. Ю.Н. Пронин, Кожитов С.Л., Дорогова Л.В. Использование поисковой системы ПК BRADFORD для организации перевода открытого технического словаря eOTD ECCMA //Труды V Российско-Японского семинара “Оборудование, технологии и аналитические системы для материаловедения, микро- и наноэлектроники” Том 2, 2007 г., с. 1016-1026//

12. Ю.Н. Пронин, Кожитов С.Л., Давидюк Н.В. Разработка и ведение российской версии открытого тезнического словаря eOTD ECCMA при помощи специализированного программного комплекса BRADFORD//Труды V Российско-Японского семинара “Оборудование, технологии и аналитические системы для материаловедения, микро- и наноэлектроники” Том 2, 2007 г., с. 1027-1039//

Соискатель  С.Л. Кожитов