УДК 004.89 + 004.91

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В СИСТЕМАХ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА

©2010 Н. А. Гладких

аспирант каф. программного обеспечения
и администрирования информационных систем
e-mail: nikolai.gladkikh@gmail.com

Курский государственный университет

Рассмотрены возможности интеллектуального анализа данных в целях оптимизации рабочего процесса систем электронного документооборота. Описываются методы интеллектуального анализа данных в рассматриваемой предметной области.

Ключевые слова: интеллектуальный анализ данных, система электронного документооборота, информационная система, база знаний.

Для современных информационных систем характерны колоссальные объемы данных и скачкообразные изменения или неопределенность факторов внешней среды. В этих условиях применение традиционных статистических подходов (таких, как аппарат математической статистики, аппарат аналитической обработки данных – OLAP и т.п.), как правило, является неоптимальным с точки зрения эффективности.Это актуализирует проблему обеспечения организаций средствами автоматизированного анализа потоков данных, извлечения и накопления знаний [Хаммер, Чамли 2006].

Перед каждой организацией стоит задача анализа внутренних и внешних факторов, влияющих на уровень ее производства. Значительно снизить временные и ресурсные затраты, повысить уровень качества делопроизводства позволяет комплексное обеспечение организации средствами централизованного управления данными, в частности средствами электронного документооборота.

Система электронного документооборота (СЭД) – это автоматизированная информационная система, предназначенная для реализации процесса удаленного обмена большими массивами форматированной информации. Преимуществами электронного документооборота являются возможность параллельного выполнения операций над документами, непрерывность движения документа; обеспечение высокоэффективного поиска документов, обеспечение отчётности по различным статусам и атрибутам документов.

Эффективность процесса автоматизированного документооборота может быть в значительной степени снижена под действием ряда факторов: возникновение в процессе эксплуатации нештатных ситуаций, не предусмотренных на этапе проектирования системы электронного документооборота; ошибки конечного пользователя (оператора, администратора системы) и др. Последствиями таких ошибок являются потеря или искажение данных, рассогласования и противоречия в работе различных структур организации. В связи с этим возрастает необходимость разработки систем интеллектуального анализа данных (СИАД), задачами которых являются извлечение, накопление и обработка знаний, сценариев, моделей поведения информационных систем организаций.

Интеллектуальный анализ данных (ИАД) является междисциплинарной областью науки, возникшей и развивающейся на базе достижений прикладной статистики,

распознавания образов, искусственного интеллекта, теории баз данных и других областей современной науки. ИАД представляет собой спектр технологий и методик выявления статистических взаимосвязей, закономерностей, правил в наборах данных. В настоящее время реализация и развитие данного подхода ведется корпорацией Oracle, компанией SAS, компанией Google Inc. и рядом других.

Задачи ИАД сводятся к реализации алгоритмов классификации, кластеризации, ассоциации, прогнозирования, построения продукций, регрессионного и корреляционного анализа. Основными задачами СИАД являются: анализ структуры баз данных, запросов, журналов транзакций, поиск закономерностей и взаимосвязей между данными, построение продукционных моделей, эмпирических моделей, деревьев решений и семантических сетей [Wang 2003].

В основе современного ИАД лежит принцип построения шаблонов гипотез. Таким образом, осуществляется формирование наборов гипотез на основе выборок данных, а не поиск выборок данных, удовлетворяющих изначально сформулированным гипотезам.

База данных информационной системы (ИС) представляет собой набор таблиц, содержащих записи о свойствах различных объектов. Эти записи содержат данные, отражающие тот или иной факт в рамках предметной области, но лишенные смысловой нагрузки.

Массивы данных значительного объема, накопленные за определенный временной промежуток, могут служить источниками информации о закономерностях поведения ИС, взаимосвязях между данными, сведений для построения сценариев функционирования системы. Получаемая таким образом информация находит свое применение в областях прогнозирования, стратегического антикризисного управления, в анализе рисков [Rabunal, Dorrado 2006].

Выделяют следующие типы закономерностей в данных, выявляемые с помощью ИАД:

- ассоциация – оценка степени связи событий друг с другом;
- секвенция – оценка вероятности возникновения последовательности событий;
- классификация – оценка степени принадлежности некоторого объекта к отдельному классу с определенным набором свойств;
- кластеризация – выделение классов объектов со сходственными наборами параметров.

На данный момент разработано и используется значительное количество методов ИАД, которые вполне применимы в целях оптимизации электронного документооборота. Среди них можно выделить следующие:

1) регрессионный, дисперсионный, корреляционный анализ;
2) построение эмпирических моделей;
3) реализация нейросетевых алгоритмов ИАД;
4) построение деревьев решений;
5) построение кластерных моделей;

Метод регрессионного, дисперсионного, корреляционного анализа подразумевает исследование: а) степени влияния нескольких независимых атрибутов на зависимый атрибут (регрессионный анализ); б) степени влияния атрибутов друг на друга (корреляционный анализ), в) степени влияния факторов, определяющих атрибут, на его значение (дисперсионный анализ).

Построение эмпирических моделей предполагает накопление сведений об опыте функционирования системы. Как правило, отдельно взятая эмпирическая модель применяется для анализа данных в строго конкретной предметной области (например, финансовом анализе).

Метод построения деревьев решений предоставляет не всегда оптимальные, но всегда наглядные решения. Этот подход также является довольно популярным в силу своей простоты и наглядности. Деревья решений представляют собой иерархическую структуру вопросов на «да» и «нет». Листьями в данном случае являются утверждения об отнесении объекта к соответствующему классу (категории) или о наступлении того или иного события. Недостатком данного подхода является его ограниченность в построении правил логического вывода – решение задачи сводится к формированию цепочки последовательного пересмотра свойств объектов, а не нахождению настоящих и логически полных закономерностей поведения объектов системы.

Построение кластерных моделей представляет собой объединение сходственных объектов в группы (классы) на основании сходных значений в наборе данных.

Особого внимания заслуживает реализация нейросетевых алгоритмов ИАД, основанная на применении аппарата искусственных нейронных сетей (ИНС). Преимуществом этого направления является одно из основных свойств ИНС – способность к обучению. Обучение ИНС состоит в «тренировке» на ранее полученных наборах, включающих входные и соответствующие им выходные данные. С помощью различных алгоритмов обучения веса межнейронных связей устанавливаются таким образом, чтобы для любых наборов входных данных получаемые ответы были максимально близки к правильным ответам. Недостатками являются необходимость в большой обучающей выборке данных; сложность анализа структуры ИНС (человеческий мозг не способен интерпретировать набор из тысяч значений весов связей внутри сети). Данный подход широко используется в задачах ИАД.

На уровне реализации целью процесса ИАД является построение модели, отражающей поведение системы и позволяющей прогнозировать поведение системы в зависимости от изменения внешних и внутренних факторов.

Предлагаемая нами методика основана на использовании регрессионного, дисперсионного, корреляционного анализа данных, нейросетевых, генетических моделей. Реализация СИАД предполагает разработку программного обеспечения – набора библиотек утилит, расширяющих функциональные возможности СЭД.

Внедрение предлагаемой методики связано с решением таких проблем, как минимизация информационных рисков, согласование порядка эксплуатации СИАД с законодательством РФ, приведение программного и аппаратного обеспечения в соответствие с возникающими в процессе функционирования системы интеллектуального анализа данных характером задач и уровнем нагрузки. Наряду с перечисленными, следует отметить важнейшие вопросы обеспечения конфиденциальности, безопасности, а также предотвращения некорректного использования и искажения данных.

Использование комплекса методов ИАД позволяет системно исследовать вопросы влияния внешних и внутренних факторов на эффективность производственных процессов. Среди основных преимуществ ИАД применительно к электронному документообороту следует особо отметить возможность осуществления ситуационного анализа с целью оптимизации производственных процессов.

Безусловным преимуществом ИАД является его универсальность, то есть возможность применения в самых различных областях современных информационных технологий. Широкое внедрение систем интеллектуального анализа данных создает предпосылки для перехода к качественно новому уровню управления производственными процессами, в том числе работой систем электронного документооборота. Это дает основание рассматривать разработку систем интеллектуального анализа данных в качестве магистрального направления развития вычислительной техники в свете стратегии модернизации науки и производства.

Библиографическийсписок

Реинженеринг корпорации: Манифест в бизнесе / Майкл Хаммер, Джеймс Чамли; пер. с англ. Ю. Е. Корнилович. М.: Манн, Иванов и Фербер, 2006.

Artificial Neural Networks in Real-Life Applications / Rabunal J. R., Dorrado J. (Eds.). Hershey – London – Melbourne – Singapore: Idea Group Publishing, 2006.

Data Mining: Opportunities and Challenges / Wang J. (Ed.). Hershey – London – Melbourne – Singapore – Beijing: Idea Group Publishing, 2003.