8. СИНТОЛ // Сборник переводов по вопросам информационной теории и практики, № 10. М.: ВИНИТИ, 1968. 177 с.

9. Ермаков А.Е., Плешко В.В. Синтаксический разбор в системах статистического анализа текста // Информационные технологии. 2002. № 7. С. 30-34.

10. Гладкий А.В., Мельчук И.А. Грамматики деревьев I. Опыт формализации преобразований синтаксических структур естественного языка: сб. «Информационные вопросы семиотики, лингвистики и автоматического перевода». Вып. 1. М., 1971. С. 16-41.

11. Ахо А., Хопкрофт Дж., Ульман Дж. Структуры данных и алгоритмы / Пер. с англ. : М.: ИД «Вильямс», 2003. 384 с.

Ольга Витальевна ЖЕЛУДКОВА —
доцент кафедры информационной безопасности
кандидат технических наук
ozheludkova@utmn.ru

Василий Александрович БЕЛЬКОВИЧ —
аспирант кафедры информационной безопасности
belkovichva@mail.ru

Марина Михайловна АКИМОВА —
ассистент кафедры информационной безопасности
a_mar@mail.ru
Институт математики и компьютерных наук
Тюменский государственный университет

УДК 004.056.2

*МОДЕЛЬ ОЦЕНКИ ПРОФЕССИОНАЛЬНОГО РЕЙТИНГА КАК СРЕДСТВО ОБЕСПЕЧЕНИЯ КАЧЕСТВА ДАННЫХ В СЛОЖНЫХ КОРПОРАТИВНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ**[*]

ASSESSMENT MODEL OF PROFESSIONAL RATING AS A MEANS OF ENSURING DATA QUALITY IN COMPLEX ENTERPRISE INFORMATION SYSTEMS

АННОТАЦИЯ. *В работе рассматриваются проблемы, связанные с обеспечением качества данных в корпоративной информационной среде. Наряду с традиционными средствами поддержки качества предлагается использовать автоматически вычисляемую оценку профессиональной репутации пользователей корпоративных информационных систем (КИС).*

SUMMARY. *The paper discusses the problems associated with ensuring data quality in corporate IT environment. Along with the traditional means of quality support the authors suggest to use automatically computed assessment of the users' professional reputation of enterprise information systems (EIS).*

[*] Работа выполнена при поддержке Министерства образования и науки РФ «Проведение научных исследований в области экологии языка и смежных наук» (ГК № 02.740.11.0594).

КЛЮЧЕВЫЕ СЛОВА. Данные, качество, профессиональный рейтинг.
KEY WORDS. Data, quality, professional rating.

Введение

В любой организации, деятельность которой связана с использованием корпоративной информационной системы (КИС), вопросам качества данных уделяется особое внимание.

В работах [1], [2], [3], [4], [5], [6], [7] обсуждаются вопросы, связанные с внутренней поддержкой качества данных за счет использования таких технологий, как интеллектуальный ввод, контроль целостности данных и мониторинг достоверности, корректности, непротиворечивости. Все эти технологии актуальны, если в системе присутствует единый регулятор. Однако когда система представляет некий общий информационный ресурс, участники которого конкурируют между собой, данные технологии не всегда успешно справляются с поставленной задачей. Например, такой системой является общий картографический фонд кадастровых инженеров, в который они выкладывают свой карт-материал, или скачивают карт-материал других пользователей. Очевидно, что автоматически проверить корректность и актуальность таких данных, используя стандартные технологии, сложно.

В данной работе мы рассмотрим подход, основанный на вычислении профессионального рейтинга для каждого пользователя (оператора) КИС с использованием экспертных оценок и статистики [9]. В качестве экспертов могут выступать как непосредственно все пользователи, так и специально наделенные полномочиями эксперта люди (например, администраторы системы). Предлагаемый метод направлен на обеспечение качества информации, относящейся к внешней (по отношению к системе) стороне.

Вычисление профессионального рейтинга

Суть метода состоит в том, что для каждого оператора КИС вычисляется его профессиональный рейтинг, позволяющий при принятии решения о доверии тем или иным данным опираться не только на внутрисистемные критерии качества, но и на рейтинг оператора, внесшего или изменившего эти данные. Иначе говоря, если у оператора низкий рейтинг, значит, высока вероятность ошибки. Прежде чем использовать в работе его данные, их лучше перепроверить. Так же, сравнив рейтинги операторов, можно определить тех, кто вносит в систему наибольшее количество некорректной информации, и в дальнейшем проводить с ними необходимые административные или организационно-учебные мероприятия. Общая модель вычисления рейтинга основывается на количестве данных, внесенных пользователем в систему и количестве допущенных им ошибок. При первоначальной настройке экспертами и администраторами определяется важность тех или иных данных для системы и бизнес-процессов в целом. Важность вносимых или изменяемых данных также учитывается при вычислении рейтинга. Последним параметром, влияющим на рейтинг, является чистота ошибок, допускаемых пользователем и их класс. Классы ошибок также описываются при первоначальной настройке системы.

Первая практически апробированная модель разрабатывалась нами для системы совместной работы с электронными документами [8] (система под-

готовки межевых дел для сообщества кадастровых инженеров), когда от корректности данных, введенных одними пользователями, зависел результат работы других пользователей. Для этой системы были построены конкретные критерии качества данных:

— полнота информации — структура документа была четко регламентирована, поэтому проверка полноты сводилась к сверке документа с этой структурой;

— корректность информации — основным критерием корректности являлось наличие отметки о государственной постановке на учет обрабатываемого участка. Также рассматривались такие критерии, как количество картографического материала на конкретный участок работ и наличие смежных участков;

— достоверность информации — в нашем случае проверка темпорального фактора, т.е. то, что не существует более поздних сведений для данного участка работ с той же степенью корректности информации.

Построенные критерии являются необходимыми, но недостаточными. Процесс постановки оформленного участка на учет требует около месяца. И все это время подготовленные документы доступны другим пользователям, однако, в соответствии с критерием, считаются некорректными. В этом случае решение о доверии таким документам пользователь принимает, опираясь на профессиональный рейтинг оператора, внесшего этот документ. Экспертами в данной предметной области было решено, что рейтинг оператора в основном зависит от двух переменных — это корректность предоставляемых материалов и их значимость для сообщества.

Обозначим через $f(x_n, y_n) \in [0;1]$ функцию, характеризующую рейтинг для оператора n.

x_n среднее арифметическое оценок экспертов для оператора n, за все документы. Данная переменная характеризует корректность предоставляемых материалов.

$$x_n = \frac{\sum_{i=1}^{E_n} \frac{\sum_{j=1}^{R_{in}} r_{jin}}{R_{in}}}{E_n},$$

где $r_{jin} \in [0;1/4;1/2;3/4;1]$ — оценка, выставленная экспертом за конкретный документ, сформированный оператором n. Если экспертом является пользователь системы, то r_{jin} умножается на его рейтинг ($j = 1...R_{in}$); E_n — общее количество документов e_i, сформированных оператором n ($i = 1...E_n$). R_{in} — количество оценок, выставленных за конкретный документ, сформированный оператором n; y_n — среднее арифметическое просмотров документов, сформированных оператором n. Данная переменная характеризует значимость материалов для сообщества.

$$y_n = \frac{\sum_{i=1}^{E_n} \frac{z_{in}}{Z}}{E_n}$$

где z_{in} — количество просмотров конкретного документа, сформированного оператором n. Учитывается только первый просмотр каждого пользователя. Не учитываются просмотры самого оператора n.
Z — количество операторов в системе.

Чтобы социальный фактор не влиял на выставляемые оценки, документы обезличиваются. За один документ каждый может выставить только одну оценку, при этом свои документы оценивать нельзя. Если экспертами являются сами пользователи, то их оценки умножаются на их рейтинг. Рейтинг администраторов системы, выступающих в роли экспертов, считается равным единице.

Так как параметры x_n и y_n не зависят друг от друга, то изначально вид функции $f(x_n, y_n)$ не определен. Для определения вида функции $f(x_n, y_n)$ была собрана статистика по параметрам x_n и y_n. Затем эксперты в данной предметной области, опираясь на эту статистику, «расставили операторов по рейтингу». Проанализировав решение экспертов и статистические данные, была выведена функция $f(x_n, y_n)$.

$$f(x_n, y_n) = x_n * \alpha + (1 - \alpha) * y_n$$

α — параметр, задаваемый в настройках. Определяет меру влияния экспертных оценок и количества просмотров на формируемый рейтинг.

Приведенная модель является конкретным приложением общей модели для обозначенной информационной системы. В настоящее время для информационной системы [8] опытным путем определен параметр $\alpha = 2/3$. Планируется обобщение модели и ее адаптация к работе в информационных системах без привязки к конкретной предметной области. Таким образом, мы получим еще один настраиваемый инструмент обеспечения качества данных.

СПИСОК ЛИТЕРАТУРЫ

1. Шахгельдян К.И. Проблемы качества данных и информации в корпоративной информационной среде вуза // Информационные технологии 2007. № 6. С. 71-80.

2. Science Magazine. Complex Systems. 1999. Vol. 284. № 5411. Pp. 1-212. P. 24.

3. Wand, Y., Wang, R. Anchoring Data Quality Dimensions in Ontological Foundations // Communications of the ACM. 1996. November. Pp. 86-95.

4. Price, R., Shanks, G. A Semiotic Information Quality Framework // Proc. IFIP International Conference on Decision Support Systems (DSS2004): Decision Support in an Uncertain and Complex World, Prato. 2004.

5. Wang, R., Storey, V., Firth, C. A framework for analysis of data quality research // IEEE Trans. on Knowl. Data Eng. 1995. 7, 4. Pp. 623-640.

6. Wang, R., Ziad, M., Lee, Y.W. Data Quality. Kluwer. 2001. P. 167.

7. Madnick, S. Wang, R. and Xian, Xiang. The Design and Implementation of a Corporate Householding Knowledge Processor to Improve Data Quality//Journal of Management Information Systems. 2004. Vol. 20. № 3. Pp. 41-69.

8. Свидетельство о государственной регистрации программ для ЭВМ № 2010615437 от 24.08.2010 АИС «Межевик» (Роспатент).

9. Грищенко В.С. Метрики репутации: построение открытых информационных сред: Дис. … канд. физ.-мат. наук. Екатеринбург, 2007. 124 с.

Александр Анатольевич ЗАХАРОВ —
зав. кафедрой информационной безопасности,
доктор технических наук, профессор
azaharov@utmn.ru

Ольга Андреевна НЕСТЕРОВА —
аспирантка кафедры информационной безопасности
o-nesteroma@mail.ru

Евгений Александрович ОЛЕННИКОВ —
доцент кафедры информационной безопасности,
кандидат технических наук
olennikov@utmn.ru

Институт математики и компьютерных наук
Тюменского государственного университета

УДК 004.852

*АЛГОРИТМ ИНФОРМАЦИОННОГО ПОИСКА В МЕДИЦИНСКИХ АРХИВАХ НА ОСНОВЕ КОНТЕКСТНО-ВРЕМЕННОЙ ОНТОЛОГИИ**

ALGORITHM OF INFORMATION RETRIEVAL IN MEDICAL RECORDS BASED ON CONTEXT-TEMPORAL ONTOLOGY

АННОТАЦИЯ. *Рассматривается использование онтологии для решения проблемы информационного поиска медицинских данных, предлагается алгоритм поиска на основе разработанной семантической модели поиска для поддержки медико-биологических исследований.*

SUMMARY. *The use of ontology for the solution of medical data retrieval problem is considered, the algorithm of semantic retrieval is proposed for medical-biological research support.*

КЛЮЧЕВЫЕ СЛОВА. *Информационный поиск, медицинская информационная система, временная семантика, онтология, контекст*

KEY WORDS. *Information retrieval, medical information system, temporal semantics, ontology, context.*

Повышение качества и доступности медицинской помощи — один из приоритетов государственной политики, благодаря которой развиваются такие направления, как скрининговая, восстановительная и доказательная медицины [1], [2]. Эти направления основываются на результатах медико-биологических

* Работа выполнена при поддержке гранта Министерства образования и науки РФ «Проведение научных исследований в области экологии языка и смежных наук» (ГК № 02.740.11.0594).