DOI: 10.25728/mlsd.2023.0789

НЕЧЕТКИЙ ЛОГИКО-ЛИНГВИСТИЧЕСКИЙ АНАЛИЗ КАЧЕСТВА ДАННЫХ В ФИНАНСОВЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ

Ломазов А.В.
Белгородский национальный исследовательский университет, Белгород, Россия
alomazov@yandex.ru
Ломазов В.А., Аничин В.Л.
Белгородский государственный аграрный университет, Белгород, Россия
vlomazov@yandex.ru, vladislavanichin@rambler.ru
Петросов Д.А.
Финансовый университет при Правительстве РФ, Москва, Россия
DAPetrosov@fa.ru

Аннотация. Предложенный подход к анализу качества данных основан на использовании аппарата нечеткого логико-лингвистического анализа Л. Заде. Показатели представлены в виде лингвистических переменных, семантика которых задается экспертами. Построенная иерархия показателей позволяет повысить научную обоснованность решений при обеспечении качества данных в финансовых системах.

Ключевые слова: финансовая система, качество данных, нечеткий логико-лингвистический анализ, иерархия показателей.

Введение

В настоящее время развитие финансового сектора национальной экономики происходит в условиях высокой волатильности рынков и воздействия большого числа различного рода внешних факторов, приводящих к необходимости учета тесных взаимосвязей с внешней средой, что является одной из важнейших особенностей крупномасштабных систем [1], общая методология управления которыми используется данной работе.

Все более широкое применение современных цифровых технологий в деятельности финансовых и кредитных организаций породило (наряду с несомненными положительными аспектами) увеличение рисков информационной безопасности. Нормативные требования к управлению риском информационных систем и к организации управления качеством данных в них определены Банком России [2]. При этом, как правило, методика оценки качества данных в информационных системах, обеспечивающих критически важные финансовые технологические процессы, разрабатывается кредитной организацией самостоятельно. В этой методике должны быть регламентированы показатели качества данных, а также методы их измерения (расчета).

В настоящее время показатели качества данных носят, как правило, числовой характер [3-5]. Целью настоящей работы является построение системы показателей качества данных в финансовых информационных системах в виде иерархии лингвистических переменных, что в наибольшей степени соответствует применению методологии экспертного оценивания и позволяет производить логико-лингвистический анализ качества данных.

1. Классификация показателей качества данных

В соответствии с [6] выделим 7 основных групп показателей качества данных финансовой информационной системы и рассмотрим показатели, входящие в эти группы.

Группа показателей, отражающих точность и достоверность данных (IND_1):

- Ind_{11} – показатель корректности записей (доля значений, не соответствующих корректным, выявленных по результатам обработки инцидентов, связанных с качеством данных);
- Ind_{12} – показатель соответствия эталонным данным/первоисточникам (доля значений, не соответствующих первоисточникам).

Группа показателей, отражающих полноту данных (IND_2):

- Ind_{21} – показатель заполнения полей документов (доля незаполненных полей, обязательных к заполнению);
- Ind_{22} – показатель полноты учета объектов (доля объектов учета, данные о которых не включены в информационную систему, в общем объеме объектов учета);
- Ind_{23} – показатель избыточности записей (доля записей с дублированием значений атрибутов).

Группа показателей, отражающих актуальность данных (IND_3):

- Ind_{31} – показатель относительной актуальности данных (доля актуальных данных в общем объеме данных);
- Ind_{32} – показатель готовности данных (пребывания в актуальном/нормативном состоянии к времени пребывания в неактуальном или неопределенном состоянии);
- Ind_{33} – показатель относительной длительности обработки данных (отношение времени проверки ошибок, согласования и внесения данных к общему времени функционирования информационной системы).

Группа показателей, отражающих согласованность данных (IND_4):

- Ind_{41} – показатель нестандартных обозначений (доля использования нестандартных альтернативных обозначений к общему количеству обозначений);
- Ind_{42} – показатель нестандартных наименований объектов учета (доля нестандартных наименований объектов учета в общем объеме объектов учета).

Группа показателей, отражающих доступность данных (IND_5):

- Ind_{51} – показатель времени доступности данных (отношение времени доступности к общему времени функционирования информационной системы);
- Ind_{52} – показатель времени простоя информационной системы (отношение времени простоя к общему времени функционирования информационной системы).

Группа показателей, отражающих контролируемость данных (IND_6);

- Ind_{61} – показатель контроля источников данных (доля данных с нефиксированным источником в общем объеме данных);
- Ind_{62} – показатель контроля изменения данных (доля данных с нефиксированным изменением в общем объеме данных).

Группа показателей, отражающих восстанавливаемость данных (IND_7):

- Ind_{71} – показатель восстанавливаемости после сбоев (доля невосстановленных данных в общем объеме данных после их утраты, повреждения или изменения в результате сбоев информационной системы);
- Ind_{72} – показатель восстанавливаемости после ошибок персонала (доля невосстановленных данных в общем объеме данных после их утраты, повреждения или изменения в результате ошибок или иных непредусмотренных действий персонала).

Следует отметить, что приведенный перечень показателей качества данных в финансовых информационных системах может быть расширен, исходя из специфики технологических процессов конкретной кредитной организации, в рамках ее самостоятельности при утверждении соответствующей методики.

2. Нечеткое логико-лингвистическое представление показателей качества данных

В рамках логико-лингвистического [7] представления показателей качества данных в финансовых информационных системах рассмотрим лингвистическую переменную *DQI* (Data Quality Indicator), формально представимую в виде пятикомпонентного кортежа

$$< DQI, D_{DQI}, TB_{DQI}, SintR_{DQI}, SemR_{DQI}> \quad (1)$$

где *DQI* – наименование индикативной лингвистической переменной; D_{DQI} – диапазон числовых значений рассматриваемого показателя (с учетом нормировки можно считать, что $D_{DQI}=[0, 1]$); TB_{DQI} – упорядоченный набор базовых индикативных термов (будем полагать TB_{DQI} = {*very low, low, medium, high, very high*}); $SintR_{DQI}$ – набор синтаксических правил, позволяющих порождать наименования индикативных термов из наименований элементов TB_{DQI} и тем самым расширяющих множество базовое множество термов TB_{DQI} до множества термов $T_{DQI} = \{t_1, t_2, \ldots, t_K\}$. $SemR_{DQI}$ – набор семантических правил устанавливающих соответствие между термами и нечеткими подмножествами D_{DQI}. В дальнейшем ограничимся синтаксическими правилами формирования наименований термов в виде логических связок (дизъюнкций и конъюнкций) базовых термов и/или их отрицаний. Семантика таких термов определяется с использованием аппарата T-норм и S-конорм [7]. При этом семантика базовых термов задается экспертами на основе их знаний о предметной области проекта. В дальнейшем полагается, что функции принадлежности нечетких подмножеств являются трапециевидными, как показано на рисунке 1.

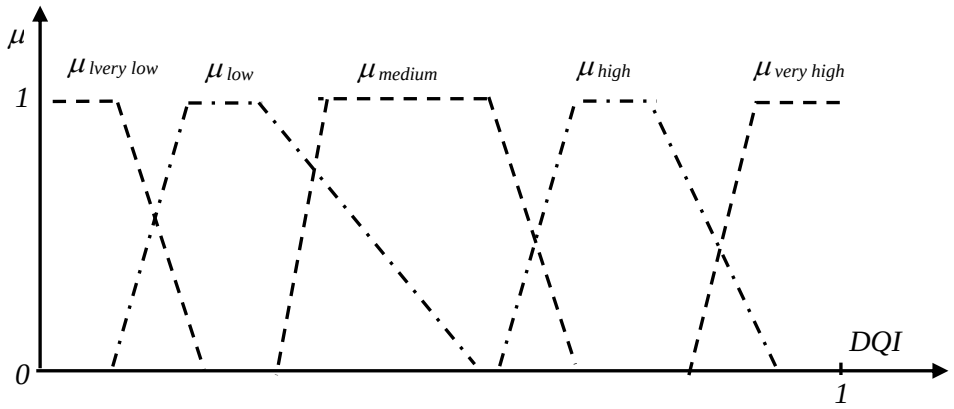

Рис. 1. Графическое представление функций принадлежности $\mu_{very\ low}$, μ_{low}, μ_{medium}, μ_{high}, $\mu_{very\ high}$, отражающих семантику термов very low, low, medium, high, very high лингвистической переменной DQI

В рамках трапециевидной формы функций принадлежности базового терма эксперты задают четыре числовых значения в диапазоне, отвечая на вопросы:

- Каково числовое значение показателя, начиная с которого эти значения соответствуют терму?
- Каково числовое значение показателя, начиная с которого эти значения перестают соответствовать терму?
- Каково минимальное числовое значение показателя, полностью соответствующее терму?
- Каково минимальное числовое значение показателя, полностью соответствующее терму?

Первые два вопроса определяют носитель функции принадлежности (нижнее основание трапеции), а вторые два вопроса – ядро функции принадлежности. Исключением из приведенного правила является построение первого и последнего термов упорядоченного перечня TB_{DQI}. Как видно из рисунка 1, этим термам соответствуют полутрапеции (левая и правая части трапеций) и поэтому для построения первой из них достаточно два числовых значения показателя (второй и третий вопросы экспертного листа), а для построения второй – также два числовых значения показателя (первый и четвертый вопросы экспертного листа)

3. Логико-лингвистическая иерархия показателей качества данных

В дальнейшем будем рассматривать введенные показатели качества данных в качестве лингвистических переменных, построенных в соответствии с представлением (1). Дополним этот набор показателей качества интегральными показателями Ind_i , $i=1,2,...,7$, соответствующими общим оценкам качества данных по группам IND_i , $i=1,2,...,7$ и представленными в виде лингвистических переменных в соответствии с (1). Будем полагать, что интегральные показатели связаны с частными показателями своей группы зависимостями, представимыми в виде системы нечетких продукционных правил:

$$IF\ F_{ip}\ (Ind_{ij} - t_k\ ;\ j=1,2,...,J_i,\ k=1,2,...,K)\ THEN\ (Ind_i - bt_p) \quad (2)$$
$$i=1,2,..,7,\ p=1,2,...,5$$

где $(Ind_{ij} - t_k)$ – нечеткие высказывания относительно соответствия значений индикаторов Ind_{ij} термам $t_k \in T_{DQI}$; J_i – количество показателей в группе IND_i; K – количество термов в T_{DQI}; F_{ip} – функции исчисления нечетких высказываний, соответствующие высказываниям относительно соответствия значений индикаторов Ind_i термам $bt_p \in TB_{DQI}$.

Дополним построенную систему показателей еще и общим показателем качества данных *Ind*, в целом характеризующий уровень управления данными в финансово-кредитной организации. Полагая, что *Ind* также может быть описан в рамках логико-лингвистического представления (1), представим его зависимость от интегральных показателей Ind_i , $i=1,2,...,7$ в виде системы нечетких продукционных правил:

$$IF\ F_p\ (Ind_i - t_k\ ;\ i=1,2,...,7,\ k=1,2,...,K)\ THEN\ (Ind - bt_p) \quad (3)$$
$$p=1,2,...,5$$

где наряду с обозначениями, использованными при построении соотношений (2), полагается, что $(Ind_i - t_k)$ – нечеткие высказывания относительно соответствия значений индикаторов Ind_i термам $t_k \in T_{DQI}$; F_p – функции исчисления нечетких высказываний, соответствующие высказываниям относительно соответствия значений индикатора *Ind* термам $bt_p \in TB_{DQI}$. При этом для построения нечетких высказывательных функций F_p (как и для построения функций F_{ip}), определим сначала их четкие аналоги. Для этого предложим экспертам заполнить таблицы значений при всех значениях аргументов рассматриваемых функций, после чего построим для таблично заданных логических функций нормальные дизъюнктивные формы (НДФ). Нечеткие аналоги построенных НДФ определят вид функций F_p.

Таким образом, сформирована нечеткая логико-лингвистическая иерархия показателей качества данных в финансово-кредитной организации:

- верхний ярус (вершина иерархического дерева) – общий показатель *Ind*;
- средний ярус – интегральные показатели Ind_i , $i=1,2,...,7$;
- нижний ярус (листья иерархического дерева) – частные показатели Ind_{ij} , $j=1,2,...,J_i$, $i=1,2,..,7$.

Общая процедура определения значения общего показателя качества данных, в рамках построенной иерархии, содержит следующие этапы:

- определение частных показателей Ind_{ij} , $j=1,2,...,J_i$, $i=1,2,..,7$;
- фаззификация частных показателей с использованием представлений (1);
- нечеткий логический вывод лингвистических значений интегральных показателей Ind_i , $i=1,2,...,7$ с использованием системы нечетких продукционных правил (2);
- нечеткий логический вывод лингвистических значений общего показателя *Ind* с использованием системы нечетких продукционных правил (3);
- акумуляция заключений и дефаззификация значения общего показателя качества данных *Ind*.

Для реализации этапа нечеткого логического вывода целесообразно использовать основные шаги метода Мамдани (агрегирование подусловий, активизация подзаключений), отличающегося от подобных алгоритмов (Tsukamoto, Sugeno, Larsen) высоким уровнем интерпретируемости.

4. Использование логико-лингвистической иерархии показателей для поддержки принятия решений при управлении качеством данных

Построенная иерархия показателей может быть использована для поддержки принятия решений при управлении системой обеспечения качеством данных в информационной системе финансово-кредитной организации.

Оценка значения общего показателя качества данных *Ind* позволяет сделать вывод о возможности оставления системы управления качеством данных в существующем виде (сохранения *Sav*) или необходимости внесения незначительных изменений (модификации, *Mod*) или даже коренного преобразования (реинжиниринга, *Rein*) этой системы. При этом возможно применение двух подходов:

- традиционный подход, основанный на использовании числовых значений показателей качества;
- логико-лингвистический подход, основанный использовании лингвистических значений показателей качества.

В рамках первого подхода выбор стратегии *Strat* из множества $STRAT = \{Sav, Mod, Rein\}$ осуществляется с использованием решающего правила:

$$Strat = \begin{cases} Sav, & 0 \le Ind < a \\ Mod, & a \le Ind < b \\ Rein, & b \le Ind \le 1 \end{cases}$$

где a, b – задаваемые экспертами границы интервалов выбора стратегий ($0 < a, b < 1$).

В соответствии со вторым подходом *Strat* понимается как лингвистическая переменная с базовым терм-множеством $\{Sav, Mod, Rein\}$ и решающие правила имеют вид нечетких продукционных правил:

$$IF\ F_{Sav}\ (Ind - bt_p\ ,\ Fact - bt_s;\ p,s=1,2,...,5)\ THEN\ (Strat - Sav)$$

$$IF\ F_{Mod}\ (Ind - bt_p\ ,\ Fact - bt_s;\ p,s=1,2,...,5)\ THEN\ (Strat - Mod)$$

$$IF\ F_{Rein}\ (Ind - bt_p\ ,\ Fact - bt_s;\ p,s=1,2,...,5)\ THEN\ (Strat - Rein)$$

где F_{Sav}, F_{Mod}, F_{Rein} – нечеткие пропозициональные функции, а *Fact* – задаваемые в виде логической переменной внешние факторы, влияющие на принятие решений.

В рамках логико-лингвистического подхода решение находится нечеткого логического вывода и имеет вид

$$< (Sav, \mu_{Sav}), (Mod, \mu_{Mod}), (Rein, \mu_{rein})>$$

где значения μ_{Sav}, μ_{Mod}, μ_{rein} , лежащие в интервале [0,1], соответствуют мере предпочтительности каждой из рассмотренных стратегий. При этом окончательный выбор стратегии остается лицу, принимающему решение.

Иерархическое представление системы критериев качества данных поддерживает также принятие решений в рамках большей детализации. Например, после того, как принято решение о модификации системы обеспечения качества данных может быть рассмотрен вопрос определения (выбора) подсистемы, в которую следует внести изменения. Тогда (полагая, что подсистемы соответствуют классам показателей) рассмотрим лингвистической переменную *Subsist*, базовое терм-множество которой содержит термы $Subsist_1$, $Subsist_2$,…, $Subsist_7$.

В этом случае решающие правила выбора будут иметь вид:

$$IF\ F_{Subsist,i}\ (Ind_s - bt_p;\ ,s=1,2,...,7,\ p=1,2,...,5)\ THEN\ (Subsist - Subsist_i)$$

$$i = 1,2,...,7$$

Решение по выбору модифицируемой подсистемы, получаемое на основе процедуры нечеткого логического вывода, будет состоять из пар

$$(Subsist_i, \mu_{Subsist,i}),\ i = 1,2,...,7$$

где значения $\mu_{Subsist,1}$, $\mu_{Subsist,2}$, …, $\mu_{Subsist,7}$, лежащие в интервале *[0,1]*, соответствуют мере необходимости в модификации каждой из рассмотренных подсистем обеспечения качества данных. При этом возможен выбор нескольких подсистем подлежащих модификации.

Дальнейшая детализация решений по обеспечению необходимого уровня качества данных может быть связано с выбором конкретных мероприятий по модификации выбранной подсистемы. Например, после того, как принято решение о модификации подсистемы $Subsist_1$ возможен выбор для последующей реализации одного или нескольких мероприятий из совокупности $\{Event_1, Event_2,..., Event_M\}$. В этом случае решающие правила выбора, принимаемого на основе значений частных показателей (нижний ярус иерархии показателей) будут иметь вид:

$$IF\ F_{Event,m}\ (Ind_{ij} - bt_p;\ ,i=1,2,...,7,\ j=1,2,...,\ J_i\ ,\ p=1,2,...,5)\ THEN\ (Event - Event_m)$$

$$m = 1,2,...,M$$

Решение по выбору способов (мероприятий), направленных на модификацию выбранной подсистемы, получаемое на основе процедуры нечеткого логического вывода, будет состоять из пар

$$(Event_m, \mu_{Event,m}),\ m = 1,2,...,M$$

где значения $\mu_{Event,1}$, $\mu_{Event,2}$, …, $\mu_{Eventt,M}$, лежащие в интервале *[0,1]*, соответствуют мере необходимости в модификации каждой из рассмотренных подсистем обеспечения качества данных. При этом возможен выбор нескольких мероприятий, обеспечивающих модификацию.

Таким образом, если иерархическое дерево показателей качества данных строилось по принципу обобщения (снизу вверх), то дерево решений строится по принципу последовательной детализации решений (сверху вниз).

5. Заключение

Предложенный (в рамках применения методов искусственного интеллекта для управления данными в финансовых информационных системах) подход к анализу качества данных основан на использовании аппарата нечеткого логико-лингвистического анализа Л. Заде. Показатели представлены в виде лингвистических переменных, семантика которых задается экспертами. Построена иерархия показателей качества данных, нижний ярус которой составляют показатели, допускающие непосредственное измерение, в то время как показатели последующих уровней являются обобщениями (интеграцией) показателей предыдущего уровня. Переход между ярусами производится на основе процедуры нечеткого логического вывода с использованием системы нечетких продукционных правил, описывающих закономерности предметной области и задаваемых

экспертами. Построенная система показателей дает возможность повысить научную обоснованность управленческих решений при управлении качеством данных в финансовых информационных системах. Одним из направлений дальнейших исследований является использование построенных показателей в рамках задачи структурно-параметрического синтеза системы управления качеством данных [10].

Литература

1. *Цвиркун А. Д.* Управление развитием крупномасштабных систем в новых условиях // Проблемы управления. 2003. N 1.– С. 34–43.
2. Положение Банка России от 08.04.2020 № 716-П «О требованиях к системе управления операционным риском в кредитной организации и банковской группе» https://cbr.ru/faq_ufr/dbrnfaq/doc/?number=716-П
3. *Woodall P., Oberhofer M., Borek A. A* Classification of data quality assessment and improvement methods// International Journal of Information Quality. 2014. №3 (4).– Pp. 298–321.
4. *Woodall P., Borek A., Parlikad A.* Data quality assessment: the hybrid approach// Information & Management. 2013. № 50 (7). – Pp.. 369–382.
5. *Baškarada S., Koronios A.* A critical success factors framework for information quality management // Information Systems Management. 2014. № 31 (4). – Pp. 1-20.
6. *Савицкая М., Рощупкин И.* Как оценить качество данных в информационных системах по Положению № 716-П и зачем это нужно // Внутренний контроль в кредитной организации. 2023. №1 (57).– С.66-81.
7. *Zadeh LA.* Generalized theory of uncertainty (GTU)–principal concepts and ideas // Computational Statistics & Data Analysis. 2006. 51(1). – Pp. 15-46.
8. *Russell S.J. , Norvig, P.* Artificial intelligence artificial intelligence: a modern approach.– New Jersey: Prentice Hall, 2020. – 1136 p.
9. *Mamdani E.H.* Applications of fuzzy algorithms for control of simple dynamic plant // Proceedings of the Institution of Electrical Engineers. 1974. 121(12).– Pp. 1585–1588.
10. *Petrosov D.A., Lomazov V.A., Lomazova V.I., Glushak A.V.* Applications of parallel computations in the problems of structural-parametric synthesis of discrete systems based on evolution methods // Journal of Advanced Research in Dynamical and Control Systems. 2018. 10 (10 Special Issue).– Pp.1840-1846.