УДК 519.766.4

ПРОГНОЗУВАННЯ ПОВЕДІНКИ ЯКОСТІ ІНФОРМАЦІЇ

М.М. КОНОВАЛЮК

Розглянуто показники якості інформації (метрики Data Quality). Запропоновано підхід та інформаційна технологія до оцінювання та прогнозування показника якості, що описує достовірність інформації. Проведено короткий аналіз робіт, що присвячені різним підходам до визначення метрик Data Quality. Нелінійний характер показника достовірності інформації дає можливість прогнозувати його майбутню поведінку із застосуванням моделі стохастичної волатильності (МСВ), в якій для оцінювання параметрів застосовано програмно реалізований алгоритм Гіббса. Запропоновано розроблену для прогнозування волатильності валютного курсу інформаційну технологію застосувати для прогнозування майбутньої поведінки міри невизначеності показника достовірності інформації. Прогнозування показника достовірності інформації має ключовий вплив на процес прийняття рішень.

ВСТУП

Теорія якості інформації (Data Quality) формується на основі різних дисциплін та різних підходів. Актуальність напряму пов'язана з появою комп'ютерів, сховищ даних з великими об'ємами інформації та з необхідністю обробки цього масиву інформації. Вивчення непрямо пов'язаних аспектів Data Quality тривало багато років у різних дисциплінах, серед яких: теорія інформації, семіотика, теорія прийняття рішень, теорія оптимального керування та ін. В окрему дисципліну Data Quality сформувалось у середині 1990-х років. Суттєвий вклад у формування Data Quality в окрему дисципліну зробили такі дослідники, як Ванг Р., Лі Ю., Балоу Д. та інші, отримавши значні теоретичні та практичні результати [1, 2].

Зусилля дослідників були зосередженні на формуванні основних понять та означень Data Quality. Проте досі немає єдиної загальноприйнятної системи означень метрик Data Quality.

Мета роботи — застосування інформаційної технології, розробленої для прогнозування волатильності валютного курсу, для прогнозування майбутньої поведінки міри невизначеності показника достовірності інформації, важливого при прийнятті рішень.

ОГЛЯД РОБІТ

Метрики Data Quality

Data Quality є одним із найменш досліджених напрямів серед дисциплін, пов'язаних з інформацією. Основним питанням є визначення метрик Data Quality. Порівняння деяких підходів до визначень метрик Data Quality, які і зараз є актуальними, проведено в роботах [3] та [4]. Розглянемо підходи, описані в [3]:

- Ванд Ю., Ванг Р. [7]. Метрики Data Quality визначаються на основі функцій, що мають своє відображення в інформаційній системі. Запропоновано 5 метрик: повнота, точність, своєчасність, несуперечливість, надій-

ність. Метрика неточності інформації визначається як представлення стану реального світу відмінним від того, яким його слід було б відобразити. Метрика повноти визначається як втрати під час відображення станів реального світу у стани інформаційної системи.

- Ванг Р., Стронг Д. Цими вченими метрики Data Quality сформовано на основі емпіричного підходу. Автори виділили 15 різних метрик із 179 розглянутих.
- Редман Т. Ним метрики Data Quality згруповано у три групи: група концептуального виду інформації, що складається з 5-и метрик; група розміру інформації з 4-х метрик; група формату інформації з 8-и метрик.
- Жарке М. запропонував специфічні метрики Data Quality, які класифікуються відповідно до ролі користувачів у середовищі бази даних, а саме: 6 метрик для якості проектування та адміністрування; 6 метрик для якості програмного забезпечення; 5 метрик для якості використання інформації та 5 метрик для якості зберігання інформації.
- Автори: Бовее М., Срівастава М. та Мак Б., пропонують 4 метрики, а саме: відкритість для доступу (accessibility); можливість інтерпретувати (interpretability); доцільність (relevance); вірогідність (credibility).
- Науманном Ф. було розглянуто 21-у метрику, які згруповано з точки зору інтегрованих Web-інформаційних систем у 4 групи: має відношення до змісту; технічна; інтелектуальна; має відношення до створення екземплярів.

Порівняння метрик DataQuality

Порівняльну таблицю згаданих вище метрик Data Quality наведено в [3]. В табиці введено такі позначення: (–) — в роботах різних авторів використовується та сама назва для метрик з різними значеннями; (+/–) — різні пропозиції використовують ту саму назву для метрик зі схожими значеннями; (+) — ті самі назви й ті самі значення для характеристик у пропозиціях різних авторів.

Таблиця. Відповідність між означеннями метрик якості інформації з однаковими назвами

Автор / Показник якості інформації	**Ванд, Ванг (1996)**	**Ванг, Стронг (1996)**	**Редман (1996)**	**Джарке (1999)**	**Бовее (2001)**	**Науманн (2002)**
Точність	+	+	+	+	+	+
Повнота	+	–	+	+/–	+	+
Несуперечність	+		+/–	+	+	
Репрезентаційна несуперечність		+	+			+
Своєчасність	+	+		+	+	+
Вжиток	+		+	+	+	
Змінність	+			+	+	
Можливість інтерпретувати		+	+/–	+/–	+	+
Можливість зрозуміти		+				+
Достовірність	–			–		
Вірогідність				–	–	
Правдоподібність		+				+

Продовження таблиці

Репутація		+				+
Об’єктивність		+				+
Доцільність		+	+		+/–	+
Відкритість для доступу		+		+	+	
Безпека/ Безпечний доступ		+		+		+
Величина доповнень		+				+
Лаконічність репрезентації		+				+
Обсяг інформації		–	–			–
Доступність				+		+
Портативність			–	–		
Чутливість				+		+

Метрики, визначені у всіх пропозиціях — це точність та повнота. Метрики, що мають відношення до несуперечності та до часу, беруться до уваги у більшості пропозицій. Також більшістю пропозицій розглядається можливість інтерпретувати. Кожна з інших метрик включається меншістю пропозицій. В деяких випадках існує повна неузгодженість у визначенні специфічних метрик, таких як достовірність.

Подібність між різними назвами метрик із схожими (+/–) або однаковими значеннями (+) представлено на рисунку.

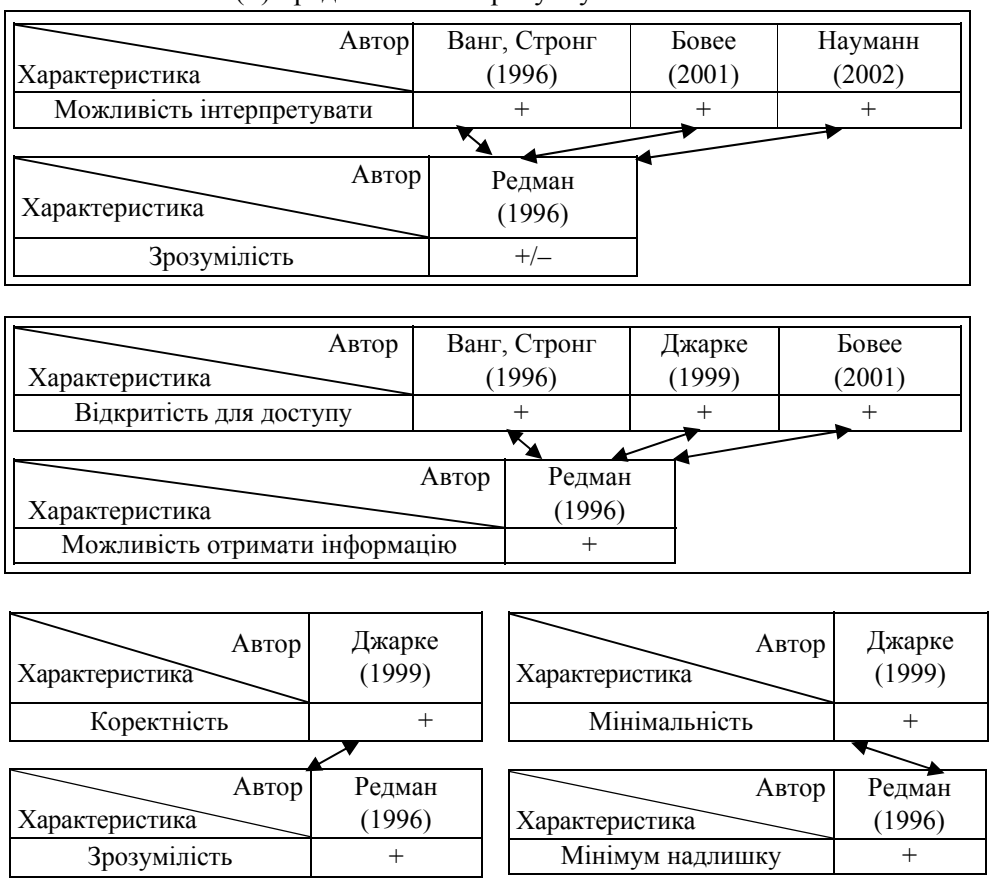

Рисунок. Відповідність між метриками інформації з різними назвами

Виходячи з більшості пропозицій стосовно метрик, Data Quality базується на таких метриках, як: *точність, повнота, несуперечність, своєчасність, можливість інтерпретувати та відкритість для доступу*. Найбільше уваги приділено повноті, достовірності та своєчасності інформації. Всі інші запропоновані метрики або мають другорядні властивості, або є більш залежними від контексту (тобто дуже специфічні).

Застосування метрик повноти, достовірності та своєчасності інформації на основі нечітких множин розглядається в роботі [5]. Продемонстровано застосування показників Data Quality з точки зору прийняття рішень. Наведено механізм визначення допустимого періоду часу на прийняття рішення та класифікацію ситуацій за умов нечітких множин. Демонструється зростання у часі метрики достовірності. За реальних обставин інформація, що надходить, може бути недостовірною, тобто значення метрики достовірності у часі може зменшуватись. Отже, метрика достовірності має невизначений характер (може зростати або спадати) та потребує детального дослідження.

ПРОГНОЗУВАННЯ МЕТРИКИ ДОСТОВІРНОСТІ ІНФОРМАЦІЇ

Серед розмаїття метрик виділимо 3 основних, поведінка яких залежить від часу — це метрики повноти, достовірності та своєчасності інформації.

З моменту отримання перших відомостей величина показника *повноти* зростає. Його поведінка може мати нерівномірний характер, але загальна тенденція не змінюється [6].

Величина показника *своєчасності* з часом зменшується. Так само як і зміна показника повноти, спадання величини показника своєчасності може мати нерівномірний характер.

Величина показника *достовірності* інформації з плином часу може випадковим чином збільшуватися або зменшуватися, оскільки інформація, що надходить, може бути як достовірною, так і недостовірною. Проаналізуємо поведінку показника достовірності, оскільки саме він може мати остаточний вплив у процесі прийняття рішення.

Поведінка показника достовірності має невизначений, тобто випадковий характер, і може бути представлена моделлю, яка описує змінну волатильність показника достовірності. Мірою невизначеності є дисперсія, поведінку якої описують різноманітні авторегресійні моделі. Серед великої кількості авторегресійних моделей виділимо нелінійні моделі, а саме, модель стохастичної волатильності (МСВ), яка здобула популярність завдяки високій адекватності опису умовної дисперсії досліджуваних процесів.

МСВ описує процес, що має нелінійний нестаціонарний характер, який і має показник достовірності. Оцінка параметрів МСВ показника достовірності дозволить прогнозувати його майбутню поведінку, а саме дисперсію показника достовірності інформації. Прогнозування не дає відповіді щодо напряму зміни показника достовірності, але дає відповідь про його майбутню стаціонарність, що може вплинути на процес прийняття рішення.

ПОБУДОВА МОДЕЛІ

Розглянемо поведінку показника достовірності інформації на прикладі валютного курсу. Позначимо P_t та P_{t+1} величину курсу валют в момент часу

t та $t+1$ відповідно, а значення P_t^* та P_{t+1}^* — очікувані значення валютного курсу в момент часу t та $t+1$ відповідно. Нехай N — загальна кількість вимірювань за період часу T, n_i — кількість вимірювань за певний період часу t_i, n_i^* — кількість вимірювань точних значень, що співпадає з прогнозованими $(P_t = P_t^*)$ за період часу t_i, $\left(N = \sum_{i=1}^{k} n_i, T = \sum_{i=1}^{k} t_i \right)$. Величиною показника достовірності за період часу t_i є відношення $I_{\text{accuracy},t_i} = \frac{n_i^*}{n_i}$. Набори вибірок величини показника достовірності I_{accuracy,t_i} утворюють вибірку на множині значень N, а саме $I_{\text{accuracy},t_1}, I_{\text{accuracy},t_2}, \ldots, I_{\text{accuracy},t_k}$. Іншими словами, формується числовий ряд.

У процесі зростання величини показника достовірності інформації, відношення $\frac{I_{\text{accuracy},t+1}}{I_{\text{accuracy},t}} > 1$, а при зменшенні — $\frac{I_{\text{accuracy},t+1}}{I_{\text{accuracy},t}} < 1$. Показник достовірності залишається незмінним при $\frac{I_{\text{accuracy},t+1}}{I_{\text{accuracy},t}} = 1$. Відповідно логарифми показників достовірності інформації $\ln\frac{I_{\text{accuracy},t+1}}{I_{\text{accuracy},t}} > 0$, $\ln\frac{I_{\text{accuracy},t+1}}{I_{\text{accuracy},t}} < 0$ та $\ln\frac{I_{\text{accuracy},t+1}}{I_{\text{accuracy},t}} = 0$. Позначимо $y_t = \ln\frac{I_{\text{accuracy},t+1}}{I_{\text{accuracy},t}}$.

Може мати місце значна зміна показника достовірності інформації. Суттєва поступова зміна величини y_t, на відміну від суттєвої точкової зміни у момент t, майже не впливає на остаточний результат. Тому доцільно зменшити вплив точкових збурень, які сильно впливають на точність результатів моделювання. З цією метою усереднимо значення y_t:

$$y_t = \ln\frac{I_{\text{accuracy},t+1}}{I_{\text{accuracy},t}} - \frac{1}{N}\sum_{i=1}^{N} \ln\frac{I_{\text{accuracy},i+1}}{I_{\text{accuracy},i}},$$

де N — обсяг вибірки.

Поведінку часового ряду y_t можна представити наступною математичною моделлю:

$$y_t = \mu + \sigma_t u_t, \quad u_t \sim \mathrm{N}(0,1),$$

де μ — середнє значення, σ_t — середньоквадратичне відхилення, тобто корінь з дисперсії, u_t — нормально розподілена випадкова величина. Середнє значення y_t дорівнює 0, тобто модель матиме вигляд:

$$y_t = \sigma_t u_t, \quad u_t \sim \mathrm{N}(0,1).$$

В результаті отримали умовно-гаусівську модель, де σ_t — середньо квадратичне відхилення або волатильність. Поведінку волатильності σ_t можна описати різними математичними моделями, як лінійними так і нелінійними.

Поведінка дисперсії величини показника достовірності на основі моделі АРУГ:

$$\begin{cases} y_t = \sigma_t u_t, \\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2. \end{cases}$$

Поведінка дисперсії величини показника достовірності на основі моделі УАРУГ. У цій моделі волатильність залежить як від показника y_t, так і від σ_t:

$$\begin{cases} y_t = \sigma_t u_t, \\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2, \end{cases}$$

де $\alpha_i > 0$, $\beta_j > 0$.

Поведінку дисперсії також можна представити на основі моделі стохастичної волатильності (МСВ), що була запропонована Тейлором [7]:

$$\begin{cases} y_t = \sigma_t u_t, \\ \ln \sigma_{t+1}^2 = \mu + \phi(\ln \sigma_t^2 - \mu) + \eta_t, \end{cases} \quad u_t \sim \mathrm{N}(0,1), \ \eta_t \sim \mathrm{N}(0, \sigma_\upsilon),$$

де σ_t — волатильність у момент часу t; u_t, υ_t — два незалежних гаусівських процеси білого шуму з дисперсіями 1 та σ_υ, відповідно; μ, ϕ, σ_υ — параметри моделі, значення яких повинні бути меншими по модулю за 1. Поведінка дисперсії являє собою ланцюг Маркова, оскільки наступне значення залежить тільки від поточного значення та не залежить від минулих. Для оцінювання параметрів моделі процес має бути стаціонарним, тобто значення параметрів моделі мають бути меншими за модулем за одиницю.

ОЦІНЮВАННЯ ПАРАМЕТРІВ МОДЕЛІ

Для оцінювання параметрів МСВ застосуємо алгоритм Гіббса [8], який складається з таких кроків: 1 — ініціалізація h_0 та μ, ϕ, σ_υ^2; 2 — моделювання h_t з $h_t \mid h_{\setminus t}, y, \mu, \phi, \sigma_\upsilon^2$, $t = 1, \ldots, n$; 3 — моделювання $\sigma_\upsilon^2 \mid y, h, \phi, \mu$; 4 — моделювання $\phi \mid h, \mu, \sigma_\upsilon^2$; 5 — моделювання $\mu \mid h, \phi, \sigma_\upsilon^2$; 6 — перехід до 2-го кроку.

Однією ітерацією алгоритму є виконання пунктів 2–5. Моделювання потребує виконання декількох тисяч ітерацій алгоритму Гіббса S для генерування вибірки. На початку моделювання значення оцінок параметрів далекі від стаціонарних, тому оцінки параметрів, отримані на перших ітераці-

ях алгоритму, не потрібно враховувати. Позначимо через S_0 порядковий номер ітерації алгоритму, після якого починається формування векторів параметрів. Розмірність векторів є різницею кількості ітерацій алгоритму Гіббса та величини, після якої починається формування векторів параметрів: $n_{\text{iteration}} = S - S_0$. У результаті моделювання отримуємо вектори оцінок параметрів:

$$\overline{\mu} = (\mu_1, \mu_2, \ldots, \mu_{n_{\text{iteration}}}),$$

$$\overline{\phi} = (\phi_1, \phi_2, \ldots, \phi_{n_{\text{iteration}}}),$$

$$\overline{\sigma} = (\sigma_1, \sigma_2, \ldots, \sigma_{n_{\text{iteration}}}).$$

Значення, які додаються у вектори — це результат функціонування алгоритму Гіббса на кожній ітерації. Більш детально алгоритм описано в роботі [9], яку присвячено реалізації алгоритму Гіббса для оцінювання параметрів МСВ на мові програмування Java. Цей алгоритм можна застосувати для прогнозування майбутньої поведінки показника достовірності інформації. Особливістю описаного та реалізованого в роботі [9] алгоритму є те, що програму розроблено таким чином, що користувачу необхідно мати лише файл з вибіркою величин показника достовірності за певний період часу.

Для прогнозування поведінки метрики достовірності інформації також можна застосувати розроблені та програмно реалізовані алгоритми оцінювання параметрів МСВ методом «змішаного зсуву» на основі процедури Монте-Карло для марковських ланцюгів із використанням алгоритму Гіббса та оцінювання параметрів моделі УАРУГ. Описаний в роботі [10] метод оцінювання параметрів МСВ відрізняється від описаного вище процедурою оцінювання волатильності процесу. Оцінювання параметрів моделі УАРУГ реалізовано з використанням процедури моделювання «адаптивного відбраковування за Метрополісом», що описана у роботі [11].

ФУНКЦІЇ ПРОГНОЗУВАННЯ

Метою оцінювання параметрів моделі є прогнозування. Об’єктом прогнозування є волатильність (дисперсія) величини метрики достовірності інформації. Модель, що описує зміну значення волатильності у часі, дає можливість прогнозувати на один крок вперед. Для цього потрібно у відповідне рівняння підставити поточне значення волатильності та відповідні параметри. Для прогнозування на декілька кроків вперед слід побудувати функцію прогнозування. В роботі [12] описано процес отримання функції прогнозування на декілька кроків вперед на основі МСВ:

$$\hat{h}(t+k) = \mu(1-\varphi)(1+\varphi+\varphi^2+\ldots+\varphi^{t-1}) + \varphi^t h(k).$$

У роботі [13] представлено функцію прогнозування на декілька кроків вперед на основі моделі УАРУГ:

$$h(t+k) = \ln E_t[\sigma_{t+k}^2] = \sigma^2 + (\alpha_1+\beta_1)^{k-1}(\sigma_{t+1}^k - \sigma^2).$$

ТЕХНІЧНА РЕАЛІЗАЦІЯ

В роботі [14] зображено програмну систему, розроблену на описаних вище підходах, яку можна модифікувати та застосувати для прогнозування майбутньої поведінки метрики достовірності інформації. Розроблена система функціонує на основі валютного курсу, значення якого зберігаються у базі даних. Модифікація полягає у необхідності зберігати у базі даних значень показника достовірності інформації.

Етапи проектування системи та технології, що використані для її реалізації, детально описані в [14]. Розглянуті всі модулі та структура системи, наведені концептуальна та фізична модель бази даних, описано інтерфейс користувача та надано інструкції з користування системою.

ВИСНОВКИ

В роботі запропоновано підхід по прогнозуванню волатильності (дисперсії) величини показника достовірності інформації на основі авторегресійних моделей, зокрема моделі стохастичної волатильності. Поведінка метрик повноти та своєчасності інформації є передбачуваною, а поведінка метрики достовірності інформації носить невизначений характер. Запропоновано розроблену для прогнозирования волатильності валютного курсу інформаційну технологію застосувати для прогнозування на короткий термін майбутньої поведінки міри невизначеності показника достовірності інформації, важливого під час прийняття рішень.

Запропонований підхід відкриває подальший напрям дослідження, який полягає у визначенні оптимального часового інтервалу, на основі якого необхідно формувати значення показника достовірності інформації.

ЛІТЕРАТУРА

1. *Ballou D. P., Wang R. Y., Pazer H. & Tayi G. K.* Modeling information manufacturing systems to determine information product quality // Management Science. — 1998. — № 44 (4). — P. 462–484.
2. *Lee Y.W., Strong D.M., Kahn B.K. & Wang R.Y.* AIMQ: A methodology for information quality assessment // Information and Management. — 2002. — № 40(2). — P. 133–146.
3. *Catarci T., Scannapieco M.* Data Quality under the Computer Science Perspective // Archivi&Computer. — 2002. — № 2. — P. 1–12.
4. *Wang R.Y., Storey V.C. and Firth C.P.* A Framework for Analysis of Data Quality Research // IEEE Transaction on Knowledge and Data Engineering. — 1995. — **7**, № 4. — P. 623–640.
5. *Згуровський М.З., Панкратова Н.Д.* Основи системного аналізу. — К.: Видавнича група BHV, 2007. — 544 с.
6. *Batini C. and Scannapieco M.* Data Quality Springer-Verlag Concepts, Methodologies and Techniques. Data-Centric Systems and Applications. — Berlin: Springer-Verlag. — 2006. — P. 19–30.
7. *Taylor S.J.* Modeling Financial Time Series. — John Wiley, Chichester, 1986. — 268 p.

8. *Kim S., Shephard N., Chib S.* Stochastic volatility: likelihood inference and comparison with ARCH models // Review of Economic Studies. — 1998. — **65**. — P. 361–393.
9. *Бідюк П.І., Коновалюк М.М.* Оцінювання параметрів моделі стохастичної волатильності з використанням алгоритму Гіббса // Регіональний міжвузівський збірник наукових праць «Системні технології». — Дніпропетровськ, 2011. — Вип. 6 (77) — С. 12–27.
10. *Бідюк П.І., Коновалюк М.М.* Оцінювання моделей стохастичної волатильності та УАРУГ на Java, Наукові праці. Комп'ютерні технології. — Миколаїв: ЧДУ ім. Петра Могили, 2012. — Вип. т. 191, 179. — С. 14–20.
11. *Gilks W.R., Best N.G., Tan K.K.* Adaptive rejection metropolis sampling within Gibbs sampling // Applied Statistics. — № 44. — P. 455–473.
12. *Бідюк П.І., Коновалюк М.М.* Прогнозування волатильності валютного ринку за нелінійними моделями. Вісник нац. унів. «Львівська політехніка». — 2013. — № 751. — С. 257–265.
13. *Engle R.F., Bollerslev T.* Modeling the persistence of conditional variance. // Econometric Reviews. — 1986. — **5**. — P. 1–50.
14. *Бідюк П.І., Коновалюк М.М.* Інформаційна система для прогнозування волатильності валютних курсів // Науково-теоретичний журнал «Искусственный интеллект». — 2012. — № 4. — С. 292–302.

Надійшла 30.12.2013