ruleDQ: Ein Regelsystem zur Datenqualitätsverbesserung medizinischer Informationssysteme

Mihael Gorupec und Gregor Endler
Lehrstuhl für Informatik 6 (Datenmanagement)
Friedrich-Alexander-Universität Erlangen-Nürnberg
m.gorupec@gmail.com

Abstract: ruleDQ realisiert ein Regelsystem zur automatisierten Messung und Verbesserung von Datenqualität. Anwendern ohne informatisches Fachwissen wird es ermöglicht, eigenständig Datenqualitätsregeln aufzustellen. Die Regeln werden von ruleDQ regelmäßig ausgewertet und erlauben eine Quantifizierung und Analyse von Datenqualitätsmängeln. ruleDQ folgt dabei den Prinzipien eines kontinuierlichen Datenqualitätsmanagements um nachhaltige Verbesserungen zu ermöglichen.

1 Einführung und Motivation

Im Gesundheitswesen besteht der gegenwärtige Trend des Zusammenschließens mehrerer individueller Praxen zu medizinischen Versorgungszentren [HAE09]. Dies hat die Erhöhung der Konkurrenzfähigkeit zum Ziel, die Versorgungszentren können durch die Zusammenlegung von Ressourcen unter anderem Kosten einsparen und Kunden einen vielfältigeren Service anbieten [EBL13].

Diese Zusammenschlüsse führen zu der Notwendigkeit einer zentralen administrativen Instanz, den Praxismanagern. Praxismanager sind für die Ressourcenplanung und Unternehmenssteuerung verantwortlich. Außer an das Personal, ergeben sich auch neue Herausforderungen an die informationstechnologische Infrastruktur. Praxismanager benötigen für ihre Aufgaben einen einheitlichen Blick auf alle Daten der einzelnen Verbundspartner. Durch die Zusammenführung von heterogenen Daten aus unterschiedlichen Quellen ergeben sich aber Herausforderungen bei der Einhaltung von gemeinsamen Datenqualitätsstandards.

Erschwerend kommt hinzu, dass es in der informationstechnologischen Landschaft von medizinischen Versorgungszentren zu ständigen Änderungen kommt. Durch die Hinzunahme von neuen Mitgliedern, durch Änderungen von Budgets und Verträgen oder durch neue Gesetzgebung müssen Datenqualitätsanforderungen ständig neu evaluiert und angepasst werden [End12].

2 Anforderungen und Ziele

Das Ziel dieser Arbeit war es, ein Konzept zur Messung und Verbesserung von Datenqualität zu schaffen und mit Hilfe eines Regelsystems zu realisieren.

Die späteren Anwender sind Domänenexperten, die in der Lage sind die vorhandenen Daten zu interpretieren und Regeln zu erkennen. Allerdings mangelt es an informatischem Fachwissen, um die gefundenen Regeln etwa in Form von SQL umzusetzen. Die Lösung muss es Anwendern daher erlauben Regeln auf möglichst einfache und leicht verständliche Weise zu formulieren.

Zur Umsetzung dieser Anforderungen wurde die Verwendung von regelbasierten Systemen als geeignet gesehen. Regelbasierte Systeme sind Applikationen, die Problemlösungs-Know-How automatisieren, indem sie Expertenwissen mit Hilfe von Regeln abbilden. Regeln bestehen in regelbasierten Systemen aus Wenn-Dann-Sätzen. Dies beruht darauf, dass menschliche Experten ihre Problemlösungs-Techniken für gewöhnlich in Form von Situations-Aktions-Regeln ausdrücken. Der Wenn-Teil einer Regel besteht aus einer Bedingung, welche definiert wann die Regel ausgelöst wird. Der Dann-Teil beschreibt die Aktion, welche ausgelöst wird wenn die Bedingung der Regel erfüllt ist [HR85].

Die geschaffene Lösung musste zudem den Prinzipien eines kontinuierlichen Datenqualitätsmanagements folgen, welches für eine nachhaltige und effektive Datenqualitätssteigerung nötig ist. Punktuelle Datenreinigungen haben nur einen kurzfristigen Effekt, die dadurch erzielten Verbesserungen gehen gerade bei sich häufig ändernden Daten schnell verloren [Red97]. Das Thema Datenqualität darf daher nicht als eine einmalige Aktion betrachtet werden. Um Datenqualität effektiv zu verbessern, bedarf es ganzheitlicher Methoden, die Daten über ihren gesamten Lebenszyklus hinweg betrachten um ein definiertes Niveau an Qualität zu garantieren [BCFM09].

3 Ergebnis

Eine im Zuge der Arbeit durchgeführte Analyse von Referenzunternehmen zeigte, dass keine vollständig automatisierten Methoden zur Messung von Datenqualität oder Behandlung von Mängeln in den Versorgungszentren existierten. Zudem gaben Verantwortliche an, häufig unter verschiedenen Datenqualitätsproblemen zu leiden. Daher wurde im Rahmen dieser Arbeit prototypisch eine Applikation zur automatisierten Messung und Verbesserung von Datenqualität entworfen und implementiert (*ruleDQ*). Die Applikation erlaubt es den Anwendern eigenständig einfache boolesche Regeln für ihre Daten zu formulieren. Die Regeln werden von der Applikation anschließend in SQL übersetzt und kontinuierlich ausgewertet. Als Ergebnis dieser Evaluierung von Regeln stehen quantitative Messwerte, welche Rückschlüsse über das Niveau von Datenqualität ziehen lassen. Die Applikation identifiziert regelverletzende Datensätze und erlaubt dem Nutzer so die Analyse der aufgetretenen Mängel.

ruleDQ wurde dabei nach dem Vorbild eines regelbasierten Systems konzipiert. Regelbasierte Systeme fordern die Trennung von Regeln, Daten und Prozessen. Geschäftsregeln

werden unabhängig vom Ausführungscode formuliert und verwaltet. Nach [Los02] besteht ein regelbasiertes System aus folgenden drei Modulen:

1. Einem Benutzerinterface zur Erstellung und Verwaltung von Regeln
2. Einer persistenten Regelbasis
3. Einem Modul zur Ausführung von Regeln

ruleDQ folgt diesem Aufbau und bietet zusätzlich dazu ein Modul zur Analyse der Ergebnisse der Regelauswertung. Das theoretische Konzept hinter ruleDQ lehnt sich an das von Wang geschaffene Total Data Quality Management (TDQM) an [WZL01]. TDQM basiert auf der Betrachtung von Daten ähnlich zu Produkten in der Fertigungsindustrie. Die in diesem Bereich weitverbreitete Qualitätssicherungs-Methodik des Demingkreises [Dem86] wurde dabei auf Daten übertragen. Der Demingkreis beschreibt eine iterative Problemlösungsmethodik die aus vier Phasen besteht: Planen, Umsetzen, Überprüfen, Handeln. TDQM definiert analog zum Demingkreis vier Phasen welche kontinuierlich durchlaufen werden müssen, um eine nachhaltige Datenqualitätsverbesserung zu erreichen. Die Phasen des TDQM gliedern sich in Definition, Messung, Analyse und Verbesserung.

In der Definitions-Phase müssen relevante Datenqualitätsdimensionen ausgewählt und entsprechende Anforderungen an diese definiert werden. In ruleDQ geschieht dies in Form von Geschäftsregeln. Will ein Anwender zum Beispiel sicherstellen, dass bestimmte Medikamentengruppen bei einer Diagnose nicht verschrieben werden, so kann er mit Hilfe von ruleDQ zu diesem Zweck eine Regel anlegen, etwa in der Form `(Diagnose = 'R50.80') AND (Medikation != 'Placebo')`.

Nach der Formulierung der Anforderungen folgt deren Anwendung auf die Daten. Es wird untersucht inwiefern die Daten den Anforderungen genügen. Dazu dienen im Falle von ruleDQ quantitative Metriken. Die vorher formulierten Regeln werden in ein SQL-Statement geparst und auf Daten in einer relationalen Datenbank angewandt. Ein mögliches Ergebnis beim obigen Beispiel wäre etwa, dass 120 von 2400 Tupeln in der Datenbank die Regel verletzten und somit bei 5% der Diagnosen unerwünschte Medikamentengruppen verschrieben werden.

In der Analyse-Phase werden Ursachen mangelnder Datenqualität festgestellt. Die Ursachen können vielfältig sein, unter anderem können fehlerhaften Anwendungen, menschliche Fehler oder schlecht gestaltete Prozesse verantwortlich sein. Zur Unterstützung in dieser Phase können mit ruleDQ regelverletzende Datensätze analysiert werden. Der Anwender kann also prüfen in welchen Fällen die Regel verletzt wurde und in obenstehendem Beispiel welche Ärzte die unerwünschten Medikamente benutzt haben.

In der Verbesserungs-Phase sollen die Fehler und deren Ursachen nachhaltig beseitigt werden. Dazu gilt es permanent qualitätssichernde Maßnahmen zu implementieren, wie etwa durch Prozess-Redesign. Einmalige Datensäuberungsmaßnahmen können davor initial zur Anwendung kommen. ruleDQ bietet für diese Phase die Versendung von Warnungen bei einem Absinken unter ein spezifiziertes Datenqualitätsniveau an. So kann automatisiert eine E-Mail als Warnung verschickt werden, wenn z.B. bei mehr als 1% der Diagnosen die

unerwünschte Medikamentengruppe benutzt wird. Nach Abschluss aller Phasen erfolgen stets eine neue Iteration und die erneute Definition von Anforderungen [BCFM09].

4 Fazit

Der Kontext dieser Arbeit liegt in der medizinischen Domäne. Da ruleDQ jedoch weder Annahmen über das Schema der Daten noch über deren Semantik trifft, kann es auch in anderen Domänen eingesetzt werden. ruleDQ bietet den Anwendern in allen Phasen des TDQM Unterstützung und trägt so zu einem kontinuierlichen Datenqualitätsmanagement bei. Durch die fortlaufende Überwachung wird ein erneutes unbemerktes Absinken der Qualität verhindert und eine nachhaltige Verbesserung der Datenqualität ermöglicht. Zukünftige Verbesserungen sind z.B. mit der Erweiterung von ruleDQ um Kontextsensitivität zu erreichen, um etwa die möglichen Optionen zur Regelformulierung in Abhängigkeit des Datentyps eines ausgewählten Feldes zu beschränken oder zu erweitern.

Literatur

[BCFM09] Carlo Batini, Cinzia Cappiello, Chiara Francalanci und Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3):16:1–16:52, Juli 2009.

[Dem86] W Edwards Deming. Out of the crisis. Cambridge, MA: Massachusetts Institute of Technology. *Center for Advanced Engineering Study*, Seite 6, 1986.

[EBL13] Gregor Endler, Philipp Baumgärtel und Richard Lenz. Pay-as-you-go data quality improvement for medical centers. In E. Ammenwerth, A. Hörbst, D. Hayn und G. Schreier, Hrsg., *Proceedings of the eHealth2013*, Seiten 13–18, 2013.

[End12] Gregor Endler. Data quality and integration in collaborative environments. In SIGMOD/PODS und ACM, Hrsg., *SIGMOD/PODS 2012 PhD Symposium*, New York, NY, USA, 2012.

[HAE09] W. Hellmann, T. Antwerpes und S. Eble. *Gesundheitsnetzwerke managen: Kooperationen erfolgreich steuern*. MWV Medizinisch Wiss. Verlag-Ges., 2009.

[HR85] Frederick Hayes-Roth. Rule-based systems. *Commun. ACM*, 28(9):921–932, September 1985.

[Los02] David Loshin. Rule-based data quality. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, Seiten 614–616, New York, NY, USA, 2002. ACM.

[Red97] Thomas C. Redman. *Data Quality for the Information Age*. Artech House, Inc., Norwood, MA, USA, 1st. Auflage, 1997.

[WZL01] Richard Y. Wang, Mostapha Ziad und Yang W. Lee. *Data Quality*, Jgg. 23 of *Advances in Database Systems*. Kluwer, 2001.