Flexible Datenintegration und Migration in ein Sensornetzwerk – ein Beispiel mit Twitter-Analysedaten

Nikolai Bock und Klaus Böhm

i3mainz – Institut für Raumbezogene Informations- und Messtechnik, Mainz · nikolai.bock@hs-mainz.de

Short paper

Zusammenfassung

Nicht nur auf Basis von Konzepten wie „Internet of Things“ erlebt der Bereich der Sensornetzwerke ein großes Wachstum. Zunehmender Bedarf entsteht auch durch immer mehr alternative Messkonzepte und Datenquellen. Hieraus ergeben sich die unterschiedlichsten Problemstellungen, u. a. bei der Integration verschiedenartiger Daten. Dieser Beitrag soll ein technisches Konzept aufzeigen, welches verschiedene Anforderungen berücksichtigt. Eine prototypische Umsetzung dieses Konzepts liefert als Ergebnis die Integration von Twitter-Analyse-Daten in einen Sensor Observation Service (SOS).

1 Einleitung

Neben den klassischen in situ Sensormesssystemen entstehen immer mehr alternative Messkonzepte. Hierzu zählen auch Konzepte, welche Daten, bzw. Informationen sammeln, die von Nutzern stationär und vor allem mobil erfasst werden. Diese Sensorsysteme lassen sich nach RESCH (2013a) in drei Konzepte klassifizieren:

1. People as Sensors,
2. Collective Sensing,
3. Citizen Science.

Bei „People as Sensors“ handelt sich um ein Messmodell, in dem Individualpersonen neben physikalischen Messungen, z. B. durch Fitnessarmbänder, auch subjektive „Messwerte“, wie Sinneseindrücke, Empfindungen oder persönliche Beobachtungen sammeln (RESCH et al. 2011). In der Literatur findet man im gleichen Kontext auch die Begriffe „Humans as Sensors“ (FORREST 2010) oder „Citizen as Sensors“ (GOODCHILD 2007). „Collective Sensing“ ist eine Methode, in der Analysen vielfältiger, anonymer Daten genutzt werden. Diese können u. a. auch aus „Social Media“-Quellen, wie Twitter oder Flickr, stammen (siehe RESCH 2013b). Die Nutzung von Mobilfunkdaten, um z. B. Bewegungsmuster zu analysieren, zeigen SAGL et al. (2014). Ein weiteres Beispiel zeigt die Kombination sozialer Medien (Flickr) mit Mobilfunkdaten (siehe SAGL et al. 2012). Im Gegensatz hierzu ist „Citizen Science“ ein Konzept, bei dem Expertenwissen von Bürgern zur Verfügung gestellt wird.

AGIT – Journal für Angewandte Geoinformatik, 1-2015, S. 448-453. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-557-7, ISSN 2364-9283, eISSN 2509-713X, doi:10.14627/537557062.

Ein Ziel für umfassende Analysen ist es, Messungen aus den unterschiedlichen Messkonzepten zu verknüpfen, um sie im Optimalfall annähernd in „Echtzeit“ analysieren zu können. Mit dem Sensor Web Enablement (SWE) bietet das Open Geospatial Consortium (OGC) unterschiedliche Standards für Sensornetzwerke. Aktuelle Forschungsarbeiten im Bereich „Sensor Web“ adressieren mehrere Gebiete. Aus der hohen Heterogenität der Messinformation entsteht der Bedarf an einem „Semantic Sensor Web“ (CORCHO & GARCÍA-CASTRO 2010). Einen Ansatz zur Nutzung von „Social Media“-Quellen zeigen SCHADE et al. 2012. Ein weiterer Bereich ist die Bereitstellung der Messinformation in einer entsprechenden Sensornetzwerkstruktur. Aus den verschiedenen Anforderungen an die Aufbereitung oder aus der Datenkomplexität ergibt sich der Bedarf eines flexiblen Prozesssystems.

2 Anforderungen

Dieser Bedarf resultiert in folgenden technischen Anforderungen:

Einfache Erweiterung (A1): Flexible Erweiterung unterschiedlicher Datenquellen, sowie Verarbeitungsprozesse. Die Erweiterung von Prozessen spielt vor allem bei der Verarbeitung komplexerer Eingangsdaten eine Rolle. So können mehrere Datenoptimierungsschritte oder das Qualitätsmanagement hohe Anforderungen besitzen.

Hohe Flexibilität, bzw. Skalierbarkeit (A2): Dies bedeutet, dass alle Prozessschritte stark voneinander gekoppelt und theoretisch auch räumlich voneinander getrennt ablaufen können sollen. Dies bekommt vor allem bei der Verarbeitung größerer Datenmengen („Big Data“) eine sehr hohe Gewichtung.

Konfigurierbarkeit (A3): Die Prozessketten sollen einfach definiert und ausgeführt werden können. Zudem soll die Steuerung der Prozessschritte über Parameter möglich sein.

Performance (A4): Größere Datenmengen können eine höhere Parallelisierbarkeit erfordern. Gegebenenfalls müssen auch Prozessschritte mit höheren Ressourcenanforderungen separat auf potenterer Hardware durchgeführt werden.

Diese Anforderungen werden durch das nachfolgend beschriebene Konzept realisiert.

3 Konzept

Um die Anforderungen abzudecken, wurde ein Konzept auf Basis von Spring XD (http://projects.spring.io/spring-xd) entwickelt. Spring XD ist ein aktuelles System der Spring Familie, worüber die Entwicklung von Big Data Applikationen erleichtert werden soll. Durch die Adaption von Spring Integration setzt es hierbei auch die „Enterprise Integration Patterns“ (siehe HOHPE & WOOLF 2003) um. Hier wird mit Adaptern, die über „Messaging-Systeme“ kommunizieren, gearbeitet. Somit lassen sich die einzelnen Komponenten sehr stark entkoppeln, was eine hohe Flexibilität, bzw. Skalierbarkeit erlaubt (A2). Zur Verarbeitung der Daten werden sogenannte Datenströme („Streams“) definiert, welche mindestens aus einer Datenquelle und einem Datenspeicher bestehen. Zwischen ihnen können mehrere Prozesse geschaltet werden. Weiterhin ist es möglich, sich in bestehende „Streams“ mit weiteren „Streams“ einzuhängen und so mehrere Prozessketten auf demselben Datenstrom durchzuführen.

Das entwickelte Konzept basiert, wie in Abbildung 1 zu sehen ist, auf Modulen, welche für die einzelnen Prozessschritte bereitgestellt und flexibel organisiert werden können. Es gibt drei Modultypen. Ein „Stream“ beginnt immer mit einer Datenquelle („Source“-Modul), die Weiterverarbeitung geschieht durch „Processor“-Module und das Ergebnis wird zum Schluss mithilfe eines „Sink“-Moduls abgelegt.

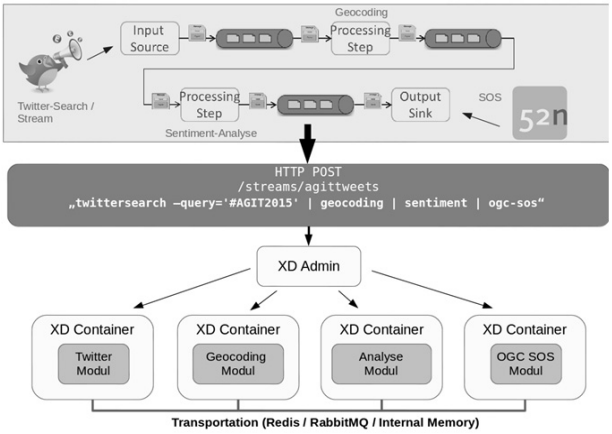

Abb. 1: Systemkonzept mit Beispielprozesskette

Durch die Entwicklung weiterer Module ist es möglich, das System beliebig um Datenquellen, Prozesse und Datenspeicher zu erweitern (A1). So lassen sich beispielsweise unterschiedliche Messinformationen bereitstellen. Zudem kann mit weiteren Prozessschritten ein komplexer Integrationsprozess abgebildet werden. Hierzu können neben den für das Beispiel entwickelten Prozessen, auch solche zur Optimierung der Sensordatenqualität zählen.

Stehen alle benötigten Module dem System zur Verfügung, lässt sich über eine REST-API des „Admin-Servers“ ein „Stream“, wie in Abbildung 1 zu sehen ist, definieren und ausführen. Diese Definition ist dabei einem Unix-Shell-Kommando ähnlich. Die „Pipes“ zwischen den Modulen repräsentieren die Kanäle. Je nach Konfiguration bestehen diese Kanäle aus Messaging-Systemen, wie Redis (http://redis.io) oder RabbitMQ (http://rabbitmq.com), oder stehen im internen Speicher bereit. Dies erlaubt es, sowohl Prozessschritte flexibel zu konfigurieren (A3), als auch so zu verteilen, dass Datenströme, wenn möglich, parallel verarbeitet werden. Zudem können Prozessschritte mit hohen Anforderungen auf entsprechender Hardware ausgeführt werden (A4).

4 Prototyp

Zur Überprüfung des Konzeptes wurde ein Prototyp entwickelt, welcher Twitter-Daten analysiert und die Ergebnisse als Sensordaten bereitstellt. Abbildung 1 zeigt diesen konkreten Prozessablauf und die benötigten Module. Diese Module sind als *Quelle* Twitter (**Source**), als *Verarbeitungsschritte* ein Geocoding (**Processor**) und die Sentiment Analyse (**Processor**) und als *Speicher* der Sensor Observation Service (**Sink**).

Ein Modul setzt sich, wie in Abbildung 2 zu sehen ist, aus einer Konfigurationsdatei, einer Properties-Datei und Java-Klassen, bzw. -Bibliotheken zusammen. Ergänzt wird dies durch

ein „Input Gateway“ (bei Sink), ein „Output Gateway“ (bei Source) oder beides (bei Processor) über welche die „Message“ empfangen bzw. gesendet wird. Eine „Message“ besteht aus einem „Header“ und dem „Payload“. Während der „Header“ die Metainformationen über die „Message“ bereithält, wird im „Payload“ der Inhalt abgelegt. Die Konfigurationsdatei sorgt mithilfe der Beschreibung des Ablaufs für die zentrale Steuerung des Moduls. In den „Properties“ können die flexiblen Parameter beschrieben werden. Für die eigentliche Verarbeitung werden Adapter genutzt, welche durch die Java-Klassen bereitgestellt werden.

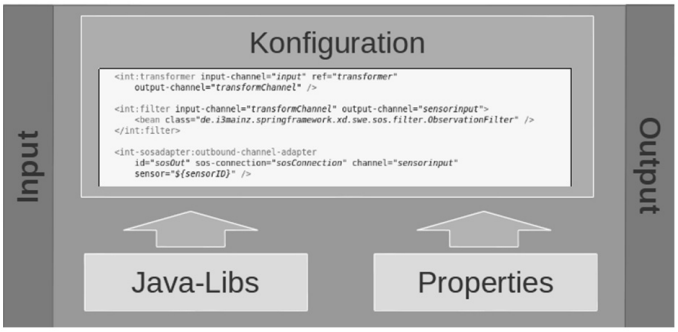

Abb. 2:
Aufbau eines Moduls

Spezielle Module für die Anbindung von Twitter, welche die Streaming-, bzw. Search-API nutzen, stehen mittlerweile in der Spring XD Basiskonfiguration bereit. Für den Prototyp wurden folgende Module neu entwickelt:

Geocoding: Dieses Modul erhält als „Input“ ein „JSON“-Objekt (hier ein Tweet), sowie die Information, über welche Attribute primäre, bzw. sekundäre räumliche Informationen gewonnen werden können. Als „Output“ produziert es ein „GeoJSON“-Objekt, welches an die bestehende „Message“ gehängt wird. Werden primäre Informationen gefunden, wird der Geocoding-Prozess übergangen und direkt das „GeoJSON“-Objekt erzeugt. Für das Geocoding von sekundären Informationen wird der OSM-Dienst „Nominatim“ (http://nominatim.openstreetmap.org) verwendet.

Sentiment-Analyse: Dieses Modul erhält als „Input“ einen Text (hier der Tweet). Als „Output“ wird das Analyseergebnis angehängt. Die Sentimentanalyse ist eine Variante des „Natural Language Processing“ (NLP) und erkennt die Stimmung bzw. eine positive oder negative Haltung in einem Text. Für die erste prototypische Umsetzung greift das Modul auf die Funktionen des „AlchemyAPI“-Dienstes (http://www.alchemyapi.com) zurück. Dieser ermittelt zunächst die Sprache des Textes und berechnet abschließend einen „Score“ zwischen -1 (negativ) und 1 (positiv).

OGC-SOS: Dieses Modul erhält als „Input“ ein Objekt mit räumlichen und zeitlichen Information (hier Tweet mit Analysedaten und Koordinaten). Es erzeugt aus den Daten eine „FeatureOfInterest“ (FOI) und eine „Observation“ und fügt diese mithilfe eines Adapters, welcher die transaktionalen Operationen der SOS-Richtlinie umsetzt, dem angegebenen SOS hinzu. Um eine flexiblere Erweiterung und Nutzung zu ermöglichen, greift der Adapter hierfür auf ein entwickeltes „Template“ zurück, welches die notwendigen Operationen abdeckt. Zudem kann das Modul optional bei der Initialisierung den Sensor im SOS erzeugen, wenn dieser noch nicht existiert.

5 Visualisierung

Zur Visualisierung wurden für die Portalsoftware Liferay (http://www.liferay.com) Portlets entwickelt, welche eine raumzeitliche Navigation durch Daten ermöglichen. Abbildung 3 zeigt das Karten-Portlet mit einer einfachen symbolischen Darstellung der Tweets. Zur zeitlichen Navigation dient der „Time-Slider“, welcher in Abbildung 4 zu sehen ist.

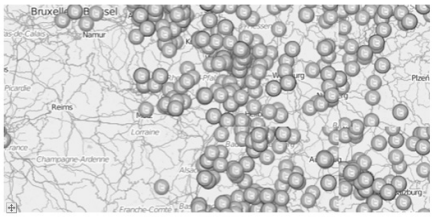

Abb. 3:
Kartenanwendung mit Tweets

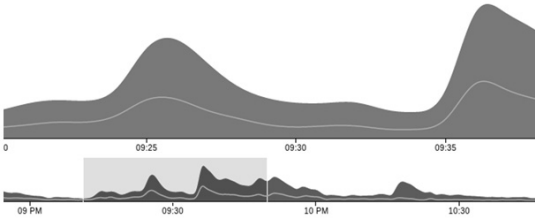

Abb. 4:
Liniendiagramm inkl. zeitlicher Navigation

Der „Time-Slider“ besteht aus zwei Diagrammen, wobei im oberen Diagramm der im unteren Teil selektierte Zeitausschnitt herangezoomt wird. Die zwei verschiedenen Linien zeigen den gesamten räumlichen Bereich (Fläche), sowie nur den räumlichen Ausschnitt, welcher in der Karte zu sehen ist, folgend dem Prinzip des „Linking and Brushing“ (siehe KEIM 2002).

6 Fazit und Ausblick

Der Beitrag zeigt ein flexibles Konzept zur Integration von Daten in ein Sensornetzwerk. Die derzeitig entwickelten Module ermöglichen das Speichern von Sentiment-Analyse-Daten basierend auf Tweets aus der Twitter Search- bzw. Streaming-API und zeigen erfolgreich die Umsetzbarkeit des Konzeptes.

Aufbauend auf dem gezeigten prototypischen Ansatz sind Weiterentwicklungen in zwei Schwerpunkten geplant. Zum einen sollen weitere Datenquellen aus den unterschiedlichen Messkonzepten erschlossen werden. Im Fokus stehen hierbei Sensorinformationen im Kontext von persönlicher Belastung, wie Feinstaub, Pollen, UV oder Ozon mit ihren Wechselwirkungen. Die heterogenen Daten sollen in ein Sensornetzwerk integriert werden, um u. a. für die Berechnung von persönlichen Belastungsmustern inkl. medizinischer Handlungsempfehlungen genutzt zu werden.

Zum anderen soll vor allem an den bisherigen Defiziten beim Datenqualitätsmanagement bei der Integration komplexerer Datenquellen gearbeitet werden. So ergeben sich mit den derzeit bereitstehenden Modulen Einschränkungen. Beim Geocoding hängt z. B. die Qualität von der Genauigkeit des automatisierten Geocoding-Prozesses ab. Des Weiteren kann die Auswahl eines einfachen Filters gegebenenfalls nicht ausreichend für die Selektion der für den Kontext benötigten „Tweets“ sein. Einige Problemlösungen hierfür zeigt das Pro-

jekt „Twitris“ (PUROHIT et al. 2013; SHETH et al. 2013). Im Zuge der weiteren Entwicklung sollen vor allem Konzepte mit kontextbezogenen Inhalten evaluiert und umgesetzt werden.

Literatur

CORCHO, O. & GARCÍA-CASTRO, R. (2010), Five challenges for the Semantic Sensor Web. Semantic Web-Interoperability, Usability, Applicability, 2010.

FORREST, B. (2010), Humans As Sensors. LbxJournal. http://www.lbxjournal.com/articles/humans-sensors/260057 (18.04.2015).

GOODCHILD, M. F. (2007), Citizen as sensors: the world of volunteered geography. GeoJournal, 69(4).

HOHPE, G. & WOOLF, B. (2003), Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

KEIM, D. (2002), Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and computer graphics, 100-107.

PUROHIT, H., HAMPTON, A., BHATT, S., SHALIN, V., SHETH, A. & FLACH, J. (2013), An Information Filtering and Management Model for Twitter Traffic to Assist Crises Response Coordination. Technical Report, Kno.e.sis Center, 2013.

RESCH, B., MITTLBÖCK, M., KRANZER, S., SAGL, G., HEISTRACHER, T. & BLASCHKE, T. (2011), „People as Sensors“ mittels personaliserten Geo-Trackings. In: STROBL, J. et al. (Hrsg.), Angewandte Geoinformatik 2011. Wichmann Verlag, Berlin/Offenbach.

RESCH, B. (2013a), Live Geography – Kombination von menschlichen und technischen Sensoren im Bereich der Geomedizin. Fachaustausch Geoinformation 11/2013, http://www.geonet-mrn.de/fileadmin/user_upload/Veranstaltungen/Fachaustausch_Geoinformation_2013/FFIII_03_Resch_.pdf (18.04.2015).

RESCH, B. (2013b), People as Sensors and Collective Sensing-Contextual Observations Completing Geo-Sensor Network Measurements. Progress in Location-Based Services – Lecture Notes in Geoinformation and Cartography, 2013, 391-406.

SAGL, G., RESCH, B., HAWELKA, B. & BEINAT, E. (2012), From Social Sensor Data to Collective Human Behavior Patterns – Analysing and Visualising Spatio-Temporal Dynamics in Urban Environments. In: Jekel, T. et al. (Eds.), GI_Forum 2012: Geovisualization, Society and Learning, 2012, Wichmann Verlag, Berlin/Offenbach, 54-63.

SAGL, G., DELMELLE, E. & DELMELLE, E. (2014), Mapping collective human activity in an urban environment based on mobile phone data. Cartography and Geographic Information Science, 41 (3).

SCHADE, S., OSTERMANN, F., SPINSANTI L. & KUHN W. (2012), Semantic Observation Integration. Future Internet, 4 (3), 807-829.

SHETH, A., JADHAV, A., KAPANIPATHI, P., LU, C., PUROHIT, H., SMITH, G., WANG, W. (2013), Twitris – a System for Collective Social Intelligence. Encyclopedia of Social Network Analysis and Mining.