Exposé

Similarity-based query answering for medical information systems

Bachelor-Thesis

Jero Mario Schäfer
Matrikelnr.: 21552103
Angewandte Informatik

Georg-August-Universität Göttingen

Abgabedatum: 01.04.19

1 Einleitung

Medizinische Informationssysteme dienen der Speicherung, Auswertung und Verarbeitung patientenbezogener Daten. So könnte zum Beispiel eine einfache Datenbank wie folgt aussehen. In der Tabelle 1 sind die persönlichen Informationen der Patienten wie z.B. Name und Alter gespeichert. Die Patienten besitzen zudem einen eindeutigen Identifikator, der zufällig bestimmt werden kann. Des Weiteren gibt es zusätzlich zu den persönlichen Daten noch die Tabelle 2, die sowohl Auskunft über die Erkrankungen geben, an denen die zu behandelnden Patienten leiden, als auch bspw. Daten über die durchgeführten bzw. durchzuführenden Behandlungen enthalten könnte:

ID	Name	Address	Age
1951	John Miller	New street 5, 44579 Newtown	28
4431	Mary Smith	Green way 15, 23240 Abcity	47
0056	Taylor Banks	On the line 90, 11005 Nington	71

Tabelle 1: Patienten Informationen

ID	PatientID	Disease	Treatment
92840	1951	Cough	Antitussive
12748	4431	Influenza	Antiviral Medication
12091	0056	Bronchitis	Bronchodilator
75634	1951	Asthma	Inhalator

Tabelle 2: Erkrankungen

Dabei referenziert ein Tupel der Tabelle über Erkrankungen jeweils einen Patienten mittels seines Identifikators und stellt somit den Zusammenhang zwischen den beiden Relationen her. Die Primärschlüssel sind in den Tabellenabbildungen durch Unterstreichungen und die Fremdschlüssel durch Überstreichungen gekennzeichnet.

Ein solches Informationssystem ist zum Einen nützlich und sogar notwendig für die behandelnden Ärzte und Krankenpfleger innerhalb eines Krankenhauses, aber zum Anderen auch wichtig für die administrativen Aufgaben, die in einem Krankenhaus anfallen. Außerdem könnten auch Hausärzte oder Fachärzte außerhalb eines Krankenhauses durch Zugriff auf die Daten über die Erkrankungen ihrer Patienten profitieren. Da die anfallenden Datenmengen mitunter sehr groß sein können, bietet es sich an nicht nur eine einzelne Datenbank zu haben sondern ein verteiltes Datenbanksystem zu nutzen, das die großen Datenmengen durch die physikalische Aufteilung der Daten auf verschiedene Server, welche alle eine Datenbank bzw. einen Teil der gesamten, verteilten Datenbank beherbergen und verwalten, handhaben kann. Um die Daten hierbei sinnvoll aufzuteilen, kann man bspw. eine horizontale Fragmentierung, also eine zeilenweise Aufteilung der Tabellen auf

verschiedene Server, anwenden. Die dadurch resultierende Partitionierung der Relationen stellt eine Unterteilung der Daten in kleinere Teilmengen dar, die letztendlich über das Netzwerk verteilt werden können.

Ziel dieser Arbeit ist es, anhand des Beispiels eines medizinischen Informationssystems ein verteiltes Datenbanksystem zu implementieren, welches aufgrund eines Maßes für Ähnlichkeit von Krankheiten ein *Clustering* der Erkrankungsrelation bestimmt, um darauf aufbauend eine horizontale Fragmentierung der Daten zu erhalten und infolgedessen eine abgeleitete Fragmentierung der persönlichen Daten der Patienten zu ermitteln, die bewirkt, dass die Tupel, welche Namen, Adresse, Alter, etc. enthalten, zusammen mit den Informationen über die Erkrankungen eines Patienten abgespeichert werden. Dies ermöglicht eine intelligente Anfragebearbeitung, bei der Informationen aus den beiden Relationen mittels eines Joins lokal verknüpft werden können, da die Tupel nicht erst über das Netzwerk von einem zum nächsten Datenbankserver gesendet werden müssen, wie es der Fall sein könnte, wenn die Daten nicht zusammen abgelegt werden. Des Weiteren soll dieses System im Falle einer Anfrage mit einer Selektionsbedingung auf dem Krankheitsattribut der Tabelle 2, sofern sie kein exaktes Ergebnis liefern kann, dem Nutzer ermöglichen, die Anfrage *flexibel* mit Hilfe der Ähnlichkeit zwischen Erkrankungen zu beantworten. Dabei wird die Anfrage des Nutzers generalisiert, wodurch ggf. Antworten von der Datenbank zurückgeliefert werden können, sodass der Nutzer zu der ursprünglich gestellten Anfrage ähnliche Antworten erhalten kann, die aus seiner Sicht nützliche und wertvolle Informationen darstellen könnten.

Die Implementation soll auf Basis einer Apache Ignite™ Datenbank (siehe `https://ignite.apache.org/`) erfolgen. Die beispielhaften Patientendaten werden randomisiert erzeugt und gespeichert, wobei die Krankheiten auf dem in der Medical Subject Headings (MeSH©) (siehe `https://meshb.nlm.nih.gov/search`) Datenbank der U.S. National Library of Medicine festgelegten Vokabular basiert. Zur Bestimmung der Ähnlichkeit zweier Terme, die Erkrankungen beschreiben, wird ein Tool des UMLS::Similarity Moduls (siehe `http://umls-similarity.sourceforge.net/`) herangezogen. Das Modul unterstützt dabei die Ontologien, wie bspw. MeSH, des Unified Medical Language System (UMLS). Die Arbeit basiert maßgeblich auf dem Artikel Wiese [2014], in dem die theoretischen Grundlagen sowie praktische Aspekte zu dem Thema untersucht werden.

Literaturverzeichnis

The Apache Software Foundation. Open source memory-centric distributed database, caching, and processing platform - apache ignite™, 2018. URL https://ignite.apache.org/. Accessed 21th Dec 2018.

U.S. National Library of Medicine. Medical subject headings, Nov 2018. URL https://www.nlm.nih.gov/mesh/. Accessed 13th Dec 2018.

Ted Pedersen. Umls::similarity. URL http://umls-similarity.sourceforge.net/. Accessed 21th Dec 2018.

U.S. National Library of Medicine. Unified medical language system (umls), April 2016. URL https://www.nlm.nih.gov/research/umls/. Accessed 21th Dec 2018.

Lena Wiese. Clustering-based fragmentation and data replication for flexible query answering in distributed databases. *Journal of Cloud Computing*, 3 (1):18, Oct 2014. ISSN 2192-113X. doi: 10.1186/s13677-014-0018-0. URL https://doi.org/10.1186/s13677-014-0018-0.