Pattern Mining in Keilschriftzeichnungen

,
bg.bartek@gmail.com
Universität Heidelberg, Deutschland

,
hubert.mara@iwr.uni-heidelberg.de
Universität Heidelberg, Deutschland

Keilschrifttafeln gehören zu den ältesten Textzeugen, die im Umfang mit den Texten in lateinischer und alt-griechischer Sprache vergleichbar sind. Da diese Tafeln aus dem gesamten Alten Orient über beinahe viertausend Jahre in Verwendung waren (Soden 1994), lassen sich damit viele interessante Fragestellungen zur Entwicklung von Religion, Politik, Wissenschaft, Handel bis hin zu Klimaveränderungen (Kaniewski et al. 2013) beantworten. Die aus Ton geformten Tafeln, bei denen Zeichen (Borger 2010) als keilförmige Abdrücke mit einem eckigen Stylus eingedrückt wurden, erfordern neue informationstechnische Methoden zu der Dokumentation und Analyse als die in Archiven üblichen Flachwaren.

Keilschrifttafeln werden mit Hilfe verschiedenster Methoden digitalisiert und in verschiedene, untereinander nicht kompatible Formate übertragen. Sie werden photographisch mit wechselnden Lichtverhältnissen aufgezeichnet, handschriftlich oder digital abgezeichnet oder mit Hilfe eines 3D-Scanners aufgenommen (Mara et al. 2010; Mara / Krömker 2013). Jede dieser Repräsentationen erfordert ein eigenes Tool-Set zur Analyse und die textuelle Analyse ist auf die jeweilige Repräsentation beschränkt.

Die Initiative für eine digitale Keilschriftdatenbank (Cuneiform Digital Library Initiative - CDLI) stellt mehr als 300.000 Keilschrifttafeln je nach Verfügbarkeit in Form von handgefertigten Abschriften, Photographien oder Umschriften zur Verfügung. Diese Datenbank besitzt keine Möglichkeit Keilschrifttafeln nach den Keilsymbolen zu durchsuchen.

In unserer bisherigen Arbeit (Bogacz / Massa et al. 2015) stellten wir Verfahren und einen Ablauf zur Homogenisierung von den drei gängigsten Datenquellen vor. Keilschriftabdrücke wurden handschriftlichen Zeichnungen, digital abgezeichneten und 3D-gescannten Tafeln entnommen. Die Datenquellen wurden zuerst, falls nötig, in das SVG Format (Scalable Vector Graphics) vektorisiert. SVG Dateien sind ein offener Standard zur Beschreibung von Vektorgrafiken, der sich den XML Standard zu nutze macht.

Die Nutzung dieses Dateiformates ermöglicht uns Wörter in den digitalen Abzeichnungen mit ihrer Übersetzung zu Annotieren und als XML-Tags zu den Grafikpfaden, die den Wörtern entsprechen, in den SVG Dateien selbst abzuspeichern. Wir nutzten diese Annotationen, um die Genauigkeit unserer Worterkennung zu überprüfen (Bogacz / Gertz et al. 2015).

Auf Grundlage der homogenisierten Datenbasis führten wir eine minimale und einheitliche Beschreibung von Keilabdrücken mit Hilfe von Merkmalsvektoren ein. Die Abdrücke einer Keilschrifttafel in dem jeweiligen Datenformat werden erkannt und extrahiert. Bei der Extraktion werden die einzelnen Keile durch mehrere verschiedene, sich ausschließende, Merkmalsvektoren modelliert. Die abschließend gewählte Untermenge von Keilmodellen für die gegebenen Keile einer Tafel ist eine global optimale Zuordnung von Keilmodellen zu den jeweiligen Keilabdrücken. Dieser Ansatz wurde gewählt, da die Abdrücke oft beschädigt oder nicht eindeutig identifizierbar sind.

Die reduzierte Darstellung als Merkmalsvektoren ermöglicht eine Analyse der Daten mit gängigen Methoden aus dem Bereich des maschinellen Lernens, wie der Priniciple Component Analysis (PCA) Dimensionsreduktion, dem k-Means Algorithmus oder auch einem Entscheidungsbaum (Mohri et al. 2012), und das Abspeichern der Keilabdrücke und der Keilschrifttafeln in austauschbaren XML Dateien zur weiteren Analyse oder in einer effizienten Suchstruktur als Grundlage für einen Suchalgorithmus.

In dieser Arbeit stellen wir ein Verfahren zur vollständig automatisierten Suche von Keilschriftsymbolen vor. Wir übernehmen die Idee von "Query Words" und adaptieren sie für geometrische Symbole. Anstatt ausschließlich Übersetzungen von Keilschrifttafeln zu durchsuchen und nicht übersetzte Tafeln auszulassen, können wir alle homogenisierten Tafeln nach Keilkonfigurationen durchsuchen. Eine beliebige geometrische Anordnung von Keilen im Merkmalsvektor Repräsentation wird als Query (Abfrage) genutzt, nach welcher Tafeln abgesucht werden können.

Unser Verfahren baut eine Suchstruktur auf, die danach mit Keilkonfigurationen abgesucht werden kann. Zuerst wird durch eine Radial Basis Function (RBF) Kernel-PCA Dimensionsreduktion (Schölkopf 1997) der Merkmalsraum der Merkmalsvektoren reduziert. Es gibt nur wenige Keiltypen und diese werden durch die hochdimensionalen (12 Merkmale pro Keil) Merkmalsvektoren überspezifisch beschrieben. Danach wird ein k-Means Clustering (Kanungo et al. 2002) durchgeführt, um die einzelnen Keiltypen automatisiert zu erkennen. Die gefunden Gruppierungen bilden die Basis für ein Wörterbuch an bekannten Keilkonfigurationen. Dieses Wörterbuch wird nun erweitert indem ein spatiales Frequent Pattern Mining (Han et al. 2007) der Tafeln durchgeführt wird. Häufig vorkommende und dicht zusammen liegende Keiltypen werden zu neuen Einträgen zusammengefasst. Keilschrifttafeln werden somit anhand der Positionen von im Wörterbuch vorhandenen Keilkonfigurationen beschrieben.

Ein Keilschriftszeichen wird gesucht, indem es in im Wörterbuch bekannte Keilkonfiguration unterteilt wird. Dazu werden die Merkmalsvektoren des Zeichens mit gelernten PCA reduziert und dem gelernten k-Means klassifiziert. Danach werden bekannte Konfigurationen im gesuchten Zeichen durch erneutes spatiales Frequent Pattern Mining identifiziert. Nun wird eine Schnittmenge von bekannten Konfigurationen im gesuchten Zeichen mit der Menge an bekannten Konfigurationen auf der Tafel gebildet. Übereinstimmungen werden durch ein genaueres Verfahren verglichen (Bogacz / Gertz et al. 2015).

Unser Verfahren Pattern Mining a Dictionary of Complex Structures (PDCS) macht sich die geringe Anzahl von Keiltypen (Winkelhaken, stehender Keil und liegender Keil) und häufig vorkommende Keilkonfigurationen zu nutze, um den Suchraum zu reduzieren. Zusammenfassend basiert es auf der Annahme, dass sich das zu durchsuchende Objekt in bekannte und grundlegende Formen, Keile der Keilschrift, zerlegen lässt, und die gesuchte Form eine geometrische Anordnung dieser Grundform ist. Dafür erweitern das Konzept des Frequent Pattern Minings indem wir die Geometrie der häufig vorkommenden Muster beachten.

Die k-Means Gruppierung der Keiltypen hat gegenwärtig eine Fehlerrate von 10%. Wir planen die Fehlerrate zu reduzieren indem wir die Parameter der PCA Dimensionreduktion automatisiert lernen und optimieren. Das Bilden der geometrischen Schnittmenge ist ein zeitaufwändiger Prozess. Wir arbeiten an einer Methode diesen Algorithmus zu beschleunigen indem wir Keilkonfigurationen aus dem Wörterbuch entfernen, die nicht zur Suche beitragen. Weitere mögliche Anwendungsbereiche für unser Verfahren sind Chinesische Zeichen, Heraldik, Maya Schriftzeichen und die kodikologische Untersuchung der Anordnung von Textpassagen eines Keilschrifttextes.

Bibliography

Bogacz, Bartosz / Gertz, Michael / Mara, Hubert (2015): "Character Retrieval of Vectorized Cuneiform Script", in: *Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR)*, Nancy, France.

Bogacz, Bartosz / Massa, Judith / Mara, Hubert (2015): "Homogenization of 2D & 3D Document Formats for Cuneiform Script Analysis", in: *Proceedings of the 2015 Workshop on Historical Imaging and Processing*, Nancy, France 115-122.

Borger, Rykle (22010): *Mesopotamisches Zeichenlexikon* (= Alter Orient und Altes Testament – Veröffentlichungen zur Kultur und Geschichte des Alten Orients und des Alten Testaments 305). Münster: Ugarit-Verlag.

Han, Jiawei / Cheng, Hong / Xin, Dong / Yan, Xifeng (2007): "Frequent pattern mining: current status and future directions", in: *Data Mining and Knowledge Discovery* 15, 1: 55-86.

Kaniewski, David / Van Campo, Elise / Guiot, Joel / Le Burel, Sabine / Otto, Thierry / Baeteman, Cecile (2013): "Environmental Roots of the Late Bronze Age Crisis", in: *PLoS One* 8, 8 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0071004 [letzter Zugriff 07. Februar 2016].

Kanungo, Tapas / Mount, David M. / Netanyahu, Nathan S. / Piatko, Christine D. / Silverman, Ruth / Wu, Angela Y. (2002): "An efficient k-means clustering algorithm: Analysis and implementation. Pattern Analysis and Machine Intelligence", in: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7: 881-892.

Mara, Hubert / Krömker, Susanne (2013): "Vectorization of 3D-Characters by Integral Invariant Filtering of High-Resolution Triangular Meshes", in: *Proceedings of 12th International Conference on Document Analysis and Recognition (ICDAR / IAPR)*, Washington D.C., USA 62–66.

Mara, Hubert / Krömker, Susanne / Jakob, Stefan / Breuckmann, Bernd (2010): "GigaMesh and Gilgamesh - 3D Multiscale Integral Invariant Cuneiform Character Extraction", in: *Proceedings of VAST10 - International Symposium on Virtual Reality, Archaeology and Cultural Heritage*, Palais du Louvre, Paris, France 131-138.

Mohri, Mehryar / Rostamizadeh, Afshin / Talwalkar, Ameet (2012): Foundations of Machine Learning (= Adaptive Computation and Machine Learning series). Cambridge, Massachusetts: MIT Press.

Schölkopf, Bernhard / Smola, Alexander / Müller, Klaus-Robert (1997): "Kernel Principal Component Analysis", in: Gerstner, Wulfram / Germond, Alain / Hasler, Martin / Nicoud, Jean-Daniel (eds.): *Artificial Neural Networks*. Proceedings of the 7th International Conference Lausanne, Switzerland (ICANN'97). Berlin / Heidelberg: Springer 583-588.

Soden, Wolfram von (1994): *The ancient Orient*. An introduction to the study of the ancient Near East. Michigan: Wm. B. Eerdmans Publishing Co.