Fachhochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Dimensionsreduktion kategorialer Daten zur Erzeugung von Themenlandkarten

Jan Dikow

Bachelorarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 20.08.2015

Hintergrund & Aufgabenstellung

Die Firma mapegy erzeugt für ihre webbasierte Analyse- und Visualisierungssoftware mapegy.scout verschiedene Visualisierungen auf Basis mehrerer Datenquellen wie beispielsweise Patentdaten und wissenschaftliche Publikationen. Eine der Visualisierungen ist eine Patentlandkarte, welche auf Grundlage der benutzerabhängigen Eingabe eine Gruppierung der Patente durchführt (Clusteranalyse) und diese Gruppen auf einer Karte darstellt (Dimensionsreduktion), sodass ähnliche Patente nahe zusammen liegen und unterschiedliche weiter auseinander. Dieser Prozess soll grundlegend überarbeitet werden, damit

1. verschiedene Typen von Dokumenten (auch z.B. News und wissenschaftliche Publikationen) anhand ihrer Zuordnung zu bestimmten Kategorien verarbeitet werden können,
2. der Prozess besser skalierbar und insgesamt schneller wird,
3. erste Ergebnisse schnell bereitgestellt werden (z.B. durch eine Vorschau, Vorprozesse oder Sampling),
4. ein Ausgabedatenmodell entsteht, das verschiedene Darstellungen im Front-End möglich macht.

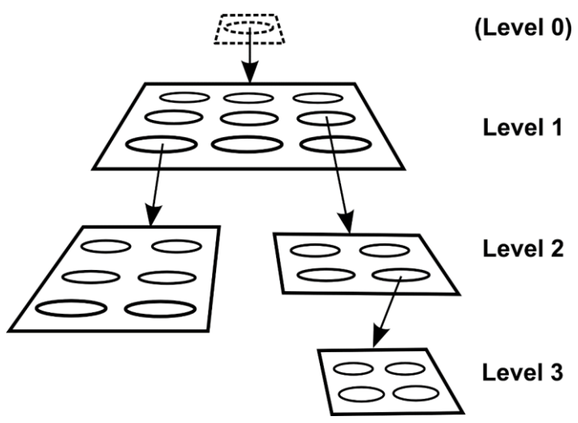

Abb. 1: Der Aufbau einer GHSOM-Karte: Unterschiedlich große Karten repräsentieren einzelne Neuronen übergeordneter Karten. Das Level 0 ist symbolisch für die Menge aller Eingabedaten. (Abb. nach [RMD02])

Konzept

Zur Erzeugung der Themenlandkarten wurde eine GHSOM (Growing Hierarchical Self-Organizing Map) gewählt (Abb. 1), deren einzelne Teilkarten aus einer Menge von Neuronenmodellen bestehen, die sich an die Trainingsdaten anpassen und somit Clustering und Dimensionsreduktion gleichzeitig realisieren. Dabei vermeiden sie die für viele andere Verfahren notwendige Erzeugung einer Distanzmatrix der zu verortenden Dokumente, welche eine quadratische Laufzeitkomplexität aufweist. Auch die Größe der einzelnen Karten und die hierarchische Struktur der gesamten GHSOM passen sich an die Trainingsdaten an. Damit lässt sich die Themenlandkarte in unterschiedlich detaillierte Level unterteilen. Höhere Ebenen können als Vorschau angezeigt werden und ab Level 2 ist die Teilkartenerzeugung parallelisierbar, wobei diese Cluster-Expansionen auch durch Nutzerinteraktion ausgelöst werden könnten.

Umsetzung

Die Implementierung des Prozesses geschah in RapidMiner (einer Java-basierten Umgebung für maschinelles Lernen und Data-Mining). Dafür wurde ein neuer RapidMiner-Operator programmiert, der die Teilkarten erzeugt. Durch geringfügige Änderungen des Prozesses ist es damit problemlos möglich zusätzliche Cluster-Eigenschaften zu generieren, um unterschiedliche Eigenschaften der Dokumente visualisieren zu können.

Visualisierung

Die testweise Visualisierung der Ergebnisse fand mittels PHP, JavaScript, HTML und auf Grundlage von Testdaten der Firma mapegy statt. Abb. 2 zeigt eine solche Visualisierung. Andere Visualisierungen ermöglichen zudem ein exploratives Navigieren durch die verschiedenen Teilkarten.

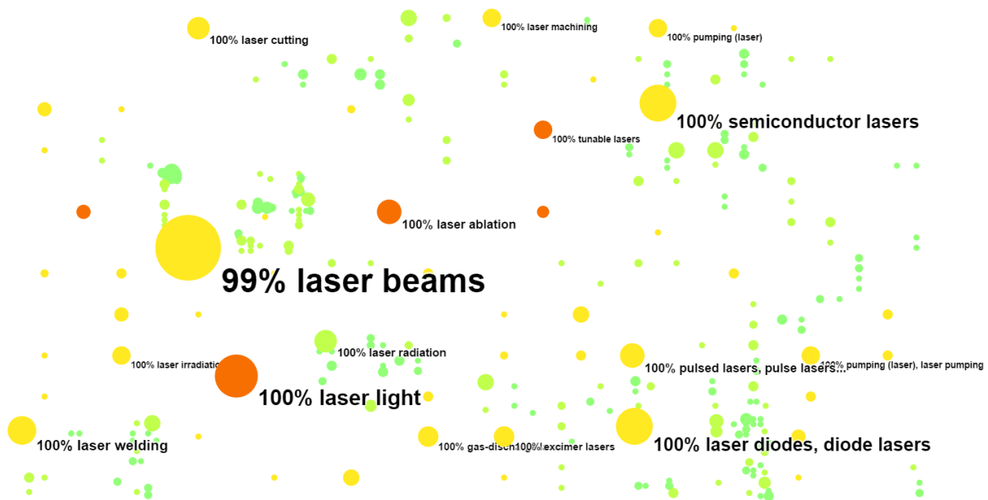

Abb. 2: „laser"-Suche in Testdaten: Dargestellt sind nur Blattknoten der entstandenen Hierarchie, beschriftet mit den häufigsten Kategorien. Die Level sind aufsteigend durch Farben von Rot nach Grün markiert.

Ergebnis & Ausblick

Es konnte in dieser Arbeit erfolgreich ein neuer Landkartenprozess entwickelt werden, der alle neuen Anforderungen berücksichtigt. Dabei wurde deutlich, dass die GHSOM für die Erzeugung von Themenlandkarten gut verwendet werden kann. Ihre Anpassung an die Daten ist jedoch nicht völlig unabhängig von Voreinstellungen und macht weitere Optimierungen nötig. Verbesserungsbedürftig sind insbesondere die Kriterien zum Kartenwachstum und zur Cluster-Expansion. Zudem verschlechtert sich die Qualität der Ergebnisse und die Berechnungsdauer bei zu vielen Dimensionen (Kategorien) der Eingabedaten (siehe Abb. 3), was den Prozess stark von der Auswahl der zu verarbeitenden Daten abhängig macht.

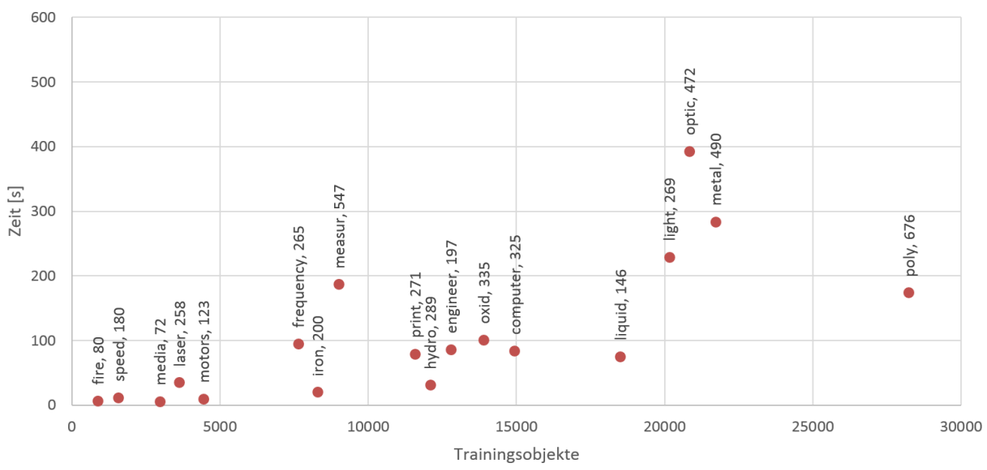

Abb. 3: Gesamtlaufzeit des Landkartenprozesses für verschiedene Mengen von Trainingsobjekten. Die Punkte sind mit dem Namen der Suche und der Anzahl der verarbeiteten Dimensionen beschriftet.

Quellen

[RMD02] Rauber, A. ; Merkl, D. ; Dittenbach, M.: The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data. In*: IEEE Transactions on Neural Networks* 13 (2002), Nr. 6.

Betreuung Dipl.-Inform. Ingo Boersch • Fachhochschule Brandenburg • Uwe Kuehn, M.Sc. • mapegy GmbH, Berlin