Kohonennetze für Information Retrieval mit User Feedback

Georg Ruß
Otto-von-Guericke-Universität Magdeburg

06.02.2003

Zusammenfassung

Richtig eingesetzt, sind selbstorganisierende Karten (SOM) ein probates Mittel, um große Datenmengen mittels unüberwachtem Lernen zu visualisieren. Durch ihre topologieerhaltende Abbildung von hochdimensionalen Eingaberäumen in niedrig– (typischerweise zwei–) dimensionale Ausgaberäume sind sie geeignet, Ähnlichkeitsbeziehungen in Daten darzustellen und ein intuitiv verständliches Resultat herzustellen. In dieser Arbeit werde ich kurz auf das zugrundeliegende Lernparadigma eingehen, dessen Ideen darlegen sowie auf den Begriff des Information Retrieval eingehen. In den desweiteren erwähnten Software–Implementationen spielt die Datenvorverarbeitung eine wichtige Rolle.

1 Einleitung

In Zeiten stark wachsender Datenbestände, ausgelöst u.a. durch Vereinfachung der Datensammlungs–Möglichkeiten sowie zunehmender digitaler Kommunikation, erscheint es mehr und mehr notwendig, Datenverarbeitungsmethoden zu entwickeln, die mit den typischerweise hochdimensionalen und nicht ohne weiteres zu visualisierenden Daten umgehen können. Dabei sollte ein Produkt entstehen, das intuitiv verständlich ist, möglichst fehlerfrei arbeitet und automatisch abläuft, möglichst ohne Intervention des Nutzers. Selbstorganisierende Karten weisen diese drei Eigenschaften auf (siehe Abschnitt 2), sind relativ gut erforscht, werden aber trotzdem kommerziell kaum eingesetzt. In Abschnitt 3 werde ich Ansätze aus dem Information Retrieval vorstellen, die die unstrukturierten Text–Datenmengen in eine für die SOM verwertbare Form bringen. Im darauffolgenden Abschnitt 4 geht es darum, wie die SOM auf diesen Eingaben arbeitet und wie die Visualisierung erfolgt. Die dritte geforderte Restriktion ("automatisch") wird bewußt etwas gelockert, um die oft gute Intuition des Benutzers mit einzubeziehen. Im letzten Teil dieser Arbeit werde ich den Nutzen darlegen, der durch die Verwendung von SOM entstanden ist und einen kurzen Ausblick geben.

2 Self–Organising Maps

Dieses zuerst von T.Kohonen in [3] beschriebene Paradigma des unüberwachten Lernens bewerkstelligt eine topologieerhaltende Abbildung vom hochdimensionalen Eingaberaum in eine zweidimensionale Karte (siehe Abbildung 1), wobei Ähnlichkeitsbeziehungen zwischen Eingabedaten durch Nachbarschaft der Ausgabedaten dargestellt werden. Eine geeignete Nachbarschaftsfunktion ist ebenfalls in Abbildung 1 zu sehen. Die beschriebene Eigenschaft begründet die intuitive Verständlichkeit der Methode.

Desweiteren sollten die SOM möglichst fehlerfrei auf Eingabedaten arbeiten, um nicht aus gültigen

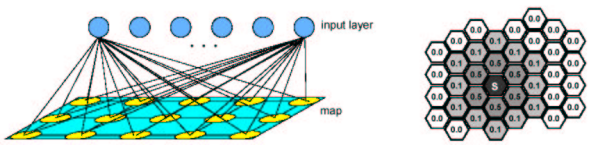

Abbildung 1: Topologieerhaltende Abbildung und Nachbarschaftsfunktion

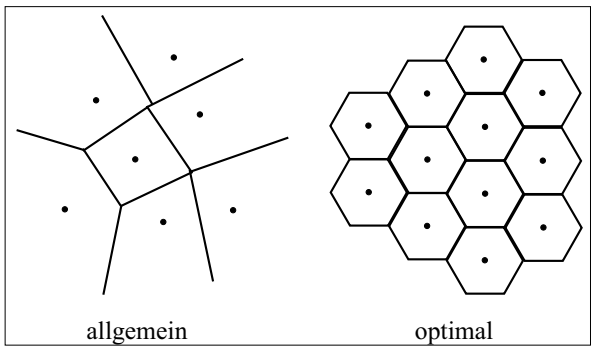

Abbildung 2: Voronoi–Zerlegung

Daten dem Benutzer falsche Schlußfolgerungen nahezulegen. Dieser Punkt ist ebenfalls gewährleistet, da die zugrundeliegende Voronoi–Zerlegung des Eingaberaums in eine endliche Anzahl von Zuständigkeitsgebieten als fehlerminimierend zu bezeichnen ist. Es hat sich herausgestellt, daß im zweidimensionalen Raum ein hexagonal aufgeteiltes Grid ein Optimum zwischen lückenlos und fehlerminimal darstellt (siehe Abbildung 2).

Auch zur Erfüllung der dritten gewünschten Eigenschaft sind die SOMs bestens geeignet. Dazu genügt es bereits, sich den Begriff "selbstorganisierend" vor Augen zu halten und zu wissen, daß die SOMs eine modellhafte Nachbildung neuronaler Strukturen darstellen. Auch der Mensch lernt automatisch und (wenn auch nicht ausschließlich) unüberwacht. Aber auch von der Implementations–Seite her läßt sich diese Eigenschaft begründen, da den SOMs nur recht einfache mathematische Vektor–Operationen wie die Berechnung der Ähnlichkeit (1) und die Neuberechnung der Gewichte des Gewinnerneurons ("best matching neuron", (2)) zugrundeliegen.

$$S(D_i, D_j) = sum_{k=1}^{m}(w_{ik}, w_{jk}) \quad (1)$$

$$\forall i : w_{s^i} = w_s + \theta(c, i) * \delta * (w_s - w_i) \quad (2)$$

3 Datenvorarbeitung und Information Retrieval

Für den Begriff des Information Retrieval existieren viele unterschiedliche Definitionen, in deren Schnitt aber mit Sicherheit die Gewinnung und Speicherung von Information enthalten ist. Wie bereits in Abschnitt 2 beschrieben, arbeiten die SOMs auf hochdimensionalen Eingaben. Zum Verarbeiten großer Datenmengen müssen die Daten aber erst einmal in diese Form gebracht werden. Dieser Prozeß wird Datenvorverarbeitung genannt und spielt in jedem Information–Retrieval–System eine wichtige Rolle, da alle anfallenden Daten nicht von vornherein diese Form besitzen, man denke z.B. an e–mail–Sammlungen oder auch Multimedia–Objekte. Am Beispiel von ASCII–Text–Sammlungen besteht die Datenvorverarbeitung aus drei sukzessiven Schritten, die in Abbildung 3 exemplarisch dargestellt sind.

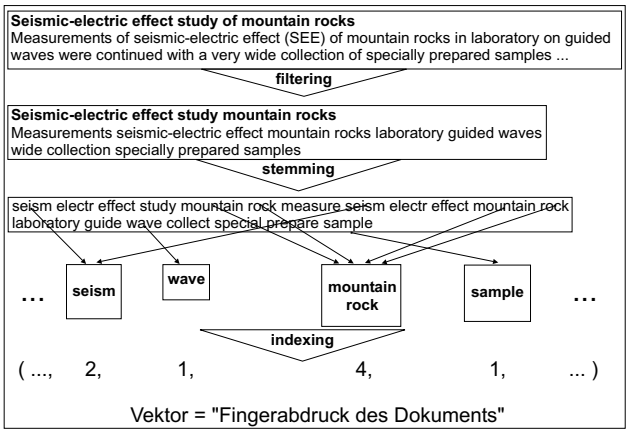

Abbildung 3: Beispiel für Datenvorverarbeitung

3.1 Filtering

Jeder Text enthält große Mengen von Wörtern, die nicht zur Unterscheidung zweier verschiedener Texte beitragen. Nach dem Shannon'schen Informationsbegriff ist der Informationsgehalt eines Zeichens (Wortes) umgekehrt proportional zur Auftretenswahrscheinlichkeit, sodaß sehr häufige Wörter entfernt

werden können, was z.B. Präpositionen, Konjunktionen, Artikel und dergleichen sind. Aber auch sehr seltene Wörter (z.B. *flahsc*–geschriebene) können entfernt werden. In der Praxis des Information Retrieval geschieht das Filtern meist über vordefinierte Stop–Lists, in denen diese sogenannten Stop–Words enthalten sind. Siehe auch [1].

3.2 Stemming

Desweiteren erscheint es sinnvoll, eine weitere Vereinfachung der Texte anzustreben: das Bilden der Wortstämme. Von den übrig gebliebenen Wörtern werden Präfixe und Suffixe entfernt, sodaß die informationstechnisch relevanten Wortstämme übrigbleiben. Eine Möglichkeit der Realisation ist die Angabe einer Grammatik mit Produktionsregeln, was durch den "Porter Stemmer" [2] realisiert wird.

3.3 Indexing

Nachdem nun die Eingabetexte bereits stark auf relevante Daten reduziert wurden, bietet es sich an, gleich zwei SOMs zu erstellen: Eine Wortkarte und eine Dokumentenkarte. Beide Karten werden sukzessive aus einer Initial–Karte aufgebaut.

3.3.1 Wortkarte

Hier liegt die Annahme zugrunde, daß Wörter, die in ähnlichem bzw. gleichem Kontext auftauchen, auch selbst zueinander ähnlich sind. Dazu wird jedem Wort ein Zufallsvektor zugewiesen, der 90 Dimensionen besitzt [4]. Nun werden für jedes Wort die Erwartungswert–Vektoren $w_{vorgänger}$ und $w_{nachfolger}$ berechnet, aus denen sich der Kontextvektor $w_c = (w_{vorgänger}, w, w_{nachfolger})$ ergibt. Der entstehende 270–dimensionale Kontextvektor kann als Eingabe für die SOM verwendet werden. Nach dem Lernparadigma der SOMs ist davon auszugehen, daß ähnliche Worte in der Wortkarte benachbart sein werden.

3.3.2 Dokumentenkarte

Ähnlich wie in 3.3.1 liegt hier die Annahme zugrunde, daß Dokumente, die inhaltlich zueinander ähnlich sind, ähnliche Vektoren besitzen und damit in der aufzubauenden Karte benachbart sein sollten. Als dritter Schritt der Dokumentverarbeitung wird nun das eigentliche Indexing durchgeführt, was im einfachsten Fall ein Einsortieren der Wörter in Buckets und das anschließende Summieren der Wörter pro Bucket darstellt. Die Summen werden in hochdimensionalen Vektoren als Komponenten verwendet, sodaß nun der "Fingerabdruck des Dokuments" als Eingabevektor für die SOMs vorliegt. Zur besseren Verständlichkeit sei hier noch einmal auf Abbildung 3 verwiesen.

4 Visualisierung

Nachdem klar ist, wie die Eingabevektoren für die SOM zustandekommen und wie die prinzipielle Verarbeitung dieser Vektoren erfolgt, bietet sich ein Beispiel zur Visualisierung des Lernprozesses an, dazu Abbildung 4.. Ausgehend von einer Initialisierung mit 3×3 Clustern kommen in jedem Lernschritt neue Cluster hinzu und die Karte erweitert sich. Nachdem dem neuronalen Netz alle Vektoren präsentiert worden sind, stellt sich bei Betrachtung der Karte heraus, daß sich größere Agglomerationen von Clustern gebildet haben, die z.B. größere Wissensgebiete (im Falle von wissenschaftlichen Publikationen) darstellen können.

Abbildung 4: Wachsende Selbstorganisierende Karte

Es bleibt noch die Frage offen, wie Dokumente einsortiert werden, die offensichtlich in mehrere Cluster gut passen würden. Ein wissenschaftlicher Aufsatz zum Information Retrieval könnte in die Cluster "Data Mining" oder "Neuronale Netze" gut passen, da fachgebietsübergreifende Methoden verwendet werden. An dieser Stelle sollte bewußt der Automatismus des unüberwachten Lernens aufgeweicht und die gute Intuition des Benutzers einbezogen werden, um

die Exaktheit der SOMs zu gewährleisten. Denkbar ist, daß an den Nutzer eine Abfrage gestellt wird, die ihm die Auswahl zwischen den angebotenen Clustern läßt und das Dokument entsprechend der Wahl des Nutzers einsortiert. In der SOM sind dazu Anpassungen des gewünschten Ähnlichkeitsmaßes nötig, was zur Änderung von Prioritäten einzelner Features führt.

5 Nutzen und Ergebnisse

Durch die Datenverarbeitung und –visualisierung mittels selbstorganisierender Karten bieten sich neue Möglichkeiten, mit den Daten umzugehen und innerhalb der Daten zu navigieren. Natürlich bleibt die bisherige Suche nach Schlüsselwörtern bestehen, da sie eine recht gut geeignete Basis für die weitere Verarbeitung und Auswahl bildet und Daten vorselektiert. Darüberhinaus ist nun aber eine visuelle Suche auf der Dokumentenkarte und der Wortkarte möglich, d.h. es werden die Treffer der Schlüsselwort–Suche auf den Karten angezeigt (ggf. farblich codiert) und es kann navigiert werden.

Auf der Wortkarte können somit neue Schlüsselwörter zur Verfeinerung der Suchanfrage (boolesche Verknüpfungen) gefunden werden, da sich diese in unmittelbarer Nachbarschaft zu den angezeigten Treffern befinden. Auf der Dokumentenkarte können ebenfalls in Nachbarschaft der Treffer ähnliche Dokumente gefunden werden, die relevant sein könnten.

Eine noch weitergehende Möglichkeit ist die inhaltsbasierte Suche bzw. Klassifikation und Suche anhand eines Beispiels. Dazu liegt ein Beispiel–Dokument vor, zu dem ähnliche Dokumente gefunden werden sollen. Um die Suche zu ermöglichen, wird die in Abschnitt 3 beschriebene Datenvorverarbeitung auf das Beispieldokument angewandt, der Vektor wird der SOM präsentiert und die Lage des Dokuments wird angezeigt. Das Ergebnis sind ähnliche Dokumente, die in der Nachbarschaft des Treffers liegen.

Denkbar wäre auch eine automatische Klassifikation eingehender e–mails oder Text–Dokumente. Auch auf diese würde die Datenvorverarbeitung angewandt und die Dokumente würden in Cluster (Mail–Ordner) einsortiert, die bereits ähnliche Dokumente enthalten.

Es existieren bereits seit einiger Zeit Prototypen, die die zugrundeliegende Idee der Verarbeitung von Text–Daten mittels selbstorganisierender Karten verwirklichen. Dazu sei auf WEBSOM [5] und
SOMAccess [6] verwiesen.

6 Zusammenfassung

Der Einsatz von selbstorganisierenden Karten innerhalb von großen Dokumentensammlungen bringt erhebliche Vorteile wie z.B. das automatische Lernen und Visualisieren großer Text–Datenbestände sowie eine intuitive Verständlichkeit der Implementierung, was bei kommerziellen Systemen nicht selbstverständlich ist. Es wurden neue, vielversprechende Möglichkeiten zur Suche in den Daten vorgestellt. Wünschenswert wäre die Einbeziehung des Nutzers in den Lernprozeß und eine Anwendung des beschriebenen Systems in der breiten Praxis.

Literatur

[1] W.B.Frakes, and R. Baeza–Yates, *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, New Jersey, 1992.

[2] M. Porter, An algorithm for suffix stripping, *Program*, pp. 130-137, 1980.

[3] T. Kohonen, Self–Organizing Maps, Springer Verlag, 1995.

[4] T. Honkela, Self–Organizing maps in natural language processing, Helsinki University of Technology, Neural Networks Research Center, Espoo, Finnland, 1997.

[5] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen, Newsgroup Exploration with the WEBSOM Method and Browsing Interface, Technical Report, In: Helsinki University of Technology, Neural Networks Research Center, Espoo, Finnland, 1996.

[6] A. Nürnberger, A. Klose, R. Kruse, G. Hartmann, and M. Richards, Interactive Text Retrieval based on Document Similarities, In: G. Hartmann, A. Nölle, M. Richards, and R. Leitinger (eds.), *Data Utilization Software Tools 2 (DUST–2 CD–ROM)*, Max–Planck–Institut für Aeronomie, Katlenburg–Lindau, Deutschland, 2000.