science softCon

IWS Otto-von-Guericke Universität Madgeburg

Interaktive Textsuche basierend auf Dokumentenähnlichkeiten

A. Nürnberger(1), A. Klose, R. Kruse, G. K. Hartmann(2), M. L. Richards

Motivation: Eingeschränkte Möglichkeiten der Standardsuchverfahren.

Idee: Gruppierung der Dokumente unter Verwendung eines Ähnlichkeitsmaßes. Dieses unterstützt den Benutzer in der Navigation durch die Dokumente. Die Navigation, insbesondere die Suche nach dem "Einstiegsdokument", soll durch konventionelle Schlüsselwortsuche unterstützt werden.

Realisierung: Entwicklung eines interaktiven Werkzeuges basierend auf Selbstorganisierenden Karten:

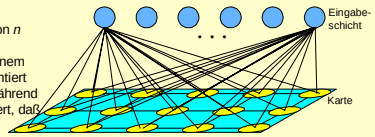
...nbestand

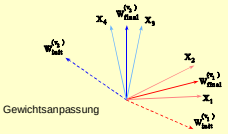
Selbstorganisierende Karten (Self-organizing maps; SOM)

Künstliches Neuronales Netz zur Projektion hochdimensionaler Daten in einen niedrigdimensionalen Datenraum (meist zwei Dimensionen), wobei Nachbarschaftsbeziehungen (Ähnlichkeiten) möglichst erhalten bleiben sollen.

Allgemeine Struktur:

Eine SOM besteht aus einer Eingabeschicht von *n* Neuronen und einer durch ein Neuronengitter definierten Karte. Der Gewichtsvektor w_i von einem Neuron *i* der Karte zur Eingabeschicht repräsentiert jeweils einen Prototypen der Eingabedaten. Während des Lernens werden die Prototypen so verändert, daß ähnliche Eingabevektoren auf benachbarte Neuronen abgebildet werden.

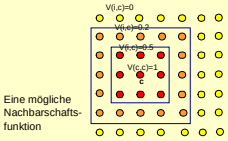




Lernverfahren (Wettbewerbslernen):

Die Gewichte (Prototypen) w_i werden zufällig initialisiert. Die Anpassung der Gewichtsvektoren erfolgt durch ein sequentielles Regressionsverfahren, wobei alle Eingabevektoren nacheinander dem Netz präsentiert werden. Sei $t = 1,2,...$ der Index der Vektoren. Für jeden Eingabevektor $x(t)$, wird zunächst der Index *c* des Gewinnerneurons bestimmt:

$$\forall i : \|w_c - x(t)\| \leq \|w_i - x(t)\|$$



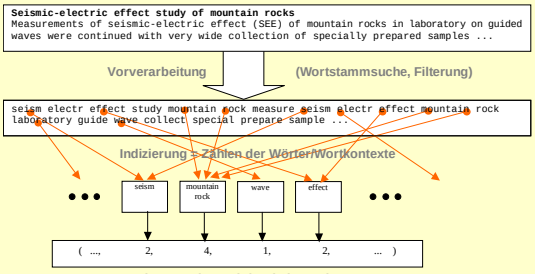
Anschließend wird der zugehörige Vektor w_c so angepasst, dass bei der nächsten Präsentation desselben Vektors ein höherer Zugehörigkeitsgrad erzielt wird. Außerdem werden alle Vektoren *i* in einer Nachbarschaft um das Siegerneuron *c* auf die gleiche Weise angepasst:

$$\forall i : w_i = w_i + v(c,i) \cdot \delta \cdot (w_i - x(t))$$

wobei δ eine Lernrate und $v(i, c)$ eine Nachbarschaftsfunktion ist, deren Ausgabewert mit zunehmendem Abstand der Neuronen *i* und *c* in der Karte kleiner wird.

Vorverarbeitung und Kodierung der Dokumente

Nach einem Vorverarbeitungsschritt (Wortstammsuche und Filterung) wird eine Liste von Körben zur Einsortierung der Wörter des Dokumentenbestandes erstellt (siehe auch „Definition der Körbe"). Basierend auf diesen Körben werden alle Dokumente als Vektoren kodiert, wobei jedes Tupel jeweils die Anzahl der Wörter in den Körben beschreibt. Diese Vektoren werden als Fingerabdrücke der Dokumente betrachtet.



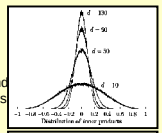
Anordnen der Dokumente: Die Dokumentenkarte (document map)

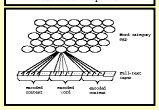
Die Fingerabdrücke der Dokumente werden als Eingabevektoren für eine zweidimensionale selbstorganisierende Karte verwendet. Nach dem Trainieren dieser Karte befinden sich alle ähnlichen Dokumente (definiert durch ihren Fingerabdruck) nahe beieinander auf der Karte, ggf. werden sie auch dem gleichen Neuron zugeordnet. Somit können, nachdem auf der Karte ein „Einstiegsdokument" gefunden wurde, durch einfache „Navigation" in der Nachbarschaft weitere ähnliche Dokumente ausgewählt werden.

Definition der Körbe (Erstellung einer Wortkategorienkarte)

a) Gruppierung ähnlicher Wörter basierend auf einem 3-Wort-Kontext

- Kodierung der Wörter als hochdimensionale Zufallsvektoren (Ritter und Kohonen, 1989)
  → Kodierung impliziert keine Anordnung: Vektoren sind "quasi-orthogonal"
- Für jedes Wort werden die Erwartungswertvektoren e_1 und e_2 über die Zufallsvektoren aller benachbarten Wörter (in allen Dokumenten) berechnet und ein Kontextvektor *v* mittels dieser Erwartungswerte und des Zufallsvektor *w* des betrachteten Wortes bestimmt: v = {e_1 *w* e_2} (Honkela et al., 1996)
  → Wörter die in ähnlichen Kontexten verwendet werden haben ähnliche Erwartungswertvektoren und somit ähnliche Vektoren *v*
- Projektion der v_i auf zwei Dimensionen mittels einer selbstorganisierenden Karte
  → Wörter die in ähnlichen Kontexten verwendet werden, werden gleichen (oder benachbarten) Neuronen zugeordnet.
- Jedes Neuronen der so bestimmten SOM (Wortkategorienkarte) wird als Korb zur Bestimmung der Fingerabdrücke (durch auszählen) verwendet.





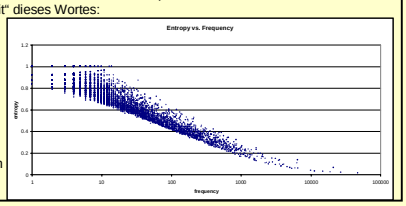
b) Selektion von Indexwörtern basierend auf ihrer Entropie

- Berechne die Entropie für jedes Wort (Lochbaum und Streeter, 1989). Dieses Maß kennzeichnet die „Wichtigkeit" dieses Wortes:

$$W(w) = 1 + \frac{1}{\ln(m)} \sum_{i}^{m} p^i(w) \cdot \ln(p^i(w))$$

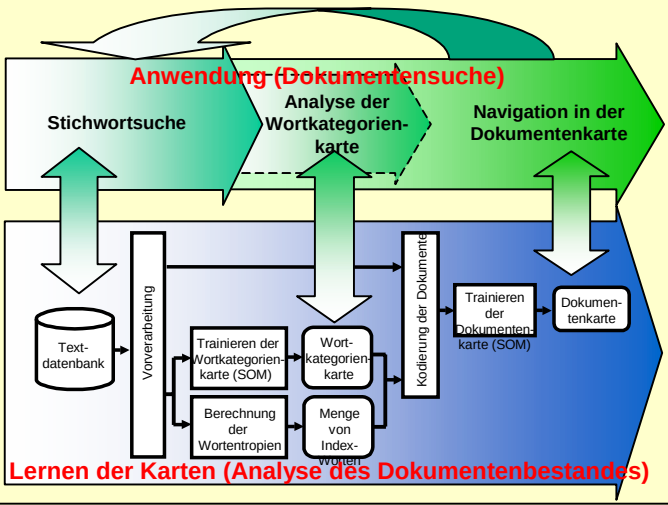
$$\text{mit } p^i(w) = \frac{n^i(w)}{\sum_{i}^{m} n^i(w)}$$

$n_i(w)$: Häufigkeit von Wort *w* im Dokument *i*
m: Anzahl Dokumente



- Wähle die Wörter, die im Vergleich zu ihrer Häufigkeit eine hohe Entropie haben
- Benutze diese Wörter als Körbe zur Bestimmung der Fingerabdrücke

Systemübersicht



Beispielsuche

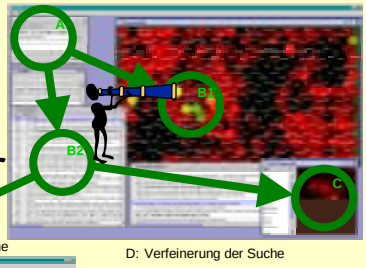
"Gibt es Dokumente über die Anwendung Neuronaler Netze im Bereich Wasserdampf?"

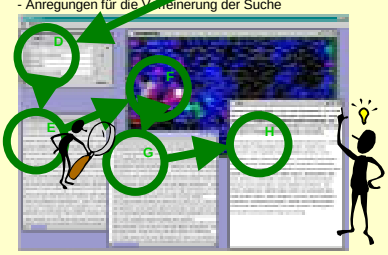
A: Stichwortsuche und Zuordnung von Farben zu den Suchbegriffen

B1: Analyse der Ergebnisse in der Dokumentenkarte:
- neural network (Grün): eingegrenzt
- water/vapour (Rot): sehr verstreut

B2: Untersuchung der Fingerabdrücke in der Wortkategorienkarte und Betrachtung der Stichwortverteilung für das Dokument

C: Analyse der Liste der Anfrageergebnisse:
- Anregungen für die Verfeinerung der Suche





D: Verfeinerung der Suche

E: Auswahl (möglicherweise) interessanter Dokumente:

F: Markierung des Treffers (oder eines Nachbarn) auf der Dokumentenkarte:

G: Liste der Dokumente die dem dem Knoten (Neuron) zugeordnet wurden.

H: Zugriff auf die relevanten Dokumente (lesen, drucken, etc.)

Kontakte:

Prof. Dr. Rudolf Kruse (1)
FIN-IWS
Universitätsplatz 2
D-39106 Magdeburg

Internet: fuzzy.cs.uni-magdeburg.de

Prof. Dr. Gerd K. Hartmann (2)
Max-Planck-Institut für Aeronomie
Max-Planck-Straße 2
D-37191 Katlenburg-Lindau, Germany

www.linmpi.mpg.de