

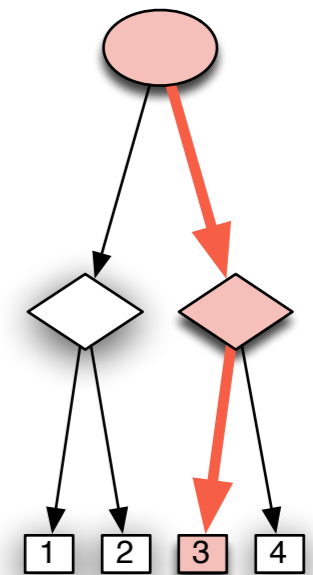
Document Clustering

Sprache und Bilder

15.12.2003

Jonas Zimmermann

jonas.zimmermann@uni-bielefeld.de



Übersicht

- Was ist das Problem?
- Vielleicht eine Lösung: Hierarchisches *Clustering*
- Knoten erzeugen Objekte
- Performanz des Modells
- Training des Modells: *Expectation-Maximization*
- Stöbern (*browsing*)
- Suchen und Finden
- Automatisches Annotieren

Was ist das Problem?

- Suchen nach Dokumenten, die bestimmte Bilder enthalten, ist schwierig
- flexible Suche ermöglichen: Bildbeschreibung, Bildbeispiel
- Datenbank so organisieren, daß intuitives *Browsing* möglich ist
- möglichst wenig Speicher und Rechenzeit verwenden:
 - Redundanz verringern

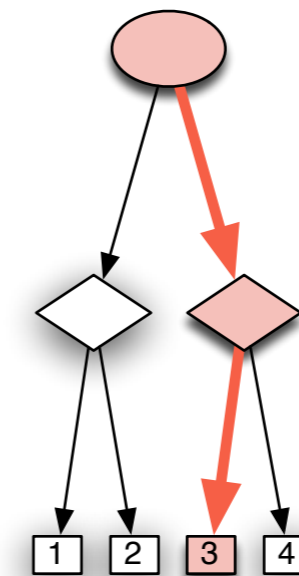
Hierarchisches Clustering

Dokument



Sonne, Himmel,
Meer, Wellen

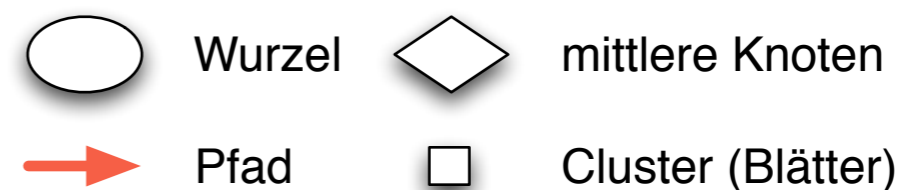
- jedes Dokument wird von genau einem *Cluster* erzeugt (*hartes Clustering*)
- ein Dokument wird mit einer gewissen Wahrscheinlichkeit von einem Cluster erzeugt (*weiches Clustering*)
- die Knoten auf dem Pfad vom Wurzelknoten zum Blatt erzeugen Wörter und Bildsegmente



hohe Knoten erzeugen allgemeine Wörter und Bildeigenschaften (z.B. „Himmel“)

speziellere Wörter und Bildeigenschaften (z.B. „Sonne“)

Blätter erzeugen dokumentenspezifische Wörter und Bildeigenschaften (z.B. „Wellen“)



Knoten erzeugen Objekte

- Wahrscheinlichkeit, daß ein Knoten ein Objekt erzeugt:

$$P(i|l, c) - i \text{ Objekt, } l \text{ Ebene (level), } c \text{ Cluster}$$

- für Wörter:
 - Tabelle, die im Training gelernt wurde; durch Häufigkeit bestimmt
- für Bildfeatures:
 - wird repräsentiert durch einen Feature-Vektor mit Informationen über Position, Größe, Farbe, Textur, Form (Blobworld Repräsentation)
 - Gaußverteilung über die Feature-Vektoren für jeden Knoten
 - hier: Unabhängigkeit der Features wird angenommen

Performanz des Modells

- Wahrscheinlichkeit, daß eine Menge von Objekten ausgehend von einem Dokument vom Modell erzeugt wird:

$$P(D|d) = \sum_c P(c) \prod_{i \in D} \left(\sum_l P(i|l, c) P(l|c, d) \right)$$

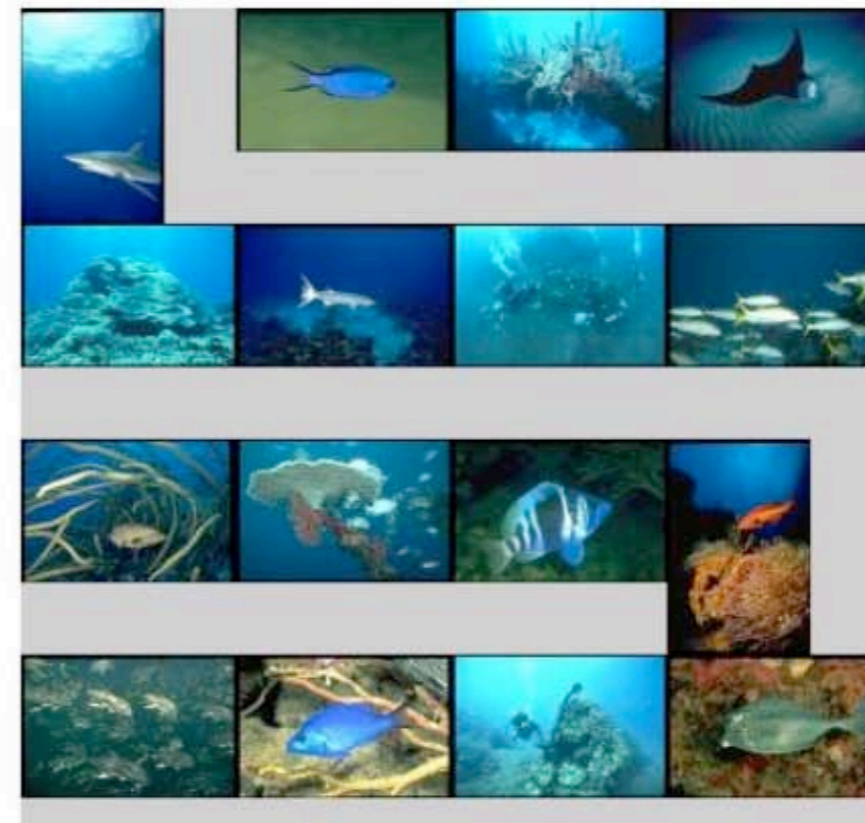
D Menge von Objekten (Wörter und Bildsegmente),
 l Ebenen, c Cluster, d Dokument

- $P(l|c, d)$ müßte für jedes neue Dokument neu berechnet werden, aber der vereinfachte Ansatz $P(l|c)$ reicht aus
- wenn die Objekte in D gerade die des Dokuments d sind, sollte $P(D|d)$ gegen 1 gehen
- wenn D zu d komplementäre Objekte enthält, sollte $P(D|d)$ klein werden

Training des Modells mit *Expectation-Maximization*

- Anfang: Menge von Dokumenten M und Startmodell (gleich- oder zufallsverteilt) $\vartheta^0 = (P(c), P(i|c,l))$
- *Expectation* Schritt:
 - Bestimmung von $P(c|d)$ und $P(l|d,c)$ mit altem Modell ϑ^m
- *Maximization* Schritt:
 - Neuberechnung der Parameter $P(c)$ und $P(i|c,l)$ von ϑ^{m+1} durch Maximierung der *Log-Likelihood* \mathcal{L}
$$\begin{aligned}\mathcal{L}(\vartheta) = \log \prod_{d \in M} P(d) &= \log \prod_d \sum_c P(c) \prod_{i \in d} \left(\sum_l P(i|l,c) P(l|c,d) \right) \\ &= \sum_d \log \sum_c P(c) \prod_{i \in d} \left(\sum_l P(i|l,c) P(l|c,d) \right)\end{aligned}$$
- bewiesen: EM führt zu (lokalem) Maximum; bereits nach wenigen Durchläufen brauchbare Ergebnisse

- EM-Algorithmus invariant unter Auswahl der Beispieldokumente
- Clustering mit Wörtern und Bildsegmenten zusammen wirkungsvoller als jeweils allein, denn
 - Wörter sorgen für semantische Kohärenz
 - Bildsegmente sorgen für visuelle Kohärenz



Stöbern

- viele Datenbanken nicht geeignet, sie ziellos zu durchstöbern
- Benutzer muß Ordnung schnell verstehen, Cluster müssen kohärent sein
- Ordnung grob → fein oder allgemein → speziell, also intuitiv verständlich
- Tests ergeben: nach diesem Verfahren gefundene Cluster werden signifikant häufiger als kohärent erkannt als zufällige

Suchen und Finden

- Suche nach Wörtern und/oder Bildeigenschaften
- berechnet Wahrscheinlichkeit, daß ein Dokument die Suchmerkmale (Q) erzeugt

$$\begin{aligned} P(Q|d) &= \sum_c P(Q|c, d)P(c|d) \\ &= \sum_c \left[\prod_{q \in Q} \left(\sum_l P(q|l, c)P(l|c, d) \right) P(c|d) \right] \end{aligned}$$

Automatisches Annotieren

- Grundlage: Wörterbuch, zu annotierendes Bild
- Berechnung der Wahrscheinlichkeit, daß ein Bild ein bestimmtes Wort generiert

$$\begin{aligned}P(w|B) &= \sum_c P(w|c, B)P(c|B) \\ &\propto \sum_c P(w|c, B)P(B|c)P(c) \\ &= \sum_c P(w|c, B) \prod_{b \in B} P(b|c)P(c) \\ &= \sum_c \left(\sum_l P(w|c, l)P(l|c, B) \right) \prod_{b \in B} \left(\sum_l P(b|l, c)P(l|c) \right) P(c)\end{aligned}$$

- $P(l|c, B)$ wieder aufwendiger zu berechnen: durch Re-Fitting, daher: benutze $P(l|c)$
- Test: Wörter, die zu einem Beispieldokument gehören, werden verglichen mit den Vorausgesagten → besser als Zufall

- Beispiel für automatische Illustration



“The large importance attached to the harpooneer's vocation is evinced by the fact, that originally in the old Dutch Fishery, two centuries and more ago, the command of a whale-ship was not wholly lodged in the person now called the captain, but was divided between him and an officer called the Specksynder. Literally this word means Fat-Cutter; usage, however, in time made it equivalent to Chief Harpooneer. In those days, the captain's authority was restricted to the navigation and general management of the vessel; while over the whale-hunting department and all its concerns, the Specksynder or Chief Harpooneer reigned supreme. In the British Greenland Fishery, under the corrupted title of Specksioneer, this old Dutch official is still retained, but his former dignity is sadly abridged. At present he ranks simply as senior Harpooneer; and as such, is but one of the captain's more inferior subalterns. Nevertheless, as upon the good conduct ...”

large importance attached fact old dutch century more command whale ship was per son
 was divided officer word means fat cutter time made days was general vessel whale
 hunting concern british title old dutch official present rank such more good american
 officer boat night watch ground command ship deck grand political sea men mast way
 professional superior

Fazit

- Dokumente werden in hierarchische Cluster eingeteilt, allgemeine Eigenschaften werden auf hohen Ebenen erzeugt
- durch Baumstruktur mit natürlicher Ordnung wird Stöbern erleichtert
- Suche nach Wörtern und Bildsegmenten gleichermaßen möglich
- automatische Annotation von Bildern oder Illustration von Wörtern möglich

Literatur

- K. Barnard, D. Forsyth, „Learning the Semantics of Words and Pictures,“ Computer Division, University of California, Berkeley
- K. Barnard, P. Duygulu, D. Forsyth, „Clustering Art,“ Computer Division, University of California, Berkeley
- T. Hofmann J. Puzicha, „Statistical Models for Co-occurrence Data,“ Massachusetts institute of technology, A.I. Memo 1635, 1998.
- C. Carson, S. Belongie, H. Greenspan und J. Malik, „Blobworld: Image segmentation using Expectation-Maximization and its application to image querying,“ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, SUB.