Multimedia-Datenbanken im SS 2010
“Distanzfunktionen II”

Dr.-Ing. Marcin Grzegorzek

08.06.2010

Inhalte und Termine

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

1. Einführung in MMDB

1.1 Grundlegende Begriffe

1.2 Suche in einem MMDBS

1.3 MMDBMS-Anwendungen 27.04.2010

2. Prinzipien des Information Retrievals

2.1 Einführung

2.2 Information-Retrieval-Modelle

2.3 Relevance Feedback

2.4 Bewertung von Retrieval-Systemen

2.5 Nutzerprofile 03.05.2010

Inhalte und Termine

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

3. Einführung in Multimedia-Retrieval

3.1 Besonderheiten der Verwaltung und des Retrievals	
3.2 Ablauf des Multimedia-Information-Retrievals	
3.3 Daten eines Multimedia-Retrieval-Systems	
3.4 Feature	
3.5 Eignung verschiedener Retrieval-Modelle	
3.6 Multimedia-Ähnlichkeitsmodell	10.05.2010

4. Feature-Transformationsverfahren

4.1 Diskrete Fourier-Transformation	11.05.2010
4.2 Diskrete Wavelet-Transformation	17.05.2010
4.3 Karhunen-Loeve-Transformation	
4.4 Latent Semantic Indexing und Singulärwertzerlegung	31.05.2010

Inhalte und Termine

5. Distanzfunktionen

5.1 Eigenschaften und Klassifikation

5.2 Distanzfunktionen auf Punkten 07.06.2010

5.3 Distanzfunktionen auf Binärdaten

5.4 Distanzfunktionen auf Sequenzen

5.5 Distanzfunktionen auf allgemeinen Mengen 08.06.2010

6. Ähnlichkeitsmaße

6.1 Einführung

6.2 Distanz versus Ähnlichkeit

6.3 Grenzen von Ähnlichkeitsmaßen

6.4 Konkrete Ähnlichkeitsmaße

6.5 Aggregation von Ähnlichkeitswerten

6.6 Umwandlung von Distanzen in Ähnlichkeitswerte und Normierung

6.7 Partielle Ähnlichkeit

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

Inhalte und Termine

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

7. Effiziente Algorithmen und Datenstrukturen

7.1 Hochdimensionale Indexstrukturen

7.2 Algorithmen zur Aggregation von Ähnlichkeitswerten

8. Anfragebehandlung

8.1 Einführung

8.2 Konzepte der Anfragebehandlung

8.3 Datenbankmodell

8.4 Sprachen

9. Zusammenfassung

Einführung

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- paarweiser Vergleich der Feature-Werte von Medienobjekten
- hier die häufigsten Distanzfunktionen analysiert nach Eigenschaften
- Eigenschaften nutzbar zur Konfiguration eines MMDBS bzgl. Suchszenario
- Distanzen auf Punkten, Binärdaten, Sequenzen und allgemeinen Mengen

Overview

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

Distanzfunktionen auf Binärdaten

Distanzfunktionen auf Sequenzen

Distanzfunktionen auf allgemeinen Mengen

Overview

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

Distanzfunktionen auf Binärdaten

Distanzfunktionen auf Sequenzen

Distanzfunktionen auf allgemeinen Mengen

Allgemeines

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- Unter Binärdaten verstehen wir hier die Erfüllung bzw. Nichterfüllung bestimmter Eigenschaften von Medienobjekten.
- Zu einer vorgegebenen Menge E von n Eigenschaften und jedem Medienobjekt bekannt ist, welche Eigenschaften erfüllt sind und welche nicht.
- Graphisch lassen sich die Feature-Daten als Eckpunkte eines n-dimensionalen Hypereinheitswürfels darstellen.

Eigenschaften und Korrespondenzen

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

Vergleicht man zwei Punkte p_1 und p_2, ergeben sich vier verschiedene Anzahlwerte, die als Grundlage für die Distanzmessung verwendet werden:

$e \in E$	e erfüllt für p_1	e nicht erfüllt für p_1
e erfüllt für p_2	$n_{1/1}$	$n_{0/1}$
e nicht erfüllt für p_2	$n_{1/0}$	$n_{0/0}$

Beispiel:

$$p_1 = (0,0,0,0,1,1,1,1)^\mathsf{T} \quad p_2 = (1,1,0,1,1,1,0,0)^\mathsf{T}$$

$$\Downarrow$$

$$n_{0/0} = 1 \quad n_{0/1} = 3 \quad n_{1/0} = 2 \quad n_{1/1} = 2$$

Minkowski-Distanzfunktion auf Binärdaten

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

In der allgemeinen Form:

$$d_{L_m}(p_1, p_2) = \left(\sum_{i=1}^{n} |p_1[i] - p_2[i]|^m \right)^{1/m}$$

Auf Binärdaten:

$$d_{L_m} = (n_{1/0} + n_{0/1})^{1/m}$$

Overview

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

Distanzfunktionen auf Binärdaten

Distanzfunktionen auf Sequenzen

Distanzfunktionen auf allgemeinen Mengen

Allgemeines

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- Sequenz-Daten bestehen aus einer Liste von Datenelementen eines Datentyps.
- Die Anzahl der Elemente zur Beschreibung von verschiedenen Medienobjekten kann unterschiedlich sein.
- Klassifikation und Beispiele:
  - keine Positionskorrespondenz: Earth-Mover-Distanzfunktion
  - Positionskorrespondenz und reelle Werte: DFT-L_2-Distanzfunktion
  - Positionskorrespondenz und nominale Werte: Editierdistanzfunktion

Earth-Mover-Distanzfunktion

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- Diese Distanzfunktion geht von dem Datentyp $\mathrm{tuple}(p_i : \mathrm{array}[1 \ldots n](\mathrm{real}) ; w_{p_i} : \mathrm{real})$ für das i-te Element einer Sequenz p aus.

- "Erdhügel und Erdlöcher"

  Um die Distanz zwischen der Sequenz p mit m Elementen und der Sequenz q mit n Elementen zu ermitteln, werden die Elemente von p als Erdhügel und die von q als Erdlöcher aufgefasst. Die Punkte p_i bzw. q_i geben die Position der Hügel bzw. der Löcher an, während w_{p_i} und w_{q_i} die Volumina der Hügel/Löcher beschreiben. Um die Distanz zwischen den Sequenzen zu ermitteln, wird nun versucht, die Erde der Hügel mit minimalen Transportkosten in die Löcher zu füllen.

Earth-Mover-Distanzfunktion

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- Ziel ist Minimierung der Transportkosten.
- Ein konkreter Transport wird durch die Angabe der Quantitätsmatrix $F = [f_{ij}]$ definiert. Der Wert f_{ij} gibt die Menge der Erde an, die vom Hügel p_i zum Loch q_j transportiert wird.
- Die Transportkosten berechnen sich aus dem Produkt der Quantitäten mit den entsprechenden Grunddistanzwerten:

$$\text{Kosten}(p, q, F) = \sum_{i=1}^{m} \sum_{j=1}^{n} d(p_i, q_j) f_{ij}$$

Earth-Mover-Distanzfunktion

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

Diese Kosten sind unter Berücksichtigung folgender Bedingungen zu minimieren:

$$f_{ij} \leq 0 \quad : \quad 1 \leq i \leq m, 1 \leq j \leq n \tag{1}$$

$$\sum_{j=1}^{n} f_{ij} \leq w_{p_i} \quad : \quad 1 \leq i \leq m \tag{2}$$

$$\sum_{i=1}^{m} f_{ij} \leq w_{q_j} \quad : \quad 1 \leq j \leq n \tag{3}$$

$$\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min \left(\sum_{i=1}^{m} w_{p_i}, \sum_{j=1}^{n} w_{g_j} \right) \tag{4}$$

Earth-Mover-Distanzfunktion

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

Der Distanzwert nach der EM-Distanzfunktion berechnet sich nun aus der Normierung der minimalen Transportkosten bezüglich der Gesamtmenge der transportierten Erde:

$$d_{\mathrm{EM}}(p,q) = \frac{\min\limits_{|f_{ij}|} \left(\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} d(p_i, q_j) f_{ij} \right)}{\sum\limits_{i=1}^{m} \sum\limits_{j=1}^{n} f_{ij}}$$

DFT-L_2-Distanzfunktion

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- Diese Distanzfunktion ist zum Vergleich von Sequenzen reeller Werte mit einer Positionskorrespondenz und fester Länge geeignet.
- Die Grundidee liegt in der Verwendung der euklidischen Distanzfunktion auf den einzelnen korrespondierenden Sequenzwerten.
- Ein typisches Beispiel für Sequenzen, bei denen diese Distanzfunktion sinnvoll angewendet werden kann, sind Zeitreihen. Z. B. können Tierpopulationen, Pegelstände, aber auch Aktienkursverläufe als zeitabhängige Werte miteinander verglichen werden.

Editierdistanz

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- Ein Beispiel für eine Distanz zwischen Sequenzen mit nominalen Werten anhand einer abgeschwächten Positionskorrespondenz.
- Die Editierdistanz misst den minimalen Aufwand, um eine Zeichenkette mittels Editieroperationen in eine andere Zeichenkette zu überführen.
- Beispiel - die Editierdistanz zwischen den Wörtern “Abend” und “Robe” beträgt 4 (Ersetzen von “A” durch “R”, Einfügen von “o”, Entfernen von “n” und “d”).

Overview

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

Distanzfunktionen auf Binärdaten

Distanzfunktionen auf Sequenzen

Distanzfunktionen auf allgemeinen Mengen

Allgemeines

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- Bisher waren die Datentypen der zu vergleichenden Objekte eingeschränkt. Jetzt soll ein Vergleich auf allgemeinen Mengen erfolgen.

- Beispiele:
  - die Bottleneck-Distanzfunktion,
  - die Distanzfunktion über das Volumen der symmetrischen Differenz,
  - die Hausdorff-Distanzfunktion, und
  - die Frechet-Distanzfunktion

Bottleneck-Distanzfunktion

5.3 DF auf Binärdaten

5.4 DF auf Sequenzen

5.5 DF auf allgemeinen Mengen

- d_B ist auf endlichen Untermengen einer Menge X mit einer gegebenen Grunddistanz $d_X : X \times X \rightarrow \mathbb{R}_0$ definiert.
- Die Kardinalität beider Untermengen muss gleich sein

$$A, B \subset X \quad \text{mit} \quad |A| = |B|$$

- Zwischen den Elementen der Untermengen existiere eine bijektive Abbildung f. Man ist an der Distanz d_X des am weitesten auseinanderliegenden Elementepaars interessiert. Das Minimum der maximalen Elementepaardistanzen über allen möglichen Bijektionen $f \in F(A, B)$ wird gesucht:

$$d_B(A, B) = \min_{f \in F(A,B)} \max_{a \in A} d_X(a, f(a))$$