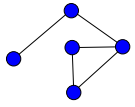
Gliederung

Peer-to-peer Systeme und Datenbanken(SS07)

- Kapitel 1: Einführung
- Kapitel 2: Beispiele
- Kapitel 3: Routing
- Kapitel 4: Schemabasierte p2p-Netzwerke
- Kapitel 5: Integrationsprobleme
  Teil 5-1: Einführung, Gleichheit
  Teil 5-2: Ähnlichkeit - 1
  Teil 5-3: Ähnlichkeit - 2
  Teil 5-4: Mappingbasierte Datenintegration
- Kapitel 6: Anonymität, Authentifikation
- Kapitel 7: Reputation

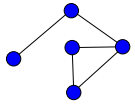
Version vom 4. Juni 2007



Kapitel 5

Integrationsprobleme

- Problembeschreibung
- Vergleiche auf Daten
- Ähnlichkeit auf Daten
  - Ähnlichkeit wegen formaler Merkmale
  - inhaltsbasierte Ähnlichkeit
- Mappingbasierte Datenintegration

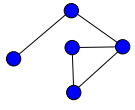


Ähnlichkeitssuche

Motivation: Gleichheit oft zu restriktiv.

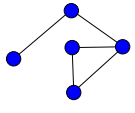
- Zahlen: Messfehler, Fehlerfortpflanzung in num. Algorithmen, instabile Algorithmen, Fehler durch Abschreiben, Ablesen, ...
- Suche in Bildern
- Einige Fehler erkennbar oder korrigierbar (fehlererkennende , -korrigierende Kodierung)- Redundanz,
  Vergleich mit theoretischen Werten.

Definition *Ähnlichkeit* verschieden möglich:
Semantische Stufe: i.a. bessere Ergebnisse
Syntaktische Stufe: formal, funktioniert auch ohne semantisches Wissen, i.a. schwächere Ergebisse.



Anwendungsszenarien

- Gesichtserkennung: markante Punkte und Verhältnisse der Entfernungen zwischen diesen, ggf. unter Beachtung von Projektionen.
- Bilder: Farbhistogramme, ...
- Klänge: Spectrum (Fourier-Analyse), ggf. zeitlicher Verlauf.
- Zeichenketten:(flexible Hilfsstruktur der Informatik)
  Soundex - Vergleich
  n-Gramm-Analyse
  edit-distance (Levenstein - Distance)



Ähnlichkeitsbegriffe

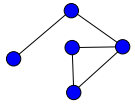
- Geometrische Ähnlichkeit: Elementargeometr. Begriff für Dreiecke, auf weitere elementargeometr. Flächen und Körper mit Hilfe der Proportionalität verallgemeinerbar.

  Diese Ähnlichkeit induziert Einteilung in Äquivalenzklassen (RST)

- Kern: Ähnlichkeit durch charakteristische Merkmale beschrieben. Auswahl kontextabhängig.

- Ähnlichkeit = „geringer Abstand “.
  erfordert Abstandsdefinition, dadurch flexibel an viele Probleme anpassbar.

  Aber i.a. nicht transitiv, d.h. keine Äquvalenzrelation
  Beispiel: a, b Elemente einer betrachteten Gesamtheit,
  Definition: Geg. $\varepsilon > 0$, dann $a \sim b \Leftrightarrow |a - b| < \varepsilon$.
  Schon im einfachsten Fall: $a, b, c \in \mathcal{N}$, $\varepsilon = 1, 1, a = 2, b = 3, c = 4$ gilt:
  $a \sim b$, $b \sim c$, **aber nicht** $a \sim c$.
  Vergleichbar: Clusterbildung um vorgegebenen Mittelpunkt.

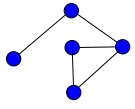


Soundex-Algorithmus

- Entstehung: Robert C. Russel - 2.April 1918 - Patent Nr. 1 261 167
- Algorithmus
  - Code $\Leftarrow$ 1.Buchstabe + 3 Ziffern
  - Streiche ab dem 2.Buchstaben alle
    a, e, i, o, u, h, w, y
    und füge 3 Ziffern nach Tabelle hinzu:

1	$\Leftarrow$	b,f,p,v
2	$\Leftarrow$	c, g, j, k, q, s, x, z
3	$\Leftarrow$	d, t
4	$\Leftarrow$	l
5	$\Leftarrow$	m, n
6	$\Leftarrow$	r

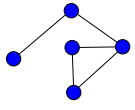
  - Beachte folgende Regeln:
    1. 2 aufeinanderfolgende Buchstaben mit demselben Kode
       $\rightarrow$ nur 1x
    2. Ergeben sich weniger als 3 Ziffern $\rightarrow$ Auffüllen mit 0 (Null).
- transitiv



Soundex-Algorithmus (2)

- Tabelle:

1	$\Leftarrow$	b,f,p,v	labial
2	$\Leftarrow$	c, g, j, k, q, s, x, z	heterogen: frikativ; plosiv,velar
3	$\Leftarrow$	d, t	plosiv, dental/alveolare
4	$\Leftarrow$	l	lateral
5	$\Leftarrow$	m, n	nasal
6	$\Leftarrow$	r	Vibranten



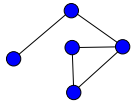
Soundex-Algorithmus (2)

- Tabelle:

1	$\Leftarrow$	b,f,p,v	labial
2	$\Leftarrow$	c, g, j, k, q, s, x, z	heterogen: frikativ; plosiv,velar
3	$\Leftarrow$	d, t	plosiv, dental/alveolare
4	$\Leftarrow$	l	lateral
5	$\Leftarrow$	m, n	nasal
6	$\Leftarrow$	r	Vibranten

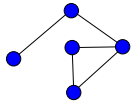
- Beispiele:

Name	Code	Bemerkung
Miller	M460	Auffüllen (Regel 1), Doppel-l (Regel 2)
Peterson	P362	3 verschiedene Konsonanten
Peters	P362	
Moskovitz	M232	
Moskowitz	M213	Fehlkodierung nichtenglischer Namen
ψ, ξ	?	z.B. $\psi \Rightarrow 12$



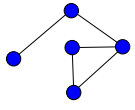
Idee hinter Soundex

- Begriffe *Phone, Phoneme* einer Sprache
- Phonemfeststellung: zwei Worte einer Sprache, die sich nur in einem Phon unterscheiden.
  Beispiel: ehren - lehren $\leftarrow$ das Phon (l) ist ein Phonem.
- Phoneme können in Klassen eingeteilt werden nach der Art und dem Ort ihrer Entstehung beim Sprechen.
- Phoneme werden in Schriftsprache durch Grapheme dargestellt - M:N - Abbildung.
- Aussprache eines Graphems kann kontextabhängig sein.
  Internationales Phonet. Alphabet (z.B. 28 a-Varianten)
- Phoneme (und zugeordnete Grapheme variieren mit Sprache), d.h. phonetische Suche muss sprachspezifisch sein.



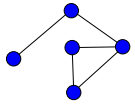
Entwicklungsschritte

- Analyse des Phonembestandes
- Klasseneinteilung der Phoneme, sinnvolle Vereinfachung des Klassensystems. Jede Reduzierung der Klassenzahl macht die Suche unschärfer, aber fehlertoleranter.
- Voraussetzung: Sprache wird durch Folgen von Graphemen beschrieben.
  Beschreibung der Zuordnung von Graphemen und Graphemfolgen zu Phonemen und Phonemfolgen.
  M:N Abbildung
  Konstruktion eines Automaten, der die Umwandlung vornimmt.



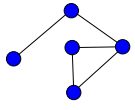
Besonderheiten des Neugriechischen

- Konsonantverdopplungen nicht hörbar bei β, λ, μ, ν, π, σ, τ, da alle Vokale kurz, aber $\gamma\gamma \mapsto$ ng.
  Weitere Kombinationen: $\alpha\upsilon$, $\varepsilon\upsilon$, $o\upsilon \mapsto$ Phonemfolgen av, af ,ev, ef bzw. u.
  ψ, $\xi \mapsto$ Phonemfolge ps, ks



Besonderheiten des Neugriechischen

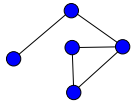
- Konsonantverdopplungen nicht hörbar bei β, λ, μ, ν, π, σ, τ, da alle Vokale kurz, aber $\gamma\gamma \mapsto$ ng.
  Weitere Kombinationen: $\alpha\upsilon$, $\varepsilon\upsilon$, $o\upsilon \mapsto$ Phonemfolgen av, af ,ev, ef bzw. u.
  ψ, $\xi \mapsto$ Phonemfolge ps, ks
- Mehrere Schreibungen ein Phonem:
  Phonem **i**: ι, η, υ, $\varepsilon\iota$, $o\iota$, $\upsilon\iota$; Phonem **e**: ε, $\alpha\iota$; Phonem **o**: o, ω



Besonderheiten des Neugriechischen

- Konsonantverdopplungen nicht hörbar bei β, λ, μ, ν, π, σ, τ, da alle Vokale kurz, aber γγ ↦ ng.
  Weitere Kombinationen: αυ, ευ, ου ↦ Phonemfolgen av, af ,ev, ef bzw. u.
  ψ, ξ ↦ Phonemfolge ps, ks
- Mehrere Schreibungen ein Phonem:
  Phonem **i**: ι, η, υ, ει, οι, υι; Phonem **e**: ε, αι; Phonem **o**: ο, ω
- Wechselwirkung der Aussprache mit Betonung, Silbenanfang:

βάκιλοι	-	v'akili
φιλοί	-	fil'i
ρολόι	-	rol'oi

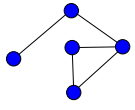


Besonderheiten des Neugriechischen

- Konsonantverdopplungen nicht hörbar bei β, λ, μ, ν, π, σ, τ, da alle Vokale kurz, aber γγ ↦ ng.
  Weitere Kombinationen: αυ, ευ, ου ↦ Phonemfolgen av, af ,ev, ef bzw. u.
  ψ, ξ ↦ Phonemfolge ps, ks
- Mehrere Schreibungen ein Phonem:
  Phonem **i**: ι, η, υ, ει, οι, υι; Phonem **e**: ε, αι; Phonem **o**: ο, ω
- Wechselwirkung der Aussprache mit Betonung, Silbenanfang:

βάκιλοι	-	v'akili
φιλοί	-	fil'i
ρολόι	-	rol'oi

- Unsicherheit bei μπ, ντ und γκ,

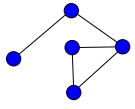


Besonderheiten des Neugriechischen

- Konsonantverdopplungen nicht hörbar bei β, λ, μ, ν, π, σ, τ, da alle Vokale kurz, aber γγ ↦ ng.
  Weitere Kombinationen: αυ, ευ, ου ↦ Phonemfolgen av, af ,ev, ef bzw. u.
  ψ, ξ ↦ Phonemfolge ps, ks
- Mehrere Schreibungen ein Phonem:
  Phonem **i**: ι, η, υ, ει, οι, υι; Phonem **e**: ε, αι; Phonem **o**: ο, ω
- Wechselwirkung der Aussprache mit Betonung, Silbenanfang:

βάκιλοι	-	v'akili
φιλοί	-	fil'i
ρολόι	-	rol'oi

- Unsicherheit bei μπ, ντ und γκ,
- Für die Grapheme δ und θ gibt es im (Hoch-) Deutschen keine adäquaten Phoneme.

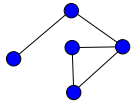


Besonderheiten des Neugriechischen

- Konsonantverdopplungen nicht hörbar bei β, λ, μ, ν, π, σ, τ, da alle Vokale kurz, aber γγ ↦ ng.
  Weitere Kombinationen: αυ, ευ, ου ↦ Phonemfolgen av, af ,ev, ef bzw. u.
  ψ, ξ ↦ Phonemfolge ps, ks
- Mehrere Schreibungen ein Phonem:
  Phonem **i**: ι, η, υ, ει, οι, υι; Phonem **e**: ε, αι; Phonem **o**: ο, ω
- Wechselwirkung der Aussprache mit Betonung, Silbenanfang:

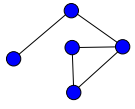
βάκιλοι	-	v'akili
φιλοί	-	fil'i
ρολόι	-	rol'oi

- Unsicherheit bei μπ, ντ und γκ,
- Für die Grapheme δ und ϑ gibt es im (Hoch-) Deutschen keine adäquaten Phoneme.
- Lautverschiebung: z.B. κτ ↦ χτ, σϑ ↦ στ.



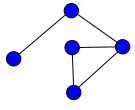
Löungsvorschlag

- Zwei-/drei-stufiges Verfahren, kann nach jeder Stufe gestoppt werden.



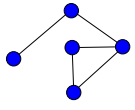
Löungsvorschlag

- Zwei-/drei-stufiges Verfahren, kann nach jeder Stufe gestoppt werden.
  1. Beseitigung der unhörbaren Varianten,
     griechische Umkodierung, Betonung bleibt erhalten:
     ca. 30 Regeln.



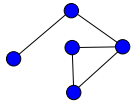
Löungsvorschlag

- Zwei-/drei-stufiges Verfahren, kann nach jeder Stufe gestoppt werden.
  1. Beseitigung der unhörbaren Varianten,
     griechische Umkodierung, Betonung bleibt erhalten:
     ca. 30 Regeln.
  2. Nachbilden der Phonetik (Lautfolgen, Zeichenfolgen),
     Reduktion (stimmhaft $\mapsto$ stimmlos), **i**-Varianten:
     ca. 60 weitere Regeln,
     Transcription auf Folgen von Phonemen aus ca. 15 Phonemklassen.
     Klassen durch Zeichen des lateinischen Alphabets beschrieben.



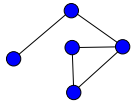
Löungsvorschlag

- Zwei-/drei-stufiges Verfahren, kann nach jeder Stufe gestoppt werden.
  1. Beseitigung der unhörbaren Varianten,
     griechische Umkodierung, Betonung bleibt erhalten:
     ca. 30 Regeln.
  2. Nachbilden der Phonetik (Lautfolgen, Zeichenfolgen),
     Reduktion (stimmhaft $\mapsto$ stimmlos), **i**-Varianten:
     ca. 60 weitere Regeln,
     Transcription auf Folgen von Phonemen aus ca. 15 Phonemklassen.
     Klassen durch Zeichen des lateinischen Alphabets beschrieben.
  3. (Soundex auf dem Ergebnis von (2) für Spezialfälle).



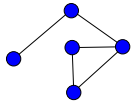
Löungsvorschlag

- Zwei-/drei-stufiges Verfahren, kann nach jeder Stufe gestoppt werden.
  1. Beseitigung der unhörbaren Varianten,
     griechische Umkodierung, Betonung bleibt erhalten:
     ca. 30 Regeln.
  2. Nachbilden der Phonetik (Lautfolgen, Zeichenfolgen),
     Reduktion (stimmhaft $\mapsto$ stimmlos), **i**-Varianten:
     ca. 60 weitere Regeln,
     Transcription auf Folgen von Phonemen aus ca. 15 Phonemklassen.
     Klassen durch Zeichen des lateinischen Alphabets beschrieben.
  3. (Soundex auf dem Ergebnis von (2) für Spezialfälle).
- Kein Weglassen der Vokale, keine Längenbegrenzung



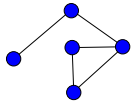
Löungsvorschlag

- Zwei-/drei-stufiges Verfahren, kann nach jeder Stufe gestoppt werden.
  1. Beseitigung der unhörbaren Varianten,
     griechische Umkodierung, Betonung bleibt erhalten:
     ca. 30 Regeln.
  2. Nachbilden der Phonetik (Lautfolgen, Zeichenfolgen),
     Reduktion (stimmhaft $\mapsto$ stimmlos), **i**-Varianten:
     ca. 60 weitere Regeln,
     Transcription auf Folgen von Phonemen aus ca. 15 Phonemklassen.
     Klassen durch Zeichen des lateinischen Alphabets beschrieben.
  3. (Soundex auf dem Ergebnis von (2) für Spezialfälle).
- Kein Weglassen der Vokale, keine Längenbegrenzung
- Kontextsensitive Regeln in jeder Stufe, Datenstrom.
  Muster: (Zeichen, nächst. Zeichen) $\mapsto$ (Aktion, Fortsetzungspunkt)



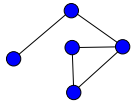
Löungsvorschlag

- Zwei-/drei-stufiges Verfahren, kann nach jeder Stufe gestoppt werden.
  1. Beseitigung der unhörbaren Varianten,
     griechische Umkodierung, Betonung bleibt erhalten:
     ca. 30 Regeln.
  2. Nachbilden der Phonetik (Lautfolgen, Zeichenfolgen),
     Reduktion (stimmhaft $\mapsto$ stimmlos), **i**-Varianten:
     ca. 60 weitere Regeln,
     Transcription auf Folgen von Phonemen aus ca. 15 Phonemklassen.
     Klassen durch Zeichen des lateinischen Alphabets beschrieben.
  3. (Soundex auf dem Ergebnis von (2) für Spezialfälle).
- Kein Weglassen der Vokale, keine Längenbegrenzung
- Kontextsensitive Regeln in jeder Stufe, Datenstrom.
  Muster: (Zeichen, nächst. Zeichen) $\mapsto$ (Aktion, Fortsetzungspunkt)
- Dreistufiger Ansatz auf andere Sprachen verallgemeinerungsfähig.
  Dritte Stufe dem Thema Wörterbuch nicht angemessen.



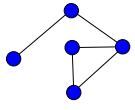
Beispiele aus den Regeln der 1. Sufe

Z	nZ	Bedingung, Bemerkung	Phonem	Code	Schritt
α	ι		e	ε	+
α	ί		e	έ	+
α	υ	nicht bearbeiten	af, av	αυ	+
α	ύ	nicht bearbeiten	af, av	αύ	+
β	β		v	β	+
ε	ι		i	ι	+
η			i	ι	
ο	ι		i	ι	+
ο	ί		i	ί	+
ο	υ	nicht bearbeiten	u	ου	+
υ			i	ι	
ύ			i	ί	
ϋ			i	ϊ	
ΰ			i	ΐ	
ω			o	ο	



2. Stufe

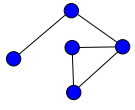
- - Buchstabenkombinationen auflösen: αυ, ευ, μπ, ντ
  - Kombinationen γγ, γκ, κκ bearbeiten
  - **i**-Allophone
  - Stimmhafte Konsonanten → stimmlose, z.B. **z** (ζ) → s
  - ψ, χ → ps, ks
  - Vokalverdopplungen entfernen



2. Stufe

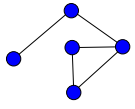
- - Buchstabenkombinationen auflösen: αυ, ευ, μπ, ντ
  - Kombinationen γγ, γκ, κκ bearbeiten
  - **i**-Allophone
  - Stimmhafte Konsonanten → stimmlose, z.B. **z** (ζ) → s
  - ψ, χ → ps, ks
  - Vokalverdopplungen entfernen
- Beispiele:

φιλοξενία	filoksenia
οινοποιείο	inopj'io → inopio



Phoneme für Konsonanten

	Artikulationsart				
	momentan		koninuierlich		
	Klusile		Frikative		Sonoranten
	stimmlos	stimmhaft	stimmlos	stimmhaft	
Labiale	p	b	f	v	m (Nasal)
Dentale	t	d	ϑ	δ	n (Nasal)
Alveolare	t^s	d^z	s	z	r (Tremulant)
Velare	k	g	(i/a)ch	γ (j)	l (Lateral)

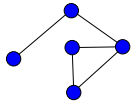


Phoneme für Konsonanten

	Artikulationsart				
	momentan		koninuierlich		
	Klusile		Frikative		Sonoranten
	stimmlos	stimmhaft	stimmlos	stimmhaft	
Labiale	p p	b p	f f	v f	m (Nasal)
Dentale	t t	d t	ϑ t?s	δ s	n (Nasal)
Alveolare	t^s ts	d^z [2] ts	s s	z s	r (Tremulant)
Velare	k k	g k	(i/a)ch c,h	γ(j) [1] j$\rightarrow$i	l (Lateral)

[1] γιορτάζω - jort'azo $\rightarrow$ jortaso iortaso

[2] τζατζίκι - dzadz'iki $\rightarrow$ tsatsiki

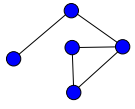


Konsonanten - Probleme

- Stimmhaftes b (μπ) und d (ντ)

  λάμπα - l'amba → lampa

  μπαμπάς - bab'as → papas

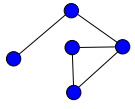


Konsonanten - Probleme

- Stimmhaftes b ($\mu\pi$) und d ($\nu\tau$)

  $\lambda\acute{\alpha}\mu\pi\alpha$ - l'amba $\rightarrow$ lampa warum nicht lapa

  $\mu\pi\alpha\mu\pi\acute{\alpha}\varsigma$ - bab'as $\rightarrow$ papas warum nicht pampas



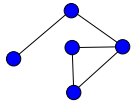
Konsonanten - Probleme

- Stimmhaftes b ($\mu\pi$) und d ($\nu\tau$)

  $\lambda\acute{\alpha}\mu\pi\alpha$ - l'amba $\longrightarrow$ lampa warum nicht lapa

  $\mu\pi\alpha\mu\pi\acute{\alpha}\varsigma$ - bab'as $\longrightarrow$ papas warum nicht pampas

- Unbetontes Phonem **i** nach Konsonanten oder vor Vokal
  Verschiedene Möglichkeiten: i, j, j-ähnlich, (schwach) (i)ch, Ausfall
  $\mapsto$ Ursache vieler Regeln



Konsonanten - Probleme

- Stimmhaftes b (μπ) und d (ντ)

  λάμπα - l'amba → lampa warum nicht lapa

  μπαμπάς - bab'as → papas warum nicht pampas

- Unbetontes Phonem **i** nach Konsonanten oder vor Vokal
  Verschiedene Möglichkeiten: i, j, j-ähnlich, (schwach) (i)ch, Ausfall
  ↦ Ursache vieler Regeln

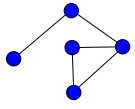
- Wort - phonetische Beschreibung M:N

  μπαμπάς - bab'as → papas und pampas

  παπάς - pap'as → papas

  για - ja → ia

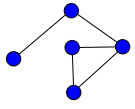
  γεια - ja → ia



Konsonanten - griechische Schreibung

	Artikulationsart				
	momentan		koninuierlich		
	Klusile		Frikative		Sonoranten
	stimmlos	stimmhaft	stimmlos	stimmhaft	
Labiale	π p p	μπ b p	φ f (α,ε,ι)υ f	β v (α,ε,ι)υ f	μ m μ(π) m(p)
Dentale	τ t t	ντ d t	ϑ t?s	δ s	ν n, ντ nt γ(γ,κ) n^k
Alveolare	τσ t^s ts	τζ d^z [2] ts	σ,ς s s	ζ z s	ρ r
Velare	κ k k	γ(α,ο,ου) g k	χ(ι,α) (i/a)ch c,h	γ(ε,ι)(j) j→i	λ l

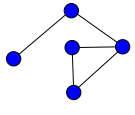
16 Phonemklassen: **p, t, k, f, s, (i)c(h), (ac)h, m, n, r, l; a, e, i, o, u**



Realisierung

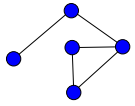
- Diplomarbeit abgeschlossen
- Prototyp:

URL http://teiresias.uni-leipzig.de



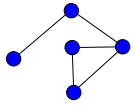
Mathematischer Abstandsbegriff

- Funktionalanalysis - Metrische Räume
  Seien $\mathcal{D}$ eine Vektorraum, ρ eine Abbildung, ρ: $\mathcal{D} \times \mathcal{D} \mapsto \mathcal{R}^+ \cup \{0\}$ mit:
  i : $\rho(x, y) \geq 0$ für $x, y \in \mathcal{D}$, $\rho = 0 \leftrightarrow x = y$.
  ii: $\rho(x, y) = \rho(y, x)$, $x, y \in \mathcal{D}$ (Symmetrie)
  iii: $\rho(x, y) + \rho(y, z) \leq \rho(x, z)$, $x, y, z \in \mathcal{D}$ (Dreiecksungleichung),
  dann heißt $(\mathcal{D}, \rho)$ metrischer Raum.
- ohne die Bedingung $\rho = 0 \leftrightarrow x = y$: Pseudometrik
- Informatik: $\mathcal{D}$ sei (nur) eine Menge.
- Zu einer Menge kann es mehrere, verschiedene Abstandsdefinitionen geben ($\rightarrow$ verschiedene Räume)
- Sei $\mathcal{B}$ ein normierter Raum mit der Norm $\|.\|$, dann ist $\rho(x, y) = \|x - y\|$ eine Metrik.
- Nicht aus Norm erzeugt: *Diskrete Metrik*: $\rho = 0 \leftrightarrow x = y, \rho = 1$ sonst.



Beispiele normierter Räume

- $\mathcal{D}$ Menge der über einem abgeschlossenen Intervall I stetigen Funktionen f:
  $$\|f\| = \max_{x \in I}(|f(x)|).$$
- L_1, L_p: $\mathcal{D}$ Menge der messbaren Funktionen über einem abgeschlossenen Intervall I mit
  $$\int_I |f|^p dx < \infty, 1 \leq p < \infty, \text{ fest.}$$
- L_∞: $\mathcal{D}$ Menge der messbaren Funktionen über einem abgeschlossenen Intervall I mit
  $$\text{ess sup}_{x \in I}(|f(x)|) < \infty.$$



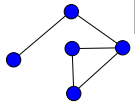
Beispiele - diskreter Fall

- l_1, l_p: $\mathcal{D}$ Menge der Folgen $\tilde{a} = \{a_i\}_{i=1}^{\infty}$ mit
  $\sum_{i=1}^{\infty}(|a_i|^p) < \infty$, $1 \leq p < \infty$, fest.
- l_∞: $\mathcal{D}$ Menge der Folgen $\tilde{a} = \{a_i\}_{i=1}^{\infty}$ mit
  $\max_i(|a_i|) < \infty$.

Endlich viele Folgenglieder: $\tilde{a} = \{a_i\}_{i=1}^{m}$

- $\mathcal{D}$ Menge der Folgen $\tilde{a} = \{a_i\}_{i=1}^{m}$ mit
  $\sum_{i=1}^{m}(|a_i|^p) < \infty$, $1 \leq p < \infty$, fest.
  $p = 1$ führt auf die Manhattan-Metrik,
  $p = 2$ auf die Euklidische.
- $\mathcal{D}$ Menge der Folgen $\tilde{a} = \{a_i\}_{i=1}^{m}$ mit
  $\max_i(|a_i|) < \infty$.

Freiwillige Übungsaufgabe: Skizzieren Sie für $m = 2$ das Aussehen des Einheitskreises in Abhängigkeit von p.

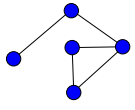


Distanzfunktionen mit Gewichten

Sei $\mathcal{A}$ eine positiv semidefinite m-reihige Matrix.
$x, y \in \mathcal{R}^m$.

- Gewichtete Distanzfunktion:
$$\rho(\mathcal{A}; x, y) = \left((x-y)^T \mathcal{A}(x-y)\right)^{1/2}$$
Anwendung: Modellierung eines Farbkreises.

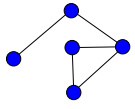
- Sonderfall
$\mathcal{A}$ hat Diagonalgestalt: Euklidische Distanz mit Gewichtung der Achsenrichtungen.



Hausdorffdistanz

Abstand zweier kompakter Mengen $\mathcal{A}, \mathcal{B}$ eines metrischen Raumes $\mathcal{R}$, Metrik $d(.,.)$ kompakte M. im metr.Raum: Grenzwert einer konverg. Folge gehört zur Menge.

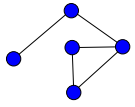
- gerichteter Abstand:
  $d1_H(\mathcal{A}, \mathcal{B}) = \max(\sup_{a \in \mathcal{A}} \inf_{b \in \mathcal{B}} d(a, b)$
- Hausdorff-Distanz:
  $d_H(\mathcal{A}, \mathcal{B}) = \max(\sup_{a \in \mathcal{A}} \inf_{b \in \mathcal{B}} d(a, b), \sup_{b \in \mathcal{B}} \inf_{a \in \mathcal{A}} d(a, b))$
- Verbal:
  Zwei Mengen haben eine HD von höchstens r voneinander, g.d.w. jeder Punkt einer Menge ist innerhalb eines Abstandes r von einem Punkt der anderen.



Hamming-Distanz

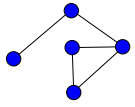
Richard W. Hamming. Error Detecting and Error Correcting Codes,
Bell System Technical Journal 26(2):147-160, 1950.

- Gegeben:
  Alphabet $\mathcal{A}$, 2 Zeichenketten $a = \{a_i\}_{i=1}^{n}, b = \{b_i\}_{i=1}^{n}$ der Länge n.
  $d_H(a,b) = \sum_{i=1, a_i \neq b_i}^{n}(1)$
  d_H ist eine Metrik auf der Menge der Zeichenketten der Länge n
- $\mathcal{A} = \{0,1\}$, Zeichenkette: Binärzahlen der Länge n
  $d_H(a,b)$ = Anzahl der 1-Zeichen in a xor b.
  Darstellung des Übergangs von a nach b als Kantenfolge in einem n-dimensionalen Hyper-Würfel.
  Beispiele: http://en.wikipedia.org/wiki/Hamming_distance
- Fehlerkorrektur - suche nach korrektem Code mit kleistem Abstand.



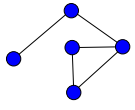
Levenstein-Distanz

- Auch: edit-distance
- **Definition:** Gegeben zwei Zeichenketten $x = \{x_i\}_{i=1}^{n}$, $y = \{y_j\}_{j=1}^{m}$.
  Grundoperationen mit Gewicht
  insert(x, c, l): fügt in Zeichenkette x das Zeichen c an der Position l ein. Gewicht g_i.
  delete$((x, l)$: löscht in Zeichenkette x das Zeichen an der Position l. Gewicht g_d.
  replace(x_l, c, l): ersetzt in Zeichenkette x das Zeichen an der Position l durch c. Gewicht g_r.
  Gesucht: eine Folge von Grundoperationen minimalen Gesamtgewichts d (= Summe der Gewichte), die x in y überführt.
  Das **Gesamtgewicht einer Minimalfolge ist die Levenstein-Distanz** von x und y.



Levenstein-Distanz (Verallgemeinerungen)

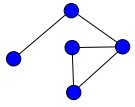
- Gültigkeit einer Dreiecksungleichung für Gewichte für Operationen an einer Position - jede Position nur einmal bearbeitet.
- Die Gewichte können abhängen vom Zeichen (sowohl dem zu ersetzenden und dem ersetzenden) (unsymmertr. Metriken, symmetrisierbar)
- Verallgemeinerung auf Baumstrukturen



Levenstein-Distanz (Berechnung)

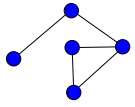
- Idee: Berechnung der Distanz aller möglichen Präfix-Paare der zwei Zeichenketten x, Y.
- $x = ua$, $y = vb$.

$$g(ua, vb) = \min \begin{cases} g_d(x, .) + g(u, vb) & - \text{ Loeschen von a} \\ g_i(., b, .) + g((ua, v) & - \text{ Einfuegen von b} \\ g_r(a, b, .) + g(u, v) & - \text{ Ersetzen a durch b} \end{cases}$$



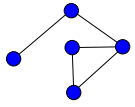
Levenstein-Distanz für Baumstrukturen

- Definition: Ein *Baum* besteht aus einem Knoten und einer daran angehängten, geordneten Folge disjunkter Bäume. Eine solche Folge heißt *Wald*.
- Grundoperationen: (jeweils mit Kosten zu versehen)
  Ersetzen eines Knotens (ändert Baumstruktur nicht)
  Einfügen eines Knotens (verschiebt den neuen Wald)
  Löschen eines Knotens (verschiebt den Wald).
- Gegeben zwei Wälder $\mathcal{F}, \mathcal{G}$. Sei $\mathcal{X}$ die Menge aller Folgen von Grundoperationen, deren Hintereinanderausführung $\mathcal{F}$ in $\mathcal{G}$ überführt. Die Editier-Distanz $d(\mathcal{F}, \mathcal{G})$ ist das kleinste Gesamtgewicht eines Elenents aus $\mathcal{X}$
- Algorithmen: Tai - 1979: $O(n^6)$, Zhang-Shasha - 1989:$O(n^4)$, Klein - 1998: $O(n \log n)$.
  Forschungsgegenstand. (2004, 2005, ...)



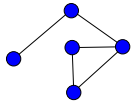
Ähnlichkeit von Vektoren

- vgl.: Math. Abstandsbegriff - normierte Räume sind Vektorräume.
- Hilberträume $\mathcal{R}$: Skalarprodukt $(.,.)$ (verträglich mit Norm)
- $d(x,y) = 1 - \frac{|(x,y)|}{\|x\| \times \|y\|}$ für $x, y \in \mathcal{R}$.
  Anschaulich im $\mathcal{R}^2$: $d(x,y) = 1 - |\cos(x,y)|$
  d.h. Abstand gering - fast gleiche Richtung.
  Verallgemeinerung: ohne Betrag
- Ähnlichkeit nach TANIMOTO
  $d(x,y) = 1 - \frac{(x,y)}{\|x\|^2 + \|y\|^2 - (x,y)}$ für $x, y \in \mathcal{R}$.
  Vergleiche von Molekülstrukturen in Bio-DB und Chemie-DB: Fingerprint - Bitkette



Ähnlichkeit von BIT-Vektoren

- Chemie, Biochemie: es existiert eine Liste von n Bestandteilen. Molekul=$(t_1, t_2, ..., t_n)$. T_j gibt an, ob das j-te Element der Liste im Molekül vorkommt (1) oder nicht (0)
- Jaccard-Koeffizient / Tanimoto-Index:
  s, t zwei Bit-Vektoren der Länge n.
  Tanimoto-Index:
  $T(s,t) = \sum_{j=1;s_j=1 \wedge t_j=1}^{n}(1) / \sum_{j=1;s_j=1 \vee t_j=1}^{n}(1)$
  Ähnlichkeit = 1 - T.
- Gemeinsames Fehlen wird ignoriert.



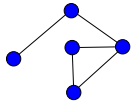
Ähnlichkeit von BIT-Vektoren II

- s, t, n wie eben,

  a Anzahl Übereinstimmungen und gleich 1,
  b Anzahl Nichtüberieinst. und $a_j = 1$,
  c Anzahl Nichtüberieinst. und $b_j = 1$,
  d Anzahl Übereinstimmungen und gleich 0.

- Auswahl:

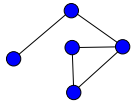
Kovarianze	$a/n - ((a+b)/n \times (a+c)/n)$	
Jaccard	$a/(a+b+c)$	Neg. Überinst. ignor.
Dice	$2a/(2a+b+c)$	Neg. Ü. ign., pos. Ü. dopp.
Russel-Rao	$a/(a+b+c+d)$	
Sokal-Sneath	$a/(a+2(b+c)$	
Normal	$(a+d)/(a+b+c+d)$	
...		



N-Gramme

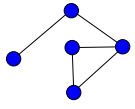
Gegeben: Zeichenketten $a, b, ...$ über einem Alphabet $\mathcal{A}$

- Ein N-Gramm x ist eine Zeichenkette über $\mathcal{A}$, die aus N Zeichen besteht. a enthält x, g.d.w. x Teilzeichenkette von a ist.
  In praxi: Bi- und Trigramme.
- Idee: Gleiche Zeicheketten = gleiche N-Gramme.
  Zwei Zeichenketten sind ähnlich, wenn sie viele gemeinsame N-Gramme haben:
  $s = 2c/(n+m)$ c Zahl der gemeinsamen Trigramme t (mit Widerholung)
  m, n Längen der Zeichenketten
- Alternativ: Virtuell zwei Leerzeichen am Anfang und Ende: bessere Bewertung dieser Stellen.
  $s = 2c/((n+2)+(m+2))$ c Zahl der gemeinsamen Trigramme
  m, n Längen der Zeichenketten



N-Gramme (2)

- Kritisch:
  Festlegung der Ähnlichkeitsschwelle problemabhängig.
  In Berechnung gehen die Längen n, m ein.
- Anwendungen:
  Wichtiges Hilfsmittel zur Zeichenketteanalyse
  - Ähnlichkeit (s.o)
  - Anzahl der Bi- und Trigramme über einem Alphabet bekannt:
    $card(\mathcal{A})^2$,$card(\mathcal{A})^3$.
    Häufigkeitsverteilungen für Sprache oder Fachsprachen charakteristisch.
    $\Rightarrow$ Erkennung der Sprache eines Textes
    $\Rightarrow$ Zuordnung eines Textes zu Fachgebiet (zusammen mit Wortanalyse)
    $\Rightarrow$ Bestandteil der Kryptoanalyse.



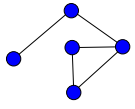
N-Gramme (3)

Reinhard Rapp: Die Berechnung von Assoziationen: ein korpuslinguistischer Ansatz. Hildesheim; Zürich; New York: Olms, 1996. ISBN: 3-487-10252-8 (Diss.)

- Trigrammähnlichkeit intuitiv gut.
- Einfache Buchstabendreher $\rightarrow$ unterbewertete Ähnlichkeit
- Bei gleicher (Trigramm-) Ähnlichkeit sollen mehrerer Worte zu einem Muster sollen häufigere Worte besser bewertet werden.

$$S = \frac{20c}{n+m} + \frac{2b}{n+m} + \frac{h}{10^8}$$

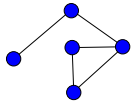
c : Anz. gemeins. Trigr.
n, m: Längen der Zeichenk.
h : Korpushäufigkeit des betrachteten Wortes.



N-Gramme (4)

Andere Ähnlichkeitsmaße für Trigramme
Gegeben: Zeichenketten a, b über einem Alphabet $\mathcal{A}$
Seien $c_a(t), c_b(t)$ die Häufigkeiten des Trigramms t in den Zeichenketten a bzw. b.

- KoKS-Projekt:
  $a = \frac{\sum_t(\min(c_a(t),c_b(t)))}{\sum_t(\max(c_a(t),c_b(t)))}$
- Beziehung zum Jaccard-Maß: Ohne Beachtung der Anzahl: Auftretenen von t ist 1, Nichtauftreten 0.
- Baldwin; Tanaka (2000):
  Ersetzt man $\sum_t(\max(c_a(t), c_b(t)))$ durch das arithm. Mittel der Längen von a, b, geht a in S über (s.o.).



Sprachprobleme

Zuordnung von Zeichenketten über Sprachgrenzen (natürl. Sprachen, Fachsprachen) - M:N.
Metawissen (Gebiet) - Ontologie d. Begriffe (Gleichheit = gl. Position in d. O.)

- (η) αναχώρηση: Abfahrt(Bahn, Schiff), Abflug (Flugzeug)
- Kontext

άγριος	:	(9 Übers.) blind (Haß, Wut), rauh (Berge, Wetter),streng (Blick)
στράβος	:	blind (nicht sehend), Syn. τυφλός
στράβος	:	blind (völlig ungebildet)
λαθραίος	:	blind (Passagier)

- Kontext *Verwandschaft* zu grob:

(ο) γαμπρός	:	Bräutigam
(ο) γαμπρός	:	Schwager (Mann der Schwester)
(ο) κουνιάδος	:	Schwager (Bruder d. Ehefrau / d. Ehemanns)

- Kontext in Fachsprachen

Quark	:	Speise; Elementarteilchentheorie
charmant, Quark	:	Elementarteilchentheorie
Farbe, Quark	:	Speise; Elementarteilchentheorie
String	:	Informatik, Kosmologie