Hochschule **RheinMain**
University of Applied Sciences
Wiesbaden Rüsselsheim

AUTOMATISCHE VERLINKUNG MIT WIKIPEDIA MITTELS C4.5-ENTSCHEIDUNGSBÄUME

Fachseminar "Machine Learning"

Letztes Update: 20. Januar 2016

Anahita Hamidi

Studienbereich Angewandte Informatik
Hochschule **RheinMain**

GLIEDERUNG

EINLEITUNG

WIKIPEDIA

→ die größte **Enzyklopädie**

→ **dicht** strukturiert

→ Millionen von **Artikeln** und **Links**

Frage:

→ Welche Begriffe in einem Artikel sollen verlinkt werden?

→ Wie werden diese Begriffe verlinkt?

Iranian POW negotiator holds talks with Iraqi ministers

The head of **Iran**'s **prisoner of war** commission met with two **Iraqi** Cabinet ministers Saturday in a bid to glean information about thousands of Iranian POWs allegedly in Iraq, the official Iraqi News Agency reported.

Iraqi Foreign Minister **Mohammed Saeed al-Sahhaf** told Abdullah al-Najafi that the two states needed to ``speed up the closure of what remains from the POW and Missing-In-Action file,'' INA said.

The issue of POWs and missing persons remains a stumbling block to normalizing relations between the two neighbors.

Iraq has long maintained that it has released all Iranian prisoners captured in the **1980**-88 **Iran-Iraq War**. The countries accuse each other of hiding POWs and preventing visits by the **International Committee of the Red Cross** to prisoner camps.

The ICRC representative in **Baghdad**, Manuel Bessler, told The **Associated Press** that his organization has had difficulty visiting POWs on both sides on a regular basis.

In April, Iran released 5,584 fied as civil law detainees in the largest exchange since **1990**.

More than 1 million people w

Baghdad

Baghdad is the capital of Iraq and of Baghdad Governorate. With a metropolitan area estimated at a population of 7,000,000, it is the largest city in Iraq. It is the second-largest city in the Arab world (after Cairo) and the second-largest city in southwest Asia (after Tehran).

open in wikipedia

aus (Milne and Witten, 2008, "Learning to Link with Wikipedia")

Abbildung: Automatische Verbindung eines Begriffs

WIKIFICATION

- → dieser Prozess heißt **Wikification**
  - → Disambiguation
  - → Detection
- → wichtiges Mittel: **Entscheidungsbäume**
- → vorheriger Ansatz **Wikify!**-System von Mihalcea und Csomai (2007)

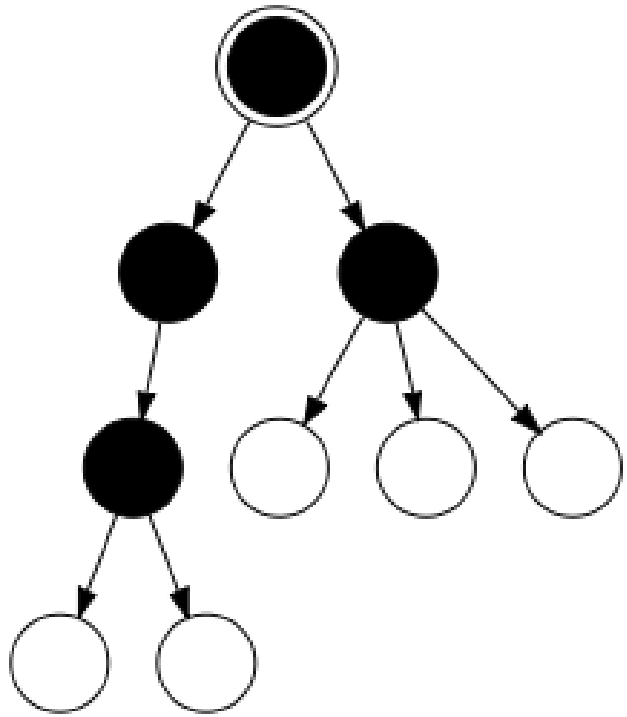
ENTSCHEIDUNGSBAUM

→ **gerichteter** Baum

→ Blätter (Klasse, Wahrheitswerte)

→ Knoten (Attribute)

→ Kanten (Attributwert)



aus Baum (Graphentheorie)

Abbildung: Ein gerichteter Baum

C4.5 ALGORITHMUS

EINLEITUNG

→ von Ross Quinlan entwickelt (1993)

→ Nachfolger von ID3

→ generiert Entscheidungsbaum aus den Trainingsdaten

Verbesserungen in C4.5:

→ sowohl **diskrete**, als auch **kontinuierliche** Attribute

→ **fehlende** Attributswerte

BEISPIEL

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

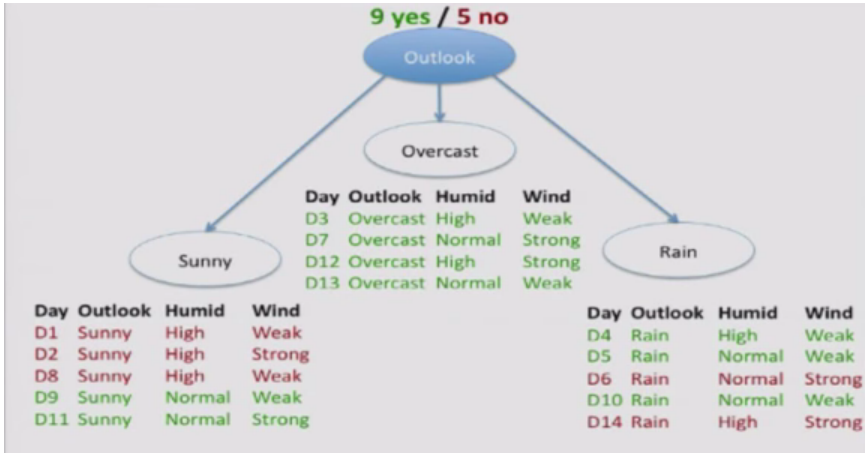


Abbildung: Trainingsdaten

BEISPIEL

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

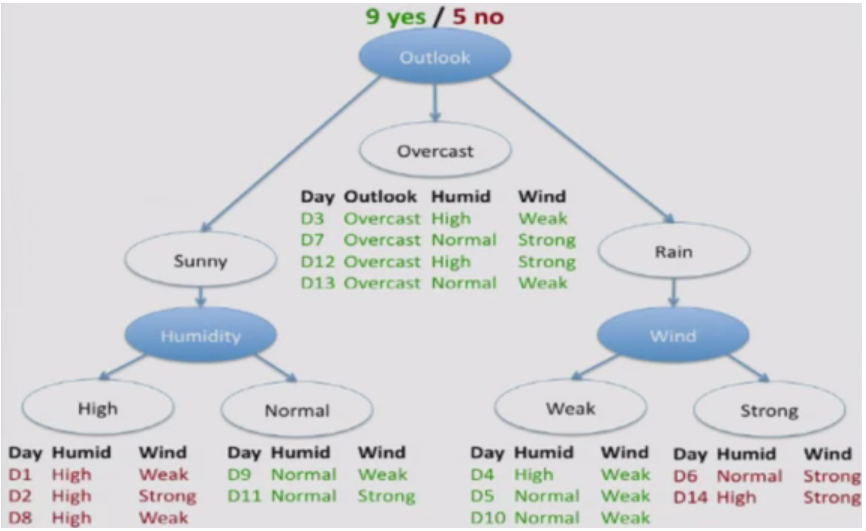


Abbildung: Trainingsdaten

BEISPIEL

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

Abbildung: Trainingsdaten

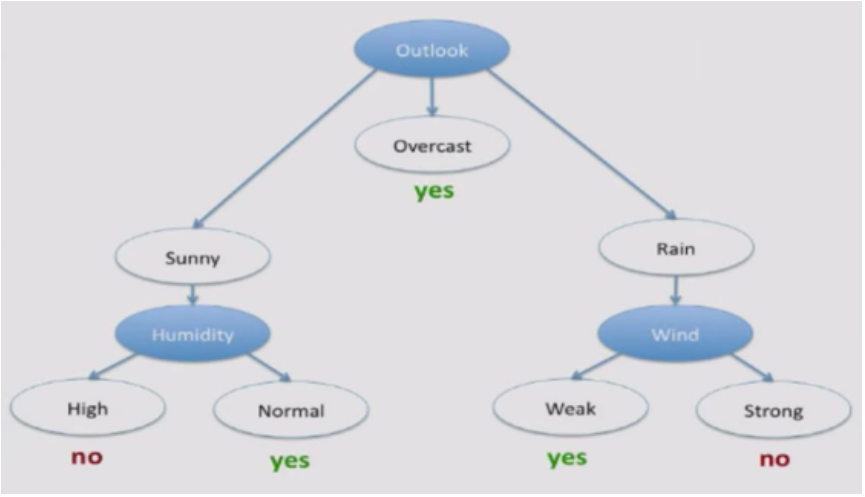


Abbildung: der erzeugte Baum

DISAMBIGUATION

BEGRIFFSERKLÄRUNG

- → wird in der Wikipedia angewendet
- → Verknüpfung eines Stichworts mit dem entsprechenden Wikipedia-Artikel
- → Merkmale
  - → commonness
  - → relatedness
  - → context quality

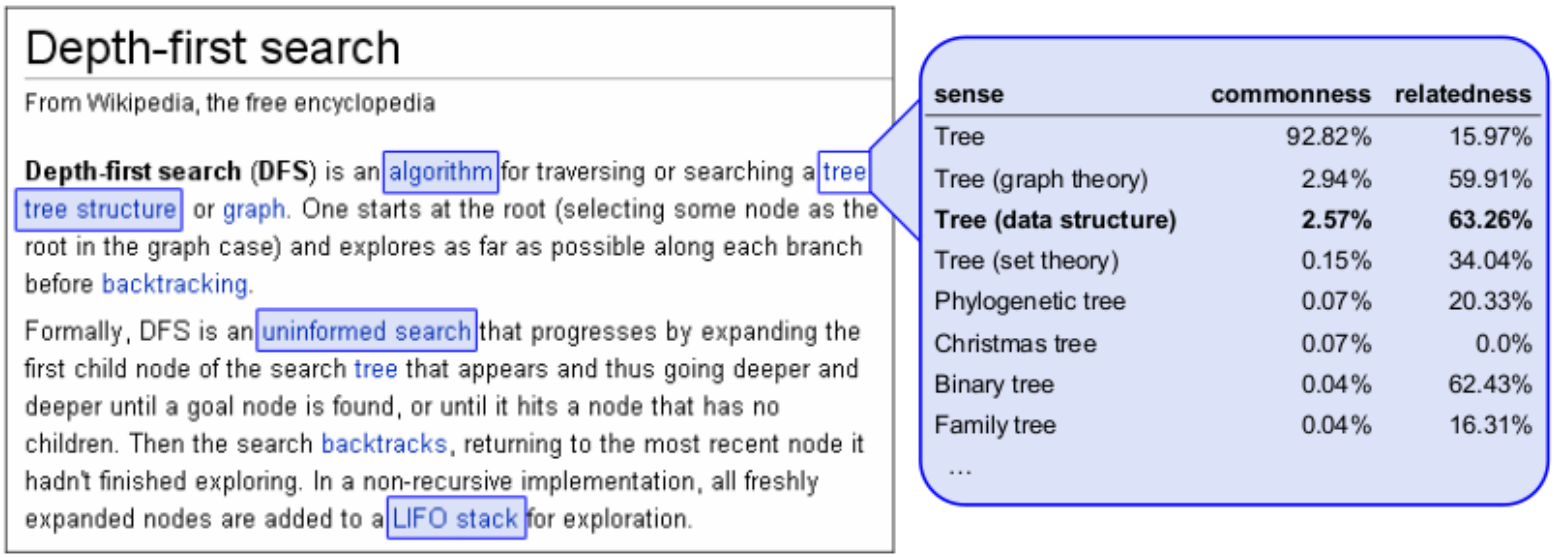
COMMONNESS

Definition: wie oft wurde ein Artikel als Ziel definiert?

$$commonness = \frac{c}{c_{\mathrm{i}}}$$

c die Anzahl der Ziel-Artikel von einem Begriff
c_{i} die Anzahl aller Artikel mit diesem Begriff
Problem: ist aber nicht immer die beste Entscheidung.
Lösung: das Stichwort im gesamten Text betrachten.

BEISPIEL



aus (Milne and Witten, 2008, "Learning to Link with Wikipedia")

Abbildung: Disambiguierung des Begriffs ”tree”

RELATEDNESS

→ der Bezug des Begriffs zum **Kontext**

→ Kontext: die eindeutigen Begriffe

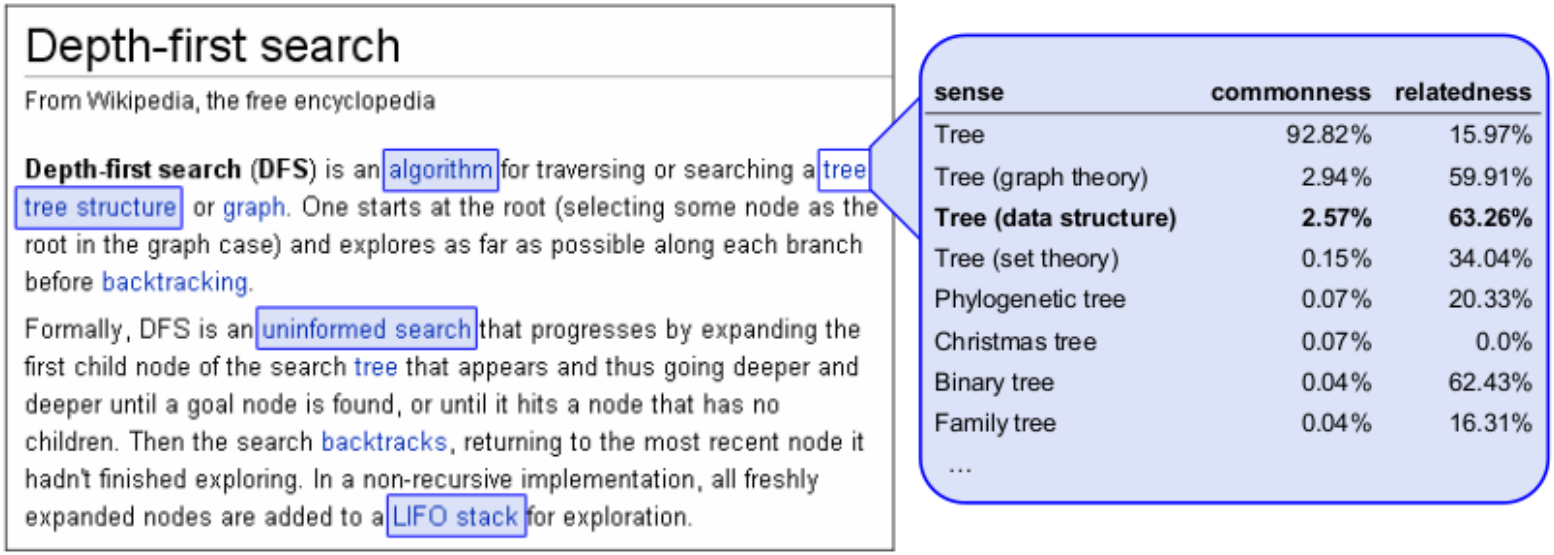
→ eindeutige Begriffe: immer auf einen Artikel

$$relatedness(a, b) = \frac{\log(max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(min(|A|,|B|))}$$

a und b: zwei Artikel

A, bzw. B: die Menge aller mit a bzw. b verlinkten Artikel

W: die Menge aller Artikel in Wikipedia



Depth-first search

From Wikipedia, the free encyclopedia

Depth-first search (**DFS**) is an algorithm for traversing or searching a tree tree structure or graph. One starts at the root (selecting some node as the root in the graph case) and explores as far as possible along each branch before backtracking.

Formally, DFS is an uninformed search that progresses by expanding the first child node of the search tree that appears and thus going deeper and deeper until a goal node is found, or until it hits a node that has no children. Then the search backtracks, returning to the most recent node it hadn't finished exploring. In a non-recursive implementation, all freshly expanded nodes are added to a LIFO stack for exploration.

sense	commonness	relatedness
Tree	92.82%	15.97%
Tree (graph theory)	2.94%	59.91%
Tree (data structure)	**2.57%**	**63.26%**
Tree (set theory)	0.15%	34.04%
Phylogenetic tree	0.07%	20.33%
Christmas tree	0.07%	0.0%
Binary tree	0.04%	62.43%
Family tree	0.04%	16.31%
…		

aus (Milne and Witten, 2008, "Learning to Link with Wikipedia")

Abbildung: Disambiguierung des Begriffs ”tree”

AUSWAHL EINES KONTEXTS

→ nicht jeder eindeutige Begriff ist nützlich (z.B. das Wort **the**)

die Link-Wahrscheinlichkeit =

→ $\frac{Anzahl(A_{\text{key}})}{Anzahl(A_{\text{W}})}$

→ A_{key} : die Artikel, die den Begriff als Link enthalten

→ A_{W} : alle Artikel, die den Begriff enthalten

Gewichtung eines Begriffs:

→ **Durchschnitt** von Link-Wahrscheinlichkeit und relatedness

CONTEXT QUALITY

context quality:

→ die Summe **der Gewichtungen** der Begriffe

$ist\ der\ Kontext\ homogen \rightarrow relatedness\ wird\ klarer$
$ist\ der\ Kontext\ gemischt \rightarrow commonness\ wird\ betrachtet$

Der C4.5 Klassifikator:

→ Trainingsdaten: 3 Merkmale

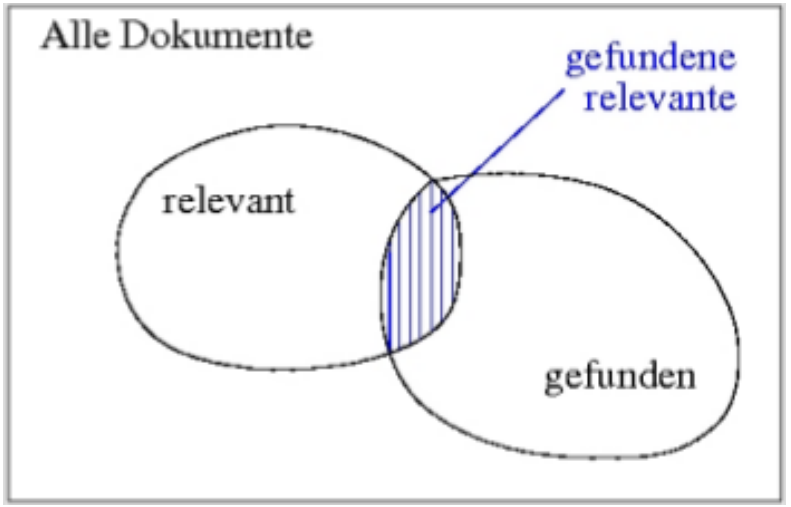
→ ergibt die Wahrscheinlichkeit für jeden Artikel

EVALUATION

$$\rightarrow precision = \frac{gefundene\ relevante}{gefundene}$$

$$\rightarrow recall = \frac{gefundene\ relevante}{relevante}$$

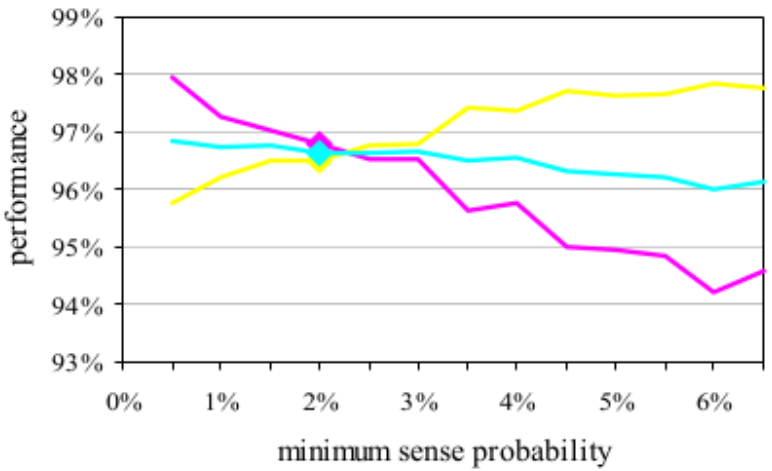
$$\rightarrow f - measure = 2 \cdot \frac{Precision \cdot recall}{precision + recall}$$



→ 100 zufällige Artikel, 11000 Anchors

→ Links mit **Threshold** von weniger als 2% ignorieren

→ bessere Ergebnisse als vorherige Ansätze



aus (Milne and Witten, 2008, "Learning to Link with Wikipedia")

	recall	precision	f-measure
Naïve Bayes	96.6	95.0	95.8
C4.5	96.8	**96.5**	96.6
Support Vector Machines	96.5	96.0	96.3

aus (Milne and Witten, 2008, "Learning to Link with Wikipedia")

DETECTION

DER LINK-ERKENNUNGSPROZESS

Definition: Identifizierung der Begriffe
n-gramm: Zerlegung eines Strings in Substrings der Länge n
Beispiel: ”Ich habe ein Brot gegessen.” in einer 2-gramm Zerlegung

→ ”Ich habe”
→ ”habe ein”
→ ”ein Brot”
→ ”Brot gegessen”

Detection:

→ alle **n-gramms** erfassen.
→ irrelevante Begriffe im Text ignorieren.(Threshold weniger als 6%)
→ alle verbleibenden Begriffe mit dem Klassifikator disambiguieren.

DER LINK-ERKENNUNGSPROZESS

Merkmale:

→ Link Probability
→ Relatedness
→ Disambiguation Confidence
→ Generality
→ Location and Spread

Der Klassifikator:

→ entscheidet welche Begriffe verlinkt werden sollen

MERKMALE

Link Probability:

→ berechnet bei der Disambiguierung

Relatedness:

→ Relatedness jedes Begriffs zum Kontext. (berechnet bei der Disambiguierung)

MERKMALE

Disambiguation Confidence:

→ wie sicher ist die Entscheidung bei der Disambiguation.

Generality:

→ unbekannte Begriffe für den Leser verlinken.

Location and Spread:

→ wie oft taucht ein Begriff in einem Dokument auf.
→ Der Begriff taucht häufig auf
  → die Link-Wahrscheinlichkeit steigt.(**frequency**)
→ an welcher Stelle taucht der Begriff auf.
  → am Anfang des Dokumentes (**first occurance**)
  → am Ende des Dokumentes (**last occurance**)
  → Abstand zwischen first und last occurance (**spread**)

EVALUATION

→ 100 zufällige Artikel

→ 9300 manuelle Links

→ alle Markierungen weg

	recall	precision	f-measure
Wikify (estimate)	46.5	49.6	48.0
Wikify (upper bound)	53.4	55.9	54.6
New link detector	**73.8**	**74.4**	**74.1**

aus (Milne and Witten, 2008, "Learning to Link with Wikipedia")

Abbildung: das Ergebnis des Link-Erkennungsprozess

QUELLEN

LITERATURANGABE

→ Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann, Amsterdam.

→ Milne, D. and Witten, I. H. (2008b). Learning to Link with Wikipedia. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'2008).

→ William B. Cavnar and John M. Trenkle. N-gram-based text categorization. In In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, pages 161–175, 1994.

→ https://en.wikipedia.org/wiki/c4.5

Vielen Dank für Ihre Aufmerksamkeit!!