

Masterarbeit

Extraktion von Entitäten aus Suchergebnissen

Denis Repkov

Mai 2015 – November 2015

Betreuer:

Professor Dr.-Ing. Norbert Fuhr
Dipl.-Inform. Sebastian Dungs



Universität Duisburg-Essen
Fakultät für Ingenieurwissenschaften
Abteilung Informatik und angewandte Kognitionswissenschaft
Fachgebiet Informationssysteme
47048 Duisburg

Kurzzusammenfassung

Moderne Suchmaschinen bieten zur Zeit keine Übersicht über die gefundene Ergebnismenge, was die Suche für den Endbenutzer komplizierter macht, da alle gefundene Dokumente angeschaut werden müssen, um einen Überblick über die gefundene Daten zu bekommen, und um Anregungen für präzisere Anfragen zu gewinnen.

Um solche Benutzerunterstützung bei der Websuche zu ermöglichen, müssen Suchmaschinen den Inhalt von Webseiten auswerten können. Eine der solchen Auswertungen ist die Erkennung und Extraktion von sogenannten Entitäten. Entität kann etwa eine Person, ein Datum, oder eine Organisation sein. Eine Entität hat außerdem verschiedene Attributen, wie Klasse, Name, kurze Beschreibung und Verbindungen zu anderen Entitäten. Eine Anreicherung von Suchergebnissen mit den Informationen über extrahierte Entitäten wäre bei der Bearbeitung von Suchergebnissen und bei der Verfeinerung von Suchanfragen sehr hilfreich.

Im Rahmen dieser Arbeit muss ein Framework entwickelt werden, der:

- Die Suchanfrage des Endbenutzers an eine konventionelle Suchmaschine weiterleitet.
- Erkennt die Entitäten auf von der Suchmaschine zurückgegebenen Suchergebnissen.
- Verlinkt die extrahierte Entitäten mit den Ontologien aus DBpedia.
- Die Suchergebnisse mit den Ontologien anreichert, und an den Benutzer als ein XML oder JSON-Dokument schickt.

Inhaltsverzeichnis

1	Einleitung	1
1.1	Motivation	1
1.2	Aufgabenstellung	1
1.3	Aufbau der Arbeit	1
2	Grundlagen	3
2.1	Grundlagen von E-Mails	3
3	Verwandte Arbeiten	5
3.1	Arbeiten zum Thema A	5
4	Das Konzept	7
5	Grundlegende Ideen	9
6	Implementierung	11
6.1	Stanford NER	11
7	Evaluierung	13
7.1	Bla bla	13
8	Zusammenfassung, Diskussion und Ausblick	15
8.1	Zusammenfassung	15
8.2	Diskussion	15
8.3	Ausblick	15
A	Daten	iii
A.1	Klassenhäufigkeiten	iii
A.2	Dokumenttypdefinition eines Tickets	iii
	Abbildungsverzeichnis	vii
	Tabellenverzeichnis	ix
	Auflistungsverzeichnis	xi

Kapitel 1

Einleitung

Im Abschnitt 1.1 wird die Motivation zu dieser Arbeit erläutert. Darauffolgend (1.2) wird die Aufgabenstellung beschrieben. Abschließend gibt Abschnitt 1.3 einen Überblick über den Aufbau der Arbeit.

1.1 Motivation

1.2 Aufgabenstellung

1.3 Aufbau der Arbeit

Kapitel 2

Grundlagen

2.1 Grundlagen von E-Mails

Abbildung 2.1 zeigt eine einfache E-Mail ?.

```
Message-ID: <480DEBCA.8040607@uni-due.de>
Date: Tue, 22 Apr 2008 15:44:42 +0200
From: Bob <bob@uni-due.de>
User-Agent: Thunderbird 2.0.0.12 (Windows/20080213)
To: alice@uni-due.de
Content-Type: text/plain; charset=ISO-8859-15
Content-Transfer-Encoding: 7bit
Subject: Frage

Hallo Alice,

wie geht es Dir?

Gruß
Bob
```

Abbildung 2.1: Beispiel für eine einfache E-Mail

Kapitel 3

Verwandte Arbeiten

3.1 Arbeiten zum Thema A

In einer vielzitierten Arbeit stellten ? das KnowMore-Projekt dar.

Außerdem wurde noch das KnowMore-Projekt (?) in die Untersuchung einbezogen.

Außerdem wurde noch das KnowMore-Projekt (?) in die Untersuchung einbezogen.

In einer vielzitierten Arbeit stellten ? das KnowMore-Projekt dar.

In einer vielzitierten Arbeit stellten ? das KnowMore-Projekt dar.

Das KnowMore-Projekt wurde bereits im Jahr ? von ? vorgestellt.

Das KnowMore-Projekt wurde bereits im Jahr ? von ? vorgestellt.

Kapitel 4

Das Konzept

Kapitel 5

Grundlegende Ideen

Hier werden die grundlegenden Ideen des eigenen Kozept beschrieben.

Kapitel 6

Implementierung

6.1 Stanford NER

Kapitel 7

Evaluierung

7.1 Bla bla

Kosten- matrix		tatsächlich				
		c_1	c_2	\dots	c_{n-1}	c_n
vorhergesagt	c_1	$cost_1^1$	$cost_1^2$	\dots	$cost_1^{n-1}$	$cost_1^n$
	c_2	$cost_2^1$	$cost_2^2$	\dots	$cost_2^{n-1}$	$cost_2^n$
	\dots	\dots	\dots	\dots	\dots	\dots
	c_{n-1}	$cost_{n-1}^1$	$cost_{n-1}^2$	\dots	$cost_{n-1}^{n-1}$	$cost_{n-1}^n$
	c_n	$cost_n^1$	$cost_n^2$	\dots	$cost_n^{n-1}$	$cost_n^n$

Tabelle 7.1: Kostenmatrix für die Klassen c_1 bis c_n

Kapitel 8

Zusammenfassung, Diskussion und Ausblick

8.1 Zusammenfassung

8.2 Diskussion

8.3 Ausblick

Anhang

A Daten

A.1 Klassenhäufigkeiten

A.2 Dokumenttypdefinition eines Tickets

```
<!ELEMENT ticket (date, from, subject, body, attachement, meta)>
<!-- ATTLIST ticket id ID #REQUIRED -->
<!ELEMENT date (#PCDATA)>
<!ELEMENT from (#PCDATA)>
<!ELEMENT subject (#PCDATA)>
<!ELEMENT body (#PCDATA)>
<!ELEMENT attachement (#PCDATA)>
<!-- ELEMENT meta (topic, supportType, priority) -->
<!ELEMENT topic (#PCDATA)>
<!-- ELEMENT supportType (#PCDATA) -->
<!-- ELEMENT priority (#PCDATA) -->
```

Aufistung A.1: Dokumenttypdefinition der Tickets im XML-Format

Verzeichnisse

Abbildungsverzeichnis

2.1	Beispiel für eine einfache E-Mail	3
-----	---	---

Tabellenverzeichnis

7.1	Kostenmatrix für die Klassen c_1 bis c_n	13
-----	--	----

Auflistungsverzeichnis

A.1 Dokumenttypdefinition der Tickets im XML-Format	iii
---	-----

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit ohne fremde Hilfe selbstständig verfasst und nur die angegebenen Quellen und Hilfsmittel benutzt habe. Ich versichere weiterhin, dass ich diese Arbeit noch keinem anderen Prüfungsgremium vorgelegt habe.

Duisburg, im November 1492

.....

Manfred Mustermann

