

Seminar « Content Management »

SommerSemester 2002

26.06.02

Referat Information Extraction

Von

Mirco Schröder

INHALT

I Überblick

1. Information Retrieval

- a) Begriffserklärung
- b) Problemfelder

2. Information Extraction

- a) Unterschiede zu IR
- b) Wie arbeitet IE Software ?

II. Arbeitsweisen von IE-Systemen

III. Anwendung

IV. Resümee

Was ist Information Retrieval?

- Mit zunehmender Menge elektronischer Daten wuchs Bedeutung
- I.R. sucht ein Set von Informationsquellen aus der Gesamtmenge
- Millionfache Nutzung z.B. im WWW
- Meist als einfache Schlagwortsuche
- In gegenwärtigen Suchmaschinen

Was ist Information Retrieval?

- Technologie mit der Zeit erweitert etwa um Thesaurus
- Kategorisierung
- Filterung
- Clustering
- Oder auf andere Medienarten angewandt : Bild, Sound

ALLERDINGS :

- Da I.R. rein wortbasiert arbeitet, ergeben sich mehrere Problemfelder

Problemfelder des I.R.

- **Synonyme**

Unterschiedliche Wörter drücken dasselbe aus - *produzieren*, *herstellen* oder *fabrizieren* verweisen gleichermaßen auf den Sachverhalt *Produktion*.

- **Polysemie**

Gleiche Wörter verweisen auf unterschiedliche Sachverhalte - *Bank* verweist gleichermaßen auf die Sachverhalte *Sitzmöbel* und *Kreditinstitut*.

- **Redewendungen/Metaphern**

Einzelne Wörter aus Redewendungen wie "*von der Arbeit erschlagen sein*" bieten an sich keine sinnvolle Indexierungsmöglichkeit ("*Stress*" wäre hier ein akzeptabler Deskriptor).

Problemfelder des I.R. (II)

- **Lokaler Kontext**

Zum Auffinden eines spezifischen Sachverhalts, bsw. "*Überfälle auf Geldtransporte*", ist ein einzelner Suchbegriff oder auch die bool'sche Verknüpfung zweier Begriffe nicht unbedingt ausreichend. In diesem Fall muß das *Ziel* des Überfalls ein Geldtransport sein. Solche Begriffsassoziationen stehen in einfachen Retrievalsystemen nicht zur Verfügung.

- **Globaler Kontext**

Einige Dokumente enthalten überhaupt keine verwertbaren Wörter als Basis für eine Indexierung - die Relevanz mancher Dokumente bezüglich eines bestimmten Sachverhalts kann sich erst aus der Betrachtung ganzer *Sätze, Abschnitte* oder gar des *gesamten Texts* ergeben.

Was ist Information Extraction?

- Im Gegensatz zu rein wortbasierten Verfahren versucht ein IE-System, diese Probleme durch kontextsensitive Textanalyse und Verwendung eines intelligenten Wörterbuchs/Regelwerks zu umgehen.
- IE überträgt Content aus inhomogenen Quellen wie *Webseiten*, *Emails* oder *Pdfs* mit Hilfe von XML in ein strukturiertes Format.

Natürlichsprachliche Texte werden also zusammengefasst und in einer einheitlichen Form präsentiert.

Wie arbeitet I.E. Software ?

1. **Identifiziert** relevante Fakten in Dokumenten
2. **Extrahiert** die gewünschten Informationen aus Unterschiedlichsten Quellen
3. Und **integriert** diese ins Outputformat bzw. ordnet sie festgelegten Kategorien zu

IE-Software (II)

Gewöhnlich für eine Domain oder ein bestimmtes Thema konstruiert, um :

- Genauigkeit zu verbessern und Entwicklung zu erleichtern
- Daher kein Anspruch auf vollkommenes Verständnis des Dokumentinhalts durch Systems
- Ausschließlich nur Fakten extrahiert, die relevante Information bezüglich des spezifizierten Wissensgebiets darstellen.

IE-Software (*ein Beispiel*)

Ein IE-System für Wirtschaftsartikel, kann z.B :

- Alle Informationen über Geschäftsberichte, Wettbewerbsbedingungen, Analystenkonferenzen und Publikationen aus Pressemitteilungen, Nachrichten oder Emails und diese in einer Datenbank zusammenfassen.
- End-Users können dann diese Datenbank durch Textabfragen nutzen: Etwa nach allen Firmenpleiten in Europa

IE-Software (IV)

Realisierung mit Einsatz von Programmen zur Analyse von :
freisprachlicher, beliebig strukturierter Dokumente

- auf grammatikalischer und syntaktischer Ebene
(sog. *Natural Language Processors*)
- in Verbindung mit Lexika, die das zu untersuchende
Themengebiet abdecken.
- Durch Trainieren des Systems mit Beispieldokumenten kann
ein Regelwerk generiert werden
- Dadurch in der Lage, Information aus Texten zu extrahieren
bzw. Texte festgelegten Kategorien zuzuordnen.

Zusammenfassung (UNTERSCHIEDE ZWISCHEN IE UND IR)

- **IR** findet **eine Reihe von Dokumenten**, die mit den Suchbegriffen übereinstimmen.
- **IE** findet **einzelne Fakten** aus verschiedenen Dokumenten.
- Der Unterschied liegt im Feinheitsgrad :

“IR is document retrieval and IE is fact retrieval“

⇒ IR und IE sind unterschiedlich, aber komplementär

INHALT

I Überblick

- 1.Information Retrieval
- 2.Information Extraction

II. Arbeitsweisen von IE-Systemen

- 1.Einfaches Template-System
- 2 Wrapper (das Programm Circus)

III. Anwendung von IE

IV. Resümee

Einfaches Template-System

- **Das System ist Textcorpus basiert**
- **Der Corpus muß compilt werden**
- **Anschließend wird manuell ein Antwortschlüssel entworfen**
- **Lexika für eine vollständige Syntaxanalyse müssen erstellt werden**
- **Das System arbeitet in 3 Schritten :**

1. Schritt

- Identifizierung von Strukturen und Wortrelationen durch lexikale Analyse
- Dabei werden insbesondere Namen gesucht :

Personen :

- Nach Titeln (Herr, Frau, Mr,...)
- Nach gängigen Vornamen (Peter, Dieter,...)
- Nach Suffixen (Snippety Smith Jr.)

Unternehmen :

- Nach ihrer Rechtsform (KG, AG, GmbH,...)
- Markanten Zusätzen wie (Berenberg Bank, Mummert + Partner, ...)

Einfaches Template-System (III)

- danach volle Syntaxanalyse zur Entdeckung von Nomengruppen, Verbgruppen
- Task-spezifisches Pattern-Matching zum Filtern der Fakten

2. Schritt :

- Integrationsphase
- Auflösung von **indirekten Beziehungen** (er, das Unternehmen) und **Interferenzen**

Sam was president. He was succeeded by Harry.

- **Fakten werden kombiniert :**

entity e1	type: person name: "Sam Schwartz"
entity e2	type: position value: "executive vice president" company: e3
entity e3	type: manufacturer name: "Hubblewhite Inc."
entity e6	type: person name: "Harry Himmelfarb"
event e7	type: leave-job person: e1 position: e2
event e8	type: succeed person1: e6 person2: e1

Schritt 3:

- Ausgabe in Templates :

```
EVENT   leave job  
PERSON  Sam Schwartz  
POSITION executive vice president  
COMPANY Hupplewhite Inc.
```

```
EVENT   start job  
PERSON  Harry Himmelfarb  
POSITION executive vice president  
COMPANY Hupplewhite Inc.
```

Fig. 3. Events extracted from Hupplewhite text.

Nachteile der Template Technik

- **Sehr hoher manueller Arbeitsaufwand**
- **Nur für einzelne Domains oder Themen machbar**
- **Mangelnde Portabilität**

WRAPPER

- Ein alternatives IE-System sind die sog. Wrapper
- Haben sich aus dem Bedarf heraus entwickelt Daten aus verschiedenen Webressourcen zu integrieren
- Diese benutzen keine handgestrickten Grammatiken, sondern statistische Methoden
- Keine volle Syntax Analyse

Ausgangssituation

Beobachtungen, wie Menschen manuell Dokumente klassifizieren:

- Einige Dokumente schwierig zuzuordnen, da thematische Überschneidungen
- Viele Dokumente schnell eingeordnet werden, da sie sofort als Themenrelevant erkannt
- Ein Satz oft ausreichend, um Dokument als relevant zu bewerten.

Ausgangssituation (II)

- ⇒ Nachdem ein (gewichtiger) Indikator für die Relevanz eines Texts erkannt wurde, kann der Rest des Dokuments ignoriert werden, ungeachtet eventuell folgender Relevanzmerkmale.
- ⇒ Auf Basis dieser Erkenntnisse wurde ein System konzipiert, das insgesamt drei Algorithmen benutzt, um Texte zu klassifizieren :

CIRCUS (*conceptual sentence analyzer*)

Prinzip der concept nodes

- Ein *concept node* repräsentiert die syntaktische Struktur eines Sachverhalts
- Zur Aufbereitung der zu analysierenden Texte wurden diese mit dem Programm CIRCUS geparkt,
- Aufgabe ist, sog. *concept nodes* aus Texten zu extrahieren.
- Für ein bestimmtes Themengebiet sind diese bereits vordefiniert und in einem Wörterbuch abgelegt.

Prinzip der concept nodes (II)

- wird aufgrund des Vorkommens bestimmter Schlüsselwörter im Text konsultiert
- Verifizierung des linguistischen Kontextes des Schlüsselworts durch den concept node
- Ggf. **Aktivierung** des concept node

d.h. relevante Informationen wurden im Kontext des concept node-Schlüsselworts gefunden.

- Wo diese zu finden sind, wird durch die Spezifikation des concept nodes vorhergesagt.

Prinzip der concept nodes (III)

- Informationen des concept nodes werden in sog. Slots gespeichert
- Aktivierte concept nodes mit erfüllten Kriterien werden instanziiert genannt
- Sind der einzige Output von CIRCUS
- Basis für die Anwendung der nachfolgend beschriebenen Algorithmen

Beispielhafte Instanzierung der concept nodes „murder“ (aktiv/passiv)

Information Extraction as a Basis for High-Precision Text Classification

Concept Nodes

Ein *Concept Node* ist abhängig von einem einzigen Wort, wird aber nur in einem bestimmten linguistischen Kontext als (output-)relevant angesehen.

“The terrorists murdered the mayor.”

“Three peasants were murdered by guerrillas.”

Name:	Murder-Active	Murder-Passive
Trigger Word:	murdered	murdered
Slots:	(perpetrator(SUBJECT)) (victim(OBJECT))	(victim(SUBJECT)) (perpetrator(PP + by))
Enabling Conditions:	(active)	(passive)

Algorithmus I - Relevancy Signatures

- *Signature* besteht aus concept node und einem Wort, das diesen aktiviert hat
- Gefundene *signature* überprüft, ob sie eine *relevancy signature*
- D.h. hohe Korrelation zu einem bestimmten Themenkomplex
- Rückschluß auf Relevanz des Dokuments bezüglich Themenkomplex

Algorithmus I - Relevancy Signatures (II)

Die Anwendung von Algorithmus I erfolgt in zwei Schritten:

a) Trainingsphase

- Trainings-Textkorpus bildet aus CIRCUS-Output *signatures*
- Statistisch bezüglich der Wahrscheinlichkeit bewertet, daß ein Dokument relevant ist, in dem die entsprechende *signature* auftaucht
- Festgelegter *Schwellwert* indiziert alle *signatures* mit einem Wahrscheinlichkeitswert darüber als *relevancy signatures*.

Algorithmus I - Relevancy Signatures (III)

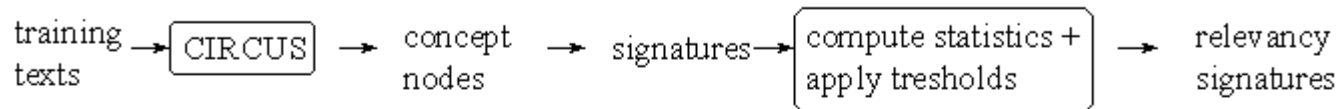
b) **Klassifikationsphase**

- *relevancy signatures* werden als Indikatoren für die Relevanz eines Dokuments benutzt
- Der verbundene *concept node* kann dann seine Slots mit Informationseinheiten des Kontext füllen

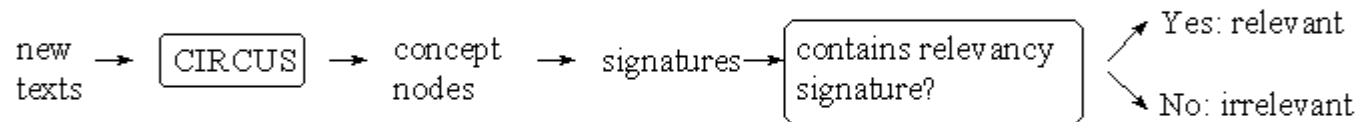
Schematischer Ablauf

Klassifizierungsprozeß

Training des Systems



Klassifizierung

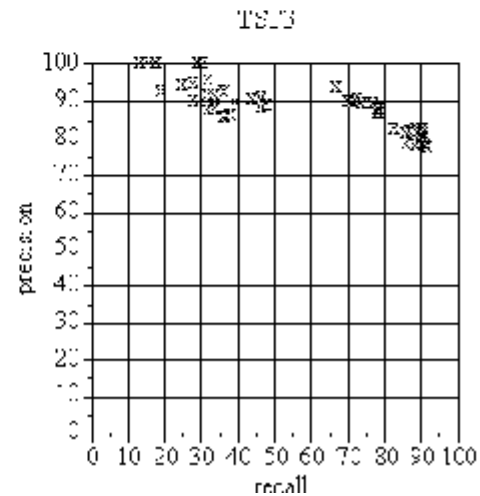


Threshold Parameters

Eine signature wird ausgewählt, wenn sie eine Wahrscheinlichkeit $P(RS) \geq R$ aufweist

UND

mindestens M-mal in der Dokumentenmenge auftaucht



Algorithmus II - Augmented Relevancy Signatures

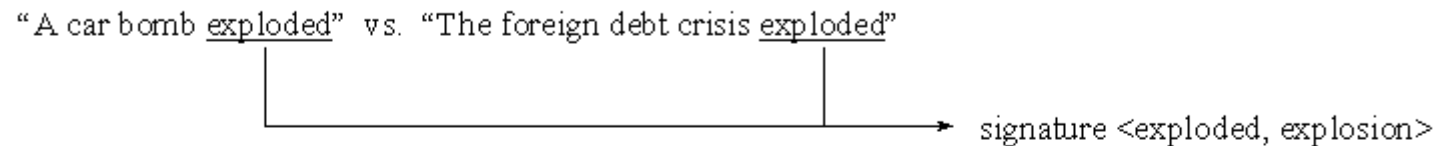
- Eine relevancy signature ist *nicht* kontextsensitiv ausgerichtet
- tritt nur in syntaktisch spezifizierten Umfeld auf (z.B. Aktiv-, Passivsatz)
- jedoch unabhängig vom semantischen Gehalt ihres Kontexts
- Daher Erweiterung der *relevancy signatures* um *Slot-Fillers*

Algorithmus II - Augmented Relevancy Signatures (II)

- Bilden den Kontext des concept nodes auf dessen slots ab
- Für jede mögliche Kontext-Kombination eines concept nodes werden Slot-Filler instanziiert
- Sogenannte *augmented relevancy signatures*
- Charakterisierung erfolgt (analog Algorithmus I) durch Ermittlung der Wahrscheinlichkeit, ob Dokument relevant ist

Instanzierung von 2 Beispielsätzen

Algorithmus II - Augmented Relevancy Signatures



Relevancy signatures können in einem irrelevanten Kontext auftauchen (z.B. in Metaphern).
 Augmented relevancy signatures berücksichtigen zusätzlich den Kontext, in dem concept nodes auftreten. Sie *instanzieren* diese concept nodes.

Die “Slots” einer relevancy signature werden gefüllt durch eine Repräsentation ihres Kontexts in der Form:

(concept node type, slot name, semantic feature)

“Mayor Heston was kidnapped by terrorists”

(kidnapping, victim, GOVERNMENT-OFFICIAL)

↓
 im Dictionary enthaltene “semantische Generalisierung”

Für jeden concept node wird ein solcher Slot-Filler entwickelt.

Algorithmus III - Case Representation

- Sätze mit instanziierten *concept nodes* werden zu *cases* zusammengefasst
- Ein *case* ist eine mögliche Kombination der *concept nodes*
- Gefundenen case-Konstrukte werden in einer Datenbank gespeichert
- gleichfalls mit Relevanz-Wahrscheinlichkeiten bewertet

Zusammenfassung

Relevancy Signatures

- haben den geringsten Informationsbedarf
- Geeignet für Dokumente mit dominanten Key Words, wo der restliche Text nicht weiter untersucht werden muß

Augmented Relevancy Signatures

- Werden angewandt, wenn der umgebene Text bedeutend ist für die Key Words

Case Representation

- Algorithmus wird benötigt, wenn das Dokument nur anhand des gesamten Kontexts eingeordnet werden kann

Zusammenfassung (II)

- Wrapper sind automatisiert
- benötigen keine Grammatik
- Keine volle Syntax Analyse
- Text wird gescannt, konzentriert sich auf Hot Spots
- Ein System wie CIRCUS ist generell portabel
- Bedarf lediglich minimalen Trainingsaufwand

INHALT

I Überblick

- 1.Information Retrieval**
- 2.Information Extraction**

II. Arbeitsweisen von IE-Systemen

- 1.Einfaches Templates-System**
- 2 Wrapper (das Programm Circus)**

III. Anwendung von IE

- 1.Beispiele**
- 2.Relevanz für Content Management**

IV. Resümee

Anwendungen von I.E.

Einsatzgebiete für Text Classification auf der Basis von Information Extraction (Beispiele)

- Gesundheitswesen**

Auswerten von medizinischen Berichten (Volltext) über Patienten, z.B. welche Therapieformen haben sich als besonders erfolgreich herausgestellt?

- Auswertung technischer Texte**

Monitoring von Zeitschriftenartikeln mit technischen Inhalten hinsichtlich erzielter Fortschritte durch Einsatz bestimmter Produktionsmethoden

- Auswertung von Wirtschaftsnachrichten**

Ein IE-System kann aus einem Text z.B. über Firmenfusionen die beteiligten Firmen, das eingesetzte Kapital, die involvierten Personen oder Beratungsgesellschaften extrahieren.

- Weitere Merkmale der IE-Technologie**

Das Erstellen von Dictionaries für das Erkennen von concept nodes kann mit hoher Erfolgsrate automatisiert werden.

SDI-Profile für Information Extraction aus dem WWW versprechen die höchste Erfolgsquote im Vergleich zu wortorientierten Systemen.

Die Algorithmen eines IE-Systems sind von ihrem Einsatzgebiet weitgehend unabhängig und können auf neue Domänen portiert werden.

Bedeutung für Content Management

PROBLEM INFORMATIONSFLUT

- Nach einer Studie von 2000 umfaßte das Web 2,5 Mrd Dokumente und wächst täglich um 7,3 mio.
- Verdoppelung also in etwas weniger als 1 Jahr

HIDDEN WEB

- zunehmend dynamische Webpages
- normale Suchmaschinen haben keinen Zugriff
- viele Informationen in relationalen Datenbanken
- IE kann auf die Informationen zugreifen

Bedeutung für Content Management (II)

- Für ein effektives Content Management wichtig Informationen schnell zu organisieren
- IE-Technologie besonders geeignet, da Integration von Daten aus verteilten Quellen
- Informationszugriffe werden verbessert, wenn es gelingt Informationen automatisch zu extrahieren und in einer einheitlichen Form zu strukturieren
- Reduziert Kosten und Arbeitszeit
- IE hat das Potenzial das Web von einer ungeordneten Linksammlung in eine Datenbank zu transformieren
- Genaueste Abfragen wären möglich anstatt von Key Word Suche mit Tausenden von ungenauen Treffern

RESÜMEE

- Skalierbarkeit und Portabilität bleiben Probleme für IE-System
- Einfache Template-Technik bietet nur Nutzbarkeit für eine Domain und ein spezielles Thema
- Sobald sich das Thema ändert müssen ganz neue Pattern eingerichtet werden
- Bsp: das Verb „to place“ ist im Zusammenhang mit einem Terrorismus-Pattern immer mit Bomben assoziiert
- Aber : Eine Vasen auf einen Tisch „stellen“ hat keinerlei Beziehung zu Terrorismus
- Diese Methode bleibt nur effektiv, solange eng eingegrenzt

RESÜMEE (II)

- Wrapper brauchen immer eine große Menge an Trainingsdaten
- Diese zu sammeln ist ebenfalls zeitintensiv
- Um diesen Aufwand zu verringern muß die Domain limitiert werden
- Bei einem Relaunch der Seite wäre der Wrapper unbrauchbar
- Wrapper sind daher leicht anfällig

ENDE