IceCite: Ein System zur Verwaltung von Wissenschaftlichen Publikationen mit Automatischer Metadaten-Extraktion

Claudius Korzen

Institut für Informatik
Albert-Ludwigs-Universität, Freiburg

31. Oktober 2011

Gliederung

Gliederung

Motivation

- Übliche Schritte einer Literaturrecherche:
  - **Literatursuche**, z.B. in Literaturdatenbanken (*DBLP*, *Medline*, etc.) mit Literatursuchmaschinen (*CompleteSearch*, *Pubmed*, etc.)
    - **Schneeballsuche:** (Iteratives) Durchsuchen von Literaturangaben in relevanter Literatur
  - **Literaturauswahl:** Entscheidung, welche der gefundenen Literatur relevant sind
  - **Literaturbeschaffung:** Suche nach entsprechenden PDF-Dateien, um Zugriff auf den Inhalt der ausgewählten Literatur zu erhalten

Motivation

- Übliche Schritte einer Literaturrecherche:
  - **Literatursuche**, z.B. in Literaturdatenbanken (*DBLP*, *Medline*, etc.) mit Literatursuchmaschinen (*CompleteSearch*, *Pubmed*, etc.)
    - **Schneeballsuche:** (Iteratives) Durchsuchen von Literaturangaben in relevanter Literatur
  - **Literaturauswahl:** Entscheidung, welche der gefundenen Literatur relevant sind
  - **Literaturbeschaffung:** Suche nach entsprechenden PDF-Dateien, um Zugriff auf den Inhalt der ausgewählten Literatur zu erhalten
- Schritte können **mühsam** und **zeitaufwändig** sein.
  $\Rightarrow$ Implementierung eines Systems, das die Schritte einer Literaturrecherche vereinfacht

Ziele

- **Titel-Extraktion:** Extrahiere den vollständigen und korrekten Titel aus der PDF-Datei einer gegebenen Publikation.

Ziele

- **Titel-Extraktion:** Extrahiere den vollständigen und korrekten Titel aus der PDF-Datei einer gegebenen Publikation.
- **Titel-Zuordnung:** Finde mithilfe des extrahierten Titels den repräsentierenden Eintrag in einer Literaturdatenbank.

Ziele

- **Titel-Extraktion:** Extrahiere den vollständigen und korrekten Titel aus der PDF-Datei einer gegebenen Publikation.
- **Titel-Zuordnung:** Finde mithilfe des extrahierten Titels den repräsentierenden Eintrag in einer Literaturdatenbank.
- **Referenzen-Extraktion:** Identifiziere das Literaturverzeichnis in einer Publikation und extrahiere jede enthaltene Referenz.

Ziele

- **Titel-Extraktion:** Extrahiere den vollständigen und korrekten Titel aus der PDF-Datei einer gegebenen Publikation.
- **Titel-Zuordnung:** Finde mithilfe des extrahierten Titels den repräsentierenden Eintrag in einer Literaturdatenbank.
- **Referenzen-Extraktion:** Identifiziere das Literaturverzeichnis in einer Publikation und extrahiere jede enthaltene Referenz.
- **Referenzen-Zuordnung:** Finde für jede extrahierte Referenz ihren repräsentierenden Eintrag in einer Literaturdatenbank.

Ziele

- **Titel-Extraktion:** Extrahiere den vollständigen und korrekten Titel aus der PDF-Datei einer gegebenen Publikation.
- **Titel-Zuordnung:** Finde mithilfe des extrahierten Titels den repräsentierenden Eintrag in einer Literaturdatenbank.
- **Referenzen-Extraktion:** Identifiziere das Literaturverzeichnis in einer Publikation und extrahiere jede enthaltene Referenz.
- **Referenzen-Zuordnung:** Finde für jede extrahierte Referenz ihren repräsentierenden Eintrag in einer Literaturdatenbank.
- **Suche:** Durchsuche alle Metadaten und Volltexte sowie Literaturdatenbanken nach einer gegebenen Suchanfrage Q.

Ziele

- **Titel-Extraktion:** Extrahiere den vollständigen und korrekten Titel aus der PDF-Datei einer gegebenen Publikation.
- **Titel-Zuordnung:** Finde mithilfe des extrahierten Titels den repräsentierenden Eintrag in einer Literaturdatenbank.
- **Referenzen-Extraktion:** Identifiziere das Literaturverzeichnis in einer Publikation und extrahiere jede enthaltene Referenz.
- **Referenzen-Zuordnung:** Finde für jede extrahierte Referenz ihren repräsentierenden Eintrag in einer Literaturdatenbank.
- **Suche:** Durchsuche alle Metadaten und Volltexte sowie Literaturdatenbanken nach einer gegebenen Suchanfrage Q.
- **Benutzerschnittstelle:** Implementiere eine Schnittstelle, die die Funktionalitäten des Systems zugänglich macht.

Benutzerschnittstelle

- Live-Demo

Gliederung

Extraktion von Titeln

Definition (Schriftauszeichnung & Schriftschnitt)

Die *Schriftauszeichnung* bezeichne die typografischen Eigenschaften (Schriftart, Schriftgröße) und der *Schriftschnitt* die Schriftstärke und die Schriftlage eines Textes.

Annahmen zur Extraktion von Titeln:

1. Der Titel einer Publikation befindet sich auf der ersten Seite.
2. Ein Titel ist durch eine andere Schriftauszeichnung gegenüber dem restlichen Text hervorgehoben, die innerhalb des Titels konsistent bleibt.
3. Die Länge eines Titels ist nicht beliebig kurz, d.h. es kann für ihn eine Mindestlänge definiert werden.

Extraktion von Titeln (2)

Gruppierung der Textzeilen der ersten Seite

Accurate Information Extraction from Research Papers using Conditional Random Fields

Fuchun Peng
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
fuchun@cs.umass.edu

Andrew McCallum
Department of Computer Science
University of Massachusetts
Amherst, MA 01003
mccallum@cs.umass.edu

Abstract

With the increasing use of research paper search engines, such as CiteSeer, for both literature search and hiring decisions, the accuracy of such systems is of paramount importance. This paper employs Conditional Random Fields (CRFs) for the task of extracting various common fields from the headers and citation of research papers. The basic theory of CRFs is becoming well-understood, but

Previous work in information extraction from research papers has been based on two major machine learning techniques. The first is hidden Markov models (HMM) (Seymore et al., 1999; Takasu, 2003). An HMM learns a generative model over input sequence and labeled sequence pairs. While enjoying wide historical success, standard HMM models have difficulty modeling multiple non-independent features of the observation sequence. The second technique is based on discriminatively-trained SVM classifiers (Han et al., 2003). These SVM classifiers can handle many non-

Extraktion von Referenzen

Definition (Referenz-Titel & Referenz-Anhang)

> [1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Pasca, and A. Soroa. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In Proceedings of NAACL-2009, pages 1927, 2009.

Ein *Referenz-Titel* bezeichne die erste Zeile der Referenz (rot markiert). Alle weiteren Zeilen der Referenz seien *Referenz-Anhang* genannt (blau markiert).

- Für jede Textzeile des Literaturverzeichnisses: Entscheidung, ob Textzeile ein Referenz-Titel oder ein Referenz-Anhang ist.
- Diese Entscheidung hängt von den Strukturmerkmalen des Literaturverzeichnisses ab.

Extraktion von Referenzen (2)

Strukturmerkmale von Literaturverzeichnissen

LV mit Ankern und ohne Einrückungen

References

[1] UNESCO publications for the World Summit on the Information Society. Twelve years of measuring linguistic diversity in the Internet: balance and perspectives. http://unesdoc.unesco.org/images/0018/001870/187016e.pdf.

[2] Xu Wei, Jun Yan. An E-learning System Architecture Based on Web Services and Intelligent Agents. Ninth International Conference on Hybrid Intelligent Systems, (2009), 173-177.

LV ohne Anker und mit Einrückungen

REFERENCES

Adzhubei,I.A. *et al.* (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.

Altshuler,D. *et al.* (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*, **407**, 513–516.

Bansal,V. (2010) A statistical method for the detection of variants from next-generation resequencing of DNA pools. *Bioinformatics*, **26**, i318–i324.

Bentley,D.R. (2006) Whole-genome re-sequencing. *Curr. Opin. Genet. Dev.*, **16**, 545–552.

Campagna,D. *et al.* (2009) PASS: a program to align short sequences. *Bioinformatics*, **25**, 967–968.

Chen,K. *et al.* (2007) PolyScan: an automatic indel and SNP detection approach to the

LV mit Ankern und mit Einrückungen

8. REFERENCES

[1] E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of NAACL-2009*, pages 19–27, 2009.

[2] K. Bellare, P. Talukdar, G. Kumaran, F. Pereira, M. Liberman, A. McCallum, and M. Dredze. Lightly-Supervised Attribute Extraction. In *NIPS Workshop on Machine Learning for Web Search*, 2007.

LV ohne Anker und ohne Einrückungen

References

Alexander, T.B., et al 1994. Corporate Business servers: An Alternative to Mainframes for Business Computing. Hewlett-Packard Journal 45 No. 3: 8-30.

Bach, M.J. 1986. The Design of the UNIX@ Operating System. Englewood Cliffs, N.J: Prentice-Hall, Inc.

Extraktion von Referenzen (3)

Beobachtungen

1. Die erste Zeile des LV ist immer ein Referenz-Titel.
2. Die Strukturmerkmale eines LV sind konsistent.
3. Referenz-Titel und Referenz-Anhänge besitzen ihre Merkmale exklusiv.
4. Das Einrückungslevel ist höchstens 1.

- Im Folgenden: Gleichzeitige Betrachtung der Textzeilen

$$l_{i-1}, \quad l_i, \quad l_{i+1}$$

mit $0 < i < n$ und $n = \#\text{Textzeilen der Publikation}$.

Extraktion von Referenzen (4)

Ist l_i ein Referenz-Titel oder ein Referenz-Anhang?

- **LV enthält Referenz-Anker:**

Extraktion von Referenzen (4)

Ist l_i ein Referenz-Titel oder ein Referenz-Anhang?

- **LV enthält Referenz-Anker:**
  - l_i ist ein Referenz-Titel, wenn l_i einen Referenz-Anker enthält.

Extraktion von Referenzen (4)

Ist l_i ein Referenz-Titel oder ein Referenz-Anhang?

- **LV enthält Referenz-Anker:**
  - l_i ist ein Referenz-Titel, wenn l_i einen Referenz-Anker enthält.
  - l_i ist ein Referenz-Anhang, wenn l_i keinen Referenz-Anker enthält.

Extraktion von Referenzen (4)

Ist l_i ein Referenz-Titel oder ein Referenz-Anhang?

- **LV enthält Referenz-Anker:**
  - l_i ist ein Referenz-Titel, wenn l_i einen Referenz-Anker enthält.
  - l_i ist ein Referenz-Anhang, wenn l_i keinen Referenz-Anker enthält.
- **LV enthält eingerückte Referenz-Anhänge:**

Extraktion von Referenzen (4)

Ist l_i ein Referenz-Titel oder ein Referenz-Anhang?

- **LV enthält Referenz-Anker:**
  - l_i ist ein Referenz-Titel, wenn l_i einen Referenz-Anker enthält.
  - l_i ist ein Referenz-Anhang, wenn l_i keinen Referenz-Anker enthält.
- **LV enthält eingerückte Referenz-Anhänge:**
  - l_i ist ein Referenz-Titel, wenn l_{i-1} oder l_{i+1} Im Vergleich zu l_i eingerückt ist.

Extraktion von Referenzen (4)

Ist l_i ein Referenz-Titel oder ein Referenz-Anhang?

- **LV enthält Referenz-Anker:**
  - l_i ist ein Referenz-Titel, wenn l_i einen Referenz-Anker enthält.
  - l_i ist ein Referenz-Anhang, wenn l_i keinen Referenz-Anker enthält.
- **LV enthält eingerückte Referenz-Anhänge:**
  - l_i ist ein Referenz-Titel, wenn l_{i-1} oder l_{i+1} Im Vergleich zu l_i eingerückt ist.
  - l_i ist ein Referenz-Anhang, wenn l_i im Vergleich zu l_{i-1} oder l_{i+1} eingerückt ist.

Extraktion von Referenzen (4)

Ist l_i ein Referenz-Titel oder ein Referenz-Anhang?

- **LV enthält Referenz-Anker:**
  - l_i ist ein Referenz-Titel, wenn l_i einen Referenz-Anker enthält.
  - l_i ist ein Referenz-Anhang, wenn l_i keinen Referenz-Anker enthält.
- **LV enthält eingerückte Referenz-Anhänge:**
  - l_i ist ein Referenz-Titel, wenn l_{i-1} oder l_{i+1} Im Vergleich zu l_i eingerückt ist.
  - l_i ist ein Referenz-Anhang, wenn l_i im Vergleich zu l_{i-1} oder l_{i+1} eingerückt ist.
  - Spezialfälle...

Gliederung

Literaturdatenbanken

Definition (Literaturdatenbank)

Eine Literaturdatenbank speichert zu wissenschaftlichen Publikationen eines bestimmten Fachgebiets ihre bibliografischen Metadaten, wie z.B. den Titel, die Autoren, das Erscheinungsjahr, den Namen des Konferenzbandes, etc.

- DBLP
  - Literaturdatenbank aus dem Bereich der Informatik
  - 1,7 Millionen Einträge
  - XML-Datei `dblp.xml` ($\sim$ 850 MB)
- Medline
  - Literaturdatenbank aus dem Bereich der Medizin
  - 19 Millionen Einträge
  - XML-Datei `medline.xml` ($\sim$ 5,3 GB)

Literaturdatenbanken (2)

Ausschnitt aus dblp.xml

```
<dblp>
   [...]
   <inproceedings key="conf/sigcomm/KrishnamurthyW00">
      <author>Balachander Krishnamurthy</author>
      <author>Jia Wang</author>
      <title>On network-aware clustering of web[...]</title>
      <pages>97-110</pages>
      <year>2000</year>
      <booktitle>SIGCOMM</booktitle>
      <ee>http://doi.acm.org/10.1145/347059.347412</ee>
      <url>db/conf/sigcomm/sigcomm2000.html#Krish[...]</url>
   </inproceedings>
   [...]
</dblp>
```

Allgemeines Prinzip der Zuordnung

- Indexbasierte Suche nach Literaturdatenbank-Einträgen
- *Invertierter Index*:
  - Prinzip vergleichbar mit Stichwortverzeichnis eines Buches:
  - Auflistung aller Positionen, an denen ein bestimmtes Wort vorkommt:

$$\begin{aligned} \texttt{Wort 1} &\mapsto 1, 4, 5, 12, \ldots \\ \texttt{Wort 2} &\mapsto 2, 3, 4, \ldots \\ \texttt{Wort 3} &\mapsto 4, 7, 8, \ldots \\ &\ldots \\ \texttt{Wort } n &\mapsto \ldots \end{aligned}$$

  - `Wort 1` kommt in den Dokumenten 1, 4, 5 und 12 vor; `Wort 2` in den Dokumenten 2, 3 und 4; etc.

Suche mit einem invertierten Index

1. Normalisiere Suchanfrage Q:

$$Q = \texttt{Extending C++ with an Object Query Capability.}$$

$$\Downarrow$$

$$\{\texttt{extending}, \texttt{object}, \texttt{query}, \texttt{capability}\}$$

Suche mit einem invertierten Index

1. Normalisiere Suchanfrage Q:

$$Q = \texttt{Extending C++ with an Object Query Capability.}$$

$$\Downarrow$$

$$\{\texttt{extending}, \texttt{object}, \texttt{query}, \texttt{capability}\}$$

2. Bestimme für jedes Einzelwort die invertierte Listen:

$$\begin{aligned}
\texttt{extending} &\mapsto 2, 5, 7 \\
\texttt{object} &\mapsto 1, 2, 4, 5 \\
\texttt{query} &\mapsto 5 \\
\texttt{capability} &\mapsto 2, 4, 5
\end{aligned}$$

Suche mit einem invertierten Index

1. Normalisiere Suchanfrage Q:

$$Q = \texttt{Extending C++ with an Object Query Capability.}$$

$$\Downarrow$$

$$\{\texttt{extending}, \texttt{object}, \texttt{query}, \texttt{capability}\}$$

2. Bestimme für jedes Einzelwort die invertierte Listen:

$$\begin{aligned} \texttt{extending} &\mapsto \texttt{2, 5, 7} \\ \texttt{object} &\mapsto \texttt{1, 2, 4, 5} \\ \texttt{query} &\mapsto \texttt{5} \\ \texttt{capability} &\mapsto \texttt{2, 4, 5} \end{aligned}$$

3. Vereinige invertierte Listen und sortiere Elemente nach ihren Häufigkeiten:

$$\{\texttt{5, 2, 4, 1, 7}\}$$

Bewertung der Kandidaten

- Zusätzliche Bewertung der ersten k Kandidaten.
- Unterscheidung, ob Q einen Titel oder eine Referenz enthält.
- Wenn Q ein **Titel** enthält:
  - Für jeden Kandidaten K: Berechnung der Distanz zwischen dem Titel von K und Q mithilfe der *Levenshtein-Distanz*.
- Wenn Q eine **Referenz** enthält:
  - Für jeden Kandidaten K: Berechnung eines Autoren-Scores $s_A(K)$, eines Jahres-Scores $s_Y(K)$ und eines Titel-Scores $s_T(K)$.
  - Titel-Score $s_T(K)$: Finden einer Teilzeichenkette von Q und dem Titel von K mit maximaler Ähnlichkeit.

Gliederung

Der Ablauf der Experimente

- Durchführung von Experimenten zur Evaluation der
  - Korrektheit der Ergebnisse
  - Effizienz der Algorithmen
- Testdatensatz:
  - 700 Publikationen aus DBLP (aus den Jahren 2000 - 2011)
  - 500 Publikationen aus Medline (aus den Jahren 2001 - 2010)
- Manuelle Bestimmung von ...
  - ... korrekten Titeln und Referenzen
  - ... Keys der repräsentierenden Literaturdatenbank-Einträge.
- Vergleich dieser Daten mit den erhaltenen Ergebnissen.
- Testumgebung:
  - Rechner mit vier Intel Xeon X5560 Prozessoren, jeweils mit 2,8 GHz Taktfrequenz und 8MB Cache-Speicher; Hauptspeicher: 36.274 MB; Betriebssystem: Ubuntu 9.10 Karmic Koala, 64Bit.

Experimente zur Titel-Extraktion & -Zuordnung

Ergebnisse der Experimente

		Korrektheit		Laufzeit	
		DBLP	Medline	DBLP	Medline
Titel-Extraktion		94,7%	89,8%	43,4ms	42,6ms
Titel-Zuordnung	k=50				
	k=500				

Experimente zur Titel-Extraktion & -Zuordnung

Ergebnisse der Experimente

		Korrektheit		Laufzeit	
		DBLP	Medline	DBLP	Medline
Titel-Extraktion		94,7%	89,8%	43,4ms	42,6ms
Titel-Zuordnung	k=50	98,1%	84,2%	4,9ms	26,3ms
	k=500				

Experimente zur Titel-Extraktion & -Zuordnung

Ergebnisse der Experimente

		Korrektheit		Laufzeit	
		DBLP	Medline	DBLP	Medline
Titel-Extraktion		94,7%	89,8%	43,4ms	42,6ms
Titel-Zuordnung	k=50	98,1%	84,2%	4,9ms	26,3ms
	k=500	99,9%	92,2%	21,4ms	56,4ms

Experimente zur Referenzen-Extraktion & -Zuordnung

Ergebnisse der Experimente

		Korrektheit		**Laufzeit**	
		DBLP	Medline	DBLP	Medline
Referenzen-Extraktion					
Ref.-Zuordnung	k=50				
	k=500				

Experimente zur Referenzen-Extraktion & -Zuordnung

Ergebnisse der Experimente

		Korrektheit		**Laufzeit**	
		DBLP	Medline	DBLP	Medline
Referenzen-Extraktion		81,6%	91,1%	141,8ms	170,7ms
Ref.-Zuordnung	k=50				
	k=500				

Experimente zur Referenzen-Extraktion & -Zuordnung

Ergebnisse der Experimente

		Korrektheit		**Laufzeit**	
		DBLP	Medline	DBLP	Medline
Referenzen-Extraktion		81,6%	91,1%	141,8ms	170,7ms
Ref.-Zuordnung	k=50	90%	83,1%	4,3ms	26,9ms
	k=500				

Experimente zur Referenzen-Extraktion & -Zuordnung

Ergebnisse der Experimente

		Korrektheit		Laufzeit	
		DBLP	Medline	DBLP	Medline
Referenzen-Extraktion		81,6%	91,1%	141,8ms	170,7ms
Ref.-Zuordnung	k=50	90%	83,1%	4,3ms	26,9ms
	k=500	94,8%	93,6%	8,4ms	37,9ms