1 PG-Thema

Intelligence Service - gezielte Informationen aus dem Internet

2 PG-Zeitraum

WS 07/08 und SoSe 08

3 PG-Umfang

jeweils 8 SWS

4 PG-Veranstalter

Prof. Dr. Katharina Morik, Dipl.-Inform. Felix Jungermann, Informatik LS 8, GB IV, Raum 123, Tel.: 5104

5 PG-Aufgabe

Ziel der PG ist das automatische Erstellen eines Pressespiegels für eine bestimmte Person (z.B. einen Politiker) oder eine bestimmte Firma aus dem Internet bzw. aus Datenbanken. Daraus sollen dann gezielt Antworten auf bestimmte Fragen extrahiert werden. Methoden zu einem solchen *Intelligence Service* werden untersucht und implementiert.

Natürlich soll die PG über das reine Datensammeln hinausgehen. Prinzipiell ist aber schon dieser Punkt erwähnenswert, da, um einen objektiven Pressespiegel für eine Person zu erstellen, möglichst viele Quellen benutzt werden müssen. Die ausschließliche Nutzung einer bestimmten Biographie-Seite kann somit auf keinen Fall ausreichen. Ein breiteres Spektrum stellt die Nutzung von Suchmaschinen zur Informationsgewinnung dar.

Allerdings ist das Spektrum der Informationen für eine einzige Anfrage hierbei zu groß– das Problem ist, die interessanten Daten zwischen den uninteressanten Daten herauszufinden. Dies ist das Problem des Information Retrieval. Der zu entwickelnde *Intelligence Service* soll natürlich über das Information Retrieval von Suchmaschinen hinausgehen.

Das grundsätzliches Problem ist, dass Suchmaschinen nicht konkrete Antworten liefern. Vielmehr wird eine Auswahl an Dokumenten geliefert, die die Antwort zu gestellten Anfrage höchstwahrscheinlich enthält. Was man aber oft möchte, ist auf eine Frage wie:

"Wann und wo findet die ICDM-07 statt?"

die Antwort: "28.-31.10.2007, Omaha, Nebraska, USA"

zusammen mit der URL, auf der die Information gefunden wurde, zu erhalten. Für solche Fragebeantwortung muss man nicht nur die relevanten Dokumente finden, sondern auch die relevanten Passagen – ein weiterer Punkt, der

von Suchmaschinen nicht erbracht wird. Wenn die Dokumente durch eine Auszeichnungssprache (XML) annotiert sind, ist die Suche in den relevanten Dokumenten erleichtert, so dass gezielt etwa nach Investitionen, Erfolgen, neuen Produkten, Börsenzahlen gesucht werden kann. Die meisten Dokumente sind aber nicht annotiert. Man muss also algorithmisch nach Entitäten eines bestimmten Typs (z.B. Person, Ort, Firma) suchen. Das Gebiet, das sich mit der Erkennung der *Entitäten* eines inhaltlichen Typs in Texten befasst, ist die *Named Entity Recognition* (NER) und verwendet statistische Verfahren und solche des maschinellen Lernens bzw. Data Mining. Somit ist die NER ein weiterer Bereich, mit dem sich die PG befassen muss.

Selbst wenn wir das Problem, die interessanten Informationen zu erfassen, einmal als gelöst betrachten, weist das Recherchieren noch mindestens ein anderes Problem auf, nämlich die strukturierte Zusammenstellung von Informationen zu einem Gesamtbild. Beispielsweise wollen Firmen oft einen Überblick über ihre Konkurrenz oder ihr eigenes Image in der Öffentlichkeit erhalten. Solche Recherchen werden oft noch von Hand durch Abfolgen von Anfragen an Suchmaschinen und das Verfolgen von *links* durchgeführt. Die Abfolge von Anfragen sollte jedoch automatisiert erfolgen, um ein allgemein nutzbares System zu schaffen. Für Politiker bietet sich hierfür beispielsweise die Internetseite *Bundestag.de* an. Hier sind zu jedem Abgeordneten die jeweiligen Biographien hinterlegt. Zusätzlich zu diesen offensichtlichen Daten kann man jedoch auch noch die digital vorliegenden Drucksachen (z.B. Anträge) und Protokolle verarbeiten. Nach durchgeführter NER über diesen Dokumenten sollen dann konkrete Fragen beantwortet werden.

5.1 Problembeschreibung

Das Auffinden von Fakten aus WWW-Seiten wurde von Tom Mitchell untersucht [CDF+00], wobei allerdings die Beschaffung der Seiten reine Handarbeit war und NER noch nicht als entscheidender Schritt zur Lösung des Problems aufgefasst wurde. Trotzdem ist sein Ziel, Wissen direkt aus dem WWW zu extrahieren, auch das Ziel dieser PG.

Lernende Suchmaschinen wie z.B. Google [BP98] werden durch *alle* ihre Benutzer trainiert. Will man die Suchergebnisse an bestimmte, eigene Interessantheitskriterien anpassen (personalisierte Suche), so müssen die Klickfolgen bestimmter Benutzer in den Suchergebnissen gespeichert und für das Lernen genutzt werden [Joa02]. Auf diese Weise wird die Suche auf das Thema und die darin interessanten Aspekte (Ereignisse in der Politik oder Wirtschaft) eingeengt.

Für intelligente Services werden Indizes gebraucht, die angeben, unter welcher URL (und an welcher Stelle im Dokument) ein Inhalt (eine entity) zu finden ist. Die Dokumente selbst werden durch einen *crawler* aufgefunden, der Querverweise verfolgt. Im Falle der Nutzung der Seite *Bundestag.de* könnte für jeden Politiker die von ihm mitverantworteten Anträge bzw. seine im Protokoll vermerkten Aussagen über einen Index zugreifbar gemacht werden. Aus den so

gewonnenen Informationen lassen sich dann gezieltere Anfragen an Suchmaschinen oder Zeitungsarchive stellen.

Will man nun innerhalb der Dokumente die Entitäten lernen lassen, so braucht man ein den Sequenz-Charakter des Textes berücksichtigendes Lernverfahren und möglichst ein Verzeichnis bekannter Eigennamen. Aktuell genutzte Verfahren zur NER sind support vector machines (SVMs) ([Bur98]), maximum entropy Markov models (MEMMs) ([MFP00]), hidden Markov models (HMMs) ([Rab89]) sowie conditional random fields (CRFs) ([LMP01]). Als Eingabe in das Lernverfahren wird eine Menge bereits annotierter Texte benötigt. Es gibt bereits eine Reihe von Datensätzen, die aus Zeitungsarchiven extrahiert wurden bzw. kompletten Zeitungsarchiven entsprechen. Am bekanntesten sind der Reuters-[RSW02] und der "Frankfurter Rundschau"-corpus. Teile des Reuterscorpus sind zu Zwecken des Text Mining für die "Conference on Computational Natural Language Learning"(CoNLL) im Jahr 2003 mit Markierungen (tags) für Kategorien (named entities(NE)) versehen worden [TKSDM03].

Damit aus neuen, noch nicht annotierten Dokumenten die relevanten Entitäten gefunden werden können, müssen die Suchergebnisse, die gefundenen relevanten Dokumente oder WWW-Seiten, von dem gelernten NER-tagger annotiert werden. Bei neuen Ereignissen können RSS-feeds von Nachrichten-Seiten überprüft werden. Hier ist zu untersuchen, inwieweit RSS-feeds nutzbar sind, und ob der Vorteil der teilweise vorhandenen Meta-Informationen den Nachteil der vielleicht zu kurzen Darstellungsform aufwiegt.

Das am Lehrstuhl 8 entwickelte Data Mining-Werkzeug Yale ([MWK+06]) erleichtert die vielen Experimente, die die PG durchführt: nicht nur sind fast alle Lernverfahren darin bereits implementiert, sondern auch viele Vorverarbeitungs- und Evaluierungsschritte. Insbesondere helfen die Text Mining Operatoren wie WordVectorTool und CRF-plugin ([Jun06]). Eigene Entwicklungen der PG lassen sich bei Bedarf als plugins leicht integrieren.

6 PG-Teilnahmevoraussetzungen

Maschinelles Lernen (M), Wissensentdeckung in Datenbanken (M),

Informationsssysteme (W), Evolutionäre Algorithmen (W), Spracherkennung (W), Java-Kenntnisse (W)

Mindestens eine der mit (M) markierten Vorlesungen muss erfolgreich besucht worden sein; wünschenswert sind mindestens zwei der mit (W) markierten Vorkenntnisse.

7 Minimalziel

Mindestens entwickelt werden sollte:

1. Die Definition und Nutzung eines Anfrageplans zur Erfassung von Informationen über einen Politiker oder eine Firma mitsamt der Indexierung gefundener Informationen – (Welche Informationen will ich von welchen

Internetseiten erfassen? Wie sehen die Wege aus, die ein *crawler* beschreiten muss, um automatisch an diese Informationen zu gelangen?)

2. Die Definition, Entwicklung und Nutzung konkreter NER-Verfahren.

3. Eine Methode zur Ausnutzung der NER für die Fragebeantwortung auf Grund einer selbst zusammengetragenen Dokumentsammlung.

8 Literatur

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual (web) search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[Bur98] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[CDF+00] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew K. McCallum, Tom M. Mitchell, Kamal Nigam, and Sean Slattery. Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118(1/2):69–113, 2000.

[Joa02] Thosten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of Knowledge Discovery in Databases*, 2002.

[Jun06] Felix Jungermann. Named entity recognition mit conditional random fields. Master's thesis, Fachbereich Informatik, Universitat Dortmund, 2006.

[LMP01] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[MFP00] Andrew McCallum, Dayne Freitag, and Fernando Pereira. Maximum Entropy Markov Models for information extraction and segmentation. In *Proc. 17th International Conf. on Machine Learning*, pages 591–598. Morgan Kaufmann, San Francisco, CA, 2000.

[MS99] Chris Manning and Hinrich Schuetze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, May 1999.

[MWK+06] Ingo Mierswa, Michael Wurst, Ralf Klinkenberg, Martin Scholz, and Timm Euler. YALE: Rapid Prototyping for Complex Data Mining Tasks. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*. ACM Press, 2006.

[Rab89] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257–286, 1989.

[RSW02] T.G. Rose, M. Stevenson, and M. Whitehead. The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 29–31, 2002.

[TKSDM03] Erik Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task - language independent named entitiy recognition. In *Proceedings of CoNLL-2003*, Edmonton, Canada, 2003. Morgan Kaufmann.

9 PG-Realisierung

Die PG beginnt in jedem Semester mit einer Seminarphase und endet mit einer Evaluation. Die Studierenden planen ihre Arbeiten selbst, entscheiden über die Werkzeuge, die sie verwenden wollen (z.B. CVS, LaTex, XML-Editor), teilen sich in kleinere teams ein – allerdings wird dies in den PG-Sitzungen von den Veranstaltern kritisch begleitet, damit es nicht zu falschen Zeitabschätzungen kommt und die PG erfolgreich innerhalb der 2 Semester abgeschlossen wird.

1. **Seminarphase:** 08.-12.10.2007

   - Lernverfahren:
     - SVMs [Bur98],
     - MEMMs [MFP00],
     - HMMs [Rab89],
     - CRFs [LMP01].
   - Methoden der Personalisierung von Suchmaschinen [Joa02]
   - Basistechniken des WWW:
     - Indexierung
     - XML (RSS)
   - Werkzeuge der PG:
     - Yale
     - Google API
     - CVS

Aufgaben des 1.Semesters:

   - Wahl eines Anwendungsszenarios, Festlegen der Fragen und der NE
   - Erstellen der Trainingsdatensätze
   - Experimente mit vorhandenen NER-Lernverfahren
   - Ansätze zur Verbesserung der NER

2. **Seminarphase:** 04.-08.02.2008 werden die Ergebnisse zusammengetragen, diskutiert und dokumentiert.

3. **Seminarphase:** Anfang April 2008

   - automatische Thesaurus-Erstellung
     - Chen et al.: Building a Web Thesaurus from Web Link Structure, 2003.
     - Chen et al.: Automatic Thesaurus Generation for an Electronic Community System, 1995.
   - automatische Fragebeantwortung

  - Pasca and Harabagiu: Answer Mining from On-Line Documents, 2001.
  - TREC Question-Answering Track Publications
- Text-Clustering
  - Zeng et al.: Learning to Cluster Web Search Results, 2004.
  - Zamir and Etzioni: Grouper: A dynamic clustering interface to web search results, 1999.
  - Wurst et al.: Localized Alternative Cluster Ensembles for Collaborative Structuring, 2006.
- Webseiten-Ranking
  - Page et al.: The PageRank citation ranking: Bringing order to the Web, 1998.
  - Kleinberg: Authoritative Sources in a Hyperlinked Environment, 1999.
  - Diligenti et al.: A Unified Probabilistic Framework for Web Page Scoring Systems, 2004.
- First Story Detection
  - Zhang et al.: Novelty and Redundancy Detection in Adaptive Filtering, 2002.
  - Allan et al.: First Story Detection In TDT Is Hard, 2000.
- Topic Tracking
  - Matsumura et al.: Discovery of Emerging Topics between Communities on WWW, 2001.
  - Matsumura et al.: Future Directions of Communities on the Web, 2001.
  - Jatowt et al.: Change Summarization in Web Collections, 2004.

Aufgaben des 2.Semesters:

- Sammeln der Daten (mittels crawling-Techniken oder z.B. clickstream-Analyse)
- Erstellung des Anfrageplans zur Erfassung von Informationen über einen Politiker oder eine Firma
- Konkrete Ausnutzung der NER (Bezug auf erstes Semester) zur Fragebeantwortung

Ergebnis des 2. Semesters: PG Abschlussbericht und -präsentation