Review zur Ausarbeitung „Product Information Extraction: Wie gewinnt man Produktinformationen aus semistrukturierten Quellen“

Max Mustermann <max.mustermann@email.xx>

Zusammenfassung

Die Ausarbeitung beschreibt zwei Methoden zur automatischen Extraktion von Produktinformationen aus semistrukturierten Quellen. Eine Einführung erklärt den Kontext der Informationsextraktion aus Internetseiten. Im zweiten Abschnitt wird der Begriff der semistrukturierten Quelle (Webseite) definiert. Nach generellen Erläuterung und Ansätzen zur Informationsgewinnung aus semistrukturierten Webseiten folgt die Beschreibung der beiden Methoden. Das Template-basierte Verfahren beruht auf der Annahme, dass Webseiten aus einer Vorlage entstehen, die mit Daten aus einer Datenbank gefüllt wird. Das MDR-Verfahren (Mining Data Records) nutzt aus, dass sich Produktinformationen oft in bestimmten Bereichen von Webseiten befinden. Außerdem lässt sich die Knotenstruktur von HTML-Dokumenten dazu benutzen, gemeinsame Elternknoten zu identifizieren, unterhalb derer relevante Informationen zu finden sind.

Positives

- Die erste Hälfte des Abstracts ist klar und gut geschrieben.
- Abb. 3 illustriert die Gegebenheiten/die Struktur sehr einleuchtend.
- Abschnitt 2 ist schlüssig und verständlich.

Fragen, Verbesserungsvorschläge und Kritikpunkte

- Titel, Abstract und Keywords weisen nicht oder nur implizit darauf hin, dass es sich um Extraktion von Informationen aus dem WWW handelt.
- Die zweite Hälfte des Abstracts ist für mich etwas verschachtelt und undurchsichtig.
- Der erste Absatz der Einführung ist ohne Quellenbeleg oder praktisches Beispiel.
- Ist der Unterschied zwischen Webseite und Website unklar (Abb. 1)?
- Im Abschnitt 3 erscheint mir der Token-Scanner unzureichend erklärt.
- Im Abschnitt 3 auf Seite 4 gibt es komplizierte Schachtelsätze.
- Im Abschnitt 3 ist der letzte Absatz auf Seite 4 zum Teil unverständlich. Was sind die flachen Verarbeitungskomponenten? Basieren die in [2] genannten Verfahren tatsächlich nicht mehr, wie suggeriert wird, auf dem Lernen von manuellen Benutzereingaben?
- Auf Seite 6, Abschnitt 4.2: Wie wird sichergestellt, ob es sich tatsächlich um eine relevante Liste handelt?
- Auf Seite 7 ist der erste Absatz von „Identifikation von Spalten und Auszügen“ unverständlich. Was sind Auszüge? Sind Spalten Spalten im Sinne von Datenfeldern oder Spalten einer Tabelle? Die Begriffe in diesem Abschnitt erscheinen unklar und behindern mein Verständnis (Was heißt es, wenn Auszüge bewertet werden? Was sind Inhaltsfeatures?).
- Auf Seite 7: „Identifikation von Reihen“ ist zum Teil verschachtelt und unverständlich. Was genau ist ein Produkteintrag? Funktioniert dieses Verfahren immer (vgl. „grundsätzlich“)?
- Zu Abschnitt 5: Inwiefern unterscheiden sich die beiden Gegebenheiten von der Template-

Annahme des anderen Verfahrens?

- Zu Abschnitt 5.1: Was ist das Durchführen von Kombinationen?
- Unter 5.2 wird im ersten Absatz zum Teil ungenau beschrieben/definiert.
- Zum Abschnitt „Bestimmen von Datenregionen“: Der Begriff der Datenregion scheint unklar. Im ersten Absatz von 5. war von nur einem Bereich, genannt Datenregion, die Rede. Der Unterschied von Datenregion und Dateneintrag ist mir unklar. Insgesamt wirkt dieser Abschnitt auf mich sehr unkonkret und diffus. Ich kann von mir nicht behaupten, dass ich nun in etwa weiß, wie das darin Beschriebene funktioniert. Zudem lässt die „Editierdistanz-Schwelle“ Einschränkungen des Verfahrens (z.B. der Genauigkeit) vermuten.
- „Product Information“ würde ich in den Keywords erwähnen.
- Das Fazit ist ohne Belege.
- Manche Rechtschreibfehler (insbesondere Interpunktion) beeinträchtigen zum Teil die Lesbarkeit und/oder Verständlichkeit.
- Da die Quellen relativ alt sind, ist für mich fraglich, ob der aktuelle Stand von Forschung bzw. Praxis in diesem Gebiet dargestellt wird (vgl. Suche nach „Product Information Extraction“ in der ACM Digital Library).
- Recherche-Tipp: „Webknox“ (z.B. http://es.csiro.au/adcs2008/proceedings/p05-urbansky.pdf), was scheinbar auch an der TU Dresden entwickelt wurde.
- In welchem Zusammenhang stehen die beiden beschriebenen Methoden? Wo wird das implizite Wissen abgeleitet und dem Benutzer zur Verfügung gestellt (vgl. Einleitung)? Wo erfolgt der Bezug oder die Spezialisierung auf Produktinformation, die der Titel vermuten lässt?