Bachelor-Thesis 2010

Untersuchung frei verfügbarer Geocoder

Autor: **Thomas Meister**

Examinator: **Prof. H.-J. Stark**

Experte: **Jean Marc Buttliger**

Untersuchung frei verfügbarer Geocoder

Geocoder sind Kernelemente wenn es darum geht, Daten in ihren räumlichen Bezug zu bringen. Dabei können Adressdaten wie auch andere wichtige Informationen im Vordergrund stehen. Als Grundlage dient dabei ein Referenzdatensatz, anhand dessen eine Übereinstimmung mit den Ausgangsdaten gesucht wird. Untersuchungen zeigten, dass bei grossen Datenmengen eine sequentielle Suche einen negativen Einfluss auf die Performance hat. Daher sind bei einem grossen Referenzdatensatz geeignete Massnahmen (z.B. Indexierung der Tabellenspalten) zu treffen. Es gibt einzelne Geocodier-Mechanismen, die bereits in offenen Paketen zur Verfügung stehen und in eigene Anwendungen integriert werden können.

Schlagworte: Geocoder, Openaddresses, PostgreSQL, Prototyp, Normalizer, Phonetik, Fuzzy, Performance,

1. Bestehende Lösungen/Komponenten

Einzelne Geocodier-Mechanismen stehen zum Teil in offenen Paketen zur Verfügung und können in eigene Anwendungen integriert werden. Die Anzahl der verfügbaren Mechanismen ist jedoch noch sehr begrenzt. Dazu kommt, dass diese Mechanismen spezifisch für einen bestimmten Referenzdatensatz konzipiert wurden und daher nur unter grossem Aufwand für eigene Anwendungen verwendet werden können. So gibt es zum Beispiel einen komplett frei verfügbaren Geocoder für Linien- und Polygondaten, der jedoch für Anwendungen mit Punktdaten als Referenzdatensatz nicht geeignet sind.

2. Untersuchungen

Einzelne Komponenten des Geocodier-Prozesses wurden in einer Testumgebung implementiert und untersucht. Als Referenzdatensatz diente dabei der komplette Datensatz von Openaddresses. Komponenten, die implementiert wurden:

- Adressnormalisierer (Extraktion von Adresskomponenten aus den Ausgangsdaten)
- Matchingverfahren nach Hausnummer, Strasse, PLZ, Ort und Land
- Interpolation von Hausnummern → Wird keine exakte eindeutige Übereinstimmung gefunden im Referenzdatensatz gefunden, besteht die Möglichkeit zwischen zwei sehr ähnlichen Übereinstimmungen zu interpolieren.
- Kölner-Phonetik Algorithmus, für Testuntersuchungen bezüglich der Sprachabhängigkeit

2.1. Performance

Bei Anwendungen, die auf eine Datenbank zugreifen, ist die Performance (dt. Leistung) ein wichtiger Indikator, welcher die Benutzerakzeptanz der Anwendung stark beeinflussen kann.

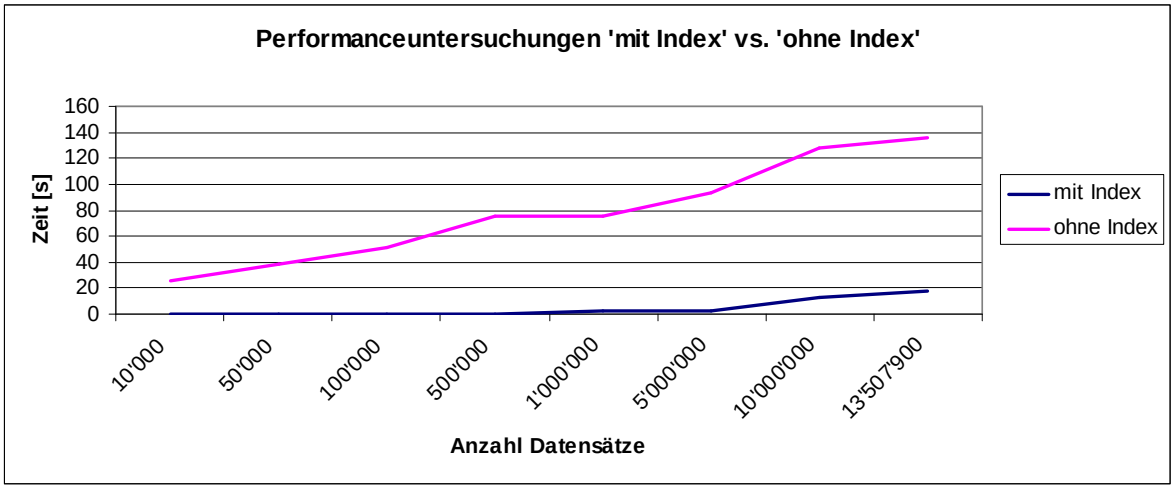

Abb 1: Performance Untersuchungen

2.2. Phonetische Algorithmen

Mit Hilfe von phonetischen Algorithmen wird versucht, über den Klang (Phonetik) Übereinstimmungen im Referenzdatensatz zu finden. Dadurch können kleinere Rechtschreibfehler in den Ausgangsdaten vernachlässigt werden. Es existieren viele unterschiedliche phonetische Algorithmen die zum Teil sprachspezifisch entworfen wurden. Diese Sprachabhängigkeit wurde mit jeweils einem Testdatensatz in Englisch und Deutsch für die beiden phonetischen Algorithmen Soundex (Englisch) und Kölner-Phonetik (Deutsch) untersucht.

Original	**Verfälscht**	**Kölner Verfahren**	**Soundex Verfahren**
		Phonetisch identisch?	
Autobahn	Autoba_n	✓	✓
Denkmal	De_kmal	⊘	⊘
Rennvelo	Rennwelo	✓	⊘
Küchenschrank	Küschenschrank	⊘	⊘
Katzenjammer	Ka_zenjammer	✓	⊘
Brückenhecke	Brueckenh_cke	✓	⊘

Abb 2: Vergleich Soundex vs. Kölner-Phonetik, Testdatensatz Deutsch

Original	Verfälscht	Kölner Verfahren	Soundex Verfahren
		Phonetisch identisch?	
highway	hig_way	✅	✅
squirrel	squir_el	✅	✅
pig	pigs	⛔	✅
basically	basical_y	⛔	✅
trolley	t_olley	⛔	⛔
forklift	forkslift	⛔	✅

Abb 3: Vergleich Soundex vs. Kölner-Phonetik, Testdatensatz Englisch

Die Verfahren sind für ihre jeweilig konzipierte Sprache geeignet und sollten daher nicht für phonetische Untersuchungen in einer anderen Sprache verwendet werden. Das Soundex Verfahren ist einfacher aufgebaut und daher auch ungenauer als das Kölner-Phonetik Verfahren. Für Anwendungen mit deutschsprachigen Zeichenketten empfiehlt es sich, das Kölner-Phonetik Verfahren zu verwenden, da dieses Ziffernfolgen berücksichtigt.

3. Kontaktbalken am Ende der Präsentation

Autor:	Thomas Meister	toeme.meister@bluewin.ch
Examinator:	Prof. H.J. Stark	hansjoerg.stark@fhnw.ch
Experte:	Jean Marc Buttliger	Jean-Marc.Buttliger@bl.ch