Stefan Wagner

Barrierefreie und thesaurusbasierte Suchfunktion für das Webportal der Stadt Nürnberg

Diplomarbeit

Diplom.de

Bibliografische Information der Deutschen Nationalbibliothek:

Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de/ abrufbar.

ISBN: 9783836607612

Stefan Wagner

Barrierefreie und thesaurusbasierte Suchfunktion für das Webportal der Stadt Nürnberg

Diplom.de

Diplomarbeit

ISBN: 978-3-8366-0761-2

Stefan Wagner

Barrierefreie und thesaurusbasierte Suchfunktion für das Webportal der Stadt Nürnberg

Diplom.de

Stefan Wagner
Barrierefreie und thesaurusbasierte Suchfunktion für das Webportal der Stadt Nürnberg

ISBN: 978-3-8366-0761-2
Druck Diplomica® Verlag GmbH, Hamburg, 2008
Zugl. Georg-Simon-Ohm-Fachhochschule Nürnberg, Nürnberg, Deutschland, Diplomarbeit, 2007

Kurzfassung

Im Internetportal der Stadt Nürnberg wurde in einer vorausgehenden Diplomarbeit eine Suchmaschine auf Basis des Produktes e:IAS der Fa. empolis GmbH realisiert. Diese Lösung soll in verschiedenen Bereichen verbessert und erweitert werden.

Es sollen aussagekräftige Logfiles generiert und ausgewertet werden, insbesondere sollen die Auswertungen mit denen der vorhergehenden Suchlösung vergleichbar sein.

Bei der Ergebnispräsentation sollen die Erfordernisse der Barrierefreiheit beachtet werden und die vorhandenen Templates entsprechende Anpassung erfahren.

Die Lösung soll um Ansätze semantischer Suche erweitert werden. Es ist angedacht die vorhandene Synonymverwendung auszubauen und um Taxonomien zu einem Thesaurus zu erweitern. Dabei sollen verschiedene Möglichkeiten untersucht werden und eine Möglichkeit, mindestens prototypisch, integriert werden.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Formelverzeichnis

1 Motivation

Das Schwerste: Immer wieder entdecken, was man ohnehin weiß.[1]

In dem Webportal der Stadt Nürnberg ist Wissen zu vielen verschiedenen Themen auf unzähligen Seiten gespeichert. Doch dieses Wissen ist nutzlos, wenn man es nicht findet. Diese Diplomarbeit soll dazu beitragen, dass die Informationen, die ein Besucher des Webportals sucht, von ihm auch gefunden werden.

Diese Arbeit baut auf der Diplomarbeit von Marek Ertel[2] auf und führt dessen Thema weiter. Neben der produktiven Inbetriebnahme der Suchmaschine auf Basis des Produktes e:IAS der Fa. empolis GmbH besteht die Arbeit aus drei Themen:

- Es sollen aussagekräftige Logfiles generiert und ausgewertet werden.
- Bei der Ergebnispräsentation sollen die Erfordernisse der Barrierefreiheit beachtet werden.
- Thesauri sollen die Lösung um Ansätze semantischer Suche erweitern.

[1] Elias Canetti (1905 - 1994), Schriftsteller spanisch-jüdischer Herkunft
[2] Siehe [Ertel2006]

2 Grundlagen

Dieses Kapitel soll einige Grundlagen klären, die für die vorliegende Diplomarbeit benötigt werden.

2.1 Textbasierte Suche

Die einfachste Suchmöglichkeit, um passende Dokumente zu finden, ist der Vergleich der Wörter in der Anfrage mit den Wörtern im Dokument; je mehr Wörter der Anfrage im Dokument vorhanden sind, desto relevanter ist es für den Benutzer. Um mehr Ergebnisse zu erzielen, können die Wörter vorher auf ihre Grundformen zurückgeführt werden, so wird z. B. „ging“ zu „gehen“ und „Häuser“ zu „Haus“. Diesen Vorgang nennt man Stemming.

Allerdings ist diese Methode zunächst nicht sehr gut geeignet, um die Relevanz eines Dokuments zu bestimmen, da beispielsweise nicht berücksichtigt wird, wie häufig das gesuchte Wort allgemein im Sprachgebrauch vorkommt. So ist z. B. „nicht“ das 16-häufigste Wort im Deutschen[3], es wird also in vielen Dokumenten der Suchbasis vorkommen und ist somit als Suchbegriff wesentlich schlechter geeignet als ein Wort, das nur selten verwendet wird. Wie relevant ein Dokument als Ergebnis einer Suche ist, hängt weiterhin sicher davon ab, wie häufig ein Suchbegriff in dem Dokument enthalten ist.

Beide Überlegungen werden mit dem Suchverfahren TF/IDF (Term Frequency / Inverted Document Frequency) verfolgt. Über die inverse Dokumenthäufigkeit (engl. IDF) bekommt ein Term, also ein Wort, das nur in wenigen Dokumenten der Dokumentenbasis vorkommt, einen höheren Wert als ein Wort, das in vielen Dokumenten der Dokumentenbasis erscheint. Formel 1 ist die dazugehörige Berechnungsfunktion, sie kann reelle Werte größer Null annehmen.

[3] Vgl. [Wortschatz]

$$idf_i = \log \frac{N}{n_i}$$

mit N Anzahl aller Dokumente

n_i Anzahl der Dokumente die Term i beinhalten

Formel 1: Inverse Dokumenthäufigkeit des Terms i

Die Termfrequenz (TF) gibt die relative Häufigkeit eines Wortes bzw. Terms in einem bestimmten Dokument an (siehe Formel 2). Sie kann Werte zwischen Null und Eins annehmen.[4]

$$tf_{i,d} = \frac{freq_{i,d}}{\max_l freq_{l,d}}$$

mit $freq_{i,d}$ Häufigkeit des Terms i im Dokument i

$max_i freq_{l,d}$ Häufigkeit des häufigsten Terms l im Dokument i

Formel 2: Termfrequenz des Terms i im Dokument d

TF/IDF ist das Produkt aus Termfrequenz und inverser Dokumenthäufigkeit und somit kann die Relevanz eines Dokuments zu einer Suchanfrage berechnet werden.

Folgendes Beispiel soll die Berechnung erläutern:

Das Dokument d_1 enthält die Wörter „Stadt Nürnberg“, das zweite Dokument d_2 „Stadt Schwabach“ und das dritte (d_3) „Landkreis Fürth“. Zuerst werden die Terme in Kleinbuchstaben umgewandelt und Umlaute durch ihre Umschreibung ersetzt. Danach erfolgt die Berechnung der IDF-Werte (siehe Tabelle 1).

[4] Vgl. [Ertel2006], Kapitel 2.4.2, S. 14 f

i	idf_i
stadt	$\log\frac{3}{2}$ = 0,176
schwabach	$\log\frac{3}{2}$ = 0,176
fuerth	$\log\frac{3}{1}$ = 0,477
nuernberg	$\log\frac{3}{1}$ = 0,477
landkreis	$\log\frac{3}{1}$ = 0,477

Tabelle 1: IDF-Werte

Ebenso werden die TF-Werte aller Terme berechnet (siehe Tabelle 2).

i / d	d1	d2	d3
stadt	1	1	0
schwabach	0	1	0
fuerth	0	0	1
nuernberg	1	0	0
landkreis	0	0	1

Tabelle 2: TF-Wert

Aus diesen Werten lassen sich dann die TF/IDF-Werte berechnen (siehe Tabelle 3), die einzelnen Spalten der Tabelle lassen sich nun auch als Gewichtsvektor des jeweiligen Dokuments lesen.

i / d	d_1	d_2	d_3
stadt	0,176	0,176	0
schwabach	0	0,477	0
fuerth	0	0	0,477
nuernberg	0,477	0	0
landkreis	0	0	0,477

Tabelle 3: TF/IDF-Werte

Wird nun eine Anfrage q mit dem Term „Stadt Nürnberg“ übermittelt, wird wieder wie oben der TF/IDF-Wert berechnet und als Anfragevektor bekommt man q = (0,176, 0,477, 0,0,0). Durch den direkten Vergleich sieht man, dass das Dokument d_2 exakt der Anfrage entspricht, die Ähnlichkeit also 1 ist, zum Dokument d_3 besteht gar keine Übereinstimmung, also eine Ähnlichkeit von 0 und zum Dokument d_1 besteht nur eine

teilweise Ähnlichkeit. Würde man die Werte in eine Berechnungsfunktion für die Größe der Ähnlichkeit einsetzten, auf die hier nicht weiter eingegangen wird (eine abgewandelte Form des Kosinusmaßes), würde man eine 35-prozentige Ähnlichkeit erhalten.[5]

Diese Form der Relevanzbestimmung wurde in einer Vorgängerdiplomarbeit in die e:IAS Suche integriert, die vorliegende Arbeit wird die Einbindung und Verwendung von Thesauri zur Relevanzbestimmung untersuchen.

2.2 Taxonomien und Thesauri

2.2.1 Was sind Taxonomien und Thesauri

Ein Thesaurus, im Sinne der Information und Dokumentation, ist nach DIN 1463-1[6] (bzw. ISO 2788) Teil eines Informationsystems. Seine wesentlichen Anwendungen lassen sich wie folgt darstellen:

Die wesentlichen Inhalte einer Wissensquelle werden mit einem Thesaurus deskribiert (erschlossen), dies geschieht durch die sogenannte Indexierung. Das Indexierungsergebnis ist eine Liste natürlichsprachiger Wörter, die nicht frei wählbar sind, sondern nach bestimmten Regeln einem Thesaurus entnommen werden müssen.

Bei dem Information Retrieval (Informationswiedergewinnung) dient der Thesaurus der Suche nach relevanten Wissensquellen (Dokumenten), indem sich der Nutzer der indexierten Wörter bedient. Da der Thesaurus diese Wörter gleichzeitig auch miteinander in Beziehung setzt, kann dieses Beziehungsgeflecht auch als Suchhilfe (Pfad) genutzt werden.

Thesauri werden immer auf Basis einer Wissenssammlung erstellt. Weltweit dürfte es mehrere Tausend Thesauri geben, die alle auf bestimmte Fachgebiete ausgerichtet sind. Der „Thesaurus Guide“ verzeichnete 1993 rund 600 aktiv genutzte Thesauri in unterschiedlichen natürlichen Sprachen[7].

Ein Thesaurus enthält ein „kontrolliertes Vokabular“, also eine eindeutige Benennung für jeden Begriff (Deskriptor oder Schlagwort) – diese Benennung kann, wenn der Thesaurus elektronisch verarbeitet wird, auch vollkommen abstrakt sein (z. B. eine

[5] Vgl. [Ertel2006], Kapitel 2.4.2 S. 16 f
[6] [DIN1463-1]
[7] [Eurobrokers1992]

laufende Nummer), deswegen spricht man hier auch von einem Konzept. Oft nimmt man aber dennoch eine natürlichsprachige Vorzugsbezeichnung, den Deskriptor. Da eine Eindeutigkeit in der natürlichen Sprache jedoch nicht gegeben ist, werden außerdem Äquivalenzrelationen eingefügt. So ist die Synonymie die Gleichheit oder auch nur große Ähnlichkeit der Bedeutung von unterschiedlichen Wörtern. Zu einem Deskriptor können also beliebig viele Synonyme in Beziehung gebracht werden. Gleiches gilt für Wörter, die unterschiedliche Schreibweisen besitzen, hier werden alle Möglichkeiten als Synonym-Beziehung angegeben; wichtig ist das insbesondere, wenn man eine Wissensbasis indexieren will, in der Dokumente in neuer und alter deutscher Rechtschreibung vorhanden sind. Auch Abkürzungen und eventuell Übersetzungen können so behandelt werden.

Schwieriger wird es mit Homonymen oder Polysemen, also Wörtern, die verschiedene Bedeutungen besitzen.[8] Hier werden die Wörter mehreren Deskriptoren zugeordnet und gleichzeitig markiert, um ihre Mehrdeutigkeit anzuzeigen. Zur richtigen Einordnung eines Dokuments muss dann der Kontext angeschaut werden, was bei der automatischen Verarbeitung Schwierigkeiten macht.

Daneben existieren noch hierarchische Relationen, um auf Hyponyme (Unterbegriffe) und Hyperonyme (Oberbegriffe) zu verweisen. DIN 1463-1 unterscheidet dabei noch zwischen generischer Relation, was als „eine hierarchische Relation zwischen zwei Begriffen, von denen der untergeordnete Begriffe (Unterbegriff) alle Merkmale des übergeordneten Begriffs (Oberbegriff) besitzt und zusätzlich mindestens ein weiteres spezifizierendes Merkmal“[9] definiert wird und partitiver Relation, was als „eine hierarchische Relation zwischen zwei Begriffen, von denen der übergeordnete (weitere) Begriff (Verbandsbegriff) einem Ganzen entspricht und der untergeordnete (engere) Begriff (Teilbegriff) einen der Bestandteile dieses Ganzen repräsentiert“[10] beschrieben wird.

Soll eine Beziehung beschrieben werden, die nicht den bisherigen Definitionen entspricht, existiert noch die Assoziationsrelation, sie ist eine „zwischen Begriffen bzw. ihren Bezeichnungen als wichtig erscheinende Relation, die weder eindeutig hierar-

[8] Besitzen beide Wörter die selben etymologische Wurzeln, spricht man von Polysemie, diese Wörter haben eine ähnliche Bedeutung (z. B. Pferd als Tier und Turngerät), andernfalls von Homonymie (z. B. Bank als Sitzmöbel oder Kreditinstitut).

[9] [DIN1463-1], Teil 1

[10] ebenda

chischer Natur ist, noch als äquivalent angesehen werden kann.“[11] Diese schwammige Definition ist auch die Problematik dieser Beziehung. Sie kann zu einem Sammelbecken geraten, in das alles hineingenommen wird, was in einem sehr weiten Sinn mit dem Ausgangsbegriff zu tun hat. Oft entstehen dadurch sehr lange Reihen solcher „verwandten Begriffe“. Ein Thesaurus sollte aber nicht versuchen, alle möglichen Zusammenhänge auszuweisen, in denen ein Begriff vorkommen kann. Der Sinn dieser Relation ist vielmehr, „zusätzlich zur hierarchischen Struktur Querbeziehungen zu anderen, für die Formulierung des Sachverhaltesmöglicherweise geeigneten Deskriptoren anzubieten“[12].

[11] ebenda

[12] [Burkart2004], Kapitel B 2.1.4.4, Seite 149