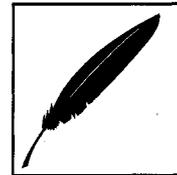


juris fährt „PASSAT“

Tilo Jung



Einleitung

Für juris-Insider beinhaltet diese Notiz keine Neuigkeit, während ein Einsteiger dahinter zunächst einen Werbegang des Wolfsburger Autoherstellers vermuten wird. Um die Spannung des interessierten Lesers nicht noch zu steigern: das Kürzel „PASSAT“ steht hier für ein Programm zur Automatischen Selektion von Stichwörtern Aus Texten, das von SIEMENS als Standardsoftware für das Betriebssystem BS 2000 angeboten wird. Das Programm wird (nicht nur von juris) auf dem Gebiet der Dokumentation und Information zur automatischen Texterschließung eingesetzt. Hauptaufgabe von PASSAT ist, diejenigen Wörter aus Texten herauszulösen, die sinntragend sind, das heißt, die Informationen über den Inhalt des Textes geben. Diese Begriffe werden dann als Deskriptoren eingestuft und ermöglichen bei der Recherche in der Datenbank das Wiederauffinden der Dokumente. Um befriedigende Ergebnisse bei der Suche in juris zu erhalten, ist es nützlich, die Aufbereitung der Dokumente zu kennen und seine Suchanfrage unter Berücksichtigung der Eigenschaften von PASSAT zu formulieren. Der folgende Aufsatz befaßt sich zunächst mit der Problematik der Texterschließung in einer EDV-Umgebung allgemein, erläutert dann in stark geraffter Form die Arbeitsweise des PASSAT-Programms und gibt zum Abschluß praktische Tipps zur Verbesserung der Recherchestrategie in juris.

*PASSAT:-
Stichwortselektions-Programm
für BS2000-Rechner*

*Sinntragende Wörter als
Deskriptoren*

Besonderheiten der deutschen Sprache

Ein Programm, das zur Aufbereitung von Texten für die Repräsentation in einer Datenbank eingesetzt wird, muß bei der Analyse eines Textes gerade auch den grammatikalischen Besonderheiten der deutschen Sprache Rechnung tragen. Zu nennen wären hier beispielsweise:

- Flexionsformen

Deklination	Konjugation
das Kind	ich renne
des Kindes	du rennst
die Kinder	wir rennen

*Flexionsformen:
Deklination und Konjugation*

- Stammvokaländerung bei Pluralbildung

Maus	-	Mäuse
Stock	-	Stöcke

Änderung des Stammvokals

- Kompositabildung

unter Auslassung von Buchstaben:

Mühle + Rad = Mühlrad

*Kompositabildung:
Endungsausfall und
Fugenmorphem*

mit Einfügung eines Fugenmorphems:

Wirtschaft + Minister = Wirtschaftsminister

- Verwendung von Mehrwortbegriffen (vor allem auch im juristischen Sprachgebrauch)

culpa in contrahendo
Treu und Glauben

Mehrwortbegriffe

Diese Formenvielfalt soll zwar bei der Ausgabe des Dokuments gewahrt bleiben, für die Formulierung einer Suchfrage, die den Volltext eines Dokuments erschließt, erscheint sie jedoch hinderlich.

Nachteile der Volltextindexierung

Sogenannte volltextindexierende Systeme, stellen – anders als bei der Erschließung mit PASSAT – alle im Text vorkommenden Worte unverändert als Suchbegriffe zur Verfügung.



Volltextsuche braucht Trunkierung

gung. Dies setzt, bei Verwendung in der deutschen Sprache, einen geschulten Benutzer voraus, der erkennt, welche Flexionsformen bei dem eingegebenen Suchbegriff im Satzzusammenhang vorkommen können, um dann durch geschickte Maskierung des Suchbegriffs korrekte Rechercheergebnisse zu erhalten. Die Maskierung des Suchbegriffs geht jedoch vor allem bei der Linkstrunkierung erheblich zu Lasten der Arbeitsgeschwindigkeit eines Systems. Außerdem können – wie oben schon aufgezeigt – die häufigen Stammvokaländerungen nur durch Hinzunahme weiterer Begriffe in die Suchfrage ausgeglichen werden. Schließlich erhält man bei volltextindexierenden Systemen, durch die immer notwendige Maskierung der Suchfrage, auch unerwünschte Treffer.

Beispiel: maskierter Suchbegriff *Held*

Treffer Kriegsheld
Heldentat
aber auch Stacheldrant

Arbeitsweise von PASSAT

Diese Fehlerquellen sollen, durch die Aufbereitung der Texte mit PASSAT, erkannt und vermieden werden. Kernstück des Programms ist die Vergleichswortliste (VWL). Diese stellt eine Art internes Wörterbuch dar, mit dem alle im Text vorkommenden Worte verglichen werden. Der Umfang der bei juris verwendeten VWL beträgt zur Zeit rund 485000 Einträge. Ein Eintrag in der VWL enthält die Stammform des Worts, die nicht mit der grammatikalischen Grundform zu verwechseln ist, sondern vielmehr den kleinsten gemeinsamen Nenner einer Wortfamilie darstellt (z.B. „renn“ für das Verb „rennen“). Weiterhin wird mittels einer Kodierung festgelegt, welche Endungen an diesen Stamm angehängt werden können und nach welchen Regeln die Kompositabildung abläuft. Zusätzlich entscheidet die Kodierung darüber, ob das Wort als freier Suchbegriff im Text vorkommt (Deskriptor), oder ob es sich bei ihm um ein sogenanntes Stoppwort handelt, das nicht der Recherche in juris zugänglich ist, wie z.B. Zahlwörter, Pronomen etc. Letztlich werden in der VWL auch die möglichen Zerlegungen von zusammengesetzten Worten – der Komposita – in ihre Einzelbestandteile festgehalten. PASSAT erstellt eine alphabetische Liste aller in einem neu aufzunehmenden Dokument vorkommenden Worte, und vergleicht sie mit der internen VWL. Entsprechend der dort abgelegten Einträge werden dann die Suchworte (Deskriptoren) für das betreffende Dokument vergeben. Gleichzeitig wird in einem Fehlerprotokoll festgehalten, ob unbekannte Zeichenfolgen, die sich oft als Schreibfehler herausstellen, in dem verarbeiteten Dokument vorkommen, beziehungsweise, ob neue Zertrennungen zusammengesetzter Begriffe von PASSAT durchgeführt werden, die eine Aufnahme des Begriffs in die VWL notwendig erscheinen lassen. Danach werden die neuen Deskriptoren in das alphabetische Register in ihrer grammatikalischen Grundform aufgenommen (Substantive im Nominativ Singular, Verben im Infinitiv), womit das Dokument dann im Datenpool zur Recherche zur Verfügung steht.

Die Vergleichswortliste:
Zur Zeit etwa 485000 Einträge

Regeln für zulässige Endungen und
Kompositabildung

Stoppworte: Nicht suchfähig

Kompositazerlegung

Problematik der Kompositazerlegung

Ein großer Vorteil bei der Verwendung von PASSAT, ist die oben schon kurz angesprochene automatische Kompositazerlegung. Sie ermöglicht dem Benutzer, sich bei der Formulierung der Suchfrage auf die Einzelbestandteile des Begriffs zu beschränken, und trotzdem auch die Dokumente zu finden, in denen die Begriffe nur in der zusammengesetzten Form vorkommen. Wenn zum Beispiel im Text der Begriff ABFALLBESEITIGUNG vorkommt, würde PASSAT automatisch eine Zerlegung in ABFALL und BESEITIGUNG durchführen, da diese als Einträge in der VWL vorliegen, und man würde das Dokument sowohl mit dem einen wie dem anderen Begriff bei der Suchanfrage als Treffer erhalten. Manchmal kann diese vollständige Zerlegung jedoch auch unerwünscht sein. So wäre die Zerlegung des Begriffs VERFASSUNGSMÄSSIGKEIT in VERFASSUNG und MÄSSIGKEIT nicht sinnvoll. In diesen Fällen wird der zusammengesetzte Begriff in die VWL aufgenommen und als Substitut nur noch VERFASSUNG vergeben. Schließlich gibt es auch Begriffe, bei denen eine Zerlegung gänzlich unsinnig wäre, wie zum Beispiel die Zerlegung von STECKDOSE in STECKEN und DOSE. Ein Begriff dieser Art wird dann ohne Substitutvergabe in die VWL aufgenommen und danach nicht mehr in seine Einzelbestandteile zerlegt.

Unerwünscht:
Die unsinnige Zerlegung



Beispiele¹

- Zerlegung

LÄRMSCHUTZ	-	LÄRM + SCHUTZ
WANDERVEREIN	-	WANDERN + VEREIN

- Substitutvergabe

ABGAS	-	GAS
KREISSÄGE	-	SÄGE

- Keine Zerlegung

ABSCHAUM	-	
SCHLÜSSELBEIN	-	

Man erkennt schnell, daß es hier, wie in allen Bereichen, bei denen mit natürlicher Sprache in einer EDV-Umgebung gearbeitet wird, keinen allgemeingültigen Zerlegungsalgorithmus geben kann, sondern daß ein sogenanntes Weltwissen Voraussetzung für das Erkennen der sinntragenden Bestandteile eines zusammengesetzten Begriffs ist. Der Bearbeiter ist bei diesem Problemfeld letztlich auf sein Sprachgefühl oder auch auf pragmatische Erwägungen angewiesen (antizipiertes Benutzerverhalten). Alle Methodik versagt schließlich bei Begriffen, deren Bedeutung sich nur aus dem Textzusammenhang ergibt. So ist zum Beispiel nur durch eine kurze juris-Recherche zu klären, ob der Eintrag ALSTERWASSER in der VWL nun im Zusammenhang über die Qualität oder den Grad der Verschmutzung eines Flusses gebraucht wird oder ob es sich um die norddeutsche Bezeichnung für ein „Radlermaß“ (saarländisch: panaché) handelt. Um die Information vollständig zu halten: der Begriff wurde tatsächlich nur in der zweiten Bedeutung in der Rechtsprechungsdatenbank von juris verwendet, wobei sich die Richter bei der Blutalkoholbestimmung mit dem Konsum eines „Biermischgetränks“ befaßten. Diesen Begriff hätte wohl niemand in seiner Suchwortliste verwendet...

*Wichtig für VWL-Bearbeiter:
Sprachgefühl*

*Bei Homonymen hilft nur noch der
Kontext.*

Tips für die Recherche

Aus dem Erläuterten ergibt sich, daß man sich an bestimmte Regeln beim Aufstellen einer Suchwortliste halten muß. Bei Substantiven muß immer der Nominativ Singular, bei Verben der Infinitiv in der Suchfrage angegeben werden. Deshalb verdienen z.B. adjektivisch verwendete Verben besondere Aufmerksamkeit. Eine Suche, die den „gestörten Gesamtschuldausgleich“ zum Ziel, hat führt nur dann zu guten Ergebnissen, wenn in der Suchwortliste auch das Verb „stören“ vorkommt. Was die Kompositzerlegung angeht, kann es durchaus nützlich sein, sowohl mit dem zusammengesetzten Begriff, als auch mit seinen Einzelbestandteilen zu suchen, und die doppelt vorliegenden Dokumente durch logische Verknüpfungen vor der Ausgabe zu entfernen. Dies sollte zwar nach dem oben Gesagten unnötig sein, Tatsache ist jedoch, daß bei juris in der VWL Substitute zu aufgenommenen Begriffen nicht immer vollständig vergeben worden sind. In dem vorstehend beschriebenen Beispiel wäre es also durchaus sinnvoll, mit GESAMTSCHULDAUSGLEICH, GESAMTSCHULD und AUSGLEICH zu suchen. Ein weiteres Beispiel aus diesem Zusammenhang (von einem juris-Nutzer bemerkt) stellte sich folgendermaßen dar: die Suche nach einem Dokument, das in juris vorhanden ist, und das sich mit der Präambel des Grundgesetzes befaßte, verlief mit dem Suchwort „Wiedervereinigung“ erfolglos. Erst als mit dem Kompositum „Wiedervereinigungsgebot“ gesucht wurde, war das Dokument bei den Treffern enthalten. Offenbar wurde der Begriff in die VWL aufgenommen, ohne die entsprechenden Substitute zu vergeben. Diese Unstimmigkeiten gehören jedoch bald der Vergangenheit an, da im Moment die VWL bei juris systematisch überarbeitet wird. So werden in absehbarer Zeit, die ganzen Fähigkeiten von PASSAT die Recherche unterstützen.

*Eine Sonderstellung:
Adjektivisch verwendete Verben*

*Beseitigung von Unstimmigkeiten:
Die VWL wird überarbeitet*

Literatur

Siemens Produktbeschreibung (BS2000-Software): Texte automatisch analysieren und erschließen mit PASSAT
 juris Briefe, 2. Jahrgang, April 1/1990
 Siemens-Schriftenreihe data praxis
 juris, Das juristische Informationssystem in der praktischen Erprobung
 Siemens Schriftenreihe: PASSAT (BS2000), Texterschließungssystem, Benutzerhandbuch

¹ Für die Beratung bei der Auswahl der Beispiele danke ich Herrn Krumbügel, der bei juris für den PASSAT-Einsatz verantwortlich ist.