In: Bekavac, Bernard; Herget, Josef; Rittberger, Marc (Hg.): Informationen zwischen Kultur und Marktwirtschaft. Proceedings des 9. InternationalenSymposiums fur Informationswissenschaft (ISI 2004), Chur, 6.-8.Oktober2004. Konstanz: UVK Verlagsgesellschaft mbH, 2004. S. 469 – 470

Effizientes Navigieren in umfangreichen wissenschaftlichen Nachschlagewerken

Alexander Eckl

Lehrstuhl für Informatik II, Universität Würzburg, Am Hubland,
D - 97074 Würzburg, eckl@informatik.uni-wuerzburg.de

Am Lehrstuhl für Informatik II der Universität Würzburg wird in Kooperation mit dem Springer-Verlag die jährliche Ausgabe von „Hagers Handbuch der Drogen und Arzneistoffe“ (www.hagershandbuch.de) auf CD-ROM entwickelt. Es handelt sich um die Standardenzyklopädie für deutschsprachige Pharmazeuten und Apotheker. Die so genannte HagerROM 2003 enthält über 10.000 Einträge zu Arzneidrogen und –stoffen. Um in der HagerROM einen Eintrag aufzurufen, kann neben einem Navigationsbaum die so genannte Schnellauswahl verwendet werden. In sie können durch Leerzeichen getrennte Präfixe eingegeben werden, wobei nach jedem Tastendruck in einer Liste die Eintragsnamen angezeigt werden, die zu jedem der Präfixe ein Wort enthalten. In Abbildung 1 wird ein Beispiel gegeben für die Trefferliste nach der Eingabe von „ec an“ bei der Suche nach „Echinacea angustifolia“.

Die Schnellauswahl hat den Vorteil, dass im Allgemeinen auch bei einer großen Anzahl enthaltener Begriffe nur wenige Zeichen eingegeben werden müssen, um ein Stichwort in einer kurzen Trefferliste zu finden. Sie ist damit für elektronische Nachschlagewerke eine effiziente Möglichkeit, um von einem Begriff zu einem zugehörigen Artikel zu gelangen. Sie wird deshalb für die HagerROM 2004 um eine große Anzahl an Synonymen und sonstigen Bezeichnungen erweitert. Dafür wurde eine neue kompakte Darstellung der hierarchischen Datenstruktur Trie [Knuth 98] entwickelt, die den schnellen Zugriff auf große statische Wortmengen ermöglicht.

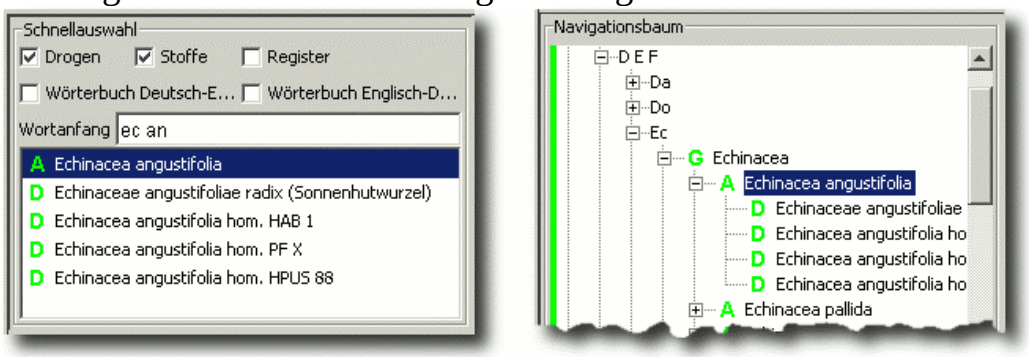

Abbildung 1: Schnellauswahl und Navigationsbaum der HagerROM 2003

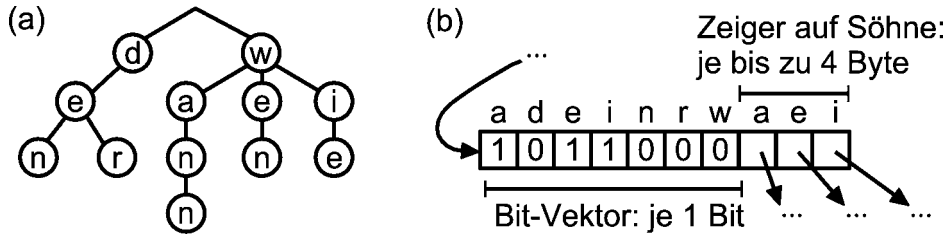

Abbildung 2: (a) Trie zu den Worten den, der, wann, wen und wie (b) Knoten zum Präfix w

Ein Trie ist ein Baum, in dem Zeichenketten von der Wurzel ausgehend abgelegt werden, siehe Abbildung 2(a). Bei der Eingabe eines Präfixes wird von der Wurzel an abgestiegen, wobei in den Knoten auf Trefferlisten verwiesen werden kann. Bei natürlichen Sprachen besitzen die Knoten nur wenige Söhne. Für Tries zu statischen Wortmengen bietet sich daher an, je Knoten einen Bit-Vektor in Alphabetgröße zu speichern, der angibt, zu welchen Alphabetzeichen ein Sohn existiert. Danach werden die Zeiger zu den Söhnen aufgeführt, siehe Abbildung 2(b), vergleiche [Purdin 90]. Bei wissenschaftlichen Werken treten jedoch durch viele Sonderzeichen große Alphabete auf, so dass dann die Bit-Vektoren den Speicherplatzbedarf dominieren.

Der neue Ansatz besteht darin, dass nicht in jedem Knoten ein Bit-Vektor zum vollständigen Alphabet abgelegt wird, sondern nur zu einem Teilalphabet von Zeichen, die an einer bestimmten Stelle im Baum auftreten können: Zum Beispiel kommen in den über 15.000 Worten der Drogen-Schnellauswahl der HagerROM 2004 bei Ignorierung der Groß- und Kleinschreibung insgesamt 44 verschiedene Zeichen vor, nach der Zeichenfolge `je` aber nur 9: `d`, `e`, `l`, `m`, `n`, `p`, `r`, `s` und `w` (Bsp.: `kajeputöl`). Indem durch die den Knoten direkt vorangehenden Zeichenfolgen Teilalphabete definiert und zusätzlich abgelegt werden, können die Bit-Vektoren entsprechend verkürzt gespeichert werden. Es wird Platz gespart, da die gleiche Zeichenfolge vor mehreren Knoten auftritt, das zugehörige Teilalphabet aber nur einmal gespeichert werden muss.

Für die oben genannten 15.000 Worte (mittlere Wortlänge 9,3 Zeichen) wurde für die Schnellauswahl mit der aktuellen Implementierung ein 129 KByte großer Trie erzeugt. Die Bit-Vektoren besaßen eine mittlere Länge von 17,6 Bit statt 44 Bit, indem Zeichenfolgen der Länge zwei zum Einschränken der Alphabete verwendet wurden. Außerdem wurden Ketten im Trie, die ohne Verzweigungen zu Blättern führten, statt als Knoten direkt als Zeichenketten abgelegte. Für die Schnellauswahl wurden zusätzlich zum Trie noch 162 KByte für Trefferlisten benötigt.

[Knuth 98] Knuth D. E.: The Art of Computer Programming, Volume 3: Sorting and Searching, Second Edition. Addison-Wesley Massachusetts, 1998.

[Purdin 90] Purdin, T. D. M.: Compressing tries for storing dictionaries. In Proceedings of the IEEE Symposium on Applied Computing, H. Berghel, J. Talburt, and D. Roach, Eds. IEEE Fayettville, Arkansas, 336–340, 1990