Povzetek

Diplomsko delo povzema osnovne načine iskanja informacij in predstavlja eno izmed možnih rešitev za izbrani problem iskanja informacij po strukturiranih podatkih. S to rešitvijo želimo izboljšati uporabniško izkušnjo s programsko opremo ter zagotoviti hitro in učinkovito iskanje po podatkih, ki ga ni mogoče doseči z uporabo iskalnih metod relacijske baze, vsaj ne tako, da bi ohranili podporo več tipom relacijskih baz in neodvisnost od jezika. Problema se lotimo z uporabo tabele pripon in standardnih metod sistemov za iskanje informacij.

V uvodu predstavimo zgodovino in osnovne pojme iskanja informacij, si na hitro ogledamo še rudarjenje ter predstavimo osnovne metode reševanja problema iskanja informacij. Nato se osredotočimo na specifičen problem in izberemo algoritme za reševanje le-tega. V sklopu teorije obdelamo drevesa trie, drevesa pripon, tabele pripon in bolj podrobno algoritem DC3 za gradnjo tabele pripon. Zaradi dinamične narave teksta, po katerem iščemo, si ogledamo tudi problem združevanja indeksov. Sledi opis rangiranja zadetkov iskanja, mimo katerega pri iskanju informacij ne moremo. Celoten sistem smo tudi implementirali, kar na kratko opišemo v predzadnjem poglavju, kjer z analizo obnašanja sistema pri različnih pogojih tudi preverimo, ali smo izpolnili dane zahteve, in podamo oceno uporabnosti dane implementacije.

Mathematics Subject Classification (2000):

68P05 Data structures
68P10 Searching and sorting
68P20 Information storage and retrieval
68U35 Information systems

Computing Classification System (1998):

H.3 Information storage and retrieval
H.3.3 Information search and retrieval
E.2 Data storage representations
F.2 Analysis of algorithms and problem complexity

Ključne besede: iskanje, iskanje informacij, iskanje po tekstu, indeksiranje, krnjenje, rangiranje, trie, drevo patricia, pripona, drevo pripon, tabela pripon, DC3, združevanje indeksov.

Keywords: searching, information retrieval, text retrieval, indexing, stemming, ranking, trie, patricia tree, suffix, suffix tree, suffix array, DC3, index merging.

Literatura

[1] Ananyan, Sergei. "Foreword to Document Warehousing and Text Mining book" *Data Mining, Text Mining and Web Mining Software: White Papers.* 02. jan. 2001. Megaputer Intelligence.
<http://www.megaputer.com/tech/wp/foreword.php3> (obiskano 15. jul. 2006)

[2] "History of information retrieval" *American Society of Indexers: About Indexing.* 14. okt. 2005. American Society of Indexers.
<http://www.asindexing.org/site/history.shtml> (obiskano 15. jul. 2006)

[3] "Information Retrieval" *GSLISWiki.* 12. jan. 2005.
<http://www.gslis.org/index.php?title=Information_Retrieval> (obiskano 15. jul. 2006)

[4] Leks, Michael. "UDT Occasional Paper 5: The seven ages of information retrieval" *IFLA: Activities & Services: Archive - Historical Material: IFLA Universal Dataflow and Telecommunications Core Activities: Publications: UDT Occasional Papers*. mar. 1996. IFLA.
<http://www.ifla.org/VI/5/op/udtop5/udt-op5.pdf> (obiskano 15. jul. 2006)

[5] Mason, Moya K. "Historical Development of Ideas Concerning Library Catalogues: Their Purpose and Organization" *Moya K. Mason: A Selection of My Papers*. 2006.
<http://www.moyak.com/researcher/resume/papers/catalogues.html> (obiskano 15. jul. 2006)

[6] Paijmans, J.J. "The Retrieval of Information from historical perspective" *Paijmans homepage*. 2004.
<http://pi0959.kub.nl/Paai/Onderw/V-I/Content/history.html> (obiskano 15. jul. 2006)

[7] Hearst, Marti A. "Untangling Text Data Mining" *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20–26, 1999.* Maryland: University of Maryland, 1999. 3–10.
<http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html>

[8] Hearst, Marti. "What Is Text Mining?" *Homepage of Marti Hearst: Research.* 17. okt. 2003. UC Berkeley.
<http://www.sims.berkeley.edu/~hearst/text-mining.html> (obiskano 15. jul. 2006)

[9] Treloar, Nathan. "Mining: Tools, Techniques, and Applications" *Knowledge technologies conference 2002*. Washington: AvaQuest, Inc., 2002.
<http://knowledgetechnologies.net/proceedings/> (obiskano 15. jul. 2006)

[10] Ólafsson, Sigurður. "Lecture notes" *Iowa state university: Siggi Olafsson: Teaching: IE 583 Knowledge Discovery and Data Mining*. 2004.
<http://www.public.iastate.edu/~olafsson/mining_schedule.html> (obiskano 15. jul. 2006)

[11] Hull, David A. *Information Retrieval Using Statistical Classification*. PhD thesis. Stanford University, Stanford, 1994.

[12] Garcia E. "Term vector models" *Mi Islita: Research Articles.* 2005.
<http://www.miislita.com/searchito/research-articles.html> (obiskano 15. jul. 2006)

[13] Baeza-Yates, Ricardo in B. Ribeiro-Neto. "Indexing and Searching" *Modern Information Retrieval*. New York: Addison-Wesley, 1999.

[14] Jackson, Peter in Isabelle Moulinier. *Natural Language Processing for Online Applications: Text Retrieval, Extraction & Categorization.* Amsterdam: John Benjamins, 2002.

[15] Dalianis, Hercules. "Improving Precision in Information Retrieval using Stemming and Spell checking" *Fifth ScandSum meeting 4–6 April 2003*. Stockholm, 2003.
<http://dsv.su.se/~hercules/scandsum/FifthScandSumAre.html> (obiskano 15. jul. 2006)

[16] Maly, Kurt in Michael Nelson. "CS 495/595, Introduction to digital libraries" *Old Dominion University: Department of computer science*. 1999.
<http://www.cs.odu.edu/~mln/dl/> (obiskano 15. jul. 2006)

[17] Rasmussen, Edie. "Course IST 2140 Information Storage and Retrieval" *School of information sciences, University of Pittsburgh: Edie Rasmussen: Courses taught*. 2001.
<http://www.sis.pitt.edu/~erasmus/courses.html> (obiskano 15. jul. 2006)

[18] Chen, Hsin-Hsi. "Course information and retrieval" *National Taiwan University: Natural Language Processing Lab: Advisor.* 2005.
<http://nlg3.csie.ntu.edu.tw/advisor.html> (obiskano 15. jul. 2006)

[19] Baeza-Yates, Ricardo. "Text Retrieval: Theory and Practice" *Proceedings of the 12th IFIP World Computer Congress, volume I*. Madrid: Elsevier Science, Sept. 1992. 465–476.

[20] Kärkkäinen, Juha in Peter Sanders. "Simple Linear Work Suffix Array Construction" *Proc. 30th International Colloquium on Automata, Languages and Programming (ICALP '03)*. Berlin: Springer, 2003. 943–955.

[21] "information retrieval", "data retrieval", "text retrieval", "document retrieval", "full text search", "data mining", "text mining", "soundex", "cyclic redundancy check", "vector space model", "inverse index", "suffix tree", "suffix array", "signature files", "trie", "radix tree" *Wikipedia*.
<http://en.wikipedia.org/> (obiskano 15. jul. 2006)

[22] Fredkin, E.H. "Trie Memory" *Communications of the ACM.* 3.9 (1960): 490–500.

[23] Morrison, Donald R. "PATRICIA – Practical Algorithm to Retrieve Information Coded in Alphanumeric" *Journal of the ACM* 15.4 (1968): 514–534.

[24] Ferragina, Paolo in Roberto Grossi. "The string B-tree: a new data structure for string search in external memory and its applications" *Journal of the ACM (JACM)*. 46.2 (1999): 236–280.

[25] Morrison, Donald R. "PATRICIA – Practical Algorithm to Retrieve Information Coded in Alphanumeric" *Journal of the ACM* 15.4 (1968): 514–534.

[26] Weiner P. "Linear pattern matching algorithm" *14th Annual IEEE Symposium on Switching and Automata Theory*. 1973. 1–11.

[27] McCreight, Edward M. "A Space-Economical Suffix Tree Construction Algorithm" *Journal of the ACM (JACM)*. 23.2 (1976): 262–272.

[28] Ukkonen, E. "On-line construction of suffix trees" *Algorithmica* 14.3 (1995): 249–260.

[29] Giegerich R. in S. Kurtz. "From Ukkonen to McCreight and Weiner: A Unifying View of Linear-Time Suffix Tree Construction" *Algorithmica* 19.3 (1997): 331–353.

[30] Farach, Martin. "Optimal suffix tree construction with large alphabets" *Proceedings of the 38th Annual Symposium on Foundations of Computer Science (FOCS '97).* 1997. 137–143.

[31] Gonnet, G. *Pat 3.1: An Efficient Text Searching System. User's Manual.* 1987. Waterloo: UW Centre for the New OED.

[32] Manber, U. in G. Myers. "Suffix arrays: a new method for on-line string searches" *SODA '90: Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*. Philadelphia: Society for Industrial and Applied Mathematics, 1990. 319–327.

[33] Manber, U. in G. Myers. "Suffix arrays: A new method for on-line string searches" *SIAM Journal on Computing* 22.5 (1993): 935–948.

[34] Ko P. in S. Aluru. "Space efficient linear time construction of suffix arrays" *In Proc. 14th Annual Symposium on Combinatorial Pattern Matching, volume 2676 of LNCS*. Springer, 2003. 200–210.

[35] Manber U. in Ricardo Baeza-Yates. "An Algorithm for String Matching with a Sequence of Don't Cares" *Information Processing Letters* 37 (1991): 133–136.

[36] Farach M. "Optimal suffix tree construction with large alphabets" *In Proc. 38th Annual Symposium on Foundations of Computer Science.* IEEE, 1997. 137–143.

[37] Kim, Dong Kyue, Jeong Seop Sim, Heejin Park in Kunsoo Park. "Linear-time construction of suffix arrays" *In Proc. 14th Annual Symposium on Combinatorial Pattern Matching, volume 2676 of LNCS.* Springer, 2003. 186–199.

[38] Salton G. in M. J. McGill. *Introduction to modern information retrieval*. New York: McGraw-Hill, 1986.

[39] Wong, Wai Yee Peter in Dik Lun Lee. "Implementations of partial document ranking using inverted files" *Information Processing and Management* 29.5 (1993): 647–669.

[40] Clark, Stephen. "Stephen Clarks's Research Interests" *Oxford University Computing Laboratory: Stephen Clark*. 2006.
<http://web.comlab.ox.ac.uk/oucl/work/stephen.clark/research.html> (obiskano 15. jul. 2006)

[41] Mariam, John. "Ranking in Information Retrieval Systems" *The University of Texas at Arlington: CSE6319 SEC 013: Data exploration and analysis in relational database.* 23. mar. 2006.
<http://crystal.uta.edu/~gdas/Courses/websitepages/spring06DBIR.htm> (obiskano 15. jul. 2006)

[42] Lin, Jimmy. "Syllabus - Boolean and Vector Space Models" *University of Maryland: Institute for advanced computer studies: Jimmy Lin: LBSC 796/INFM 718R: Information Retrieval Systems*. 13. feb. 2006.
<http://www.umiacs.umd.edu/~jimmylin/LBSC796-INFM718R-2006-Spring/> (obiskano 15. jul. 2006)

[43] Lee, Dik L., Huei Chuang in Kent Seamons. "Document Ranking and the Vector-Space Model" *IEEE Software* 14.2 (1997): 67–75.