Programiranje 1

Beleške sa vežbi

Školska 2006/2007 godina

Matematički fakultet, Beograd

Jelena Tomašević

October 9, 2006

Sadržaj

1

[1]

1.1 Reprezentacija znakovnih podataka

1.1.1 Tekst je niz karaktera

- Iako obično tekst zamišljamo kao dvodimenzioni objekat, u računarima se tekst predstavlja kao jednodimenzioni (linearni) niz karaktera.
- Potrebno je, dakle, uvesti specijalne karaktere koji označavaju prelazak u novi red, tabulator, kraj teksta i slično

1.1.2 Zapis karaktera u računaru

- Računari su zasnovani na binarnoj aritmetici
- Cele brojeve je moguće predstaviti u binarnom sistemu
- Osnovna ideja je svakom karakteru pridružiti određeni ceo broj na unapred dogovoreni način
- Ove brojeve zovemo kodovima karaktera (character codes)

1.1.3 Skupovi karaktera

- Koliko karaktera želimo da predstavimo u računarima? Tokom razvoja računarstva broj karaktera je postajao sve veći
- Pošto je u početku razvoja englesko govorno područje bilo dominantno osnovno je bilo predstaviti sledeće karaktere :
  - Velika slova engleskog alfabeta : A,B,...,Z
  - Mala slova engleskog alfabeta : a,b,...,z
  - Cifre : 0,1,...,9
  - Interpunkcijske znake : ., :; i slično
  - Kontrolne znake : kraj reda, tabulator i slično
- Standardni karakterski kodovi: Sedamdesetih godina su se pojavile tabele standardnih karakterskih kodova dovoljne za zapis pomenutih karaktera

  Najpoznatiji su

[1] Zasnovano na materijalu "Zapis tekstova u računaru" Filipa Marića

  - EBCDIC IBM-ov standard, pogodan za bušene kartice
  - ASCII Standard iz koga se razvila većina današnjih standarda

- ASCII (American Standard Code for Information Interchange)
  ASCII je sedmobitan (broj karaktera koji je njime predstavljen je $128 = 2^7$)
- Npr. kod za A je $(41)_{16}$ tj. $0x41$ što je $(65)_{10}$ tj. $(1000001)_2$,
  kod za a je $(61)_{16}$ tj. $0x61$ što je $(95)_{10}$ tj. $(1100001)_2$.
- PRIMER: transformisanje malih slova u velika
- Razmak SP se zapisuje kao $(20)_{16}$ što je $(32)_{10}$ tj. $(0100000)_2$.
- **Osobine ASCII koda**: Prvih 32 karaktera (kodovi 0x00-0x1F) i poslednji karakter (kod 0x7F) su kontrolni karakteri. To su karakteri bez grafije, kao CR (kod 0x0D), LF (kod 0x0A).
- Prvi karakter sa grafijom je blanko (kod 0x20). Njegova grafija je belina.
- Skup velikih slova A-Z (kodovi 0x41-0x5A), kao i skup malih slova a-z (kodovi 0x61-0x7A), je u alfabetskom redosledu unutar kolacione sekvencije
  ($0x41 < 0x42$, prema tome $A < B$ to odgovara alfabetskom redosledu).
- Skup cifara $0-9$ (kodovi $0x31-0x39$) je u rastućem brojčanom redosledu unutar kolacione sekvencije
  ($0x31 < 0x32$, prema tome $1 < 2$ što odgovara brojčanom redosledu).
- Skup velikih slova A-Z (kodovi $0x41-0x5A$), skup malih slova a-z (kodovi $0x61-0x7A$) i skup cifara 0-9 (kodovi $0x31-0x39$) su kontingentni unutar kolacione sekvencije (između slova A i slova Z nema drugih karaktera osim onih koji odgovaraju velim slovima engleske abecede). Sve cifre prethode svim velikim slovima, sva velika slova prethode svim malim slovima u kolacionoj sekvenciji. Specijalni i interpunkcijski znaci su izmešani između njih.
- Kod svakog velikog slova je za 32 (ili 0x20) manji od koda odgovarajućeg malog slova. Na primer, za slovo E važi da je $0x45+0x20 = 0x65$, odnosno, $01000101+00100000 = 01100101$. Prema tome, binarni kodovi velih i malih slova razlikuju se samo u jednoj cifri, onoj koja odgovara petom stepenu osnove 2.

1.1.4 Kodiranje ostalih jezika

- Razvojem računarstva se javlja potreba kodiranja tekstova i na drugim jezicima
- Kroz istoriju su postojala mnoga rešenja, od kojih su se neka zadržala, a neka su nestala

1.1.5 Osobine YUSCII koda

- ASCII kod je jedna verzija međunardonog standarda ISO 646 IRV koji predstavlja međunarodnu referentnu verziju za 7-bitni kod. Ovaj standard propisuje da se pozicijama $0x40$, $0x5B-0x5E$, $0x60$ i $0x7B-0x7E$ ne pridružuje obavezna grafija već se da se one u nacionalnim verijama standarda i u određenim aplikacijama mogu slobodno koristiti.
- Standard JUS.B1.002 koristi ovih 10 pozicija za kodiranje slova specifičnih za srpsku latinicu Ž, Š, Đ, Ć, Č.
- Yu-ASCII skup zadržava sve navedene osobine ASCII koda osim jedne, a ta je da ni velika ni mala slova nisu u alfabetskom redosledu unutar kolacione sekvencije. Naime, $0x40 < 0x41$, dakle $\check{Z} < A$ što ne odgovara alfabetskom redosledu unutar kolacione sekvencije. Nakon slova Ž slede velika slova engleske abecede, zatim slova Š(0x5B), Đ(0x5C), Ć(0x5D), Č(0x5E), ž(Ox60), zatim slede mala slova abecede, i na kraju slova š(0x7B), đ(0x7C), ć(0x7D), č(0x7E).

Zadatak 1 *Poređati slova vašeg prezimena koristeći YUSCII kodnu šemu.*

1.1.6 Kodne strane

- Pod *kodnom* stranom (Code page) tj. *skupom karaktera* (Character set, charset) podrazumevamo uređenu listu karaktera predstavljenih svojim karakterskim kodovima
- Podaci se u računarima obično zapisuju bajt po bajt
- ASCII je sedmobitni standard
- ASCII karakteri se zapisuju tao što se u svakom bajtu bit najvece težine postavi na 0
- To ostavlja prostor za novih 128 karaktera čiji binarni zapis počinje sa 1
- Ovaj prostor se može popuniti na razne načine
- Rešenje nije univerzalno, jer svakako na svetu postoji više od 256 različitih karaktera
- Postavljeni su razni standardi dopunjavanja ovih 128 karaktera
- Svim ovim kodnim stranama je zajedničko prvih 128 karaktera i oni se poklapaju sa ASCII
- Ovako napravljene kodne strane obično omogućuju kodiranje tekstova na više srodnih jezika (obicno i geografski bliskih)
- Nama su uglavnom važne kodne strane napravljene za centralno-evropske (Central European) latinice, kao i ćirilicne kodne strane

1.1.7 Kodne strane kod nas

- Najčešće korišćene kodne strane kod nas (Prve dve su delo međunarodne organizacije za standardizaciju (International Standard organization), dok su naredne dve Microsoft-ovi standardi):
  - ISO 8859-2 (Latin2)
  - ISO 8859-5 (Ćirilicna)
  - Windows 1250
  - Windows 1251 (Ćirilicna)
- **Latin 1**: Poželjno je poznavati i osnovnu kodnu stranu ISO 8859-1 (Latin1) jer je veoma često postavljena kao podrazumevana kodna strana. Ona se koristi za zapis tekstova na zapadno evropskim jezicima (Western European)

1.1.8 Višebajtni karakterski kodovi

- Iako navedene kodne strane omogućuju kodiranje tekstova koji nisu na engleskom jeziku nije moguće npr. u istom tekstu mešati ćirilicu i našu latinicu.
- Azijskim jezicima nije dovoljno 256 mesta za zapis svih karaktera.
- Zbog toga se uvode višebajtni karakterski kodovi

- MBCS: Pre svega zbog potreba istočno azijskih korisnika uvedeni su tzv. višebajtni skupovi karaktera tj. Multi-Byte Character Sets (MBCS)
- Ideja je u tome da se najčešće korišćeni karakteri zapisuju koristeći samo jedan bajt, dok se ostali karakteri zapisuju koristeći dva bajta, tj. koristi se mešavina jednobajtnih i dvobajtnih karakterskih kodova (pod UNIX-om nekad čak i trobajtnih)
- Ovo značajno otežava tumačenje podataka
- **ISO 10646** je zamišljen kao 4 bajtni standard. Pri tome se prvih 65536 karaktera koriste kao osnovni višejezični skup karaktera dok je ostali prostor ostavljen kao proširenje za drevne jezike, celokupnu naucnu notaciju i slično.
- **UNICODE**: svakom karakteru dodeljuje dvobajtni kod
- Prvih 128 karaktera se poklapaju sa ASCII standardom, dok su sledećih 128 napravljeni tako da se poklapaju sa Latin1 standardom
- UCS-2: Unicode standard u suštini predstavlja veliku tabelu koja svakom karakteru dodeljuje broj.
- Standardi koji opisuju kako se niske karaktera onda prevode u nizove bajtova se dodadno definišu
- ISO definiše UCS-2 standard koji jednostavno svaki UNICODE karakter prevodi u odgovarajuca dva bajta
- UTF:A Unicode transformation format (UTF) algoritam koji svakom UNICODE karakteru dodeljuje određeni niz bajtova čija dužina varira od 1 do najviše 6.
- UTF je ASCII kompatibilan, što znaci da se ASCII karakteri zapisuju pomoću jednog bajta, na standardni način.
- Najčešće korišćena varijanta ovog agloritma je UTF-8 koja je dovoljna za zapis svih dvobajtnih UNICODE karaktera
- Pored ovoga ISO uvodi i UTF-16, UTF-32, kao i standard UCS-4

1.1.9 Karakteri, Glifovi, Fontovi

- Vrlo često se ne pravi jasna razlika između karaktera i njihove graficke reprezentacije
- Grafičku reprezentaciju karaktera nazivamo glifovima (*glyph*) Skupove glifova nazivamo fontovima (*font*)
- Korespodencija izmedju karaktera i glifova ne mora biti jednoznacna
- Jedan glif može da predstavi više karaktera (ligature)
- Isti karakter može da se predstavlja razlicitim glifovima u zavisnosti od svoje pozicije u reči
- Za razliku od tradicionalnih fontova koji u sebi sadže glifove za karaktere jedne kodne strane, TrueType fontovi koji podržavaju WGL4 standard sadrže glifove za sve evropske karaktere

Zadatak 2 *Zapisati cifru* 3 *u ASCII kodu.*

Rešenje:
Broj 3 se zapisuje kao $(33)_{16}$ tj. 0x33 što je $(51)_{10}$ tj. $(1010001)_2$

Zadatak 3 *Zapisati reč* Fakultet *u ASCII kodu.*

Zadatak 4 *Zapisati reči* MATF *i* lišće *u kodnim stranama ISO 8859-2, Windows 1250, Windows 1251.*

Rešenje:

Reč *MATF* se zapisuje isto u kodnim stranama ISO 8859-2, Windows 1250 i Windows 1251 zato što su njeni karakteri zapravo ASCII karakteri a svim ovim kodnim stranama zajedničko je prvih 128 karaktera i oni se poklapaju sa ASCII kodovima.
Dakle, reč *MATF* se u ovim kodnim stranama zapisuje preko 4 bajta i to $(4D)_{16}$, $(41)_{16}$, $(54)_{16}$, $(46)_{16}$ a to je isto što i $(77)_{10}$, $(65)_{10}$, $(84)_{10}$, $(70)_{10}$ odnosno $(01001101)_2$, $(01000001)_2$, $(01010100)_2$, $(01000110)_2$.

Reč *lišće* sadrži u sebi karaktere š i ć koji nisu ASCII karakteri pa se različito kodiraju u svakoj od kodnih strana.
U kodnoj strani ISO 8859-2 odgovarajući kod je $(6c)_{16}$, $(69)_{16}$, $(b9)_{16}$, $(e6)_{16}$, $(65)_{16}$ a to je isto što i $(108)_{10}$, $(105)_{10}$, $(185)_{10}$, $(230)_{10}$, $(101)_{10}$ odnosno $(01101100)_2$, $(01101001)_2$, $(10111001)_2$, $(11100110)_2$, $(01100101)_2$.
U kodnoj strani Windows 1250 karakteri š i ć se kodiraju sa $(9A)_{16}$, $(E6)_{16}$.
U kodnoj strani Windows 1251 karakteri š i ć se ne mogu kodirati.

Zadatak 5 *Šta predstavlja niz kodova 138 65 111 33 u kodnoj strani ISO 8859-2? A u Latin1?*

Rešenje:

U kodnoj strani ISO 8859-2 ovaj niz kodova predstavlja |*Ao*! a u Latin1 ŠAo!