

Analiziranje podskupa Million Songs Dataset-a

Seminarski rad u okviru kursa
Istraživanje podataka
Matematički fakultet

Milana Kovačević
Ivan Ristović

jun 2018.

Sažetak

U ovom radu je opisana nova tehnika neograničene provere modela softvera za pronalazak grešaka programa pisanih u programskom jeziku C, koristeći inkrementalne SAT rešavače. Od polaznog programa se pravi *DimSpec* formula. *DimSpec* formula predstavlja konjunkciju četiri KNF formule koje kodiraju početno, krajnje i svako tranziciono stanje programa, kao i prelaze između susednih stanja programa. Proces dobijanja *DimSpec* formule od izvornog C koda se odvija u više koraka. Prvo se C kod prevodi u *LLVM* module, koji predstavljaju medureprezentaciju između C koda i mašinskog koda. Ti moduli se zatim prevode u *DimSpec* formulu korišćenjem alata *LLUMC* tako da ciljno stanje predstavlja stanje greške. Tako dobijena *DimSpec* formula se može rešiti pomoću inkrementalnih SAT rešavača ili *IC3* algoritmom za proveru invarijanti. Rešenje predstavlja dokaz neispravnosti programa jer je pronađen način da se iz polaznog stanja stigne u stanje greške. Ovakvo kodiranje proširuje funkcionalnost tradicionalne ograničene provera modela softvera jer pokriva velike i beskonačne petlje, uz održavanje pristojnih vremenskih performansi.

Sadržaj

1	Opis skupa podataka	2
2	Korišćeni alati	2
3	Preprocesiranje	2
4	Zaključak	2
	Literatura	2
A	Konverzija iz HDF5 u CSV format	4

1 Opis skupa podataka

The Million Song Dataset [6] je skup od milion slogova koji sadrže informacije o popularnim pesmama. S obzirom da ovaj skup preveliki za udoban rad, ograničićemo se na podskup od deset hiljada slogova izdvojen od strane autora originalnog skupa.

Svaki slog pomenutog skupa podataka sadrži informacije o jednoj pesmi: detalje o izvođaču, segmentima, tempu kao i ID pesme na raznim online servisima (*Echo Nest* [2], *7digital* [1], *MusicBrainz* [3] i *PlayMe* [4]). Detaljne informacije o atributima se mogu videti na slici 1.

Skup u svojoj originalnoj formi je organizovan u *HDF5* format [7]. Mi ćemo izdvojiti informacije iz datog modela i podatke organizovati u CSV format, zarad lakšeg ubacivanja u alate koje ćemo koje ćemo opisati kasnije. Izdvajanje podataka i njihova konverzija u CSV format je izvršena korišćenjem Python programa koji se mogu naći u dodatku A.

2 Korišćeni alati

Za obradu podataka ćemo koristiti alate *Knime Analytics Platform* [5] i *IBM SPSS Modeler* [8].

3 Preprocesiranje

TODO Preprocesiranje

4 Zaključak

Zaključak... TODO

Literatura

- [1] 7digital. on-line at <https://www.7digital.com/>.
- [2] The echo nest. on-line at <http://the.echonest.com/>.
- [3] Musicbrainz. on-line at <https://www.musicbrainz.org/>.
- [4] Play.me. on-line at <https://www.playme.com/>.
- [5] KNIME AG. Knime analytics platform. on-line at <https://www.knime.com/knime-analytics-platform>.
- [6] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset, 2011. on-line at <https://labrosa.ee.columbia.edu/millionsong/>.
- [7] The HDF Group. The hdf5 data model, 2011. on-line at <https://support.hdfgroup.org/HDF5/>.
- [8] IBM. Ibm spss modeler. on-line at <https://www.ibm.com/analytics/data-science/predictive-analytics/spss-statistical-software>.

Atribut	Tip podatka	Kratki opis
analysis sample rate	float	učestalost uzorkovanja
artist 7digitalid	int	7digital ID izvođača ili -1
artist familiarity	float	algoritamska aproksimacija
artist hotttnesss	float	algoritamska aproksimacija
artist id	string	Echo Nest ID izvođača
artist latitude	float	geografska širina
artist location	string	lokacija autora
artist longitude	float	geografska dužina
artist mbid	string	MusicBrainz ID izvođača
artist mbtags	array string	niz MusicBrainz tagova
artist mbtags count	array int	broj MusicBrainz tagova
artist name	string	ime autora
artist playmeid	int	PlayMe ID izvođača ili -1
artist terms	array string	niz Echo Nest tagova
artist terms freq	array float	frekvencije Echo Nest tagova
artist terms weight	array float	težina Echo Nest tagova
audio md5	string	MD5 heš kod audio zapisa
bars confidence	array float	pouzdanost takta
bars start	array float	niz početaka taktova
beats confidence	array float	pouzdanost ritma
beats start	array float	niz početaka ritmova
danceability	float	algoritamska aproksimacija
duration	float	trajanje audio zapisa (u sekundama)
end of fade in	float	vreme u odnosu na pocetak u kom prestaje fade-in efekat (u sekundama)
energy	float	algoritamska aproksimacija energije pesme od strane slušaoca
key	int	tonalitet u kojem je audio zapis
key confidence	float	pouzdanost tonaliteta
loudness	float	prosečna jačina (u dB)
mode	int	mod - dur ili mol
mode confidence	float	pouzdanost moda
release	string	ime albuma
release 7digitalid	int	7digital ID albuma ili -1
sections confidence	array float	niz pouzdanosti stihova
sections start	array float	počeci stihova
segments confidence	array float	niz pouzdanosti segmenata
segments loudness max	array float	niz maksimalnih jačina unutar segmenata (u dB)
segments loudness max time	array float	niz vremena dostizanja maksimalne jačine unutar segmenata
segments loudness max start	array float	niz jačina na počecima segmenata
segments pitches	2D array float	niz jačina po segmentima, jedna vrednost za svaku notu
segments start	array float	počeci segmenata
segments timbre	2D array float	informacije o teksturi (MFCC + PCA)
similar artists	array string	niz Echo Nest sličnih izvođača
song hotttnesss	float	algoritamska aproksimacija
song id	string	Echo Nest ID pesme
start of fade out	float	vreme u odnosu na pocetak u kom počinje fade-out efekat (u sekundama)
tatums confidence	array float	pouzdanost najmanjih elemenata ritma
tatums start	array float	niz najmanjih elemenata ritma
tempo	float	procenjen tempo (u BPM)
time signature	int	procenjen broj ritmova u taktu, npr. 4
time signature confidence	float	pouzdanost procene broja ritmova u taktu
title	string	naziv pesme
track id	string	Echo Nest ID pesme
track 7digitalid	int	ID 7digital ID pesme ili -1
year	int	godina izdavanja uzeta sa MusicBrainz ili 0

Slika 1: Atributi prisutni u *The Million Song Dataset* skupu podataka

Dodatak A Konverzija iz HDF5 u CSV format

```
1  """
2  Alexis Greenstreet (October 4, 2015) University of Wisconsin-
   Madison
3  """
4  class Song:
5      songCount = 0
6
7      def __init__(self, songID):
8          self.id = songID
9          Song.songCount += 1
10         self.albumName = None
11         self.albumID = None
12         self.artistID = None
13         self.artistLatitude = None
14         self.artistLocation = None
15         self.artistLongitude = None
16         self.artistName = None
17         self.danceability = None
18         self.duration = None
19         self.genreList = []
20         self.keySignature = None
21         self.keySignatureConfidence = None
22         self.lyrics = None
23         self.loudness = None
24         self.popularity = None
25         self.tempo = None
26         self.timeSignature = None
27         self.timeSignatureConfidence = None
28         self.title = None
29         self.year = None
30         self.artistFamiliarity = None #float
31         self.artistHottnesss = None #float
32         self.audioMd5 = None # string
33         self.endOfFadeIn = None #float
34         self.energy = None #float
35         self.key = None #int
36         self.keyConfidence = None #float
37         self.mode = None # int
38         self.modeConfidence = None #float
39         self.release = None #string
40         self.songHottness = None #float
41         self.songId = None #string
42         self.startOfFadeOut = None #float
43         self.trackId = None #string
44         self.genre = None # list of strings
```

Slika 2: Klasa korišćena za deserijalizaciju podataka.

```

1  """
2  Alexis Greenstreet (October 4, 2015) University of Wisconsin-
   Madison
3  """
4  outputFile1 = open('SongCSV.csv', 'w')
5  csvRowString = ""
6
7  csvRowString = ("SongID,AlbumID, ...")
8      csvAttributeList = re.split('\W+', csvRowString)
9  for i, v in enumerate(csvAttributeList):
10      csvAttributeList[i] = csvAttributeList[i].lower()
11  outputFile1.write("SongNumber,");
12  outputFile1.write(csvRowString + "\n");
13  csvRowString = ""
14
15  #####
16  #Set the basedir here, the root directory from which the
   search
17  basedir = "/home/m/Documents/MillionSongSubset/data"
18  ext = ".h5"
19  #####
20
21  csvRowStringTotal = ""
22
23  for root, dirs, files in os.walk(basedir):
24      files = glob.glob(os.path.join(root, '*' + ext))
25      for f in files:
26          print f
27
28          songH5File = hdf5_getters.open_h5_file_read(f)
29          song = Song(str(hdf5_getters.get_song_id(songH5File)))
30
31          song.artistID = str(hdf5_getters.get_artist_id(
songH5File))
32          # Isto za ostala polja
33
34          artistMbtags = np.array(hdf5_getters.get_artist_mbtags
(songH5File))
35          song.genre = ' | '.join(artistMbtags)
36
37          csvRowString += str(song.songCount) + ","
38          csvRowString += song.id + ","
39          # Isto za ostala polja
40
41          csvRowString += song.trackId + ","
42          csvRowString += song.genre + "\n"
43          csvRowStringTotal += csvRowString
44          csvRowString = ""
45
46          songH5File.close()
47
48  outputFile1.write(csvRowStringTotal)
49  outputFile1.close()

```

Slika 3: Uprošćena verzija programa korišćenog za konvertovanje iz HDF5 u CSV format.