Big Data Analysis - Frequent item sets assignment

2019-10-11 **Cedric Mingneau, William Thenaers**

Algorithm

Apriori with first pass counting unique author frequencies and caching `key_to_index` and `index_to_key` lists. Subsequent passes make a dictionary of each supported author from the previous pass (using only their ids as keys from `key_to_index`), and use this to lazily create n-tuples when encountering them in the baskets (only baskets with at least n items are considered). After running through the entire file, counted tuples are filtered on support threshold and passed to the next pass.

The supported tuples are chained into a new dict (for O(1) lookup) consisting of every unique author that was supported in the previous pass and these will be used to check if a newly evaluated author was supported and can be used to create new n-tuples.

Progress is indicated when reading the input file, and relevant data is printed during execution, e.g.:

```
> Run with "dblp.xml" and support 12
Start pass 1...
  Finding 1-tuples...: 7075607it [02:04, 57051.80it/s]
  Caching singles LUT...
  Created 2356436 1-tuples.
  249454 1-tuples >= support 12.
Start pass 2...
  249454 supported uniques from previous pass to find tuples.
  Finding 2-tuples...: 100%|#######| 7075607/7075607 [02:13<00:00, 53123.45it/s]
  Created 3371001 2-tuples.
  136019 2-tuples >= support 12.
Start pass 3...
...
```

Encountered problems

We first created a list of every possible combination within the current pass, but this took a tremendous amount of memory to save.

Now a hash map (as a Python dict) is kept between passes, which contains only those authors that were frequent in the previous pass (`frequency >= support_threshold`). Furthermore, authors are stored as IDs from the `key_to_index` cache. When a basket is read from the file, only authors who appeared in frequent author map are kept. From these, combinations are made and added to another dict to count their frequency. Since these newly generated n-tuple combinations consist of only those authors who were already frequent in the previous pass, they are guaranteed to have at most the same frequency, hence no other checks are needed to discard them.

Each pass, the code reads the file again using the `NTupleFrequency` strategy to calculate n-tuple frequencies. The first pass uses a simplified strategy to only count single author frequencies. It would also be possible to use a third strategy to create the baskets in memory, needing to read the file only once. This is however not used, as the point of the assignment was to not read the file into memory, although reading only the author baskets would be feasible.

Benchmarks

Captured on a system with: `Intel i7 7700k @ 4.8GHz`

The program output for the settings below can be found in the benchmarks folder.

Dataset	Support threshold	Time (mm:ss)	Memory
DBPL 50k	12	00:02	~150Mb
DBPL 50k	15	00:03	~150Mb
DBPL full	12	61:06	300-600 Mb
DBPL full	15	32:01	400-600 Mb

Example output

DBPL full dataset with support threshold 12

(only n-tuples with frequency >= 12 are kept for next pass)

Pass 1

- ('Carmen Heine',): 2
- ('Gerd Hoff',): 5
- ('Margo I. Seltzer',): 100
- ('Frank Neven',): 131
- ('Kris Luyten',): 164
- ...

Pass 2

- ('Manish Jain', 'Bo An'): 6
- ('Manish Jain', 'Milind Tambe'): 34
- ('Frank Neven', 'Jan Van den Bussche'): 13
- ('Kris Luyten', 'Karin Coninx'): 115
- ('Kris Luyten', 'Jan Van den Bergh 0001'): 25
- ...

Pass 3

- ('Manish Jain', 'Bo An', 'Milind Tambe'): 4
- ('Kun Sun 0001', 'Peng Ning', 'Cliff Wang'): 5
- ('Geert Jan Bex', 'Frank Neven', 'Jan Van den Bussche'): 1
- ('Kris Luyten', 'Karin Coninx', 'Mieke Haesen'): 17

- ...

Pass 12

- ('Julie Mullen', 'Jeremy Kepner', 'David Bestor', 'Bill Bergeron', 'Siddharth Samsi', 'Peter Michaleas', 'Michael Houle', 'Vijay Gadepally', 'Michael Jones 0001', 'Antonio Rosa', 'Matthew Hubbell', 'Albert Reuther'): 16
- ('Lech Raczynski', 'Pawel Kowalski', 'Andrzej Kochanowski', 'Artur Slomski', 'Marek Palka', 'Natalia Zon', 'Pawel Moskal', 'Szymon Niedzwiecki', 'Jakub Kowal', 'Adam Strzelecki', 'Neha Gupta-Sharma', 'Wojciech Wislicki'): 14
- ...

Pass 25

(single result)

- ('Lech Raczynski', 'Pawel Kowalski', 'Andrzej Kochanowski', 'Artur Slomski', 'Marek Palka', 'Grzegorz Korcyl', 'Natalia Zon', 'Tomasz Kozik', 'Michal Silarski', 'Pawel Moskal', 'Tomasz Bednarski', 'Piotr Salabura', 'Marcin Molenda', 'Piotr Bialas', 'Zbigniew Rudy', 'Lukasz Kaplon', 'Wojciech Krzemien', 'Jerzy Smyrski', 'Eryk Czerwinski', 'Szymon Niedzwiecki', 'Jakub Kowal', 'Adam Strzelecki', 'Neha Gupta-Sharma', 'Marcin Zielinski', 'Wojciech Wislicki'): 14