EDAMI Project

Using Non-Zero Dimensions for the Cosine Similarity among Real Valued Vectors

Tomasz Adamiec

Krzysztof Lichocki

## Task description

The cosine similarity measure is used for searching sufficiently similar vectors and it depends on the angle between vectors. Computations for cosine similarity performed on sparse high dimensional data sets are time consuming. Recent results of scientific research performed by professor Marzena Kryszkiewicz has proven that there are efficient methods to speed up the similarity estimation process.

Our team decided to analyze methods for selecting the smallest set of dimensions necessary to be shared by vectors in order to be similar. We decided to investigate how much time the computation takes on data sets when applying each method and then to compare results obtained when working on standard not optimized data.

Formulas used for candidate selection:

## Assumptions

We assume that input file has proper structure as described below.

## Form of input and output data

**INPUT**

Input file must have following structure:

vc dc ddc

a b a b a b

...

a b a b

where:

vc – vectors count, dc – dimensions count, ddc – distinct dimensions count

a – dimension number, b – value

Actual values of vc, dc, ddc are dismissed – it is only important that relevant data starts on second line. Each line describes one vector in sparse manner – that is includes all non-zero dimensions immediately followed by their values. Dimensions must be sorted in ascending order.

**OUTPUT**

Output data is recorded in three log files.

Files are named in the following manner: *processedFileName-resultType-date*.log

where *processedFileName* is name of the input file, *resultType* is “results”, “resultsInfo” or “resultsDimData”, *date* is system date.

*“results” –* each line corresponds with one vector and contains following fields: vector number, number of candidates selected for evaluation, number of neighbors and listed neighbors.

*“resultsInfo”* - in this file following information about program execution is stored: specified file name, type mode, epsilon value, time taken by specific operations and statistics on results, such as maximum, minimum and average number of candidates and neighbors.

*“resultsDimData” -*  contains following fields: dimension number and maximum, minimum and average number of occurrences of given dimension.

## Program structure

Program consists of following files:

main.cpp

Mode.cpp

Mode.h

vectorOperations.cpp

vectorOperations.h

Main is containing user interface and main program function.

Mode is containing enum structure for representation of all methods of obtaining set of vector candidates.

VectorOperations is containing essential functions necessary for determining cosine neighborhood.

Vectors that are read into the memory are normalized and stored in

std::vector<std::vector<std::pair<int, double>>\*> vectors;

Inverted index is created and stored in

std::vector<std::set<int>\*> invertedIndex;

If program is executed with “top values” or “top k values” mode another copy of vectors is stored in std::vector<std::vector<std::pair<int, double>>\*> sortedVectors;

where dimensions of vectors are sorted not ascending wrt. their values.

If program is executed with “most popular dimensions” or “least popular dimensions” mode dimensions are sorted wrt. their popularity and stored in

std::vector<int> popularDimensions;

## User guide

Our program is a console application with a textual interface. User is asked to specify following parameters:

* Input file name
* Type of vectors’ values (real/binary)
* Epsilon value
* Method of determining candidate vectors
  + Default order
  + Top values
  + Top k values
  + Using most popular dimensions
  + Using least popular dimensions
  + Test (using all above methods sequentially)

After program evaluation user has possibility of choosing if to save gathered information regarding dimension specification.

## Results for a small input data

Fragment of the file *cacmcisi.mat-results-2013-06-16\_23-22-32.log*

0;5;1:0

1;3;1:1

2;3;1:2

3;10;5:3 6 9 12 18

4;59;1:4

5;2;1:5

6;10;5:3 6 9 12 18

7;6;1:7

8;2;1:8

9;10;5:3 6 9 12 18

File *cacmcisi.mat-resultsInfo-2013-06-16\_23-22-32.log*

File name:

cacmcisi.mat

Type of vectors' values:

real

number of vectors: 3209

number of distinct dimensions: 15295

dimension number: 14410

inverted index creation: 0s

Epsilon:

0.9

mode: default order

processing time: 0.061s

max candidate number: 1152

min candidate number: 1

average candidate number: 134.615

max number of neighbors: 9

min number of neighbors: 1

average number of neighbors: 1.46432

mode: top values

vectors sorting time: 0s

processing time: 0.061s

overall time: 0.061s

max candidate number: 1152

min candidate number: 1

average candidate number: 136.601

max number of neighbors: 9

min number of neighbors: 1

average number of neighbors: 1.46432

mode: k top values

vectors sorting time: 0s

processing time: 0.061s

overall time: 0.061s

max candidate number: 1152

min candidate number: 1

average candidate number: 138.628

max number of neighbors: 9

min number of neighbors: 1

average number of neighbors: 1.46432

mode: most popular dimensions

dimensions sorting time: 0s

processing time: 0.25s

overall time: 0.25s

max candidate number: 1415

min candidate number: 1

average candidate number: 464.642

max number of neighbors: 9

min number of neighbors: 1

average number of neighbors: 1.46432

mode: least popular dimensions

dimensions sorting time: 0s

processing time: 0s

overall time: 0s

max candidate number: 1037

min candidate number: 1

average candidate number: 11.0642

max number of neighbors: 9

min number of neighbors: 1

average number of neighbors: 1.46432

Fragment of the file *cacmcisi.mat-resultsDimData-2013-06-16\_23-22-32.log*

0 : 0 : 0 : 0

1 : 3 : 3 : 3

2 : 2 : 2 : 2

3 : 2 : 1 : 1.06667

4 : 1 : 1 : 1

5 : 1 : 1 : 1

6 : 2 : 1 : 1.16667

7 : 1 : 1 : 1

8 : 2 : 2 : 2

9 : 1 : 1 : 1

## Experimental results for large data sets

File name:

cacmcisi.mat

Type of vectors' values:

real

number of vectors: 4663

number of distinct dimensions: 83181

dimension number: 14409

Time of execution [s]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ε | angle | default order | top values | top k values | most popular dimensions | least popular dimensions |
| 0.95 | 18.2 | 0.829 | 0.513 | 0.798 | 1.664 | 0.048 |
| 0.96 | 16.3 | 0.653 | 0.375 | 0.777 | 1.44 | 0.094 |
| 0.97 | 14.1 | 0.454 | 0.423 | 0.625 | 1.393 | 0.078 |
| 0.98 | 11.5 | 0.454 | 0.484 | 0.763 | 1.435 | 0.016 |
| 0.99 | 8.1 | 0.376 | 0.516 | 0.514 | 0.999 | 0.061 |
| 1 | 0.0 | 0.293 | 0.563 | 0.56 | 1.218 | 0.047 |

Average number of candidates

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ε | angle | default order | top values | top k values | most popular dimensions | least popular dimensions |
| 0.95 | 18.2 | 283.149 | 222.893 | 362.052 | 731.062 | 21.5979 |
| 0.96 | 16.3 | 265.539 | 219.767 | 339.198 | 712.449 | 18.0382 |
| 0.97 | 14.1 | 246.069 | 218.408 | 311.392 | 694.153 | 15.2981 |
| 0.98 | 11.5 | 224.093 | 217.933 | 278.617 | 669.362 | 13.4581 |
| 0.99 | 8.1 | 198.09 | 217.83 | 235.893 | 639.729 | 12.5801 |
| 1 | 0.0 | 167.289 | 217.83 | 217.83 | 594.553 | 12.3551 |

File name:

sports.mat

Type of vectors' values:

real

number of vectors: 8580

number of distinct dimensions: 1107980

dimension number: 126355

Time of execution [s]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ε | angle | default order | top values | top k values | most popular dimensions | least popular dimensions |
| 0.95 | 18.2 | 70.477 | 23.561 | 86.101 | 96.293 | 12.962 |
| 0.96 | 16.3 | 67.015 | 20.435 | 79.41 | 90.023 | 8.938 |
| 0.97 | 14.1 | 58.183 | 20.206 | 71.379 | 88.206 | 6.221 |
| 0.98 | 11.5 | 49.455 | 18.829 | 62.623 | 83.923 | 3.748 |
| 0.99 | 8.1 | 35.921 | 20.658 | 47.179 | 82.374 | 1.589 |
| 1 | 0.0 | 8.529 | 19.464 | 19.813 | 75.329 | 0.125 |

Average number of candidates

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ε | angle | default order | top values | top k values | most popular dimensions | least popular dimensions |
| 0.95 | 18.2 | 5154.9 | 1672.6 | 6378.79 | 7571.56 | 790.533 |
| 0.96 | 16.3 | 4766.72 | 1553.08 | 6025.37 | 7425.86 | 586.159 |
| 0.97 | 14.1 | 4268.76 | 1495.82 | 5531.57 | 7237.4 | 393.812 |
| 0.98 | 11.5 | 3586.87 | 1469.84 | 4762.93 | 6979.36 | 221.869 |
| 0.99 | 8.1 | 2550.47 | 1467.71 | 3431.75 | 6578.65 | 82.6691 |
| 1 | 0.0 | 669.019 | 1467.71 | 1467.71 | 5874.36 | 7.24627 |

File name:

reviews.mat

Type of vectors' values:

real

number of vectors: 4069

number of distinct dimensions: 781635

dimension number: 126354

Time of execution [s]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ε | angle | default order | top values | top k values | most popular dimensions | least popular dimensions |
| 0.95 | 18.2 | 23.018 | 4.245 | 24.991 | 29.878 | 3.773 |
| 0.96 | 16.3 | 21.646 | 3.91 | 23.195 | 29.376 | 3.064 |
| 0.97 | 14.1 | 20.076 | 3.787 | 19.857 | 29.366 | 2.004 |
| 0.98 | 11.5 | 17.738 | 3.917 | 17.762 | 27.904 | 1.142 |
| 0.99 | 8.1 | 13.61 | 3.841 | 12.524 | 25.664 | 0.516 |
| 1 | 0.0 | 1.893 | 3.664 | 3.83 | 17.384 | 0.031 |

Average number of candidates

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ε | angle | default order | top values | top k values | most popular dimensions | least popular dimensions |
| 0.95 | 18.2 | 2593.48 | 471.512 | 2763.12 | 3561.36 | 375.426 |
| 0.96 | 16.3 | 2427.41 | 431.026 | 2557.57 | 3490.58 | 279.961 |
| 0.97 | 14.1 | 2205.42 | 412.47 | 2283.34 | 3396.61 | 191.005 |
| 0.98 | 11.5 | 1899.94 | 400.867 | 1897.09 | 3238.16 | 108.789 |
| 0.99 | 8.1 | 1426.45 | 399.613 | 1292.14 | 2952.85 | 41.8575 |
| 1 | 0.0 | 211.728 | 399.613 | 399.613 | 1884.58 | 1.93831 |

Average number of candidates in binary versions of files (ε = 0.95)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | default order | top values | top k values | most popular dimensions | least popular dimensions |
| cacmcisi.mat | 256.8 | 256.8 | 256.8 | 807.042 | 13.3661 |
| sports.mat | 4866.02 | 4866.02 | 4866.02 | 7984.37 | 216.879 |
| reviews.mat | 2485.18 | 2485.18 | 2485.21 | 3714.04 | 98.35 |

## Conclusions

We have observed that the best execution time and the least average candidate number is achieved when selection of candidate vectors is based on least popular dimensions. Second best results are obtained with “top values” method, although with epsilon approaching 1 “default order” method gives slightly better results. Method “top k values” performs worse than both “top values” and “default order”, but with epsilon approaching 1 results are approaching those of “top values” method. It is because of value of k getting nearer the count of selected dimensions in “top values” method. Not surprisingly, selection based on most popular dimensions gives the worst results.

Average number of candidates in method based on least popular dimensions is in most cases more than 10 times better than in other methods. As this method doesn’t need any additional computational expensive operations we would expect similar difference in execution times. Execution time is indeed in most cases more than 10 times better than in other methods.

Data with binary valued dimensions give similar results, except that “default order”, “top values” and “top k values” methods act identically and have the same performance.

## Reference

Marzena Kryszkiewicz: “Using Non-Zero Dimensions for the Cosine and Tanimoto Similarity Search among Real Valued Vectors”