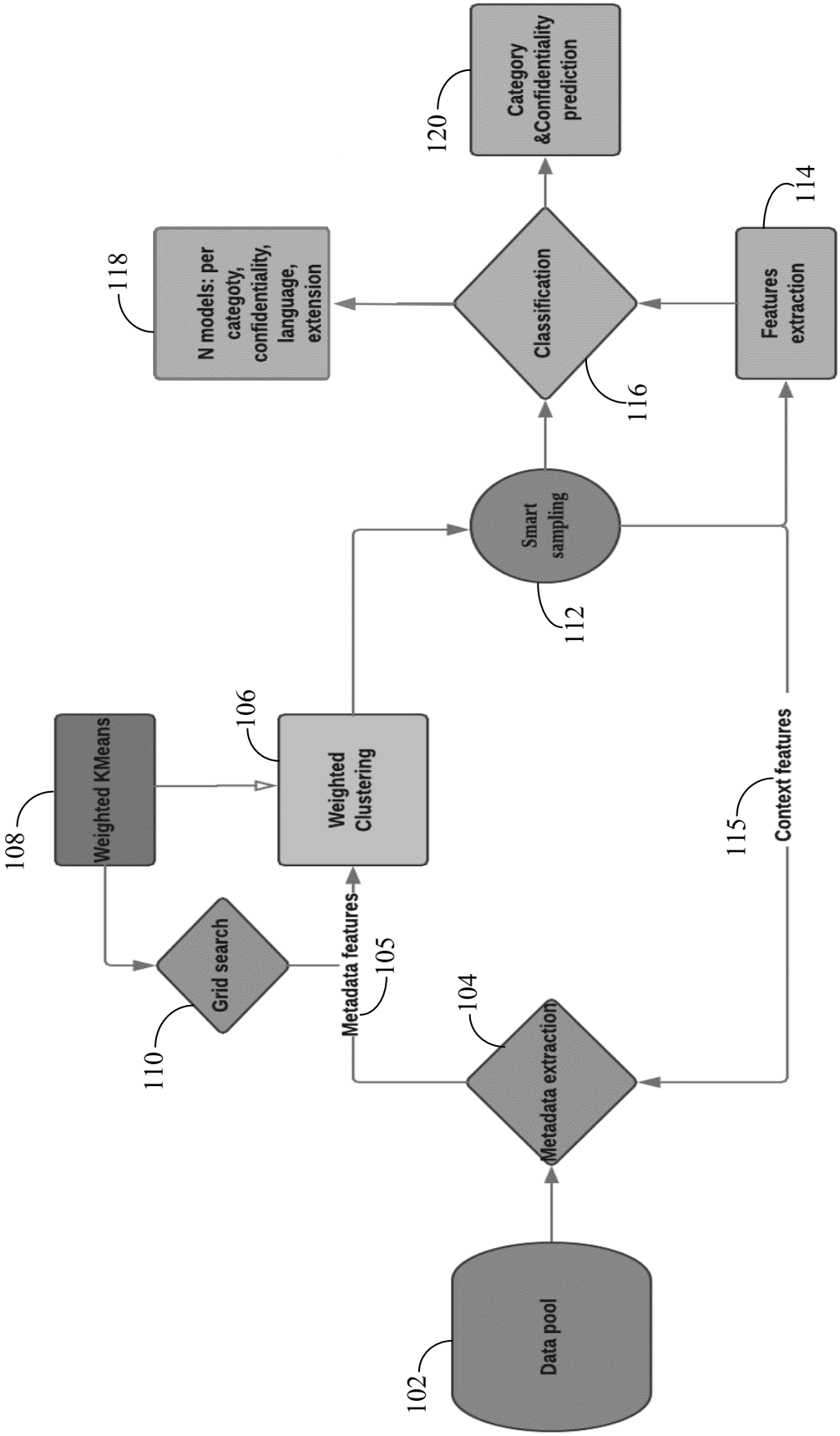
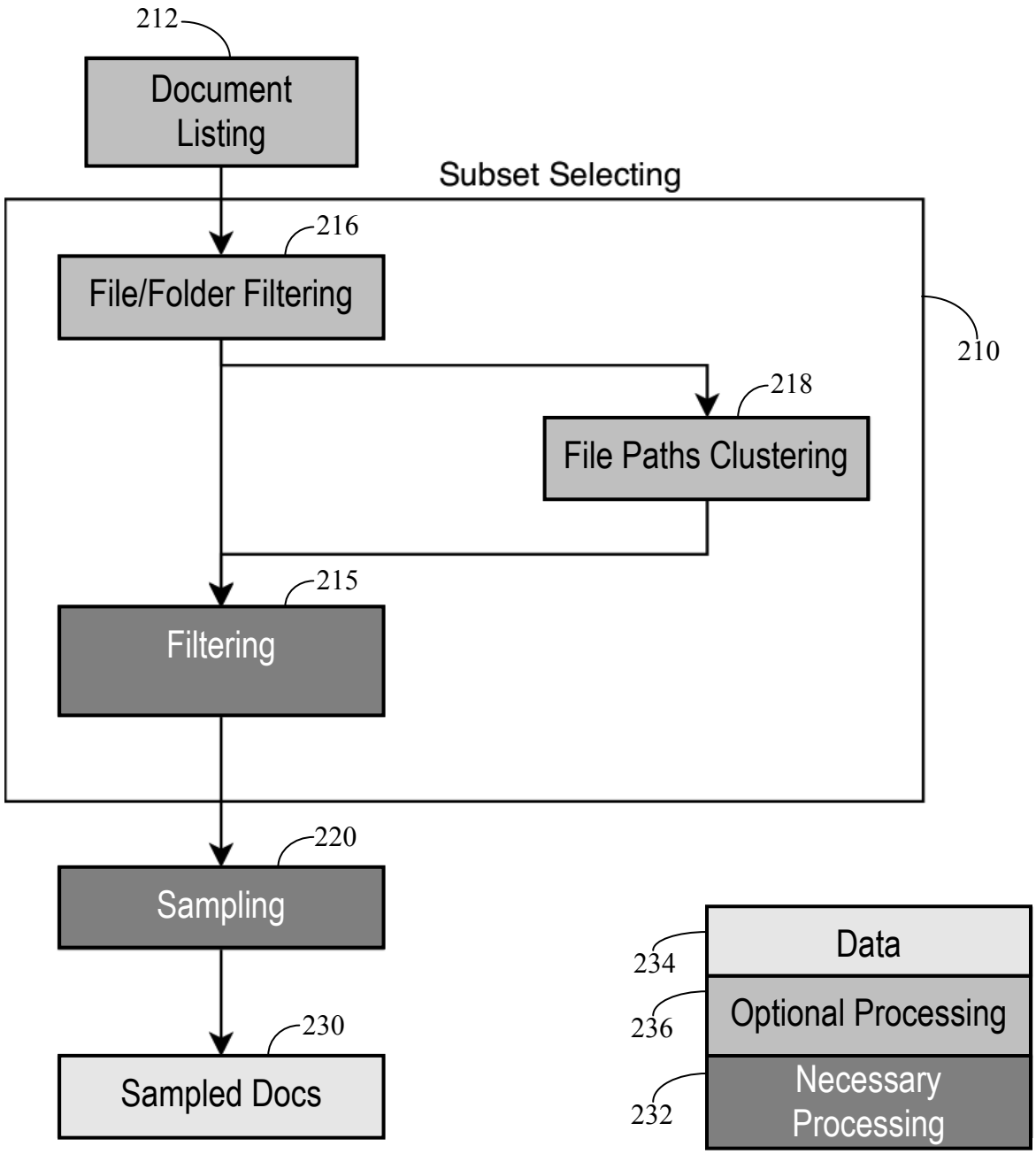

Data pool
102
Metadata extraction
104
Metadata features
105
Grid search
110
Weighted KMeans
108
Weighted Clustering
106
Smart sampling
112
Classification
116
N models: per categoty, confidentiality, language, extension
118
Category &Confidentiality prediction
120
Features extraction
114
Context features
115

100

FIG. 1


212
Document Listing
Subset Selecting
216
File/Folder Filtering
210
218
File Paths Clustering
215
Filtering
220
Sampling
230
Sampled Docs
234
Data
236
Optional Processing
232
Necessary Processing

200

FIG. 2A

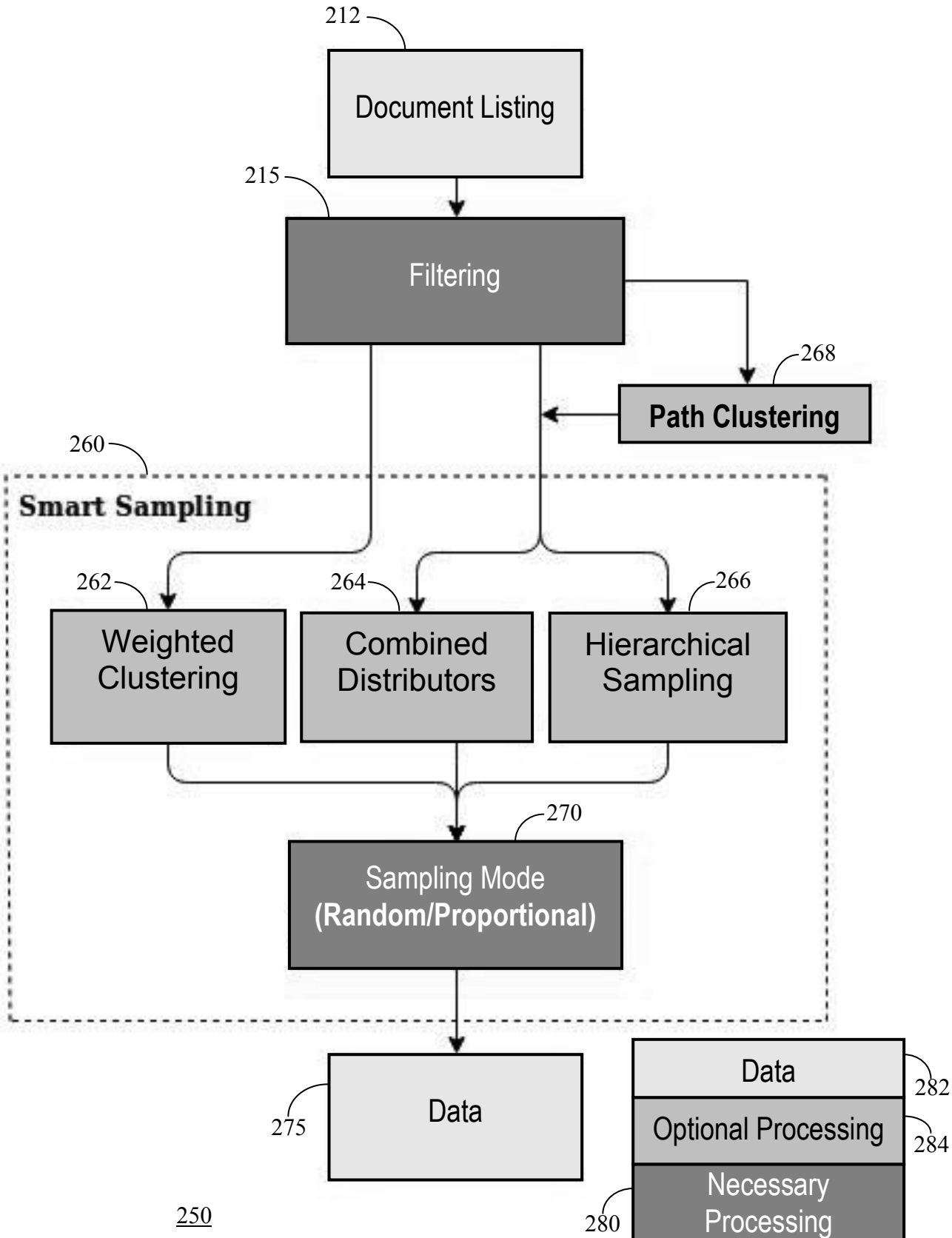
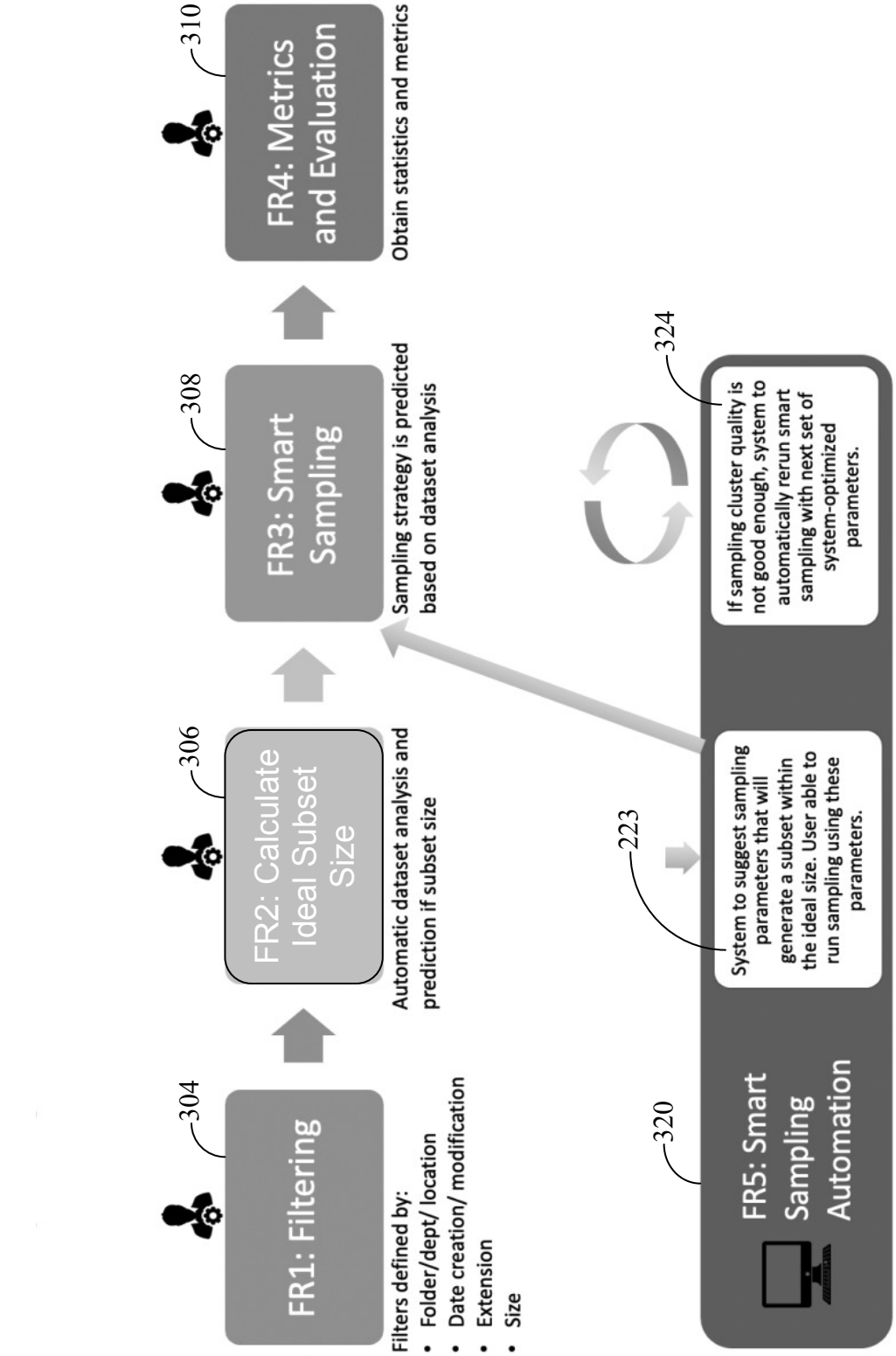

212
Document Listing
215
Filtering
268
Path Clustering
260
Smart Sampling
262
Weighted Clustering
264
Combined Distributors
266
Hierarchical Sampling
270
Sampling Mode (Random/Proportional)
275
Data
Data
282
Optional Processing
284
Necessary Processing
280
250

FIG. 2B


302
Main dataset
304
FR1: Filtering
Filters defined by:
• Folder/dept/ location
• Date creation/ modification
• Extension
• Size
306
FR2: Calculate Ideal Subset Size
Automatic dataset analysis and prediction if subset size
308
FR3: Smart Sampling
Sampling strategy is predicted based on dataset analysis
310
FR4: Metrics and Evaluation
Obtain statistics and metrics
320
FR5: Smart Sampling Automation
223
System to suggest sampling parameters that will generate a subset within the ideal size. User able to run sampling using these parameters.
324
If sampling cluster quality is not good enough, system to automatically rerun smart sampling with next set of system-optimized parameters.

300

FIG. 3A

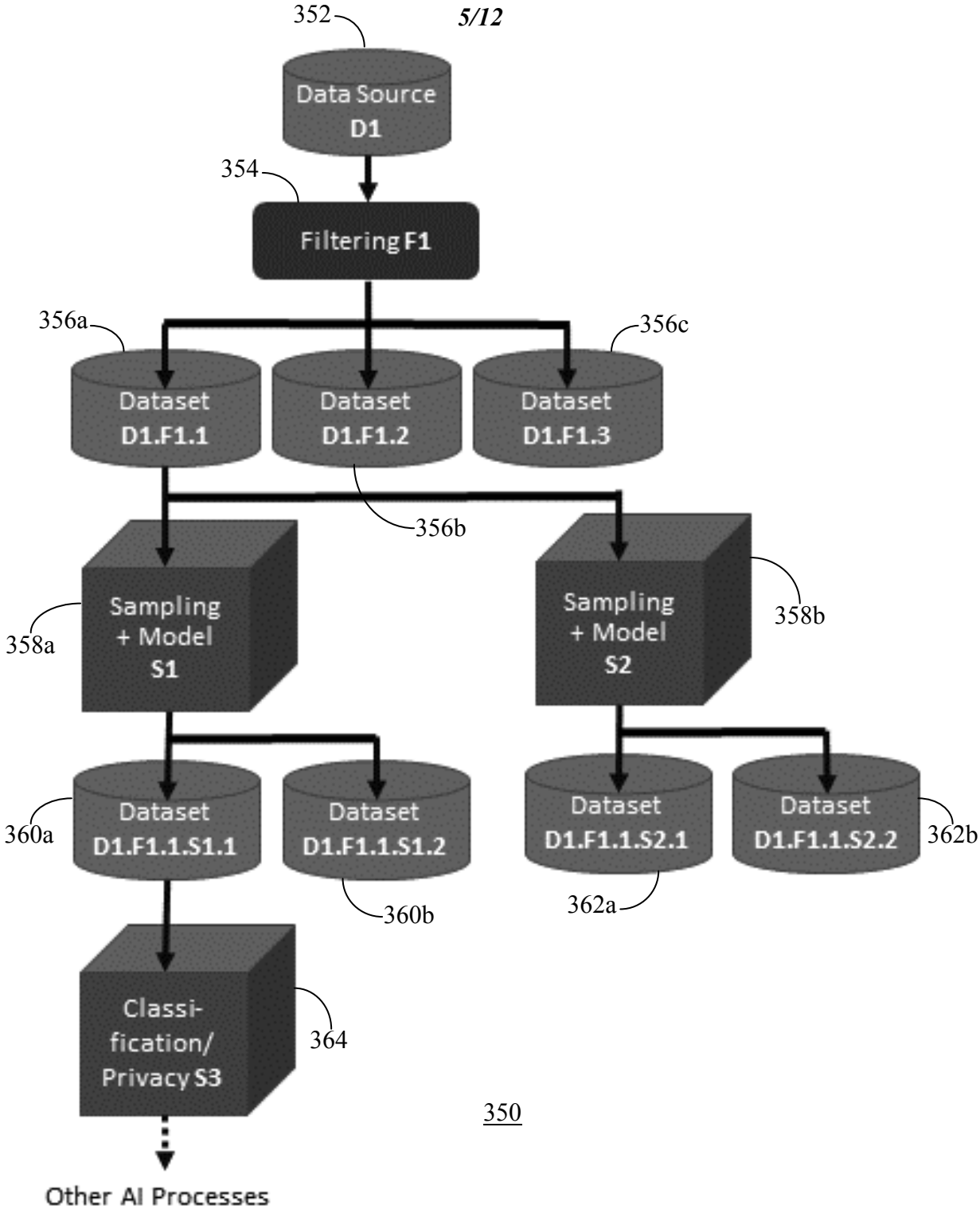

352
Data Source
D1
354
Filtering F1
356a
356c
Dataset
D1.F1.1
Dataset
D1.F1.2
Dataset
D1.F1.3
356b
Sampling
+ Model
S1
358a
Sampling
+ Model
S2
358b
Dataset
D1.F1.1.S1.1
360a
Dataset
D1.F1.1.S1.2
360b
Dataset
D1.F1.1.S2.1
362a
Dataset
D1.F1.1.S2.2
362b
Classi-
fication/
Privacy S3
364
350
Other AI Processes

FIG. 3B

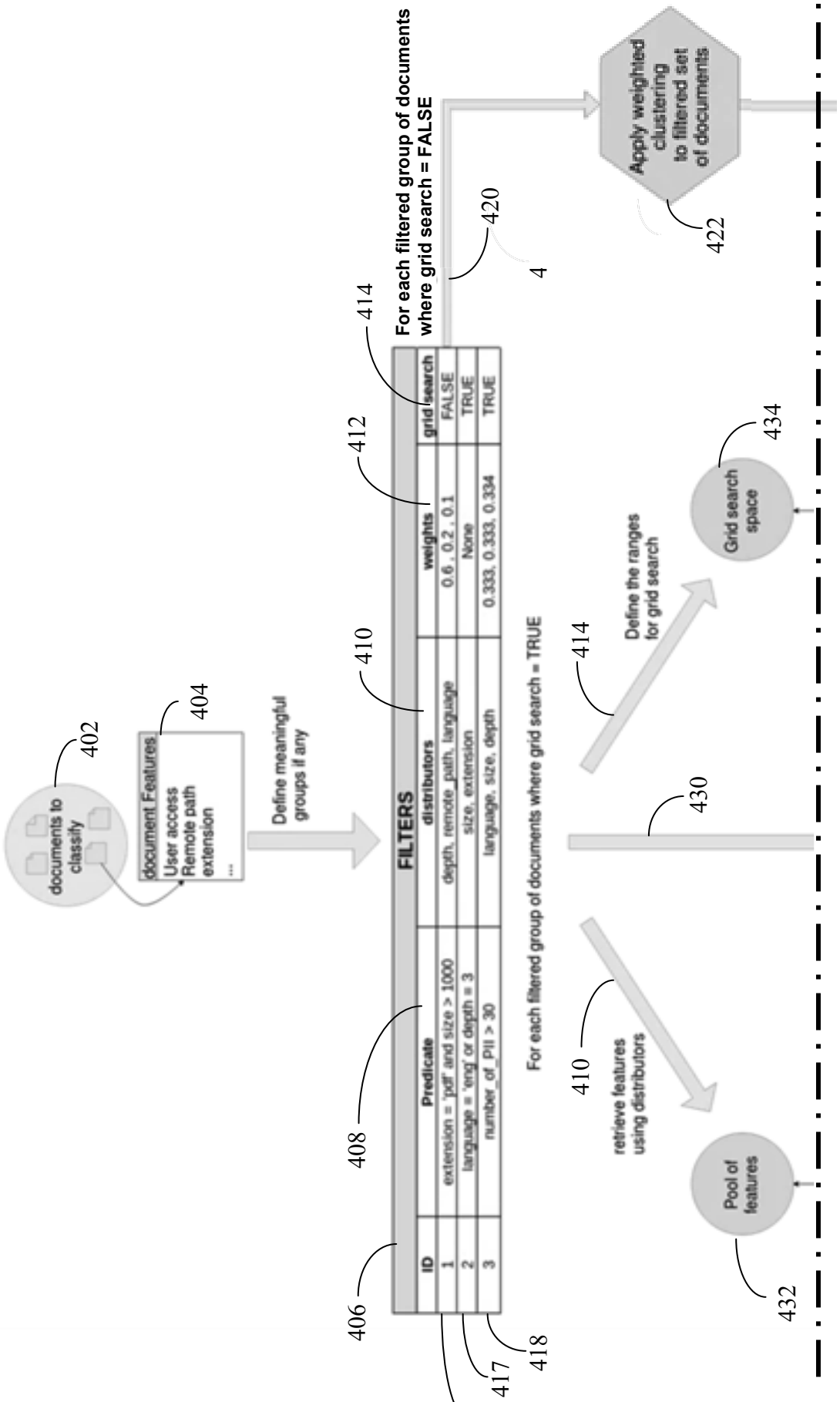

documents to classify
402
document Features
User access
Remote path
extension
...
404
Define meaningful groups if any
FILTERS
ID
Predicate
distributors
weights
grid search
1
extension = 'pdf' and size > 1000
depth, remote_path, language
0.6 , 0.2 , 0.1
FALSE
2
language = 'eng' or depth = 3
size, extension
None
TRUE
3
number_of_PII > 30
language, size, depth
0.333, 0.333, 0.334
TRUE
406
408
410
412
414
416
417
418
For each filtered group of documents where grid search = TRUE
For each filtered group of documents where grid search = FALSE
420
4
Apply weighted clustering to filtered set of documents
422
Define the ranges for grid search
414
Grid search space
434
430
410
retrieve features using distributors
Pool of features
432
400

FIG. 4A

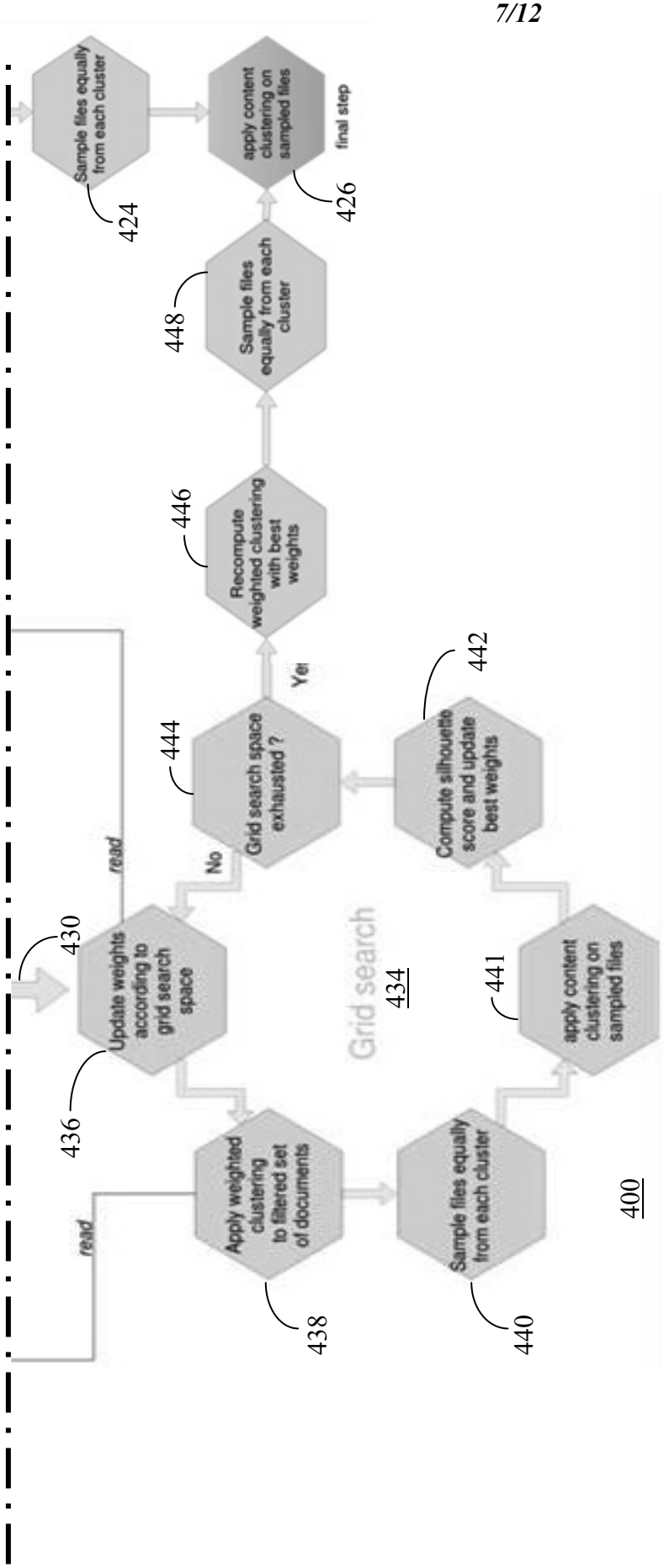

Sample files equally from each cluster
424
apply content clustering on sampled files
final step
426
448
Sample files equally from each cluster
446
Recompute weighted clustering with best weights
Yes
444
Grid search space exhausted ?
No
442
Compute silhouette score and update best weights
read
430
Update weights according to grid search space
436
Grid search
434
441
apply content clustering on sampled files
read
Apply weighted clustering to filtered set of documents
438
Sample files equally from each cluster
440
400

FIG. 4B

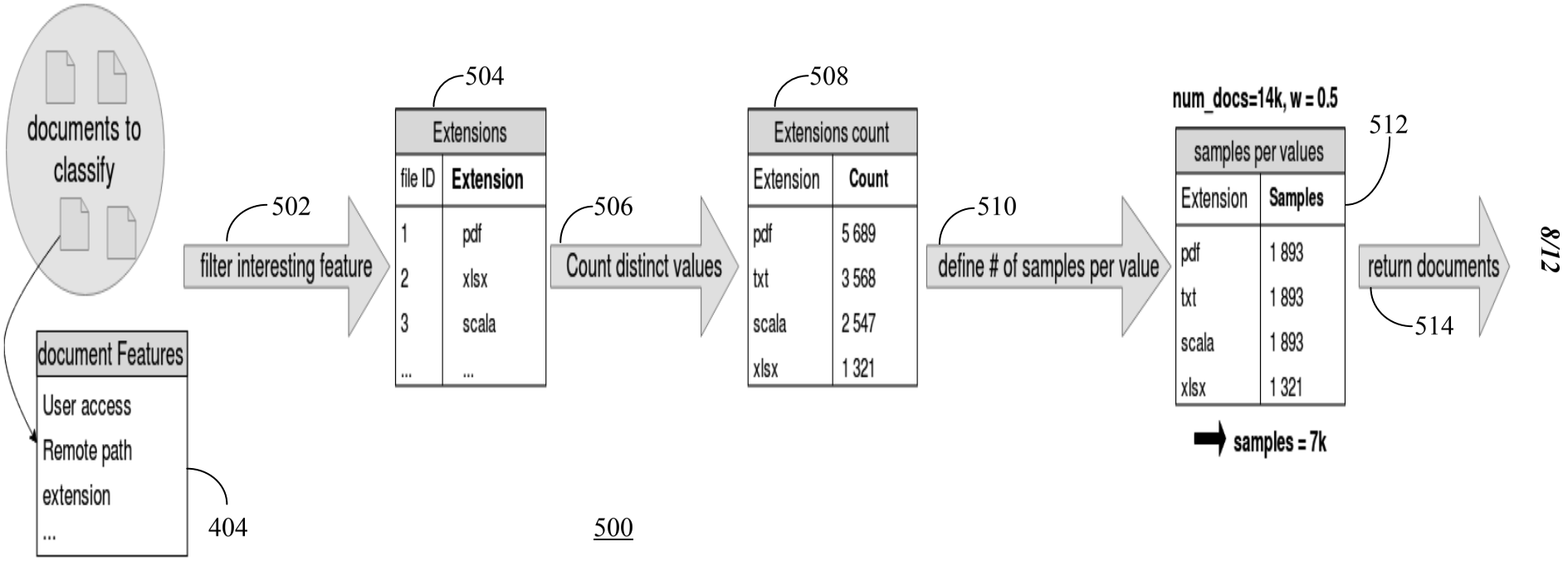

documents to classify
document Features
User access
Remote path
extension
...
404
502
filter interesting feature
504
Extensions
file ID
Extension
1
pdf
2
xlsx
3
scala
...
...
506
Count distinct values
508
Extensions count
Extension
Count
pdf
5 689
txt
3 568
scala
2 547
xlsx
1 321
510
define # of samples per value
num_docs=14k, w = 0.5
512
samples per values
Extension
Samples
pdf
1 893
txt
1 893
scala
1 893
xlsx
1 321
samples = 7k
return documents
514
500

FIG. 5

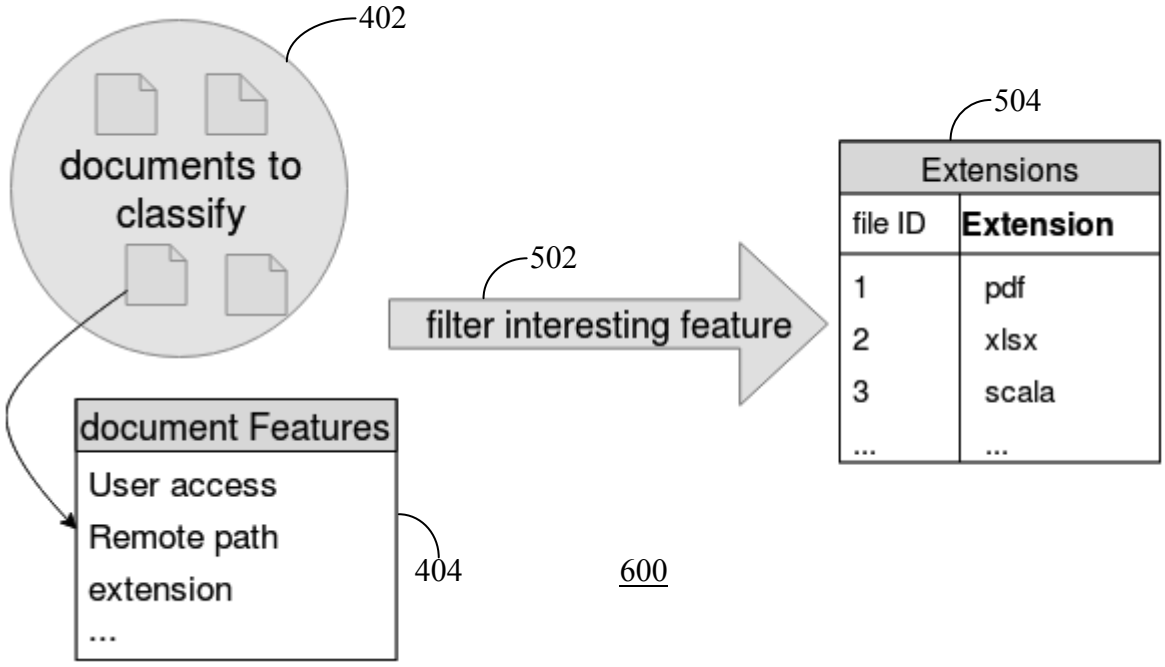
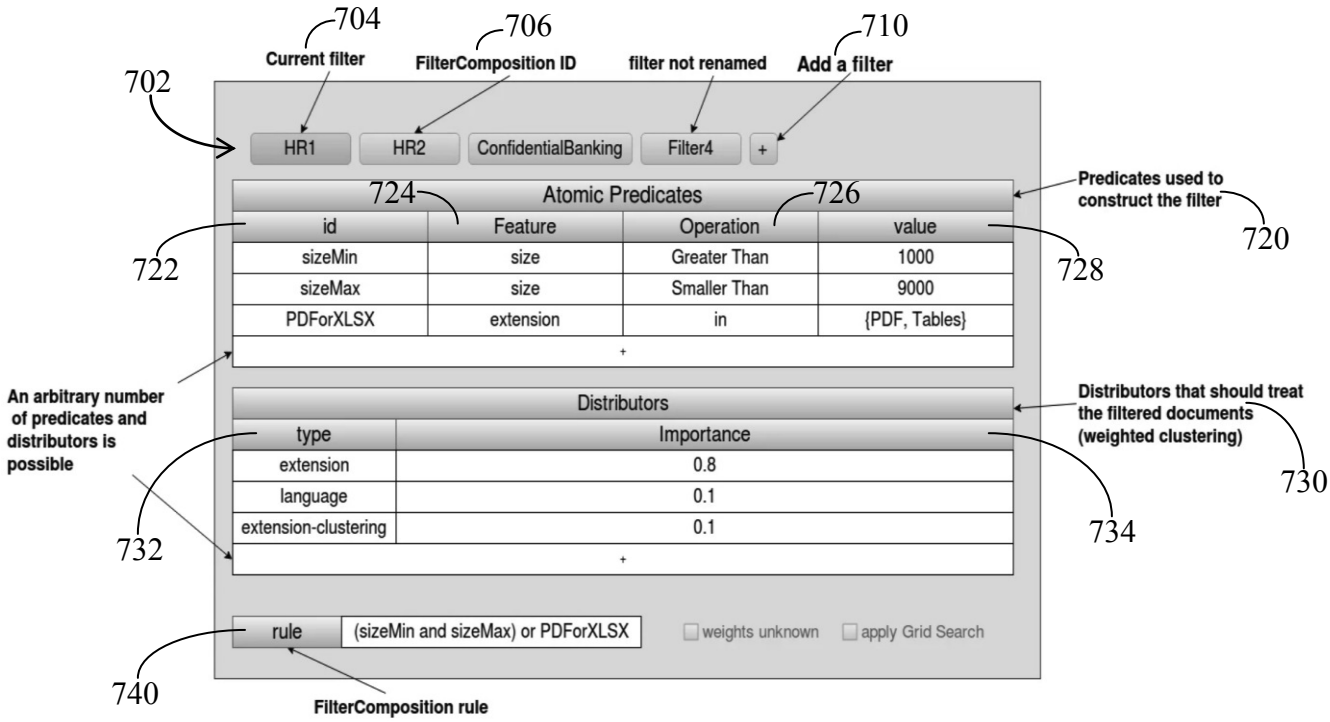
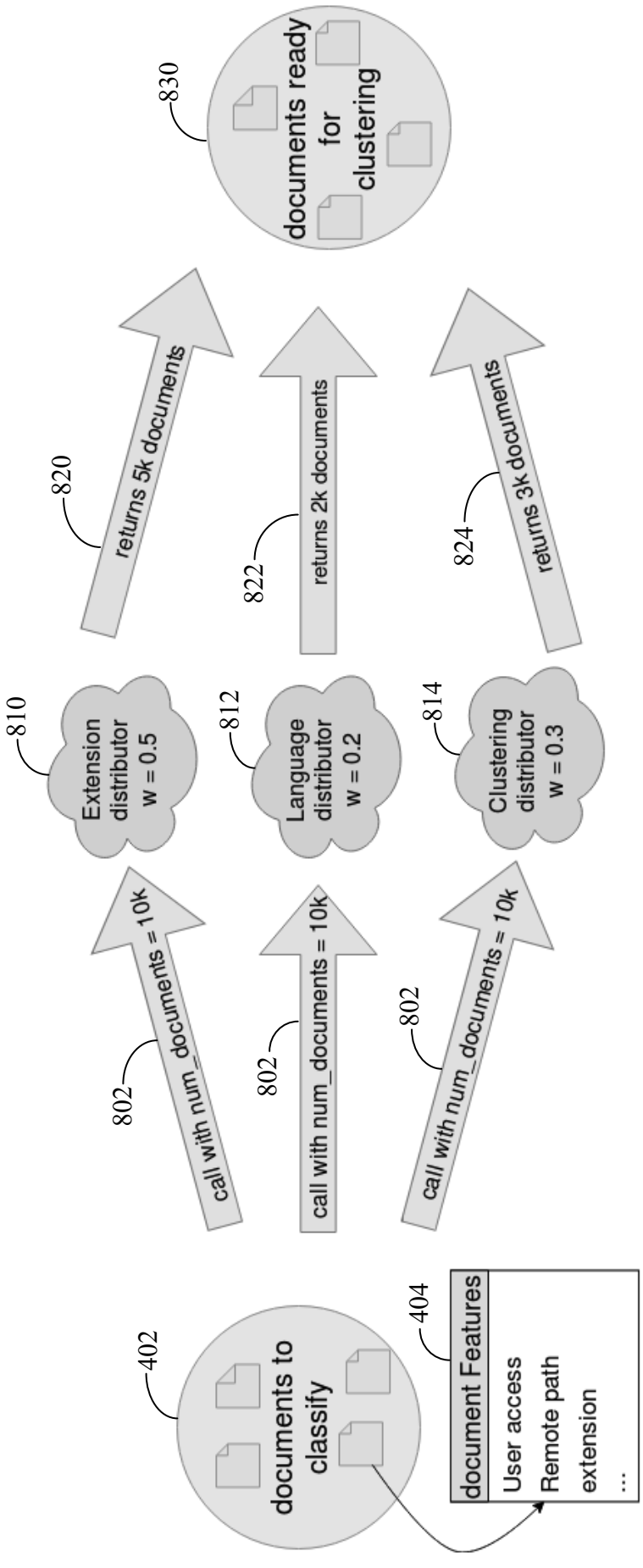


FIG. 6



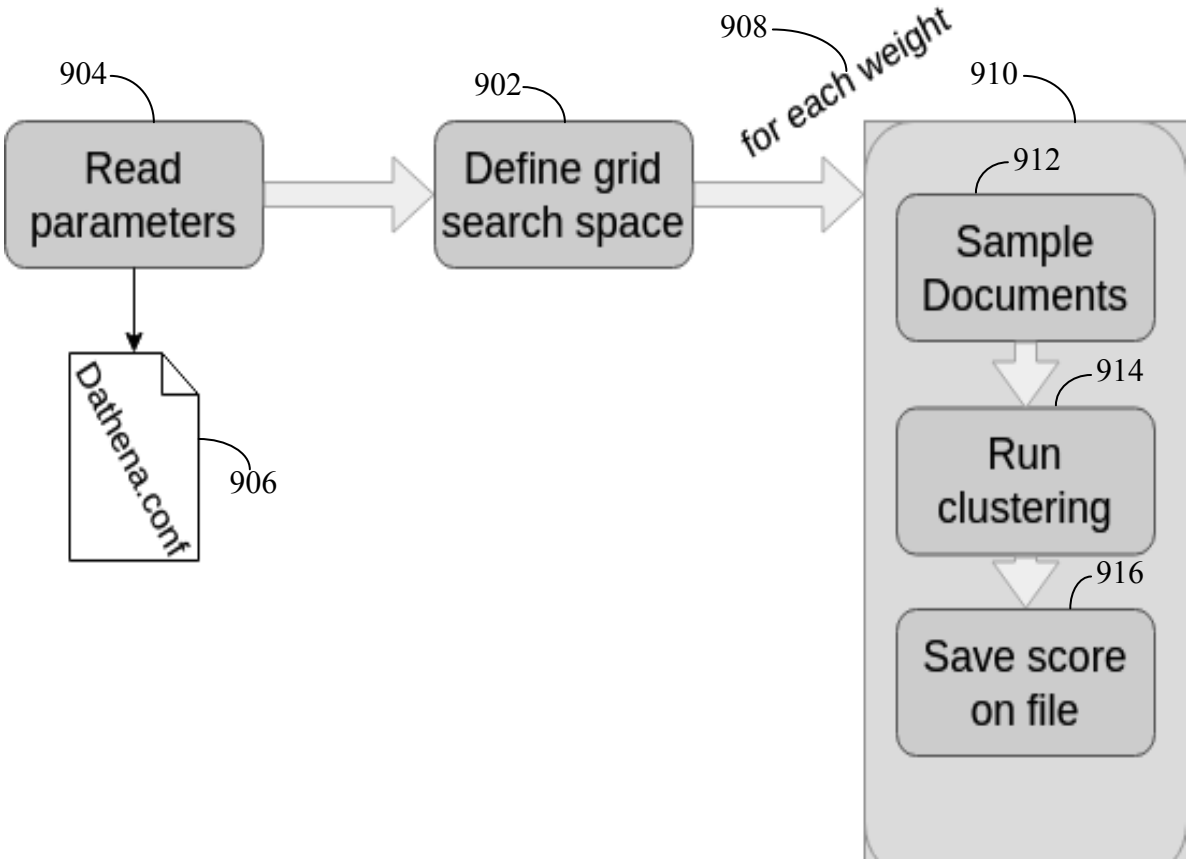
700

FIG. 7


documents to classify
402
document Features
User access
Remote path
extension
...
404
call with num_documents = 10k
802
call with num_documents = 10k
802
call with num_documents = 10k
802
Extension distributor w = 0.5
810
Language distributor w = 0.2
812
Clustering distributor w = 0.3
814
returns 5k documents
820
returns 2k documents
822
returns 3k documents
824
documents ready for clustering
830

800

FIG. 8


904
Read parameters
902
Define grid search space
908
for each weight
910
912
Sample Documents
914
Run clustering
916
Save score on file
Dathena.conf
906

900

FIG. 9

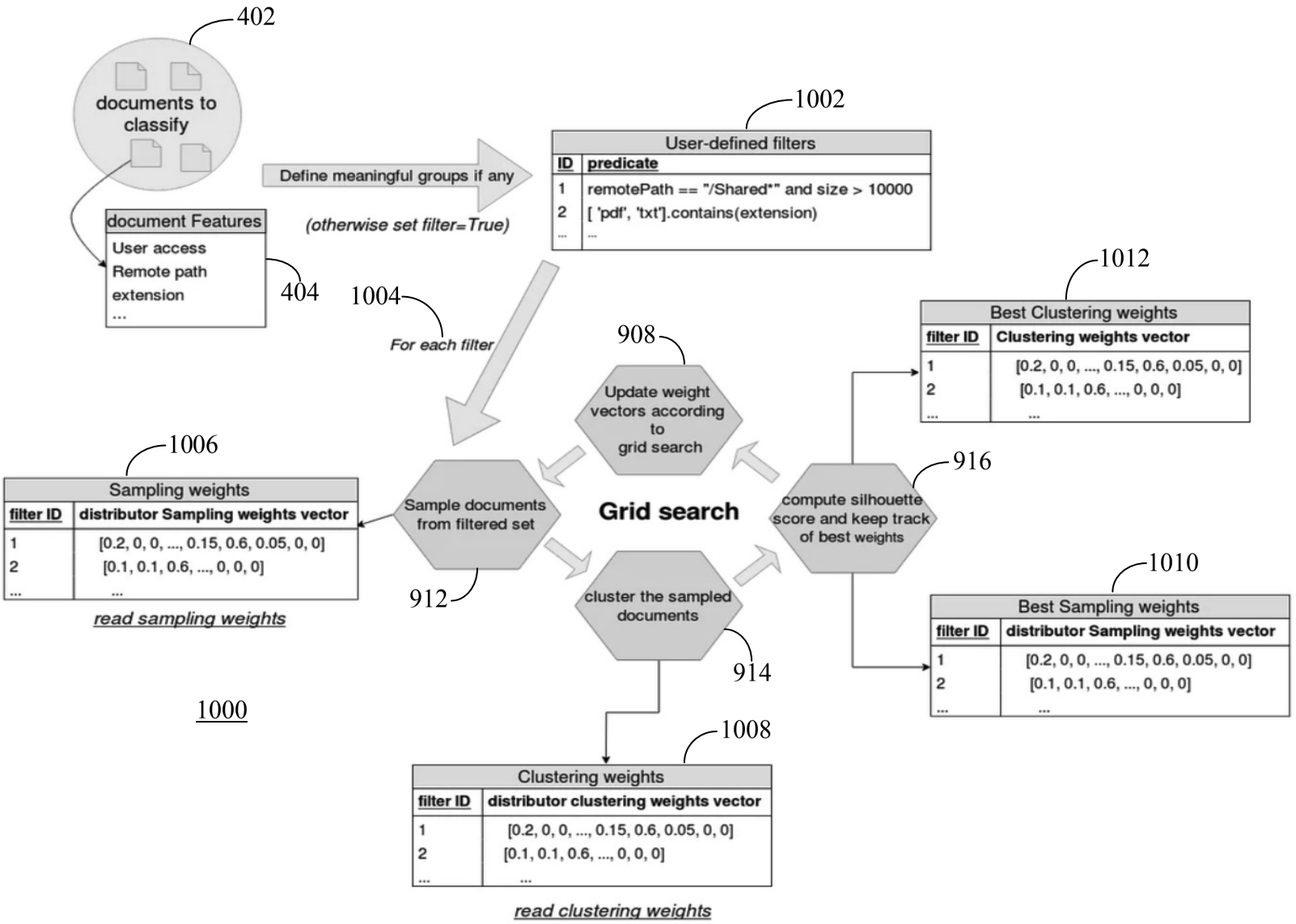

402
documents to classify
Define meaningful groups if any
(otherwise set filter=True)
1002
User-defined filters
ID | predicate
1 | remotePath == "/Shared*" and size > 10000
2 | ['pdf', 'txt'].contains(extension)
document Features
User access
Remote path
extension
...
404
1004
For each filter
908
Update weight vectors according to grid search
1012
Best Clustering weights
filter ID | Clustering weights vector
1 | [0.2, 0, 0, ..., 0.15, 0.6, 0.05, 0, 0]
2 | [0.1, 0.1, 0.6, ..., 0, 0, 0]
1006
Sampling weights
filter ID | distributor Sampling weights vector
1 | [0.2, 0, 0, ..., 0.15, 0.6, 0.05, 0, 0]
2 | [0.1, 0.1, 0.6, ..., 0, 0, 0]
read sampling weights
Sample documents from filtered set
Grid search
compute silhouette score and keep track of best weights
916
912
cluster the sampled documents
914
1010
Best Sampling weights
filter ID | distributor Sampling weights vector
1 | [0.2, 0, 0, ..., 0.15, 0.6, 0.05, 0, 0]
2 | [0.1, 0.1, 0.6, ..., 0, 0, 0]
1000
1008
Clustering weights
filter ID | distributor clustering weights vector
1 | [0.2, 0, 0, ..., 0.15, 0.6, 0.05, 0, 0]
2 | [0.1, 0.1, 0.6, ..., 0, 0, 0]
read clustering weights

FIG. 10