Fast Supervised Dimensionality Reduction Algorithm with Applications to Document Categorization and Retrieval

George Karypis

Eui-Hong (Sam) Han

University of Minnesota

Presented by Jason Keller

Contents

- Representation of documents
- Dimensionality reduction via clustering
- Interpretation as Concept Indexing
- Experimental results
- Conclusions

Vector Space Modeling of Documents

- *tf-idf* representation
  - Represent each document as a vector of term frequencies, tf_i
  - Multiply by $\log(N / df_i)$, where N is the total number of documents, and df_i is the document frequency
  - $\vec{d}_{tf-idf} = (tf_1 \log(N / df_1), \ldots, tf_n \log(N / df_n))$
  - Normalize all vectors to length 1

Vector Space Modeling of Documents

- Given a set *S* of documents and corresponding vector representations,

$$\vec{C} = \frac{1}{|S|} \sum_{\vec{d} \in S} \vec{d}$$

  is the centroid of supporting set *S*.
- Document similarity is measured by cosine correlation
- Major result: $\left\| \vec{C} \right\|_2^2$ measures pairwise similarity between documents in *S*

Unsupervised Dimensionality Reduction

- Partition the collection of documents into *k* disjoint sets
- Find the centroid for each disjoint set
- Scale the centroids to unit length
- Let *C* be the matrix in which the *i*th column corresponds to the *i*th centroid
- Apply *C* to the document vectors to project them into *k*-dimensional space

Supervised Dimensionality Reduction

- Modification of unsupervised dimensionality reduction
- Given j document classes, compute a j-way clustering using the classes
- If k dimensions are desired for the dimensionality reduction and $k > j$, clusters will be partitioned in increasing order of similarity
- Agglomerative clustering can be used if $k < j$, but may combine unlike concepts

Concept Indexing

- Largest values in centroids correspond to the most important terms associated with a concept

	re1									
cocoa	0.62 cocoa	0.40 buffer	0.29 icco	0.25 deleg	0.23 stock	0.18 rule	0.12 consum	0.11 council	0.10 ghana	0.09 compromis
grain	0.37 wheate	0.27 corn	0.27 tonne	0.24 grain	0.16 export	0.16 min	0.14 soviet	0.14 usda	0.13 maize	0.13 crop
veg	0.44 palm	0.35 oil	0.25 tax	0.24 veget	0.21 ec	0.18 tonne	0.15 fate	0.13 indonesia	0.13 olein	0.12 rbd
wheat	0.52 wheate	0.28 tonne	0.17 stg	0.16 intervent	0.16 bonu	0.16 home	0.15 market	0.15 flour	0.13 barlel	0.13 fe
copper	0.72 copper	0.21 mine	0.17 ct	0.17 cent	0.17 magma	0.16 cathod	0.14 ton	0.14 lb	0.12 noranda	0.11 miner
coffee	0.67 coffee	0.26 ico	0.26 quota	0.17 bag	0.16 export	0.15 brazil	0.14 colombia	0.14 meet	0.12 ibc	0.12 produc
sugar	0.72 sugar	0.22 tonne	0.22 white	0.15 trader	0.14 intervent	0.14 ec	0.13 tender	0.12 ecu	0.12 rebat	0.11 cargoe
ship	0.33 ship	0.27 port	0.23 strike	0.20 vessel	0.20 seamen	0.14 union	0.13 cargo	0.13 tanker	0.12 gulf	0.12 worker
cotton	0.77 cotton	0.35 bale	0.14 plant	0.12 upland	0.11 weather	0.11 crop	0.10 certif	0.08 china	0.08 exchang	0.08 pct
carcass	0.46 beef	0.34 meate	0.19 iowa	0.16 slaughter	0.15 dakota	0.15 plant	0.15 pork	0.15 citi	0.15 lockout	0.14 ufcwu
crude	0.41 oil	0.24 crude	0.24 barrel	0.21 opec	0.17 bpd	0.17 dir	0.17 min	0.14 price	0.13 bble	0.12 energi
nat	0.59 ga	0.27 natur	0.25 feet	0.21 pipelin	0.18 cubic	0.17 butan	0.12 it	0.11 flow	0.11 energi	0.11 co
meal	0.33 meal	0.30 fe	0.24 tonne	0.22 pellet	0.18 cake	0.18 compound	0.16 min	0.16 guarante	0.15 credit	0.14 fish
alum	0.59 aluminium	0.28 alcan	0.27 aluminum	0.26 smelter	0.15 alumina	0.14 lme	0.12 tonne	0.12 metal	0.12 suralco	0.11 capac
oilseed	0.49 soybean	0.24 tonne	0.21 crusher	0.21 rapese	0.21 oilsee	0.16 loan	0.16 shipment	0.15 cargill	0.14 japanes	0.12 bought
gold	0.64 gold	0.38 ounce	0.25 mine	0.22 ton	0.14 coin	0.13 feet	0.12 silver	0.12 ore	0.11 assal	0.10 reserv
tin	0.63 tin	0.28 miner	0.18 atpc	0.17 itc	0.17 strike	0.16 bolivia	0.11 comibol	0.11 bolivian	0.10 paz	0.10 hunger
livestock	0.40 beef	0.35 cattle	0.29 pork	0.23 meate	0.17 dairi	0.16 lb	0.14 head	0.13 japan	0.12 bonu	0.11 nppc
iron	0.69 steel	0.19 iron	0.13 min	0.13 industri	0.12 ore	0.12 product	0.12 coal	0.11 steelmak	0.10 tonne	0.10 plate
rubber	0.65 rubber	0.25 pact	0.24 inra	0.16 confer	0.15 consum	0.15 price	0.14 natur	0.14 kuto	0.12 agreem	0.11 adopt
zinc	0.71 zinc	0.16 pound	0.16 grade	0.15 metal	0.14 februari	0.14 januari	0.13 smelter	0.12 mint	0.12 smelt	0.11 ct
orange	0.46 orang	0.41 juice	0.31 fcoj	0.27 dutl	0.26 gallon	0.21 frozen	0.17 florida	0.16 citru	0.13 brazil	0.12 depart
pet	0.31 resin	0.28 ethylen	0.25 pound	0.19 dow	0.19 chemic	0.19 plant	0.18 polypropylen	0.17 ventur	0.16 ct	0.15 petrochem
dlr	0.65 dollar	0.32 yen	0.28 bank	0.17 dealer	0.16 japan	0.16 baker	0.15 rate	0.15 currenc	0.14 interven	0.14 parl
gas	0.59 gasolin	0.23 unlead	0.17 min	0.16 distill	0.16 tax	0.14 fuel	0.13 refin	0.13 ela	0.13 octan	0.11 compon

Table 1: The ten highest weight terms in the centroids of the classes for a subset of the Reuters-21578 text collection.

Observations

- Few high-weight terms in each centroid; these terms can act as keywords to describe concepts
- High-weight terms are synonymous or closely related to the classes they represent
- Terms in each document vector correspond to how well the document matches the information in each centroid
- Representation captures latent associations between terms that describe concepts

Experiments

- Multi-class Categorization: CI versus higher-dimensional document space
- Single-class Categorization: Compares CI with Naïve Bayes, LSI, and higher-dimensional document space representation
- Query Retrieval: CI versus LSI

Results: Multi-Class Categorization

- Measured in terms of microaveraged Precision/Recall Breakeven Point

Topic	*k*NN	CI-*k*NN	SVM	CI-SVM
earn	97.10	97.40	98.46	98.45
acq	91.00	92.60	92.89	92.35
money-fx	77.40	82.10	76.26	82.32
grain	85.40	89.20	92.66	93.83
crude	85.50	88.60	87.83	88.76
trade	74.80	81.80	76.32	80.00
interest	72.10	78.40	68.80	76.07
ship	81.30	85.60	83.79	87.20
wheat	80.30	80.00	83.33	87.14
corn	78.40	78.90	85.15	84.87
microaverage	83.13	86.10	85.15	87.62

Table 2: Precision/Recall breakeven point on the ten most frequent Reuters topics and microaveraged performance over all Reuters topics.

Results: Single-Class Categorization

- *k*-NN and C4.5 using CI performed about 3-7% better in terms of accuracy than using higher-dimensional space
- CI clearly outperformed LSI
- Was comparable in accuracy to Naïve Bayes

Results: Single-Class Categorization

	Original Space		CI Reduced Space		LSI Reduced Space				
					C4.5		*k*NN		
	C4.5	*k*NN	C4.5	*k*NN	25 Dims	50 Dims	25 Dims	50 Dims	NB
west1	85.5%	82.9%	86.2%	**86.7%**	73.7%	74.5%	83.0%	81.4%	**86.7%**
west2	75.3%	77.2%	75.3%	**78.7%**	63.8%	59.2%	75.5%	73.8%	76.5%
west3	73.5%	76.1%	74.5%	**80.6%**	57.8%	55.3%	75.5%	77.3%	75.1%
oh0	82.8%	84.4%	87.3%	**89.8%**	74.5%	72.8%	83.9%	81.9%	89.1%
oh5	79.6%	85.6%	88.4%	**92.0%**	76.5%	76.7%	87.0%	86.8%	87.1%
oh10	73.1%	77.5%	79.6%	**82.6%**	70.9%	65.5%	79.4%	77.7%	81.2%
oh15	75.2%	81.7%	84.6%	**86.4%**	67.5%	64.9%	81.3%	80.7%	84.0%
re0	75.8%	77.9%	82.3%	**85.0%**	69.1%	64.4%	79.5%	76.3%	81.1%
re1	77.9%	78.9%	80.0%	**81.6%**	59.8%	60.6%	71.2%	75.4%	80.5%
tr11	78.2%	85.3%	87.0%	**88.9%**	79.3%	80.5%	81.3%	83.0%	85.3%
tr12	79.2%	85.7%	88.4%	**89.0%**	76.2%	72.5%	80.8%	82.7%	79.8%
tr21	81.3%	89.1%	**90.3%**	90.0%	74.6%	73.1%	87.6%	88.5%	59.6%
tr31	93.3%	93.9%	94.7%	**96.9%**	90.2%	87.5%	93.0%	92.3%	94.1%
tr41	89.6%	93.5%	95.3%	**95.9%**	89.9%	87.3%	93.4%	92.4%	94.5%
tr45	91.3%	91.1%	92.9%	**93.6%**	80.3%	80.9%	91.1%	92.1%	84.7%
la1	75.2%	82.7%	85.7%	**87.6%**	76.1%	74.2%	83.4%	82.1%	**87.6%**
la2	77.3%	84.1%	87.2%	88.6%	78.2%	76.1%	85.9%	84.7%	**89.9%**
fbis	73.6%	78.0%	81.3%	**84.1%**	59.7%	56.0%	76.4%	76.3%	77.9%
wap	68.1%	75.1%	77.5%	**82.9%**	62.3%	60.2%	74.3%	76.1%	80.6%
ohscal	71.5%	62.5%	73.5%	**77.8%**	59.4%	57.5%	70.9%	69.6%	74.6%
new3	72.7%	67.9%	73.1%	**77.2%**	41.1%	43.5%	53.9%	63.1%	74.4%

Table 4: The classification accuracy of the original and reduced dimensional data sets.

Query Retrieval

- Found the *k* nearest neighbors for each document *d* in original and reduced space (LSI and CI)
- Counted the number of neighbors belonging to the same class as *d*
- Summed up counts over all documents in each class
- Compared retrieval improvements: ratio of recalled documents in reduced space to recalled documents in original space

Results: Query Retrieval

- CI improved retrieval and outperformed LSI in all classes
- CI performed well regardless of class size; LSI did worse with smaller classes

Results: Query Retrieval

re0		
Size	CI	LSI
608	1.12	1.01
319	1.31	1.11
219	1.28	1.12
80	1.89	1.30
60	1.26	0.99
42	2.17	1.14
39	1.30	1.14
38	1.38	0.82
37	1.66	1.16
20	1.54	1.06
16	1.60	1.00
15	1.32	0.76
11	1.64	0.73

re1		
Size	CI	LSI
371	1.25	1.05
330	1.18	1.06
137	1.51	1.24
106	1.23	1.13
99	1.11	1.04
87	1.11	1.04
60	1.44	1.14
50	1.94	0.90
48	1.05	0.99
42	2.13	1.01
37	1.59	1.22
32	1.33	1.19
31	1.67	1.23
31	1.72	1.26
27	1.84	1.30
20	2.01	1.06
20	1.41	1.27
19	1.81	0.93
19	2.18	0.80
18	1.69	0.97
18	3.67	1.09
17	1.49	0.83
15	3.75	0.98
13	1.40	0.80
10	2.27	0.43

fbis		
Size	CI	LSI
506	1.07	1.03
387	1.02	0.99
358	1.31	1.14
190	1.07	0.99
139	1.17	1.04
125	1.32	1.15
121	1.17	1.09
119	1.03	0.99
94	1.33	1.20
92	1.44	1.09
65	1.40	1.04
48	1.80	1.29
46	1.80	1.14
46	1.09	1.06
46	1.73	0.97
43	2.26	0.91
38	2.68	0.94

wap		
Size	CI	LSI
341	1.05	1.04
196	1.72	1.32
168	1.31	0.94
130	1.42	1.03
97	1.17	1.09
91	1.75	1.29
91	1.94	1.74
76	1.37	1.14
65	1.22	0.99
54	1.71	1.09
44	3.81	1.34
40	1.14	0.88
37	2.36	1.27
35	2.98	1.52
33	2.83	1.10
18	3.63	0.52
15	3.49	0.76
13	2.57	0.87
11	2.66	1.02
5	2.78	0.78

new3		
Size	CI	LSI
696	1.13	1.05
568	1.03	0.98
493	1.87	1.24
369	1.31	1.11
330	1.09	1.03
328	1.49	1.08
326	1.24	1.09
306	1.08	1.05
281	1.18	1.05
278	1.16	1.06
276	1.07	1.03
270	1.23	1.14
253	1.63	1.29
243	1.07	1.04
238	1.35	1.08
218	1.24	1.11
211	1.17	1.02
198	1.85	1.38
196	1.20	1.14
187	1.34	1.16
181	1.39	1.23
179	1.14	1.02
174	1.84	0.99
171	1.92	1.35
171	1.09	1.00
161	1.19	1.11
159	1.41	1.19
153	1.25	1.02
141	1.69	1.16
139	1.25	1.10
139	1.27	1.11
136	1.19	1.08
130	1.29	1.22
126	1.66	1.08
124	1.06	1.03
123	1.23	1.16
120	1.03	0.97
116	1.53	0.92
115	1.18	1.03
110	1.18	1.08
110	1.11	1.07
106	1.04	1.02
105	1.28	1.16
104	2.54	1.17

la1		
Size	CI	LSI
943	1.33	1.12
738	1.11	1.07
555	1.21	1.11
354	1.34	1.25
341	1.41	1.14
273	2.22	1.08

la2		
Size	CI	LSI
905	1.31	1.13
759	1.10	1.06
487	1.25	1.13
375	1.20	1.15
301	1.48	1.14
248	1.75	1.09

ohscal		
Size	CI	LSI
1621	1.38	1.24
1460	1.56	1.37
1297	1.37	1.19
1260	1.46	1.29
1159	1.63	1.41
1037	1.81	1.39
1001	1.85	1.53
864	1.47	1.33
764	1.78	1.35
709	1.51	1.28

Table 5: The per-class RI measures for various data sets for supervised dimensionality reduction. The first column shows the number of documents in each class.

Conclusions

- CI provides a method of dimensionality reduction which performs better in classification accuracy than traditional methods
- CI also improves recall performance
- For more information:
  - www.cs.umn.edu/~karypis
  - Karypis has a longer paper on CI:

www-users.cs.umn.edu/~karypis/publications/Papers/PDF/ci.pdf