



Concept Indexing

Improving Document Retrieval & Categorization

George Karypis & Eui-Hong (Sam) Han
Department of Computer Science & Engineering
University of Minnesota

<http://www.cs.umn.edu/~karypis>



Too Many Documents!

- There has been a tremendous growth in the volume of online text documents.
 - WWW, Intranets, Digital Libraries, *etc.*
- Documents are and will remain the dominant data type stored online.
- It is critical to provide tools that will allow us to move from *data* to *information*.
 - We need better tools to
 - Retrieve, Organize, Navigate, & Personalize



Vector-Space Model (TF-IDF)

- Each document is considered to be a vector in the term-space.

$$\vec{d} = (tf_1, tf_2, \dots, tf_n) \quad \text{where } tf_i \text{ is the frequency of the } i^{\text{th}} \text{ term}$$

- The weight of each term is scaled according to the *inverse document frequency* (IDF).
- The documents are scaled so that they have unit length, *i.e.*, $\|\vec{d}\|_2 = 1$.
- The similarity between two documents is computed using the *cosine* function:

$$sim_{\cos}(\vec{d}_i, \vec{d}_j) = \frac{\vec{d}_i \bullet \vec{d}_j}{\|\vec{d}_i\|_2 \|\vec{d}_j\|_2} = \vec{d}_i \bullet \vec{d}_j \quad \text{where } "\bullet" \text{ is the dot-product operator.}$$



Information Retrieval Problem Definition

- Given a set of *query terms*, we will like to retrieve the most relevant set of documents that discuss the *concept* encapsulated in the query terms.
- Traditional approaches rely on lexical matching.
 - This approach leads to sub-optimal results due to *synonymy* and *polysemy*.
 - They fail to retrieve documents that discuss the desired concept using synonym terms.
 - The retrieve documents that contain the query terms but in an entirely different context.
- Dimensionality reduction techniques have been shown to address some of the problems.
 - Latent Semantic Indexing (LSI) is such a technique.



Dimensionality Reduction Latent Semantic Indexing

- Computes a truncated SVD of the document-term matrix, and uses the singular vectors as the axes of the lower dimensional space.
 - The goal is that in the lower dimensional space, documents that describe the similar concepts should be brought closer together.
- Experimental evaluation of LSI has shown that it tends to improve the retrieval quality.
- The disadvantages of LSI are
 - Relatively high computational cost.
 - It cannot perform supervised dimensionality reduction.
 - It ignores any pre-existing information about class-membership.



Concept Indexing (CI)

- The idea behind CI is to first find the concepts in a document collection, and then represent each document as a function of these concepts.
 - How each concept is represented?
 - How each document is represented as a function of the concepts?
 - How do we find the concepts?



How each Concept is Represented?

- Each concept can be described by a relatively small number of terms.
 - *i.e.*, the *keywords* describing the concept.
- Each concept is represented by a vector in the term space, \vec{C}_i .
 - Each concept vector is of unit length.
 - The various terms of the concept can have different weights (*i.e.*, importance).



How each Document is Represented as a Function of the Concepts?

- Given a set of concept vectors $\vec{C}_1, \vec{C}_2, \dots, \vec{C}_k$ a document \vec{d}_i is represented by the vector

$$\begin{aligned} d_i^{CI} &= (\text{sim}(\vec{d}_i, \vec{C}_1), \text{sim}(\vec{d}_i, \vec{C}_2), \dots, \text{sim}(\vec{d}_i, \vec{C}_k)) \\ &= (\vec{d}_i \bullet \vec{C}_1, \vec{d}_i \bullet \vec{C}_2, \dots, \vec{d}_i \bullet \vec{C}_k) \end{aligned}$$

- In general k is much smaller than the original number of dimensions.
 - Dimensionality reduction



How do we Find the Concepts?

Supervised Setting

- The concepts correspond to the centroid vectors of each class:

$$\vec{C}_i = \frac{1}{|K_i|} \sum_{d_j \in K_i} \vec{d}_j$$

where K_i is the set of documents belonging to the i^{th} class

- The centroid vectors provide an effective mechanism by which to summarize the concept represented in each class.
- Finer concepts can be constructed by further clustering the documents of each class.



Examples of Centroid Vectors Supervised

	wap															
1	0.20	diana	0.17	film	0.13	showbiz	0.13	notabl	0.13	angel	0.13	annual	0.12	albert	0.12	lo
2	0.26	emmi	0.23	cb	0.22	tv	0.21	rate	0.21	nbc	0.20	adult	0.16	abc	0.14	household
3	0.19	studi	0.19	research	0.19	cell	0.18	risk	0.18	cancer	0.16	patient	0.15	diseas	0.14	women
4	0.41	newspap	0.22	editor	0.19	advertis	0.14	media	0.13	peruvian	0.13	coverag	0.12	percent	0.12	journalist
5	0.25	exhibit	0.21	auction	0.21	stolen	0.20	art	0.18	gogh	0.16	draw	0.16	sculptor	0.15	paint
6	0.38	film	0.19	box	0.16	million	0.15	star	0.14	offic	0.13	weekend	0.13	festiv	0.13	pictur
7	0.33	stock	0.21	dow	0.18	compani	0.17	percent	0.14	greenspan	0.14	industri	0.14	busi	0.14	financi
8	0.49	cable	0.21	network	0.15	fcc	0.15	rate	0.14	usa	0.13	showtim	0.13	hbo	0.12	espn
9	0.34	week	0.34	bestsell	0.26	weekli	0.25	publish	0.22	hardcov	0.19	paperback	0.19	book	0.13	nea
10	0.29	album	0.28	music	0.23	record	0.23	song	0.14	band	0.13	concert	0.12	sold	0.12	rock
11	0.39	clinton	0.27	senat	0.27	house	0.24	white	0.23	campaign	0.20	reform	0.19	republican	0.15	financ
12	0.27	game	0.17	smith	0.15	coach	0.14	season	0.13	win	0.13	championship	0.12	se	0.11	nomo
13	0.14	charact	0.13	film	0.11	david	0.11	music	0.11	product	0.10	review	0.09	michael	0.09	sound
14	0.33	internet	0.25	microsoft	0.22	comput	0.19	zdnet	0.19	wir	0.15	access	0.15	servic	0.15	reserv
15	0.37	ticket	0.28	hottest	0.28	opera	0.24	theater	0.19	broadwai	0.19	receipt	0.16	lyric	0.13	week
16	0.36	casino	0.34	farm	0.27	legion	0.20	trump	0.20	mirag	0.18	miami	0.18	aid	0.16	concert
17	0.43	internet	0.35	onlin	0.24	comput	0.18	servic	0.17	microsoft	0.16	web	0.14	america	0.13	compuserv
18	0.28	murdoch	0.16	disnei	0.15	compani	0.15	stock	0.15	usa	0.13	network	0.13	viacom	0.12	million
19	0.28	daili	0.22	hollywood	0.21	insid	0.20	front	0.18	fox	0.17	tv	0.16	film	0.14	ink
20	0.48	dvd	0.24	game	0.23	player	0.21	toshiba	0.15	emeri	0.13	typ	0.12	video	0.11	digit

Examples of Centroid Vectors Supervised

	new3															
1	0.34	waste	0.29	dump	0.26	water	0.26	pollution	0.23	sea	0.22	environment	0.20	river	0.18	radioact
2	0.44	export	0.37	cocom	0.22	russian	0.18	control	0.18	technologi	0.16	russia	0.13	missil	0.12	german
3	0.52	japan	0.35	japanes	0.23	tokyo	0.18	trade	0.14	insur	0.14	talk	0.13	kyodo	0.12	market
4	0.41	nuclear	0.41	korea	0.31	north	0.30	iaea	0.25	korean	0.18	dprk	0.17	inspect	0.14	pyongyang
5	0.41	al	0.28	palestinia	0.24	israe	0.20	arab	0.20	lebanon	0.19	hizballah	0.17	israel	0.15	abu
6	0.34	grain	0.32	agricultur	0.20	price	0.19	rice	0.18	product	0.16	percent	0.14	farm	0.14	market
7	0.37	newspap	0.26	publish	0.23	press	0.17	media	0.16	public	0.15	editor	0.13	russian	0.12	magazin
8	0.29	murder	0.18	al	0.16	kill	0.14	polic	0.12	terrorist	0.11	assassin	0.11	crime	0.10	court
9	0.52	nuclear	0.26	ukrain	0.21	korea	0.20	iaea	0.19	treati	0.16	north	0.16	dprk	0.14	weapon
10	0.55	drug	0.24	traffick	0.23	gang	0.23	polic	0.20	heroin	0.17	arrest	0.16	narcot	0.16	kg
11	0.49	nafta	0.40	mexico	0.24	job	0.23	mexicar	0.17	american	0.15	trade	0.15	worker	0.13	export
12	0.60	violenc	0.40	women	0.26	domest	0.17	crime	0.16	abus	0.15	speaker	0.15	victim	0.14	batter
13	0.33	china	0.23	trade	0.22	embargo	0.22	mfn	0.18	clinton	0.16	right	0.16	vietnam	0.14	human
14	0.56	earthquak	0.24	quake	0.22	insur	0.21	disast	0.15	california	0.15	volcano	0.14	dollar	0.12	reinsur
15	0.48	submarin	0.32	rosyth	0.26	trident	0.23	devonpc	0.21	defenc	0.19	nuclear	0.18	dockyard	0.16	refit
16	0.44	pulp	0.41	paper	0.30	price	0.24	cent	0.22	mill	0.17	newsprint	0.13	compani	0.13	cdollar
17	0.61	tax	0.29	pound	0.28	cent	0.22	vate	0.19	incom	0.18	rate	0.12	taxe	0.10	taxat
18	0.44	drug	0.30	traffick	0.28	cocain	0.26	cartel	0.17	colombian	0.16	colombia	0.15	cali	0.14	polic
19	0.36	speci	0.25	whale	0.23	endang	0.23	wolve	0.22	wildlif	0.17	hyph	0.17	blank	0.16	mammal
20	0.30	rwanda	0.25	rebel	0.24	africa	0.17	kill	0.17	hutu	0.17	kigali	0.16	unita	0.16	tutsi
21	0.38	project	0.31	dam	0.24	hydroelec	0.21	power	0.19	hyph	0.18	electr	0.15	gorge	0.15	hydropow
22	0.53	vw	0.36	lopez	0.29	gm	0.24	opel	0.21	volkswager	0.21	piech	0.19	motor	0.14	espionag
23	0.14	hous	0.13	pound	0.13	properti	0.13	home	0.12	liv	0.12	house	0.12	retir	0.12	life
24	0.35	fuel	0.32	energi	0.31	plutonium	0.27	nuclear	0.24	reactor	0.19	electr	0.17	power	0.14	coal
25	0.54	women	0.23	parti	0.22	elect	0.21	labour	0.16	vote	0.16	parliam	0.15	candid	0.14	mp
26	0.56	argentina	0.33	argentin	0.31	falkland	0.23	bueno	0.23	aire	0.17	tella	0.17	malvina	0.16	british
27	0.59	bank	0.25	imf	0.23	world	0.15	lend	0.12	develop	0.11	loan	0.11	project	0.11	monetari
28	0.29	tax	0.23	helmslei	0.22	hunter	0.18	ir	0.18	evasion	0.17	fraud	0.15	dominelli	0.14	rose
29	0.39	polic	0.30	kill	0.23	policema	0.21	offic	0.14	policemen	0.13	murder	0.13	milit	0.12	shot
30	0.42	school	0.38	educ	0.38	curriculum	0.32	teacher	0.26	test	0.19	patten	0.19	pupil	0.13	teach



How do we Find the Concepts?

Unsupervised Setting

- The documents are grouped into k groups.
 - For each cluster, the centroid vectors are used as the concepts.
- We used a recursive bisection *seed* based document clustering, based on *k-means*.
 - This class of algorithms have been shown to be very effective for document clustering.
- At each recursive step, we split the cluster with the highest aggregate dissimilarity.
 - This can be easily measured by looking at the length of the centroid vectors of the clusters.

$$\text{Aggregate Dissimilarity} = |S_i|^2 \left(1 - \|\vec{C}_i\|_2^2 \right)$$

Examples of Centroid Vectors

Unsupervised

	re1															
1	0.65	corn	0.20	acre	0.19	bushel	0.18	soybean	0.17	usda	0.17	unknown	0.16	ussr	0.16	tonne
2	0.46	ga	0.24	oil	0.22	cubic	0.21	reserv	0.20	barrel	0.20	feet	0.19	natur	0.15	drill
3	0.65	coffee	0.28	quota	0.27	ico	0.17	bag	0.16	export	0.16	brazil	0.14	colombia	0.14	meet
4	0.45	tonne	0.35	palm	0.20	import	0.18	oil	0.15	januari	0.14	rapese	0.14	beef	0.14	februari
5	0.35	copper	0.30	steel	0.20	ct	0.19	aluminium	0.16	cent	0.15	smelter	0.14	pound	0.14	lb
6	0.32	crop	0.24	grain	0.20	wheate	0.19	cotton	0.19	mln	0.19	weather	0.16	china	0.16	rain
7	0.45	bble	0.39	crude	0.31	post	0.26	ct	0.22	dlr	0.21	wti	0.20	raise	0.16	distill
8	0.45	dollar	0.28	bank	0.24	portland	0.23	yen	0.17	load	0.16	juice	0.16	ship	0.14	japan
9	0.73	sugar	0.22	tonne	0.22	white	0.15	trader	0.14	intervent	0.14	ec	0.13	tender	0.12	ecu
10	0.59	gold	0.35	ounce	0.33	ton	0.30	mine	0.14	ore	0.12	feet	0.12	silver	0.10	assai
11	0.49	ec	0.34	maize	0.24	tax	0.20	tonne	0.17	european	0.17	licenc	0.17	ecu	0.16	commiss
12	0.30	wheate	0.27	soviet	0.22	farm	0.22	lyng	0.21	bill	0.19	offer	0.18	grain	0.15	agricultur
13	0.39	cocoa	0.35	buffer	0.26	deleg	0.24	rubber	0.22	stock	0.22	icco	0.17	pact	0.17	consum
14	0.32	ship	0.24	gulf	0.22	tanker	0.22	iran	0.21	missil	0.18	vessel	0.15	attack	0.14	iranian
15	0.43	oil	0.29	tax	0.18	herrington	0.17	explor	0.16	energi	0.15	import	0.12	reagan	0.12	studi
16	0.28	credit	0.28	wheate	0.25	ccc	0.24	depart	0.22	nil	0.19	sale	0.18	commod	0.18	guarante
17	0.43	ecuador	0.27	bpd	0.27	refineri	0.25	crude	0.25	oil	0.21	pipelin	0.20	venezuela	0.13	mln
18	0.43	wheate	0.42	tonne	0.24	tender	0.24	barlei	0.22	taiwan	0.18	shipment	0.15	soft	0.14	export
19	0.48	strike	0.28	seamen	0.28	union	0.25	port	0.22	worker	0.14	employ	0.13	ship	0.12	pai
20	0.49	opec	0.31	saudi	0.27	oil	0.25	bpd	0.24	barrel	0.18	mln	0.17	price	0.15	arabia

Examples of Centroid Vectors

Unsupervised

	new3															
1	0.25	russian	0.19	russia	0.18	rwanda	0.17	moscow	0.14	soviet	0.14	rebel	0.13	nato	0.13	un
2	0.41	vw	0.30	lopez	0.24	iraq	0.23	gm	0.20	matrix	0.19	opel	0.18	inquiri	0.18	churchill
3	0.15	econom	0.15	export	0.14	percent	0.12	enterpris	0.12	russian	0.11	reform	0.11	product	0.11	economi
4	0.26	tunnel	0.19	rail	0.16	argentina	0.15	school	0.14	curriculum	0.14	eurotunnel	0.14	british	0.14	pound
5	0.39	hyph	0.29	food	0.22	blank	0.19	label	0.16	fda	0.14	fsi	0.14	speci	0.14	poultri
6	0.71	drug	0.21	patient	0.16	azt	0.14	aid	0.14	fda	0.12	addict	0.10	epo	0.09	treatment
7	0.46	korea	0.33	north	0.32	nuclear	0.31	iaea	0.28	korean	0.21	dprk	0.18	inspect	0.16	pyongyang
8	0.52	tax	0.28	bank	0.24	cent	0.23	pound	0.17	incom	0.16	vate	0.15	rate	0.12	taxe
9	0.28	japan	0.25	vietnam	0.24	china	0.23	trade	0.22	rice	0.19	japanes	0.17	gat	0.15	tokyo
10	0.59	women	0.47	violenc	0.19	domest	0.15	crime	0.13	speaker	0.12	victim	0.12	abus	0.10	batter
11	0.26	helmsle	0.24	hunter	0.20	tax	0.18	fraud	0.17	evasion	0.16	dominelli	0.15	rose	0.15	sentenc
12	0.38	al	0.24	palestinian	0.23	arab	0.22	israe	0.18	israel	0.17	islam	0.16	lebanon	0.14	kill
13	0.35	cent	0.24	compani	0.21	dollar	0.18	pound	0.16	pharmaceu	0.16	price	0.14	pulp	0.13	paper
14	0.43	kong	0.43	hong	0.22	chines	0.21	china	0.20	beij	0.18	journalist	0.16	taiwan	0.15	yang
15	0.47	grain	0.34	agricultur	0.23	price	0.19	rural	0.18	product	0.17	percent	0.15	yuan	0.15	farm
16	0.62	nuclear	0.30	pakistan	0.23	india	0.18	weapon	0.17	ukrain	0.15	plutonium	0.12	treati	0.12	prolifer
17	0.38	nafta	0.33	mexico	0.17	mexican	0.17	speaker	0.16	american	0.16	trade	0.16	gentleman	0.16	job
18	0.24	polic	0.17	kill	0.16	anc	0.15	murder	0.14	africa	0.11	offic	0.10	death	0.10	journalist
19	0.47	drug	0.34	traffick	0.25	cocain	0.20	cartel	0.20	colombia	0.17	colombian	0.17	polic	0.13	arrest
20	0.24	water	0.24	forest	0.22	environment	0.21	river	0.21	project	0.16	pollution	0.16	amazon	0.14	power



Experimental Evaluation

- The performance of the lower dimensional space computed by CI was evaluated by measuring how closely it brings together documents of the same class.
 - For each document we found the 20 nearest documents and we computed the fraction of these documents that belong to the same class.
 - This essentially measures the *precision* of a query in which the document itself acts as the query.
 - “Find documents similar to that”
 - These fractions were averaged over all the documents and they were compared with respect to the same average fractions in the original space.

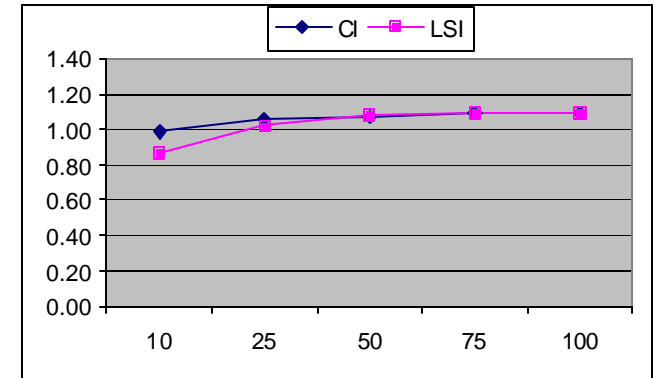
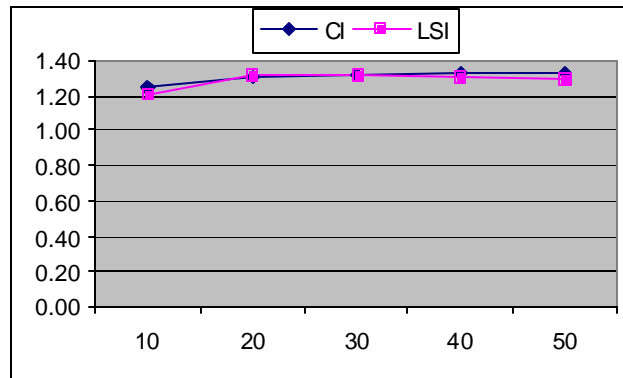
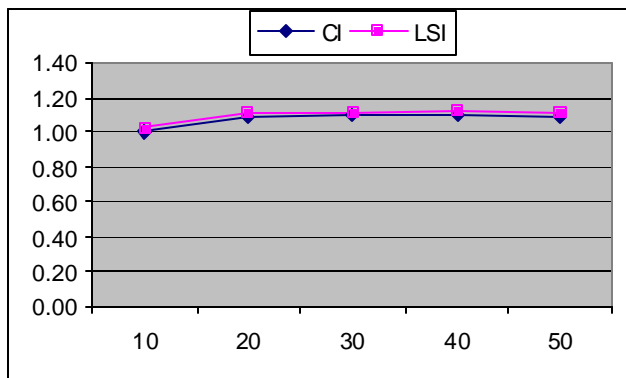
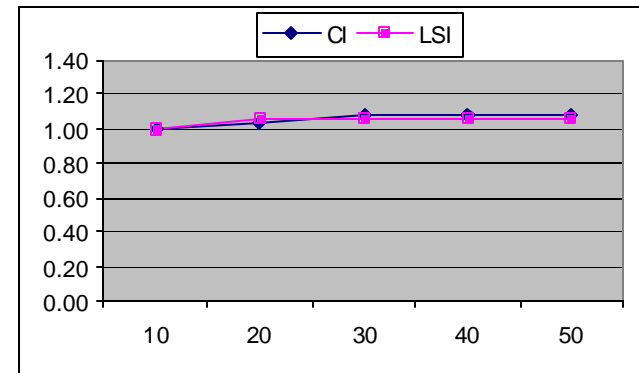
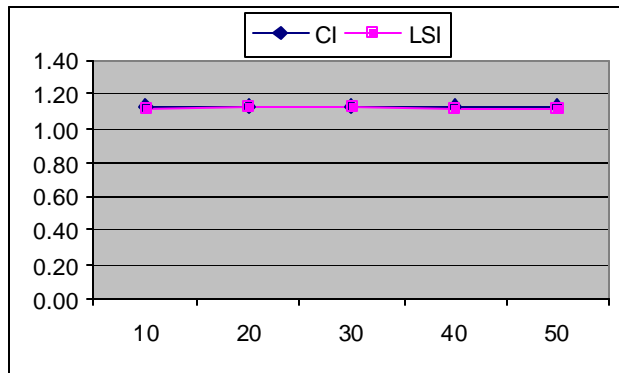
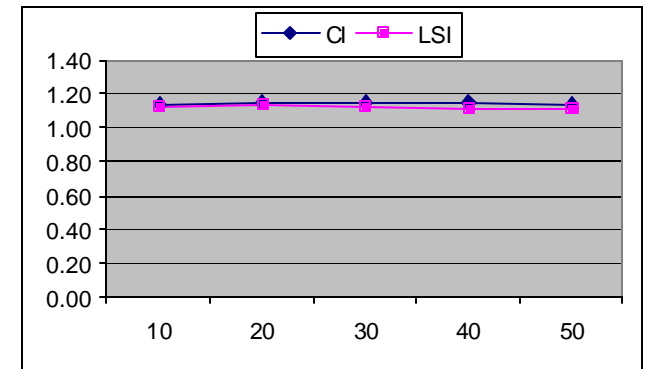
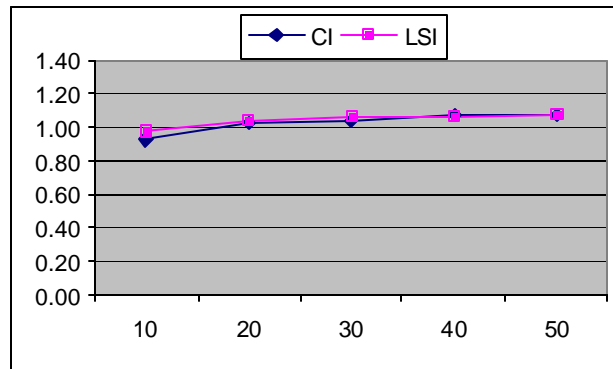
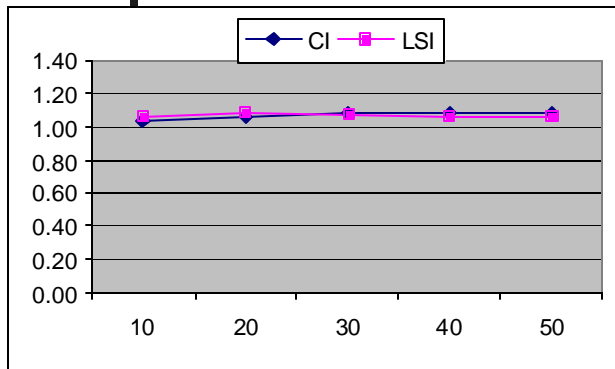


Test Data Sets

- **West Publishing Documents**
 - 3 data sets out of 149,655 documents
 - class labels include "sales", "counties", and "insurance"
- **Text Retrieval Conference (TREC) Documents**
 - documents from several sources
 - 8 data sets based on queries used in the conference
- **OHSUMED**
 - subset of MEDLINE database from 270 medical journals
 - title plus abstract of an article as a document
 - 5 data sets selected based on the mix of categories
- **Reuters-21578**
 - categories cover topics such as commodities, interest rates, and foreign exchange
 - 2 data sets based on the mix of categories
- **LA Times**
 - Categories correspond to the sections of the newspaper (*e.g.*, national, metro, financial, *etc.*).
 - 3 data sets were generated
- **WAP**
 - web document collected according to *Yahoo!* Subject hierarchy for WebACE project
 - categories include topics such as sports, entertainment, health, and politics

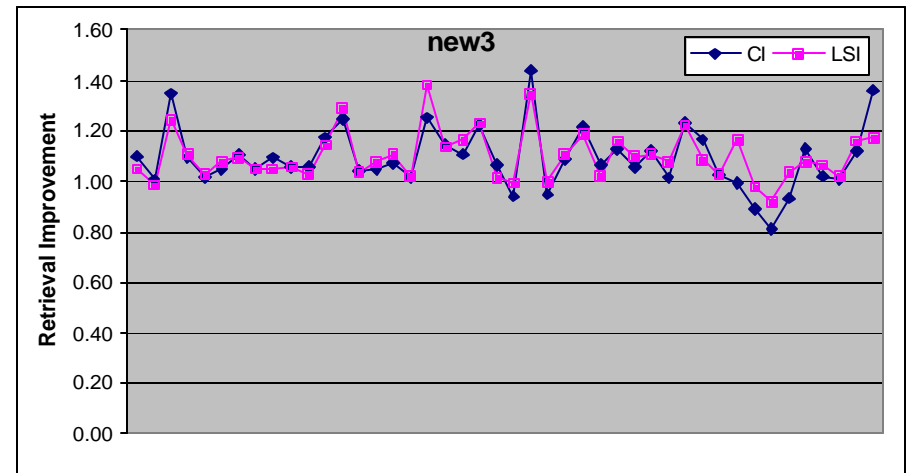
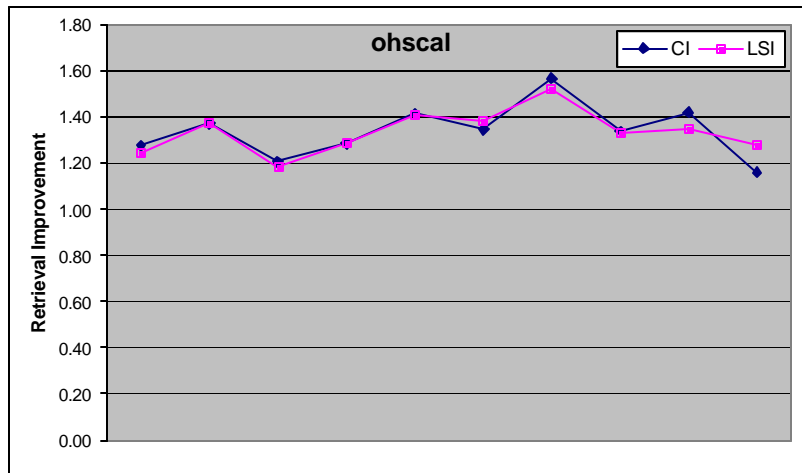
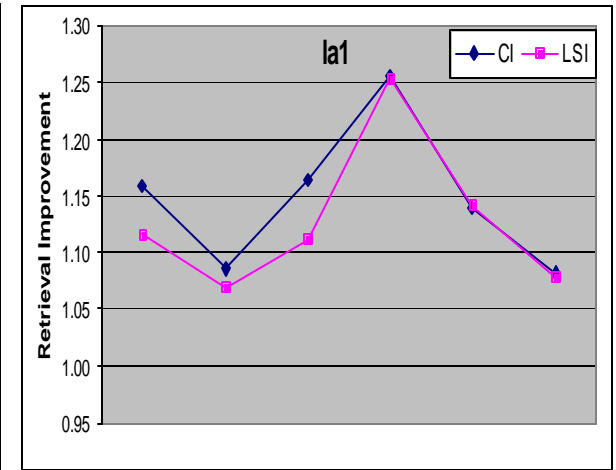
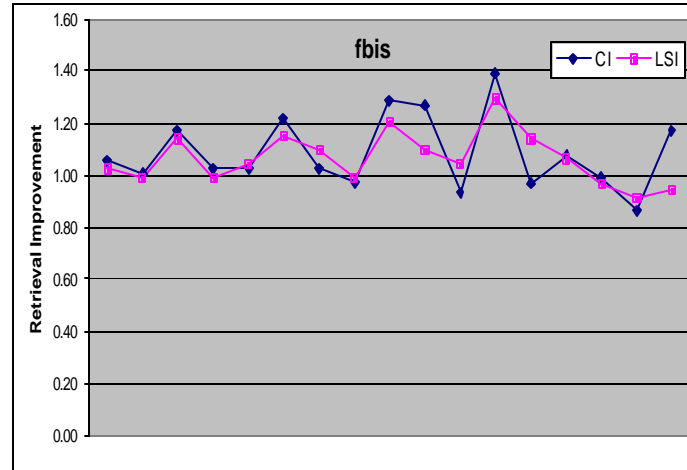
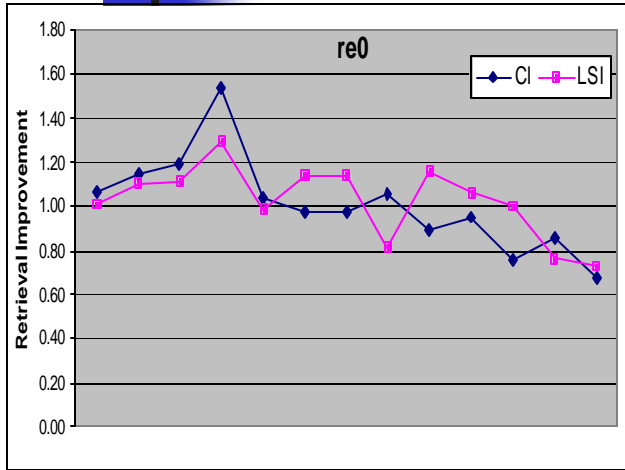
CI & LSI versus Original Space

Unsupervised



CI & LSI versus Original Space

Per-class Retrieval Improvement





CI vs LSI

Runtime (seconds)

	<i>re0</i>	<i>re1</i>	<i>la1</i>	<i>la2</i>	<i>fbis</i>	<i>wap</i>	<i>ohscal</i>	<i>new3</i>
CI	0.56	0.72	5.01	4.59	3.17	1.97	7.01	29.85
LSI	6.58	7.00	44.20	39.80	20.10	18.10	65.10	275.00

All experiments were performed on an 500Mhz Pentium II

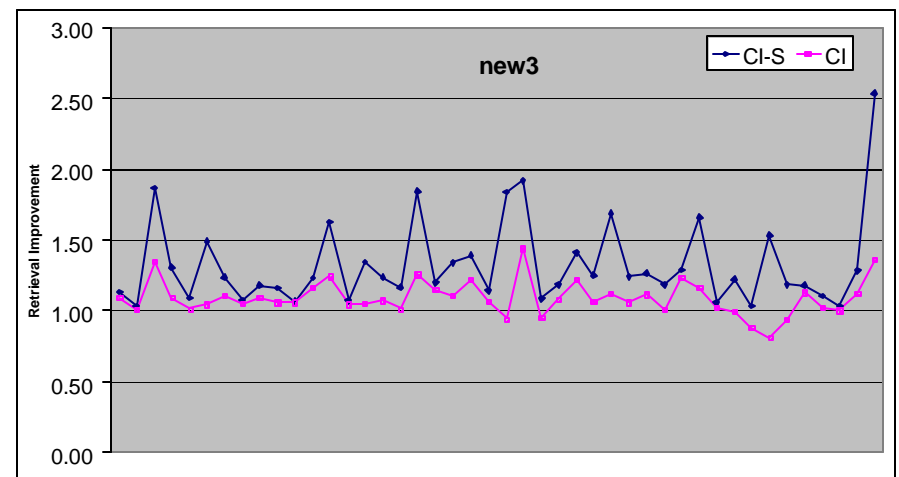
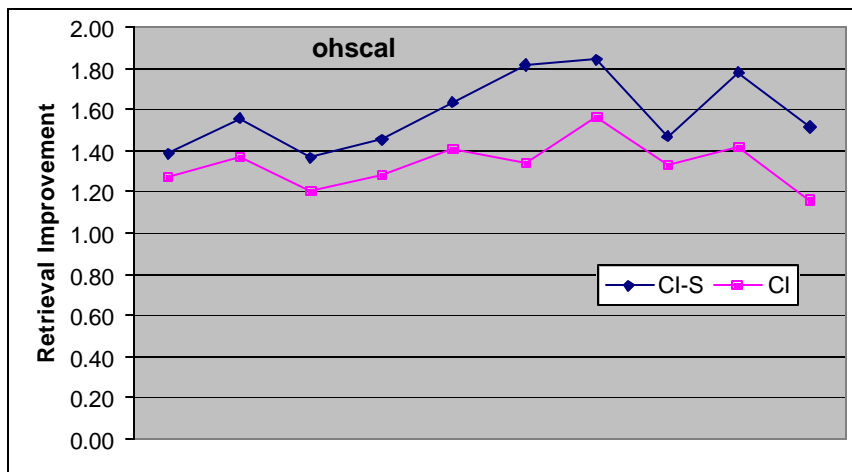
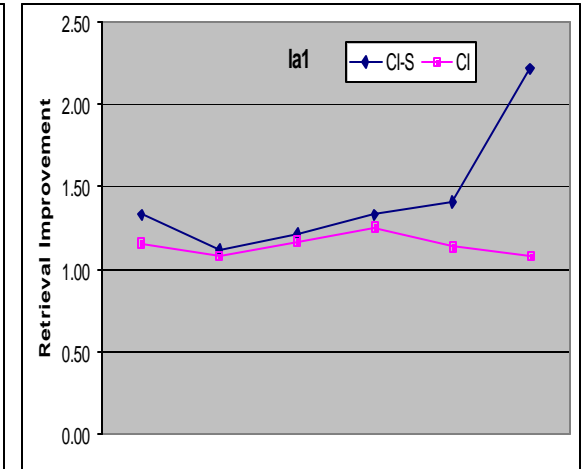
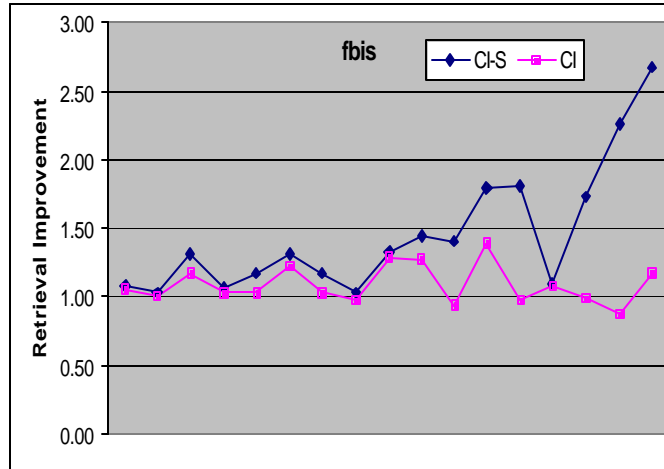
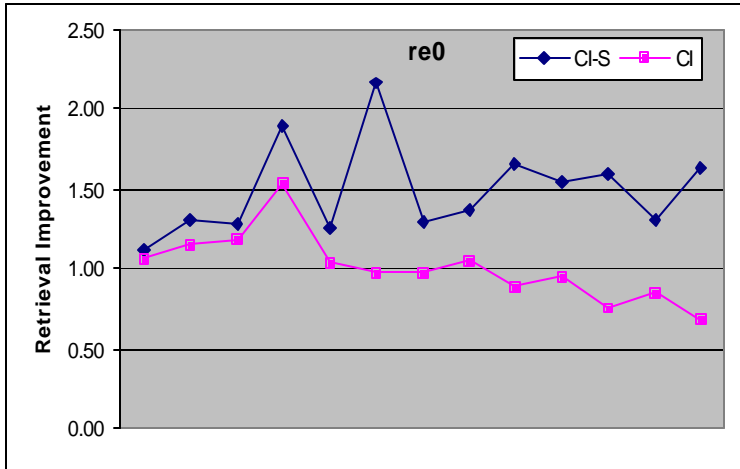
CI versus Original Space

Supervised, Per-Class Improvement

re0		re1		wap		fbis		la2	
Size	CI-S	Size	CI-S	Size	CI-S	Size	CI-S	Size	CI-S
608	1.12	371	1.25	341	1.05	506	1.07	905	1.31
319	1.31	330	1.18	196	1.72	387	1.02	759	1.10
219	1.28	137	1.51	168	1.31	358	1.31	487	1.25
80	1.89	106	1.23	130	1.42	190	1.07	375	1.20
60	1.26	99	1.11	97	1.17	139	1.17	301	1.48
42	2.17	87	1.11	91	1.75	125	1.32	248	1.75
39	1.30	60	1.44	91	1.94	121	1.17		
38	1.38	50	1.94	76	1.37	119	1.03	ohscal	
37	1.66	48	1.05	65	1.22	94	1.33	Size	CI-S
20	1.54	42	2.13	54	1.71	92	1.44	1621	1.38
16	1.60	37	1.59	44	3.81	65	1.40	1450	1.56
15	1.32	32	1.33	40	1.14	48	1.80	1297	1.37
11	1.64	31	1.67	37	2.36	46	1.80	1260	1.46
		31	1.72	35	2.98	46	1.09	1159	1.63
		27	1.84	33	2.83	46	1.73	1037	1.81
la1		20	2.01	18	3.63	43	2.26	1001	1.85
Size	CI-S	20	1.41	15	3.49	38	2.68	864	1.47
943	1.33	19	1.81	13	2.57			764	1.78
738	1.11	19	2.18	11	2.66			709	1.51
555	1.21	18	1.69	5	2.78				
354	1.34	18	3.67						
341	1.41	17	1.49						
273	2.22	15	3.75						
		13	1.40						
		10	2.27						

CI versus Original Space

Supervised, Per-Class Improvement





Improving the Classification of Traditional Algorithms

- The supervised dimensionality reduction performed by CI can improve the performance of classification algorithms when operating on the lower dimensional space.
 - Both training set is used to determine the lower dimensional space.
 - Both the training and the test set are projected into this space.
 - The classifier operates on the reduced space only.
- The same holds for LSI, however the gains may not be significant, as it cannot reduce the dimensionality in a supervised setting.

Classification Performance in the Lower Dimensional Space

					LSI Reduced Space				NB
	Original Space		CI Reduced Space		C4.5		kNN		
	C4.5	kNN	C4.5	kNN	25 Dims	50 Dims	25 Dims	50 Dims	
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%

C4.5 improves by 7%

kNN improves by 6%

kNN in the CI space outperforms Naive Bayesian by 5%!



Conclusions & Future Research

- CI is a fast algorithm for computing lower dimensional representations of document data sets.
 - It can perform supervised dimensionality reduction.
 - It can take advantage of topic hierarchies.
 - Alternate methods of finding the *concepts* can be developed.
 - If the number of concepts is large, LSI can be used to reduced the dimensionality of the CI space.