An Architecture for Logging Text and Searching Chat Messages

Master's Thesis Defense Rajan Vijayaraghavan 04.28.2003

Committee: Dr. Susan Gauch (Chair) Dr. Arvin Agah Dr. Joseph Evans



I T T C



Organization

- Motivation
- > Objectives
- > Architecture
- Testing
- Evaluation
- Future Work



Motivation

- Instant Messaging
 - Chat Rooms
 - Instant Messengers
- Growing Popularity
 - People of nearly different age groups participate
 - > All conceivable topics discussed
- Reach of Instant Messenger Services
 - Corporate users
 - Home users
 - US Marines



Motivation (cont'd)

- > So what is the problem ?
 - Parental Controls
 - Need for message archiving
 - Search Feature
 - Scalability Issues



Related Works

- Chat Track Project
- Commercial Systems (Archiving capability)
 - Iambigbrother (Keyword search)
 - Net Nanny
 - Cyber-Scoop
 - Desktop Snooper
 - I-Spy Now

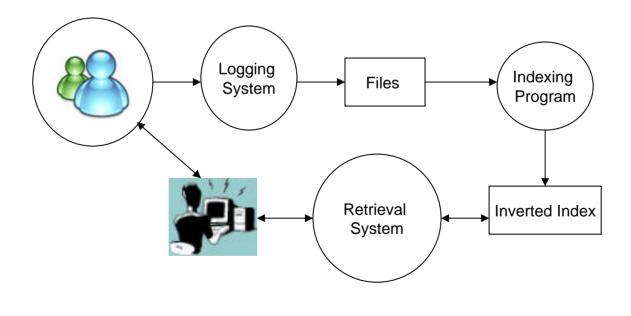


Objective

- To create a system that can log text from Microsoft MSN Messenger
- Index the text frequently
- Support variety of queries

Architecture

- Message Logging system for MSN Messenger
- Indexing System
- Retrieval System





Message Logging System

- Stand Alone Application
- Actions
 - Identify new/existing conversation window
 - Check for new message
 - Log both incoming and outgoing messages
 - Log the sign-in name of the speaker of the message
 - Keep track of entering/leaving users of a conversation window
 - Log administrative messages
 - Detect Window Closing



- Conversation window monitoring based on Window Handles
 - Each window will have a unique window handle
 - Any MSN messenger window will have " Instant Message" or
 - "- Conversation " in the title bar





- Logs messages approximately every 1 second
- > Writes message file, speaker id file and listener id file
- Adding messages is an incremental process

If (current message length > previous message length)

If Administrative Message

Log the event;

else if User Message

Log Message, Speaker Id, Listener Id(s), Add File name to "to be indexed list"; else

continue;

- Monitors administrative messages
 - Logs them as said by user 'none'
 - This helps in playing back the conversation as it happened



Example.,

Assume Window Handle 41157726; date 04.26.2003; time 15:17 PM

💒 The Real Slim Shady - Instant Message	
<u>File Edit V</u> iew <u>H</u> elp	
र्ट्टि Talk 💁 Invite 🙆 Block	
To: The Real Slim Shady <rharishnandan@hotmail.com></rharishnandan@hotmail.com>	
Never give out your password or credit card number in an instant message conversation. The Real Slim Shady says: there's a concept that works 20 million other white rappers emerge but no matter her many fish in the sea it'd be so empty without me	iow
	Send
Last message received on 4/26/03 at 3:17 PM.	1.



Message File : session000041157726/2003/04/26/151754866.mesg Speaker Id File : session000041157726/2003/04/26/151754866.uid Listener Id File :

session000041157726/2003/04/26/151754866.receiver

File Contents:

Message File: "there's a concept that works 20 million other white rappers emerge but no matter how many fish in the sea it'd be so empty without me"

Speaker Id File: "rharishnandan"

Listener Id File : "rharishnandan"

File to Index list: session000041157726 2003 04 26 151754866



Current system monitors 100 conversation windows

Practical number might be around 10 conversation windows

Calls the indexing program periodically

- If (Files to Index)
 - Call Indexer;

else

continue;



Indexing System

Indexing

- Creation of Inverted Index from the raw files.
 - Dictionary File
 - Contains one record for each unique word in the collection

Words, number of documents in which the word is present, total frequency of the word in the collection, inverse document frequency, pointer to postings record

- Postings File
 - Each record Information contains information about the word occurrence in a document.

Frequency of the word in a document, weight of the word in a document, document id for the document, pointer to next postings record.

Example

Word 'prozac' occurs in 100 documents and total number of documents 200 and occurs 120 times in total.

One dictionary record { 'prozac', 100,120, idf = $log_2(200/100)$, pointer to postings }

Postings File { 100 postings record (100 documents),

weight in document j = idfi * frequency of 'prozac' in document j;

document id = unique for a document;

pointer to next postings record; }



Indexing System (cont'd)

- Indexing
 - Batch Indexing
 - Incremental Indexing
- Creates Two sets of Inverted Index
 - One for Keyword (kyDictionary, kyPostings and kyDocuments)
 - One for Speaker Id (spDictionary, spPostings and spDocuments)



Retrieve Application

🛃 RetrieveDialog Application

Query	
Speaker Id	Search
Date Filter Year Month Date	Hour Minute
From	
Year Month Date	Hour Minute
Sort By Date	🗖 Speaker Name
Number of lines before/after match	01 💌
Show me Who heard	User Get
Number of Results Per Page	03 💌
Display Page	01 💌



- 0 ×



Scenario 1: "keyword" query

Query cricke				
Speaker Id			<u>Searc</u>	h)
Date Filter Year	Month	Date	Hour	Minute
From	Month	Date	Hour	
Year	Month	Date	Hour	Minute
To 🔽		-		
	□ <u>D</u> ate		i Speaker	Name
Display Preference Number of lines b		atch	01 💌	
Show me Who he	eard	1	🗖 User	Get
Number of Result	s Per Page	1	03 💌	
Display Page			01 💌	

ChatTrack Project	
Results for the que	ry : cricket onversations. Number of relevant conversation text found : 20
	s per page. Page number : 1 of 7
orspraying a resurt	s per page, rage number . I of 7
1 . Chat session id s	ession000096837408 on [03/25/2003 Time : 02:28]
	at he should be allowed to play. Even though he is not a role mode
	e respects, he would definitely attract children into the g
	e of cricket. We need more kids playing cricket and to improve the
	e cricket, so I think this is the right decision by the ECB.
	Irew D'cruz, London, UK
	time off - a season of village cricket should rekindle his
enthusiasm (and our	rs, his fans) and keep his a
2 . Chat session id s	ession000097716612 on [03/25/2003 Time : 02:30]
rajan_77 : a new	team to international cricket, we should have more patience to se
them having good re	
	Kenya is a new team as well, bu
	they are playing very good cricket. But Kenya has a very good pla
	ne cricket. They have a coach who has been working for them fo
	ry long time. Whereas Bangladesh changes their coach every now
and then. And it is (only recently that they have got a permanent cricket field.
3 . Chat session id s	ession000095959188 on [03/25/2003 Time : 02:26]
	h believes Warne will not find it easy when he is free to make his
comeback.	
	it of sport, particularly for a guy who loves pla
Pajan_77 :yin	
20 C	without cricket is going to be very difficult for him. But he loves
playing cricket, tha	
	and I hope he comes back." he Test series in India in Sectomber 2004 for his return
iWarne will not app	he Test series in India in September 2004 for his return. eal
Warne has gone	



Scenario 2: "key word + user name" query

Query Cr	cket			
Speaker Id	rajan_77		(<u>S</u> earc	h
Date Filter Year	Month	Date	Hour	Minute
From			Hoa	
Year	Month	Date	Hour	Minute
То		-		
Sort By	🗖 <u>D</u> ate		Speaker	Name
Display Prefere Number of lin	nces es before/after ma	atch [01 💌	
Show me Wł	io heard	r	User	Get
Number of R	esults Per Page	Į	03 💌	
Display Page		ſ	01 💌	



ChatTrack	Project Results Page.
Results fo	r the query : cricket
Searching	62040 conversations. Number of relevant conversation text found : 15
Searched	for user rajan_77
Displaying	3 results per page. Page number : 1 of 5
1. Chat see	sion id session000096837408 on [03/25/2003 Time : 02:28]
r_sothe	25h : k that he should be allowed to play. Even though he is not a role mode
	en in some respects, he would definitely attract children into the g
	_77 :ame of cricket. We need more kids playing cricket and to improve th
	of village cricket, so I think this is the right decision by the ECB.
	esh : Andrew D'cruz, London, UK
	eds some time off - a season of village cricket should rekindle his
entnusiasn	n (and ours, his fans) and keep his a
2 . Chat se	ssion id session000095959188 on [03/25/2003 Time : 02:26]
	esh : augh believes Warne will not find it easy when he is free to make his
comeback.	
	ng time out of sport, particularly for a guy who loves pla
	_77 :ying at the top level.
	? months without cricket is going to be very difficult for him. But he loves
	icket, that is what he is best
	25h : at, and I hope he comes back."
93300 33	y target the Test series in India in September 2004 for his return. Il not appeal
Warne has	gone
3 . Chat se	ssion id session000097276887 on [03/25/2003 Time : 02:29]
r_sathe	esh : should have beaten Australia!
Tom Denn	εhy, UK
It is right	for Nasser to stand down, England need a new captain, a man who knows
	i big match
🕨 rajan	
	e one of the best cricket teams, second only to Australia, in both codes of
225	e just need to stop choking when the pressures re
r_sathe game is los	25h : ally on. All too often we turn on the style when a series is over or a st.
Mike, Esse	
	ll the blame on Nasser Hussain the whole W

Scenario 3: "key word + date query" specifying "from" option

Query Cricke	et			1
Speaker Id			<u>Searcl</u>	h)
Date Filter Year	Month	Date	Hour	Minute
From 2003 -		24 💌		
Year	Month	Date	Hour	Minute
To 🔽			-	
Sort By	□ <u>D</u> ate	_	Speaker	Name
Display Preference Number of lines		atch	01 💌	
Show me Who H	heard	I	🗖 User	Get
Number of Resu	lts Per Page	I	03 💌	
Display Page		I	01 💌	

ChatTrack Project Results Page. Results for the query : cricket	
Searching 62040 conversations. Number of relevant conversation text fo Displaying 3 results per page. Page number : 1 of 7	und : 20
1. Chat session id session000096837408 on [03/25/2003 Time : 02:28]	ă.
r_satheesh : k that he should be allowed to play. Even though he is not for children in some respects, he would definitely attract children into th rajan_77 : ame of cricket. We need more kids playing cricket and to standards of village cricket, so I think this is the right decision by the EC r_satheesh : Andrew D'cruz, London, UK Warney needs some time off - a season of village cricket should rekindle enthusiasm (and ours, his fans) and keep his a	ie g improve the B.
2. Chat session id session000097716612 on [03/25/2003 Time : 02:30]	ň.
 rajan_77 : a new team to international cricket, we should have more pathem having good results. Some may say that Kenya is a new team as well, bu ▶ r_satheesh : t they are playing very good cricket. But Kenya has a vefor the future of the cricket. They have a coach who has been working for rajan_77 : r a very long time. Whereas Bangladesh changes their coach and then. And it is only recently that they have got a permanent cricket find 	ry good pla or them fo every now
3. Chat session id session000095959188 on [03/25/2003 Time : 02:26]	ä.
 r_satheesh : augh believes Warne will not find it easy when he is free comeback. "It is a long time out of sport, particularly for a guy who loves pla rajan_77 :ying at the top level. "Having 12 months without cricket is going to be very difficult for him. Bu playing cricket, that is what he is best r_satheesh : at, and I hope he comes back." Warne may target the Test series in India in September 2004 for his retriewarne will not appeal 	ut he loves
Warne has gone	



Scenario 4: "key word + date query" specifying "to" option

20

Query	cricket				
Speaker I	d L			<u>S</u> earch	
Date Filter	ar	Month	Date	Hour	Minute
From				- Tioda	
Ye	ear	Month	Date	Hour	Minute
To 20	03 💌	03 💌	27 💌	<u> </u>	
Sort By		<u>D</u> ate		Speaker I	Name
	eferences of lines be	fore/after ma	itch [01 💌	
Show m	e Who hea	ard	Г	User	Get
Number	of Results	Per Page	ſ	03 💌	
Display	Page		ſ	01 💌	

ChatTrack Project Results Page. Results for the query : cricket
Searching 62040 conversations. Number of relevant conversation text found : 20
Displaying 3 results per page. Page number : 1 of 7
1. Chat session id session000096837408 on [03/25/2003 Time : 02:28]
r_satheesh : k that he should be allowed to play. Even though he is not a role mode
for children in some respects, he would definitely attract children into the g
rajan_77 : ame of cricket. We need more kids playing cricket and to improve the
standards of village cricket, so I think this is the right decision by the ECB.
r_satheesh : Andrew D'cruz, London, UK
Warney needs some time off - a season of village cricket should rekindle his
enthusiasm (and ours, his fans) and keep his a
2. Chat session id session000097716612 on [03/25/2003 Time : 02:30]
rajan_77 : a new team to international cricket, we should have more patience to see
them having good results.
Some may say that Kenya is a new team as well, bu
r_satheesh :t they are playing very good cricket. But Kenya has a very good play for the future of the cricket. They have a coach who has been working for them fo
rajan_77 : r a very long time. Whereas Bangladesh changes their coach every now
and then. And it is only recently that they have got a permanent cricket field.
 3. Chat session id session000095959188 on [03/25/2003 Time : 02:26]
r_satheesh : augh believes Warne will not find it easy when he is free to make his comeback.
"It is a long time out of sport, particularly for a quy who loves pla
rajan_77 :ying at the top level.
"Having 12 months without cricket is going to be very difficult for him. But he loves
playing cricket, that is what he is best
r_satheesh : at, and I hope he comes back."
Warne may target the Test series in India in September 2004 for his return.
iWarne will not appeal



Scenario 5: "key word + date query" specifying "from" & "to" options

Query C	ricket			
Speaker Id	I		(<u>S</u> earc	h]
Date Filter Year	Month	Date	Hour	Minute
From 2003	▼ 03 ▼	24 •	Hour I	
Year	Month	Date	Hour	Minute
To 2003	• 03 •	27 💌		
Sort By	□ <u>D</u> ate	Ē	Speaker	Name
Display Prefer Number of li	ences nes before/after r	natch	01 💌	
Show me W	ho heard		🗖 User	<u>G</u> et
Number of F	lesults Per Page		03 💌	
Display Pag	8		01 💌	

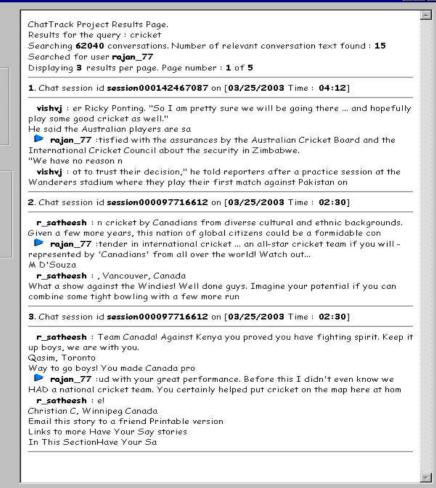
Results	nck Project Results Page. for the query : cricket
	ng 62040 conversations. Number of relevant conversation text found : 20 ing 3 results per page. Page number : 1 of 7
1 . Chat	session id session000096837408 on [03/25/2003 Time : 02:28]
for chil r a standar r_sat Warney	heesh : k that he should be allowed to play. Even though he is not a role model dren in some respects, he would definitely attract children into the g jan_77 :ame of cricket. We need more kids playing cricket and to improve the ds of village cricket, so I think this is the right decision by the ECB. heesh : Andrew D'cruz, London, UK needs some time off - a season of village cricket should rekindle his asm (and ours, his fans) and keep his a
2 . Chat	session id session000097716612 on [03/25/2003 Time: 02:30]
them ha Some m Pr_ for the rajan	
3 . Chat	session id session000095959188 on [03/25/2003 Time : 02:26]
comeba "It is a P ra "Having playing r_sat Warne i	heesh : augh believes Warne will not find it easy when he is free to make his ck. long time out of sport, particularly for a guy who loves pla jan_77 :ying at the top level. 12 months without cricket is going to be very difficult for him. But he loves cricket, that is what he is best heesh : at, and I hope he comes back." nay target the Test series in India in September 2004 for his return. will not appeal
Warne	nas gone



Scenario 6: "key word + Speaker Id + date sorted"

🛃 RetrieveDialog Application

Query cricket				
Speaker Id ra	an_77		<u>S</u> earch	
Date Filter Year	Month	Date	Hour	Minute
From			•	
Year	Month	Date	Hour	Minute
To 💌		-	-	
	✓ <u>D</u> ate		Speaker N	lame
Display Preferences Number of lines be		ich [[01 💌	
Show me Who he	ard	Г	User	Get
Number of Results	Per Page	ſ	03 💌	
Display Page		[D1 💌	





Scenario 7: "key word + date + Speaker id sorted"

🛃 RetrieveDialog Application

Query cricke	t			
Speaker Id			<u>Searc</u>	h
Date Filter Year	Month	Date	Hour	Minute
From 2003			Hoa	
Year	Month	Date	Hour	Minute
To 2003 💌	04 💌	01 💌		
Sort By	□ <u>D</u> ate	5	Speaker	Name
Display Preference Number of lines t		atch	01 💌	
Show me Who h	eard	1	🗖 User	Get
Number of Resul	ts Per Page	ļ	03 💌	
Display Page		1	01 💌	

Results	nck Project Results Page. for the query : cricket
Search	ing 62040 conversations. Number of relevant conversation text found : 20 ing 3 results per page. Page number : 1 of 7
1 . Chat	session id session000084199527 on [03/25/2003 Time : 01:59]
disorgo	_77 : ry talented side and players are full of potential, but they are nised and they must overcome this factor in order to get productive results." rthik_nt :
keen fo	raf, who is also the patron of the Pakistan Cricket Board (PCB), said he is a llower of cricket and has pinned his hopes on the Pakista _77 : n team winning the event.
"One h	is a lot of hope," he said after inaugurating the national cricket academy here.
Mushar	raf in January promised luc
2 . Chat	session id session000094205700 on [03/25/2003 Time : 02:22]
	_77 : e feels that the best part of the job is that he gets to see a lot of 5. And he has watched a lot of matches in his 30 years at the ground.
	etanush_82 :m not a cricket purist and have never played competitive cricket evel," he says.
Chris f	eels one-day internationals are more interesting as co
	_77 : mpared to Tests. "Five days for one match is too long," he says. "Tests ar I enjoy one-day cricket."
But cri	ket is nowhere even close to bei
3. Chat	session id session000097716612 on [03/25/2003 Time : 02:30]
them ho	_77 : a new team to international cricket, we should have more patience to see ving good results.
	ay say that Kenya is a new team as well, bu
	satheesh :t they are playing very good cricket. But Kenya has a very good plan future of the cricket. They have a coach who has been working for them fo
TOL ULE	

×



Scenario 8: "key word + Speaker id sorted"

Query Cricke				
Speaker Id			<u>Searc</u>	h)
Date Filter Year	Month	Date	Hour	Minute
From				0 💽
Year To	Month	Date	Hour	Minute
Sort By	Date		7 Speaker	
Display Preference			opearei	Traine .
Number of lines b	efore/after ma	atch	01 💌	
Show me Who he	ard		🗖 User	Get
Number of Result	s Per Page		03 💌	·
Display Page			01 💌	

Se	ntTrack Project Results Page. sults for the query : cricket arching 62040 conversations. Number of relevant conversation text found : 20 playing 3 results per page. Page number : 1 of 7
	Chat session id session000084199527 on [03/25/2003 Time : 01:59]
dis	ajan_77 : ry talented side and players are full of potential, but they are organised and they must overcome this factor in order to get productive results."
Mu kee	sharnaf, who is also the patron of the Pakistan Cricket Board (PCB), said he is a sharraf, who is also the patron of the Pakista of cricket and has pinned his hopes on the Pakista ajan_77 : n team winning the event.
"Oi	ne has a lot of hope," he said after inaugurating the national cricket academy here.
Mu	sharraf in January promised luc
2 . (Chat session id session000094205700 on [03/25/2003 Time : 02:22]
ma "I at Chr bor	ajan_77 : e feels that the best part of the job is that he gets to see a lot of tches. And he has watched a lot of matches in his 30 years at the ground. a "metanush_82 ::m not a cricket purist and have never played competitive cricket any level," he says. "is feels one-day internationals are more interesting as co ajan_77 : mpared to Tests. "Five days for one match is too long," he says. "Tests ar- ing. I enjoy one-day cricket." "cricket is nowhere even close to bei
3. (Chat session id session000097716612 on [03/25/2003 Time : 02:30]
the	ajan_77 : a new team to international cricket, we should have more patience to see m having good results. ne may say that Kenya is a new team as well, bu * r_satheesh :t they are playing very good cricket. But Kenya has a very good plan the future of the cricket. They have a coach who has been working for them fo



Scenario 9: "key word + date sorted"

Query Cricke	t			
Speaker Id			<u>S</u> earc	h)
Date Filter Year	Month	Date	Нош	Minute
From	Monar			
Year	Month	Date	Hour	Minute
To 💌		-		
Sort By	☑ <u>D</u> ate	Ľ	Speaker	Name
Display Preference Number of lines t		atch	01 💌	
Show me Who h	eard	1	🗖 User	Get
Number of Resul	ts Per Page	1	03 💌	
Display Page			01 💌	

	hatTrack Project Results Page. esults for the query : cricket
	earching 62040 conversations. Number of relevant conversation text found : 20
	isplaying 3 results per page. Page number : 1 of 7
1	Chat session id session000142467087 on [03/25/2003 Time : 04:12]
n	vishvj : er Ricky Ponting. "So I am pretty sure we will be going there and hopefull lay some good cricket as well."
	e said the Australian players are sa
	rajan_77 :tisfied with the assurances by the Australian Cricket Board and the
I	Are have no reason n
	vishvj : ot to trust their decision," he told reporters after a practice session at the
W	anderers stadium where they play their first match against Pakistan on
2	Chat session id session000097716612 on [03/25/2003 Time : 02:30]
	r_satheesh : n cricket by Canadians from diverse cultural and ethnic backgrounds.
G	iven a few more years, this nation of global citizens could be a formidable con
	rajan_77 :tender in international cricket an all-star cricket team if you will -
	epresented by 'Canadians' from all over the world! Watch out
	D'Souza
	r_satheesh : , Vancouver, Canada
м	/hat a show against the Windies! Well done guys. Imagine your potential if you can
	ombine some tight bowling with a few more run
3	Chat session id session000097716612 on [03/25/2003 Time: 02:30]
	r_satheesh : Team Canada! Against Kenya you proved you have fighting spirit. Keep
u	p boys, we are with you.
G	lasim, Toronto
W	/ay to go boys! You made Canada pro
	🕨 rajan_77 :ud with your great performance. Before this I didn't even know we
	AD a national cricket team. You certainly helped put cricket on the map here at hom
	r satheesh : el
C	hristian C, Winnipeg Canada
	mail this story to a friend Printable version
	inks to more Have Your Say stories
1.2	n This SectionHave Your Sa



Scenario 10: "key word + two results per page option"

Query cricket			
Speaker Id		Searc	h
Date Filter Year	Month D)ate Hour	Minute
From			
Year	Month D	ate Hour	Minute
To 🔽			
Sort By	□ <u>D</u> ate	🗖 Speaker	Name
Display Preference: Number of lines b			
reamber of mice b	croior artor matori	01 💌	
Show me Who he	eard	🔽 User	Get
Number of Result	s Per Page	02 💌	
Display Page		01 💌	

ChatTrack Project Results for the que	
	conversations. Number of relevant conversation text found : 20
	ts per page. Page number : 1 of 10
1 . Chat session id s	session000096837408 on [03/25/2003 Time : 02:28]
r_satheesh : k t	hat he should be allowed to play. Even though he is not a role model
for children in som	ne respects, he would definitely attract children into the g
	e of cricket. We need more kids playing cricket and to improve the
	ge cricket, so I think this is the right decision by the ECB. drew D'cruz, London, UK
	e time off - a season of village cricket should rekindle his
	irs, his fans) and keep his a
2 . Chat session id s	session000097716612 on [03/25/2003 Time : 02:30]
rajan_77 : a nev	v team to international cricket, we should have more patience to see
يرافيه والالترابية الأستانية	i caulta
them having good r	esuns.
Some may say that	Kenya is a new team as well, bu
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	he cricket. They have a coach who has been working for them fo
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
r_satheesh :	Kenya is a new team as well, bu
for the future of t	t they are playing very good cricket. But Kenya has a very good plar
rajan_77 : r a ve	he cricket. They have a coach who has been working for them fo
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now
Some may say that	Kenya is a new team as well, bu
r_satheesh ::	t they are playing very good cricket. But Kenya has a very good plar
for the future of t	the cricket. They have a coach who has been working for them fo
rajan_77 : r a ve	ery long time. Whereas Bangladesh changes their coach every now



Scenario 11: "key word + 4 lines before/after match"

Speaker Id Date Filter	Search
Year Month Date	Hour Minute
Year Month Date	Hour Minute
Sort By 🗖 Date	🗖 Speaker Name
Display Preferences	
Number of lines before/after match	04 💌
Show me Who heard	☐ User Get
Number of Results Per Page	
Display Page	
	02
	04 05
	06
	08
	10
	11 12 13
	14

	0 conversations. Number of relevant conversation text found : 20 ults per page. Page number : 1 of 20
1 . Chat session i	d session000096837408 on [03/25/2003 Time : 02:28]
He's been a bit of drugs cheat. Bein r_satheesh : s Let this be his w the po rajan_77 : mm performances an John, Australia I thin r_satheesh : k for children in s rajan_77 : standards of vill r_satheesh : x Warney needs so	Hawkes, Australia of a pillock but don't ruin everything for him, we all know he's not a ng off the international stage will hurt him deeply, that's punishment enough. ay of putting something back into grass-roots cricket. You never know ie kids flocking to watch him in Kent might be inspired by his id that can only be good for the English game! If that he should be allowed to play. Even though he is not a role model ome respects, he would definitely attract children into the g ame of cricket. We need more kids playing cricket and to improve the lage cricket, so I think this is the right decision by the ECB. Andrew D'cruz, London, UK ome time off - a season of village cricket should rekindle his ours, his fans) and keep his a
Reward him with Peter, Malawi	from going rusty. a pint and a pie occasionally, but not real money! st don't let him make any noney from it
touch on that ski rajan_77 : g t would not affec	ould be allowed to play. He is a terrific spinner and should not lose his ill. It has nothin o do with the cricket board anyway. Consuming the banned diuretic t the way he spin bowls. hard lesson and it is a



Scenario 12: "key word + show who listened"

Query cricket	8			
Speaker Id			<u>S</u> earc	h
Date Filter Year	Month	Date	Hour	Minute
From				
Year	Month	Date	Hour	Minute
To 🔽		-		
Sort By Display Preference:	□ <u>D</u> ate	Ľ	i Speaker	Name
Number of lines b		atch	01 💌	
Show me Who he	eard	1	🔽 User	Get
Number of Result	s Per Page		03 💌	
Display Page		1	01 💌	



Chat	Track Project Results Page.
Resu	Its for the query : cricket
Sear	ching 62040 conversations. Number of relevant conversation text found : 20
Disp	aying 3 results per page. Page number : 1 of 7
1 . Cr	at session id session000096837408 on [03/25/2003 Time: 02:28]
222 572	atheesh : k that he should be allowed to play. Even though he is not a role mode
	hildren in some respects, he would definitely attract children into the g a who heard : r_sotheestr
	rajan_77 :ame of cricket. We need more kids playing cricket and to improve the
	dards of village cricket, so I think this is the right decision by the ECB.
r_:	atheesh : Andrew D'cruz, London, UK
Wari	ney needs some time off - a season of village cricket should rekindle his
	isiasm (and ours, his fans) and keep his a
Showin	g who heard : r_satheesh
2 . Ch	at session id session000097716612 on [03/25/2003 Time : 02:30]
raj	an_77 : a new team to international cricket, we should have more patience to se
them	having good results.
	: may say that Kenya is a new team as well, bu
Showin	g who heard : r_satheesh r_satheesh :t they are playing very good cricket. But Kenya has a very good pla
fort	r_samesh it mey are playing very good cricket. But Kenya has a very good plo he future of the cricket. They have a coach who has been working for them fo
	g who heard : r_satheesh
ra	an_77 : r a very long time. Whereas Bangladesh changes their coach every now
	hen. And it is only recently that they have got a permanent cricket field.
Showin	g who heard : r_satheesh
3. Cł	at session id session000095959188 on [03/25/2003 Time : 02:26]
	atheesh : augh believes Warne will not find it easy when he is free to make his back.
"It is	a long time out of sport, particularly for a guy who loves pla
Showin	g who heard : r_satheesh
	rajan_77 : ying at the top level.
	ing 12 months without cricket is going to be very difficult for him. But he loves
	ng cricket, that is what he is best g who heard : r_satheesh
	atheesh : at, and I hope he comes back."
	ne may target the Test series in India in September 2004 for his return.
	ne will not appeal
Wari	ne has gone
	g who heard : r_satheesh

Summary of queries supported

- Keywords
- Keyword + Speaker Id
- Keyword + Date based filtering of results
- Keyword + Speaker Id + Date based filtering of results
- Keyword + Speaker Id + Date based sorting of results
- Keyword + Date based Filtering + sorting results by speaker id
- Keyword + sorting results based on Speaker id
- Keyword + sorting results based on dates
- Varying number of results per page
- Showing text around the exact document match
- Showing who (all) listened



Latest Additions

- > Queries based on Dates, Returns session ids
- > Queries based on Listener Id, Returns session ids
- > Displaying complete session as it happened



Screen Shot Showing the complete session

Under y presentation Speaker Id Speaker Id Listener Id Listener Id Date Filter Year Year Month Date Filter Year Year <th>Query presentation</th> <th>ChatTricel, Breiset Deruke Berei</th>	Query presentation	ChatTricel, Breiset Deruke Berei
Speaker Id Search Listemer Listemer Date Filter That is the ner Year Month Date Hour Month Date Hour Minute To The state Hour Month Date Hour Minute To The state Hour Minute The state Hour Month Date Hour Minute To The state Hour Minute The state Hour Minute The state Hour Minute The state Hour Minute The state Hour Number of lines before/after match OT Number of Pages OT Show me Who heard User Number of Results Per Page O3 Display Session Session Closed on 04/26/2003 Time : 15:20	Query presentation	ChatTrack Project Results Page.
Listener id Listener Listener Date Filter Year Year Year Month Date Hour Minute To Year Year Month Date Hour Minute To Year Month Date Hour Minute To Year Number of Pages Display Session Select the Drop Down List to view the complete conversation for the session	Speaker Id Search	
Date Filter Year Month Date Filter Year Month Date Hour Minute From Image: Second State Hour Year Month Date Hour Minute Year Month Year Month Date Image: Second State Hour Year Month Date Image: Second State Hour Sort By Image: Second State Hour Number of lines before/after match Image: Second State Hour Number of Pages Image: Second State Hour Show me Who heard Image: Second State Hour Display Second Second State Hour Select the Drop Down List to view the complete conversation for the session Second State Hour Select the Drop Down List to view the complete conversation for the session Second State Hour		Orspraying Dession Session 00001137728
Year Month Date Hour Minute From Image: Second State Sta	Listener id Listener	Session Started on 04/26/2003 Time : 15:17
From rejan_77 : Thanks a lot rharishnandan : sure rojan_77 : Please try to come to the presentation. rharishnandan : good luck rharishnandan : sure rojan_77 : Because I no come to the presentation. rharishnandan : sure rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s rojan_77 : Because I need to tell who "The Real Slim Shady" i s Show me Who heard Number of Results Per Page Display Session Select the Drop Down List to view the complete conversation for the session		
Year Month Date Hour Minute To Year Month Date Hour Minute To Year Month Date Financial Stresson Sort By Date Speaker Id Display Preferences Speaker Id Number of Pages OI Show me Who heard User Number of Results Per Page OI Display Session Select the Drop Down List to view the complete conversation for the session	Year Month Date Hour Minute	
Year Month Date Hour Minute To <pto< p=""> To <p< td=""><td>From</td><td></td></p<></pto<>	From	
To To Sort By Date Speaker Id Display Preferences Number of Pages OT Show me Who heard Number of Results Per Page		- rharishnandan : sure
To Sort By Sort By Display Preferences Number of lines before/after match OI Number of Pages OI Show me Who heard Number of Results Per Page Display Session Select the Drop Down List to view the complete conversation for the session	Year Month Date Hour Minute	
Sort By Date Speaker Id Display Preferences Number of lines before/after match Display Preferences Number of Pages Display Session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Page Display Session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Page Display Session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the set to the set to the conversation for the set to th		
Sort By Date Display Preferences Number of lines before/after match Number of Pages Show me Who heard Number of Results Per Page Display Session Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session		
Display Preferences Number of lines before/after match OT Show me Who heard Number of Results Per Page OT Select the Drop Down List to view the complete conversation for the session		
Display Preferences Number of lines before/after match OI Number of Pages OI Show me Who heard Number of Results Per Page Oisplay Session Select the Drop Down List to view the complete conversation for the session	Sort By 🔽 Date 🗖 Speaker Id	
Number of lines before/after match OI Number of lines before/after match Number of Pages OI Show me Who heard User Number of Results Per Page OI Display Session Select the Drop Down List to view the complete conversation for the session A set of the Drop Down List to view the complete conversation for the session		
Number of Pages Show me Who heard User Number of Results Per Page O3 Image: Select the Drop Down List to view the complete conversation for the session Select the Drop Down List to view the complete conversation for the session		
Number of Pages OI	Number of lines before/after match	
Show me Who heard Number of Results Per Page Display Session Select the Drop Down List to view the complete conversation for the session Output Image: Show me Who heard Image: Show me Who hea	Number of Dense	
Show me Who heard User Number of Results Per Page O3 Image Phorishnondon : i will make it Display Session Select the Drop Down List to view the complete conversation for the session Complete conversation for the session	Number of Pages 01 - Get	
Number of Results Per Page Image: Comparison of the session Display Session Select the Drop Down List to view the complete conversation for the session	Chamma Mithe hand To U	rharishnandan : i understand
Display Session Select the Drop Down List to view the complete conversation for the session	Show me who heard	rharishnandan : i will make it
Select the Drop Down List to view the complete conversation for the session	Number of Results Per Page 03 💌	Session Closed on 04/26/2003 Time : 15:20
Select the Drop Down List to view the complete conversation for the session	Display Session	
complete conversation for the session		
Session000041157726		
session000041157726		
	session000041157726 Show Session	



Scalability Issues

Issues

Number of Files

- Assuming a user chats 1 hour a day, creating 12 message files/minute

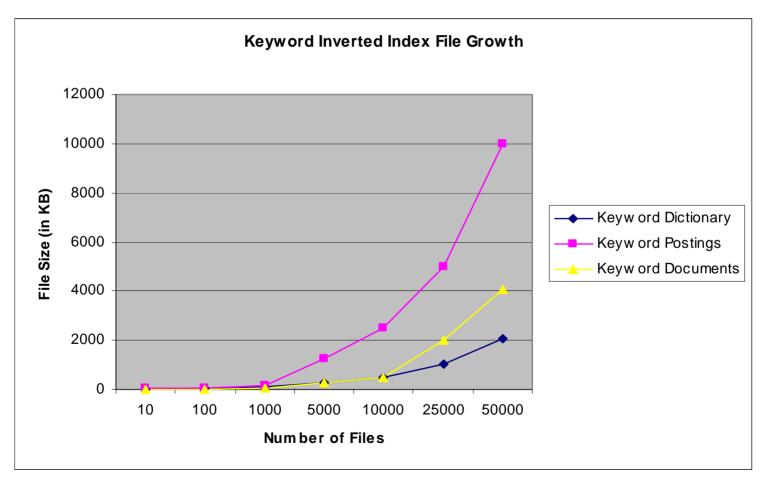
In a day, 720 message files. (Total files = 2,160) In a month, 21,600 message files. (Total files = 64,800) In a year, the total will be 259,200 files (Total files = 777,600) Windows allows up 2^{32} -1 Files = 4,294,967,295 files.

Disk usage for files

- Assuming each message file with 150 Bytes of data
 39 MB of Message data + 7 MB for storing user information per year
- Size of inverted index
- Time for retrieve

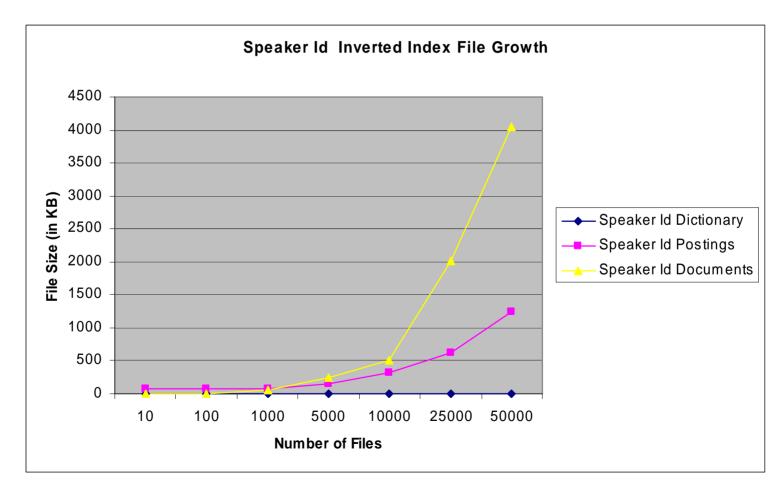


Key Word Inverted Index File Growth



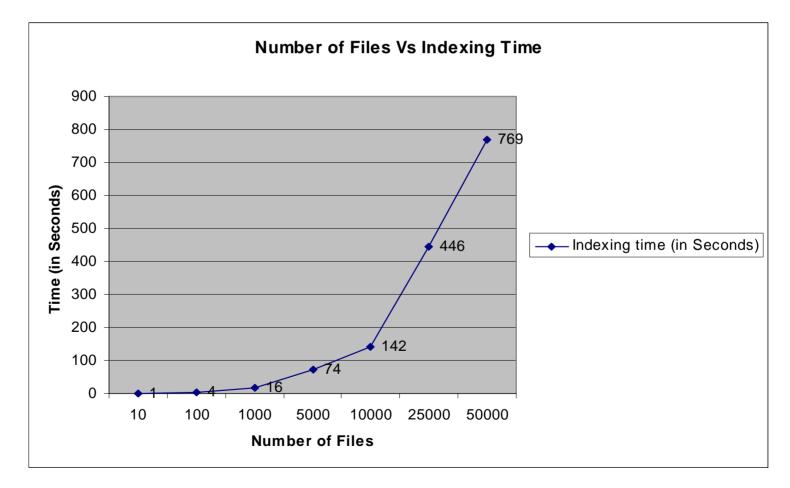


Speaker Id Inverted Index File Growth



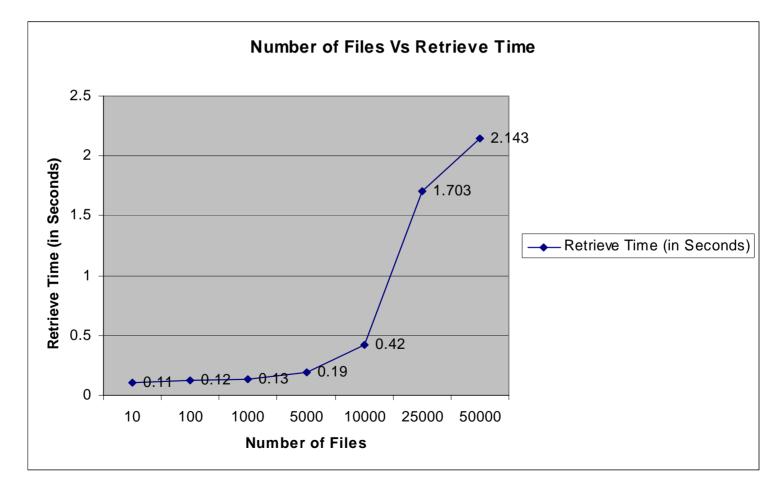


Number of Files Vs Indexing Time





Number of Files Vs Retrieval Time





System Features

Environments

- Windows NT
- Windows 2000 Professional

MSN Messenger Versions

- 5.0 and earlier



Future work

> XML

Able to be used by chat servers open architecture

- Building User Profiles
- Topic Detection and Tracking
- Segmentation
- Text Summarization

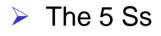


Conclusions

- An architecture for logging text, index and retrieve information in an efficient way
- > First academic system to do research with chat data
- Shares code with server-based version



Special Acknowledgements





Questions ??

