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

ITTC

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

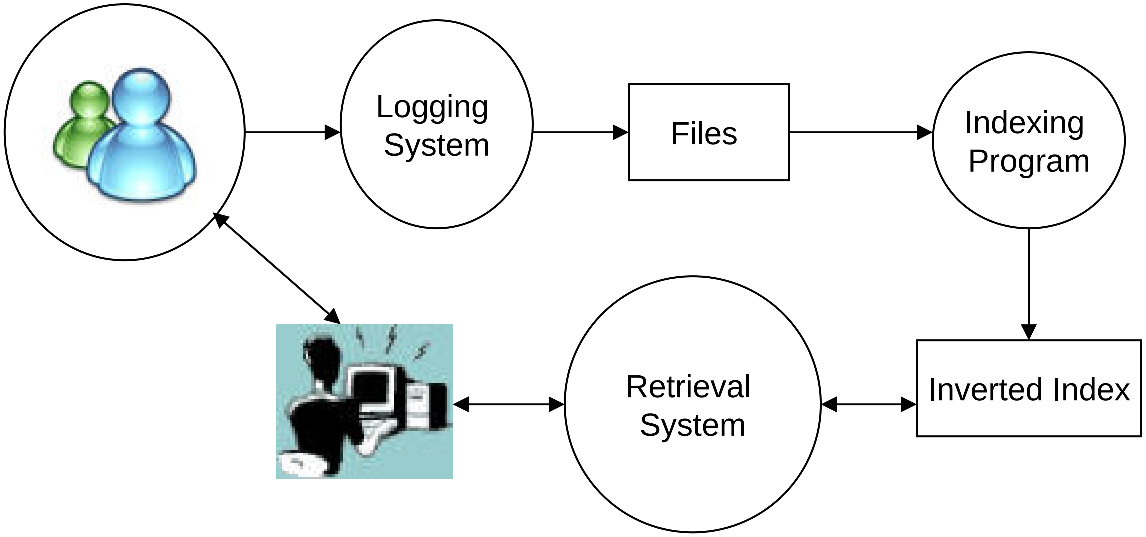
- Commercial Systems (Archiving capability)
  - Iambigbrother (Keyword search)
  - Net Nanny
  - Cyber-Scoop
  - Desktop Snooper
  - I-Spy Now

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

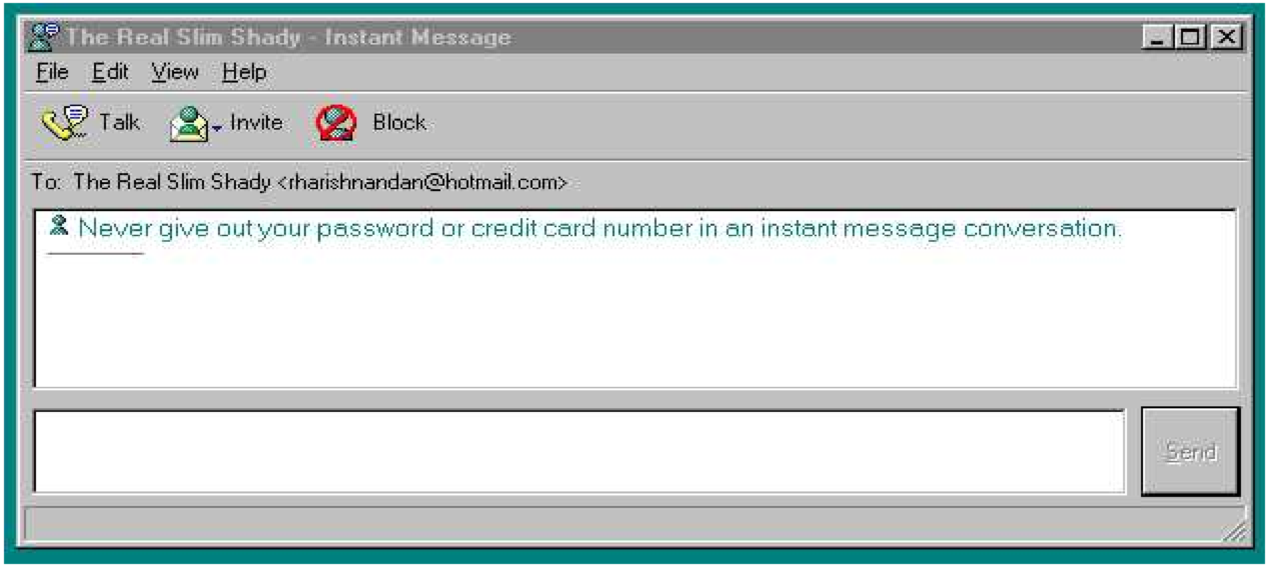


Message Logging System

- 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

Message Logging System (cont’d)

- 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



The Real Slim Shady - Instant Message

To: The Real Slim Shady <rharishnandan@hotmail.com>

Never give out your password or credit card number in an instant message conversation.

Message Logging System (cont'd)

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

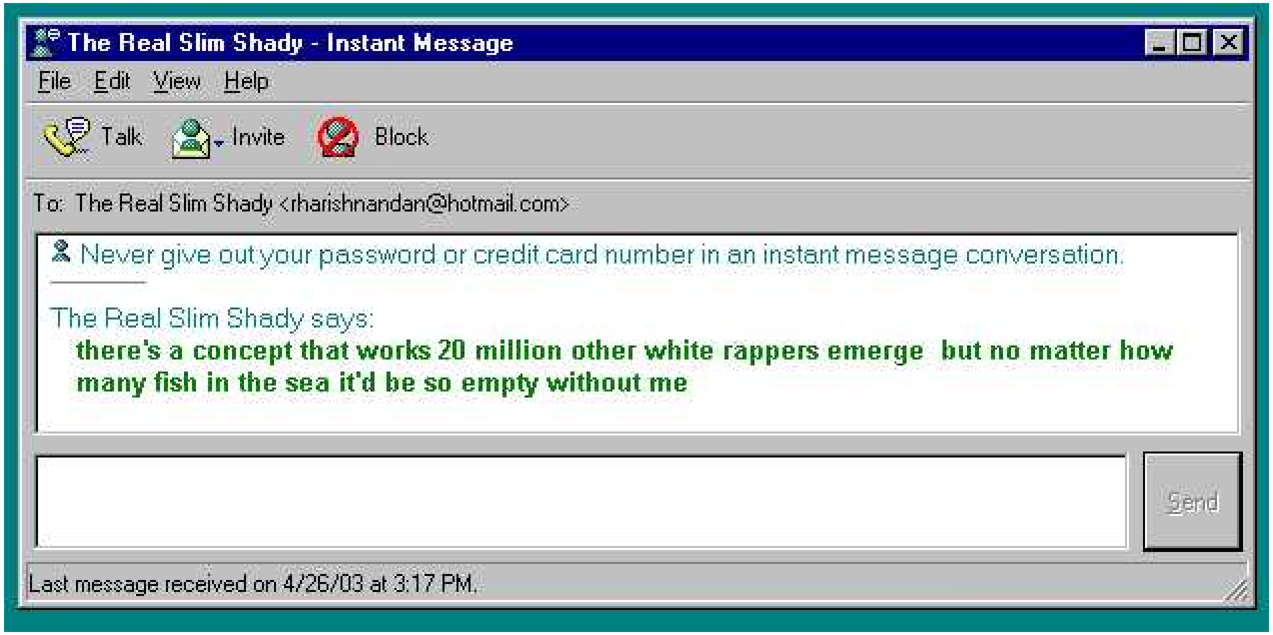
```
If (current message length > previous message length)
        If Administrative Message
            Log the event;
        else if User Message
            Log Message, Speaker Id, Listener Id(s), Add File name to "to be indexed list";
        else
            continue;
```

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

Message Logging System (cont'd)

Example.,

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



To: The Real Slim Shady <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 how many fish in the sea it'd be so empty without me

Last message received on 4/26/03 at 3:17 PM.

Message Logging System (cont’d)

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

Message Logging System (cont'd)

- 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;

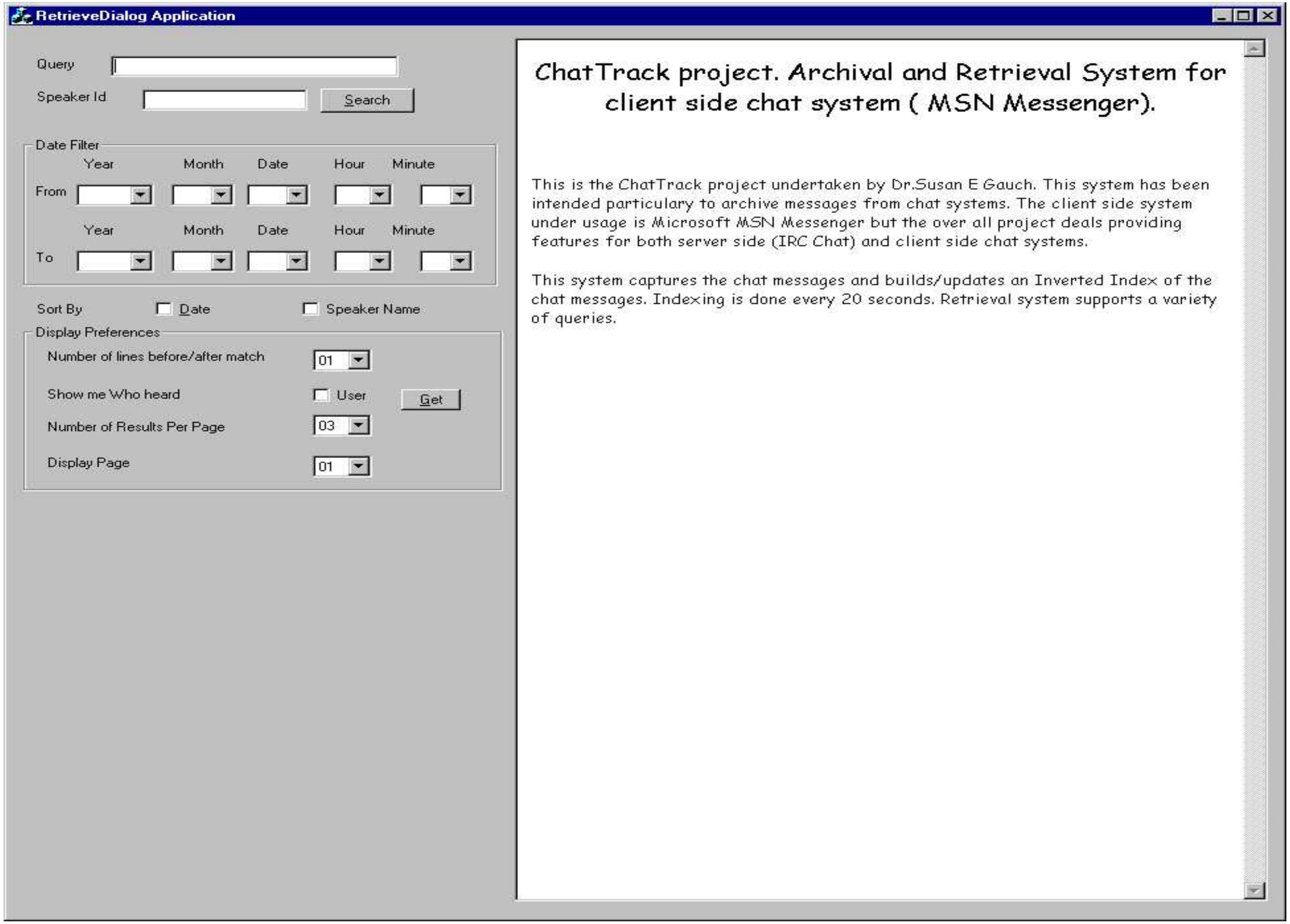
    document id = unique for a document;

    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



Number of lines before/after match: 01

Number of Results Per Page: 03

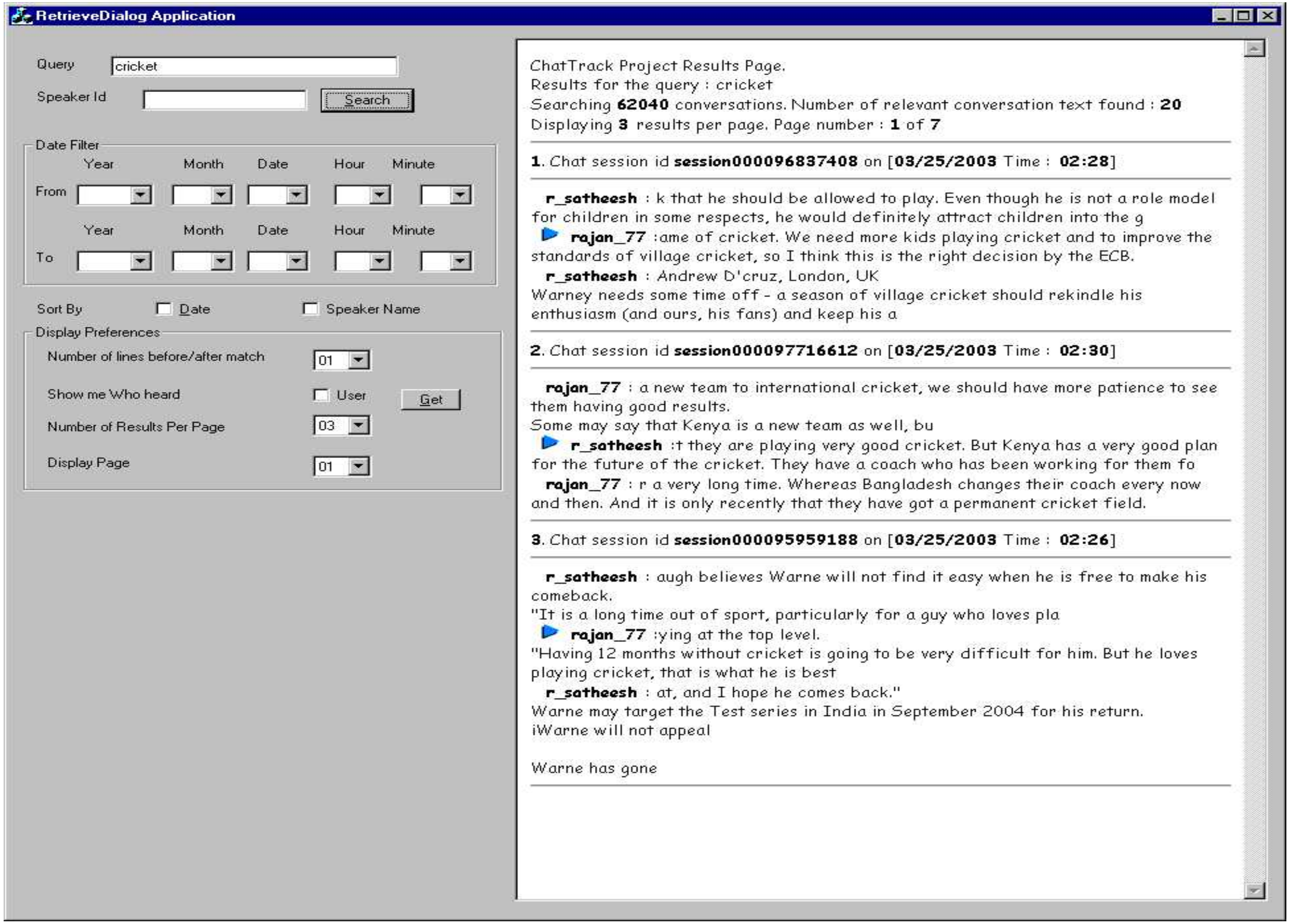
Display Page: 01

ChatTrack project. Archival and Retrieval System for client side chat system (MSN Messenger).

This is the ChatTrack project undertaken by Dr.Susan E Gauch. This system has been intended particulary to archive messages from chat systems. The client side system under usage is Microsoft MSN Messenger but the over all project deals providing features for both server side (IRC Chat) and client side chat systems.

This system captures the chat messages and builds/updates an Inverted Index of the chat messages. Indexing is done every 20 seconds. Retrieval system supports a variety of queries.

Scenario 1: “keyword” query



Query: cricket

Speaker Id:

Date Filter — From: Year, Month, Date, Hour, Minute; To: Year, Month, Date, Hour, Minute

Sort By: □ Date □ Speaker Name

Number of lines before/after match: 01

Show me Who heard: □ User

Number of Results Per Page: 03

Display Page: 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 model 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**]

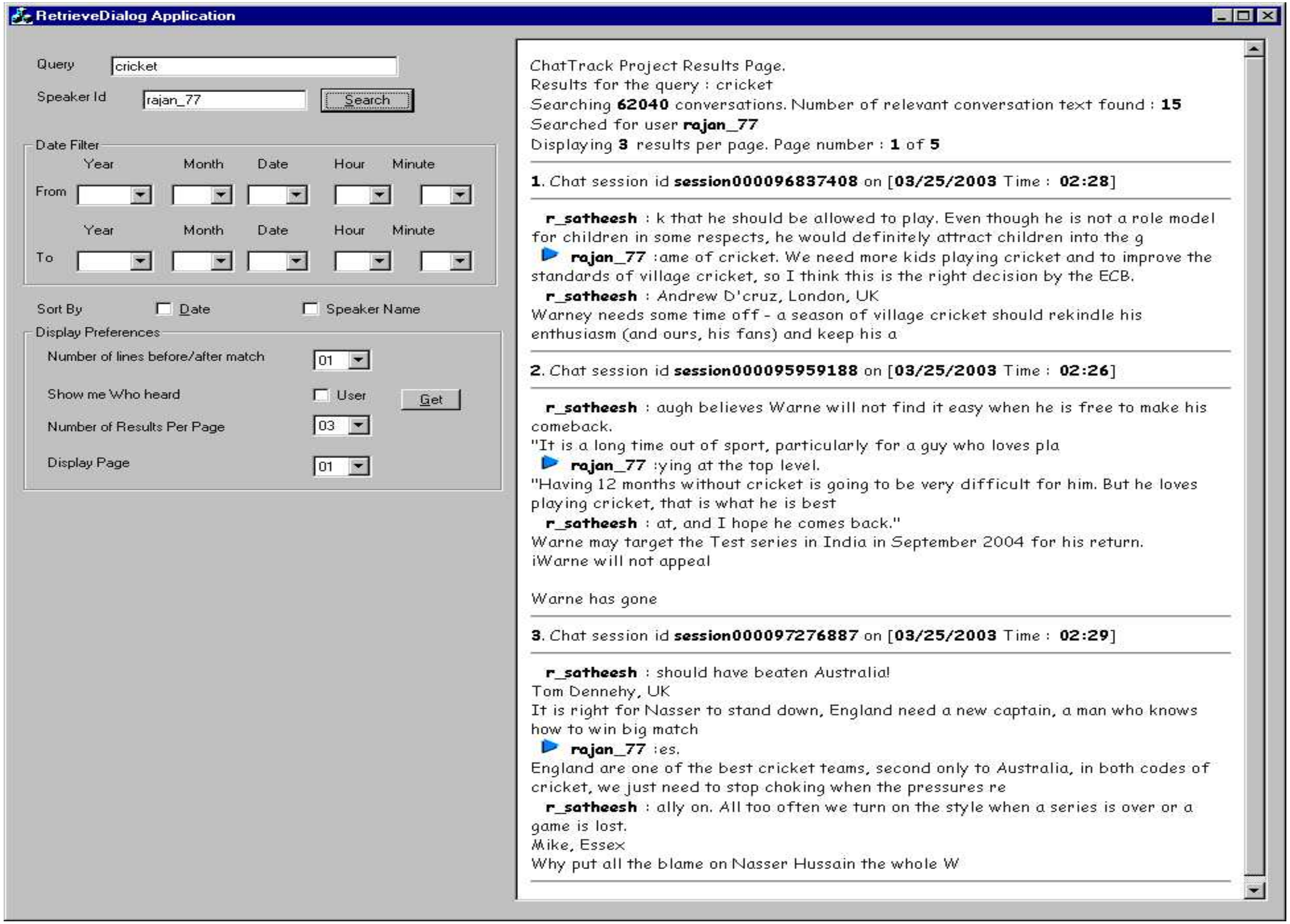
rajan_77 : a new team to international cricket, we should have more patience to see them having good results.
Some may say that Kenya is a new team as well, bu
r_satheesh :t they are playing very good cricket. But Kenya has a very good plan for the future of the cricket. They have a coach who has been 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 guy 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

Warne has gone

Scenario 2: “key word + user name” query



Query: cricket

Speaker Id: rajan_77

Date Filter

	Year	Month	Date	Hour	Minute
From					
To					

Sort By □ Date □ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard □ User

Number of Results Per Page: 03

Display Page: 01

ChatTrack Project Results Page.
Results for 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 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 model 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 **session000095959188** on [**03/25/2003** Time : **02:26**]

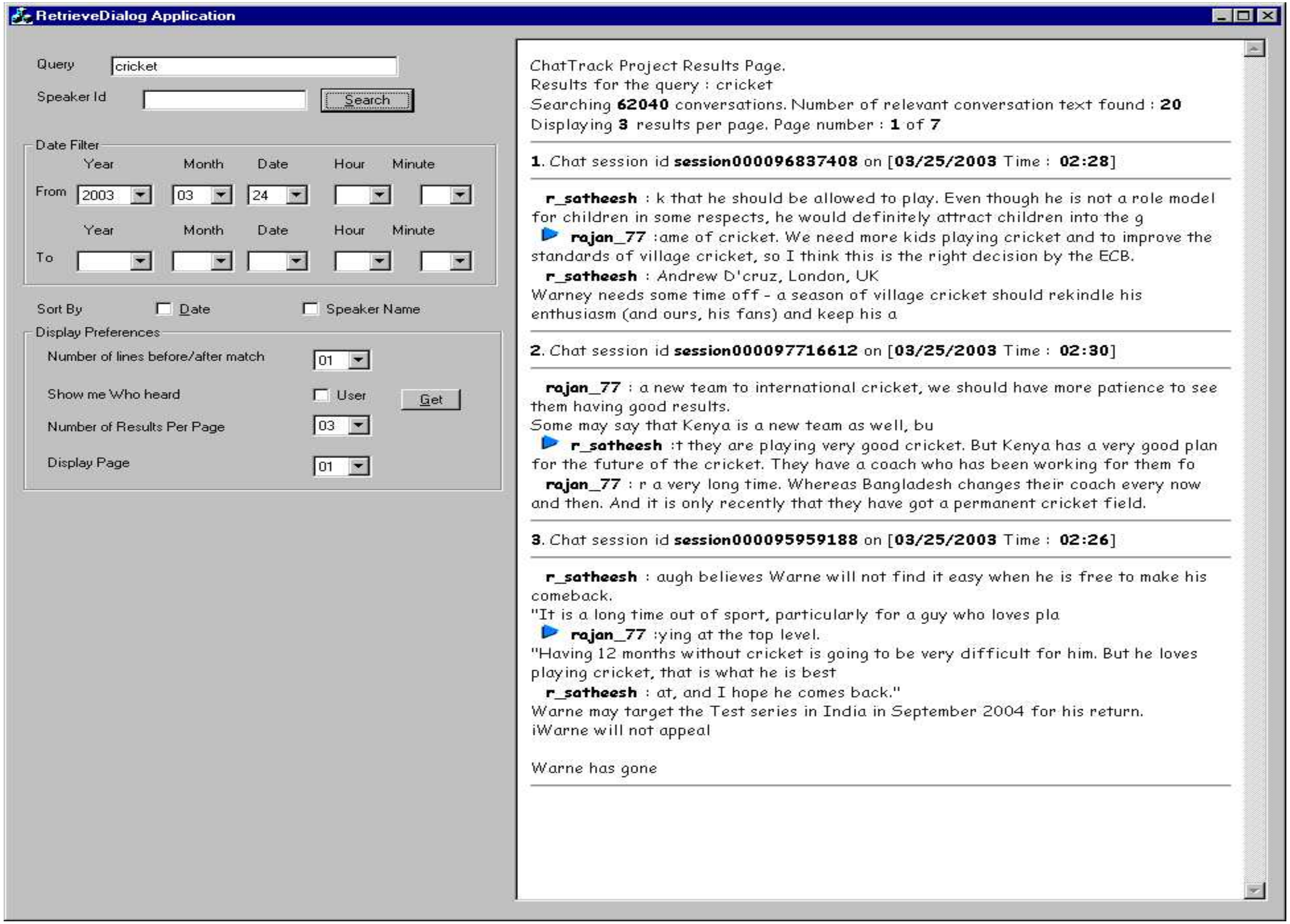
r_satheesh : augh believes Warne will not find it easy when he is free to make his comeback.
"It is a long time out of sport, particularly for a guy who loves pla
▶ **rajan_77** :ying at the top level.
"Having 12 months without cricket is going to be very difficult for him. But he loves playing cricket, 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

Warne has gone

3. Chat session id **session000097276887** on [**03/25/2003** Time : **02:29**]

r_satheesh : should have beaten Australia!
Tom Dennehy, UK
It is right for Nasser to stand down, England need a new captain, a man who knows how to win big match
▶ **rajan_77** :es.
England are one of the best cricket teams, second only to Australia, in both codes of cricket, we just need to stop choking when the pressures re
r_satheesh : ally on. All too often we turn on the style when a series is over or a game is lost.
Mike, Essex
Why put all the blame on Nasser Hussain the whole W

Scenario 3: “key word + date query” specifying “from” option



Query: cricket

Speaker Id:

Date Filter

	Year	Month	Date	Hour	Minute
From	2003	03	24		
To					

Sort By ☐ Date ☐ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard ☐ User

Number of Results Per Page: 03

Display Page: 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 model 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**]

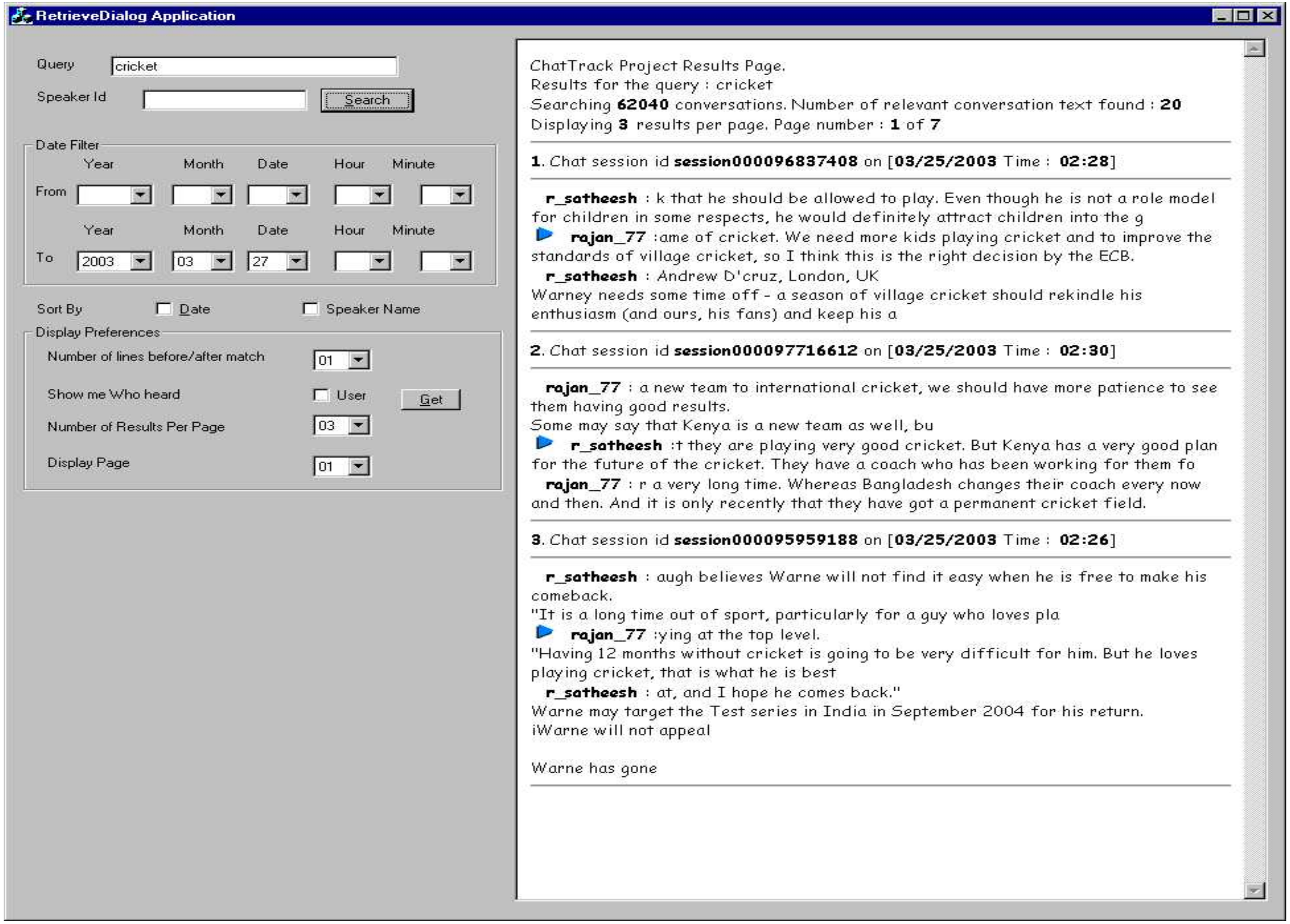
rajan_77 : a new team to international cricket, we should have more patience to see them having good results.
Some may say that Kenya is a new team as well, bu
▶ **r_satheesh** :t they are playing very good cricket. But Kenya has a very good plan for the future of the cricket. They have a coach who has 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 guy 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

Warne has gone

Scenario 4: “key word + date query” specifying “to” option



Query: cricket

Speaker Id:

Date Filter

	Year	Month	Date	Hour	Minute
From					
To	2003	03	27		

Sort By: □ Date □ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard: □ User

Number of Results Per Page: 03

Display Page: 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 model 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**]

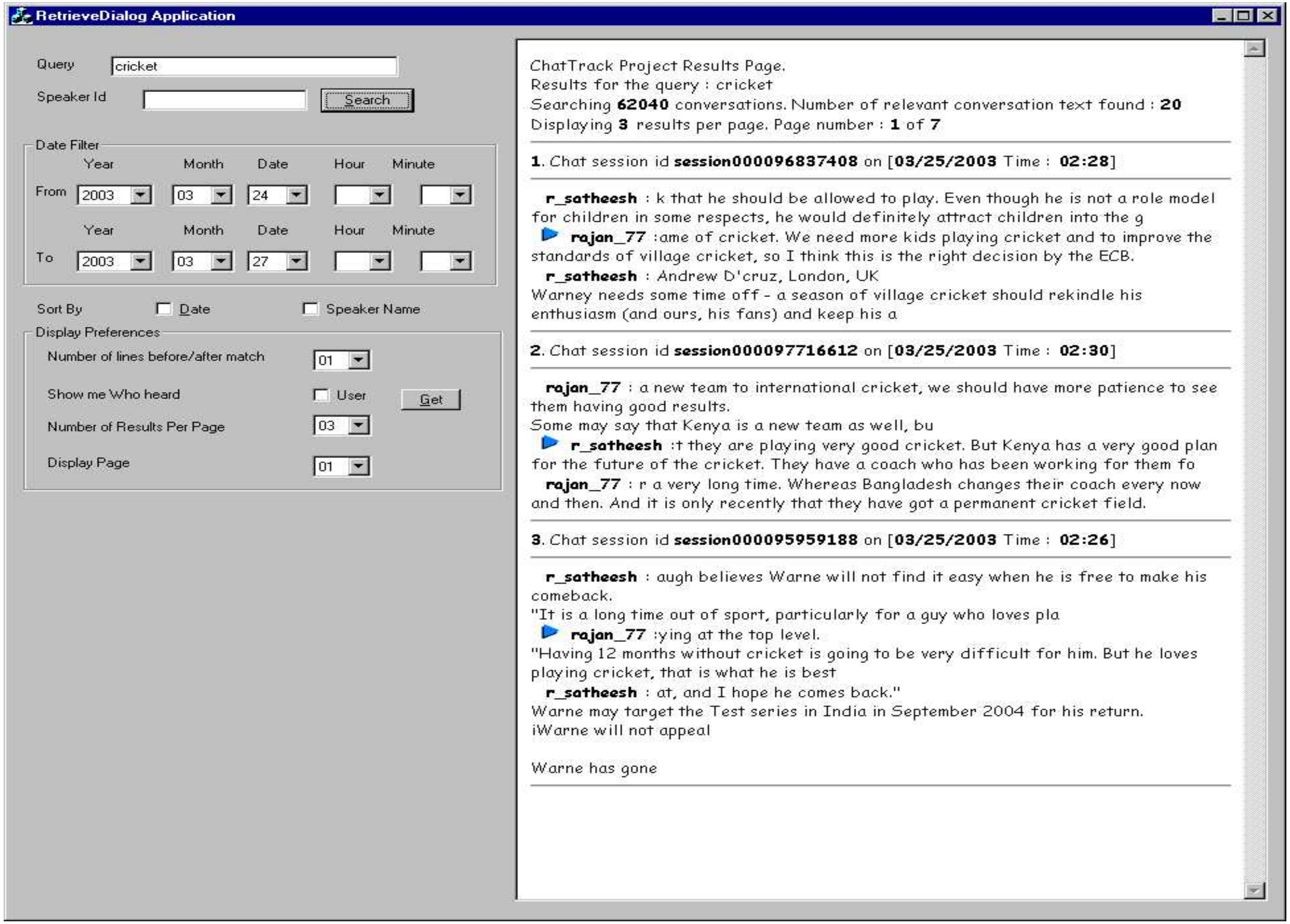
rajan_77 : a new team to international cricket, we should have more patience to see them having good results.
Some may say that Kenya is a new team as well, bu
▶ **r_satheesh** :t they are playing very good cricket. But Kenya has a very good plan for the future of the cricket. They have a coach who has 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 guy 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

Warne has gone

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



Query: cricket

Speaker Id:

Date Filter

	Year	Month	Date	Hour	Minute
From	2003	03	24		
To	2003	03	27		

Sort By ☐ Date ☐ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard ☐ User

Number of Results Per Page: 03

Display Page: 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 model 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**]

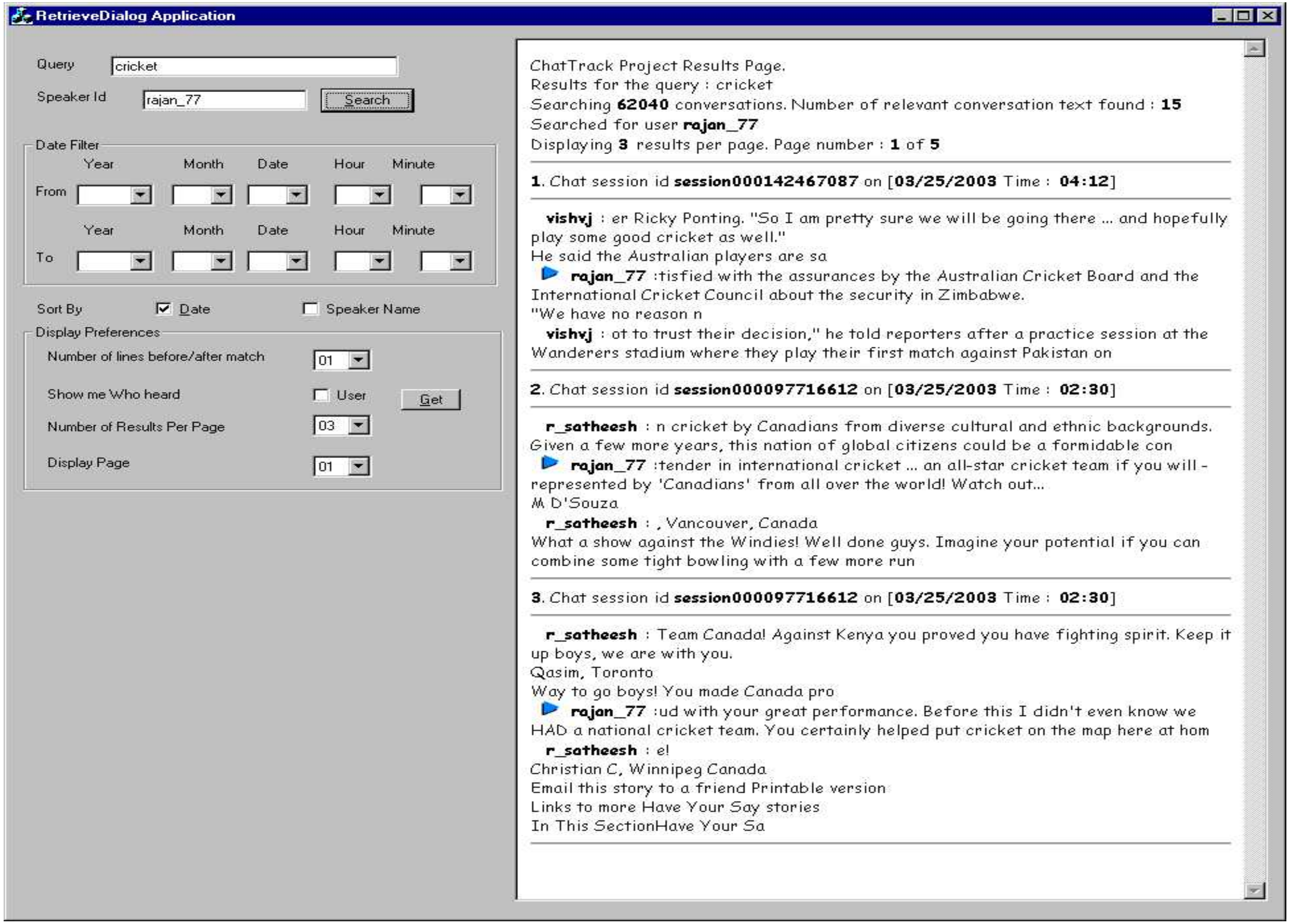
rajan_77 : a new team to international cricket, we should have more patience to see them having good results.
Some may say that Kenya is a new team as well, bu
▶ **r_satheesh** :t they are playing very good cricket. But Kenya has a very good plan for the future of the cricket. They have a coach who has 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 guy 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

Warne has gone

Scenario 6: “key word + Speaker Id + date sorted”



Query: cricket
Speaker Id: rajan_77

Sort By ☑ Date ☐ Speaker Name

Number of lines before/after match: 01
Show me Who heard ☐ User
Number of Results Per Page: 03
Display Page: 01

ChatTrack Project Results Page.
Results for 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 session id **session000142467087** on [**03/25/2003** Time : **04:12**]

vishvj : er Ricky Ponting. "So I am pretty sure we will be going there ... and hopefully play some good cricket as well."
He said the Australian players are sa
rajan_77 :tisfied with the assurances by the Australian Cricket Board and the International Cricket Council about the security in Zimbabwe.
"We have no reason n
vishvj : ot to trust their decision," he told reporters after a practice session at the Wanderers stadium where they play their first match against Pakistan on

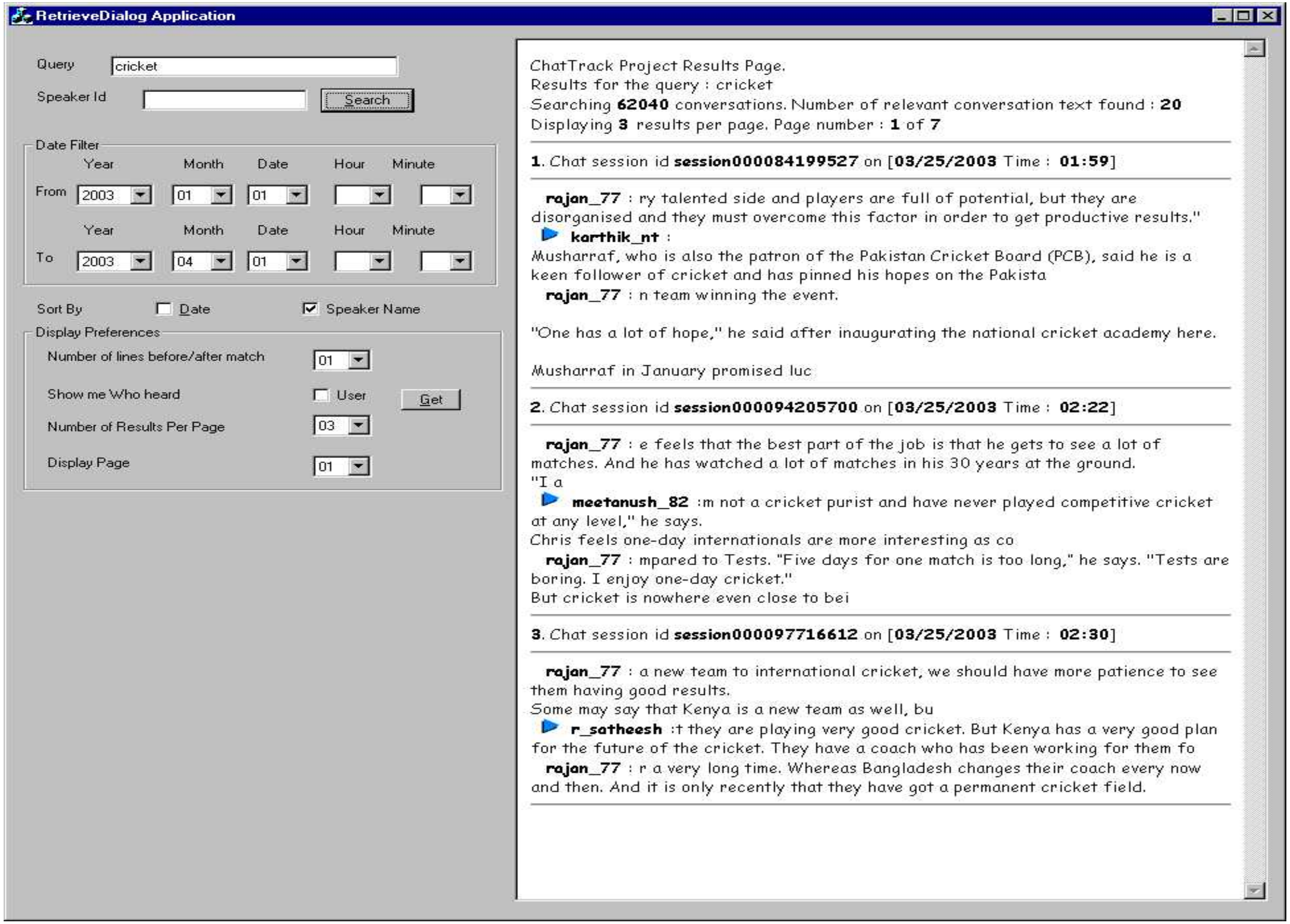
2. Chat session id **session000097716612** on [**03/25/2003** Time : **02:30**]

r_satheesh : n cricket by Canadians from diverse cultural and ethnic backgrounds. Given 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 - represented by 'Canadians' from all over the world! Watch out...
M D'Souza
r_satheesh : , Vancouver, Canada
What a show against the Windies! Well done guys. Imagine your potential if you can combine 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 it up boys, we are with you.
Qasim, Toronto
Way to go boys! You made Canada pro
rajan_77 :ud with your great performance. Before this I didn't even know we HAD a national cricket team. You certainly helped put cricket on the map here at hom
r_satheesh : e!
Christian C, Winnipeg Canada
Email this story to a friend Printable version
Links to more Have Your Say stories
In This SectionHave Your Sa

Scenario 7: “key word + date + Speaker id sorted”



Query: cricket

Speaker Id:

Date Filter

	Year	Month	Date	Hour	Minute
From	2003	01	01		
To	2003	04	01		

Sort By ☐ Date ☑ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard ☐ User

Number of Results Per Page: 03

Display Page: 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 **session000084199527** on [**03/25/2003** Time : **01:59**]

rajan_77 : ry talented side and players are full of potential, but they are disorganised and they must overcome this factor in order to get productive results."
▶ **karthik_nt** :
Musharraf, who is also the patron of the Pakistan Cricket Board (PCB), said he is a keen follower of cricket and has pinned his hopes on the Pakista
rajan_77 : n team winning the event.

"One has a lot of hope," he said after inaugurating the national cricket academy here.

Musharraf in January promised luc

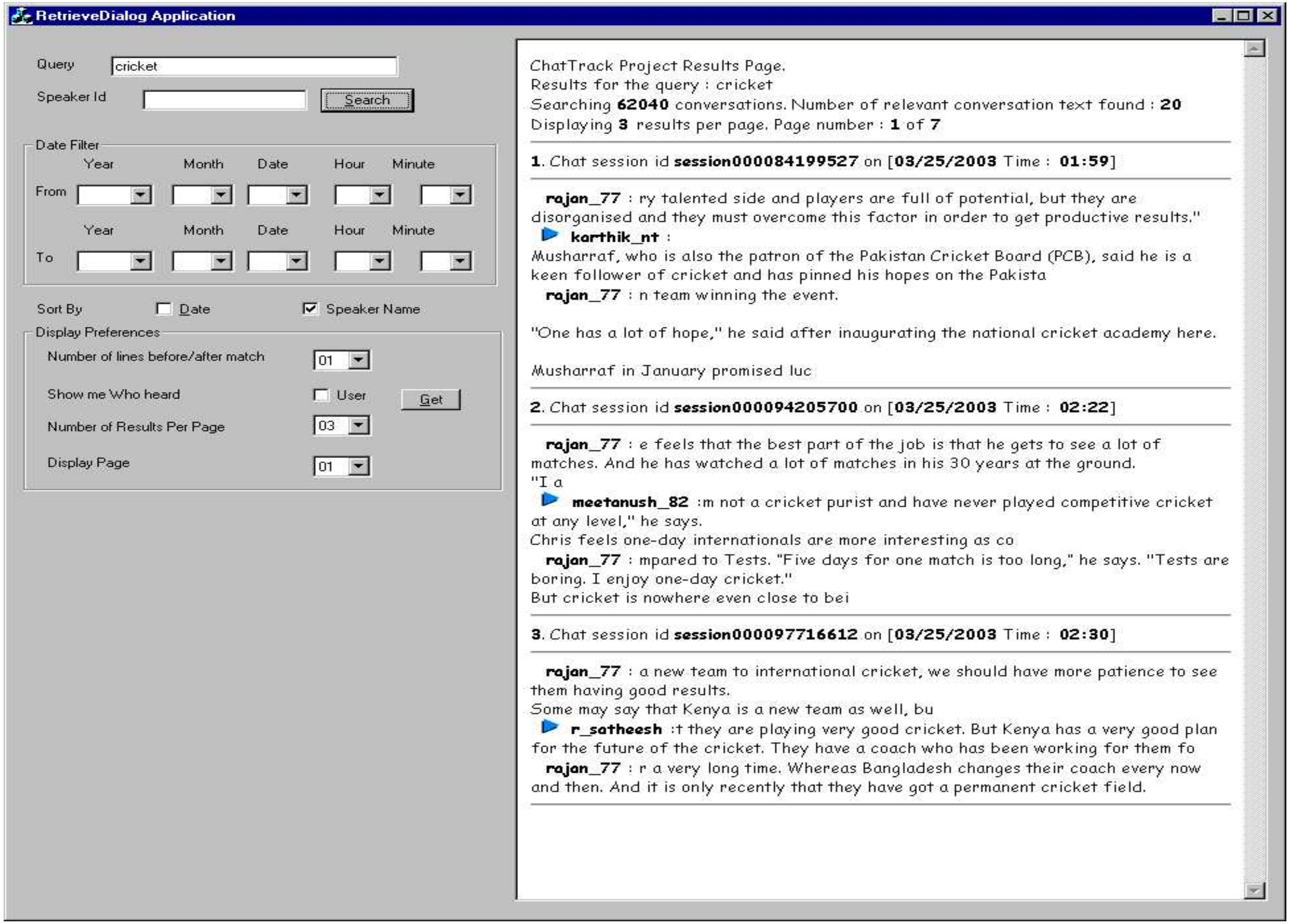
2. Chat session id **session000094205700** on [**03/25/2003** Time : **02:22**]

rajan_77 : e feels that the best part of the job is that he gets to see a lot of matches. And he has watched a lot of matches in his 30 years at the ground.
"I a
▶ **meetanush_82** :m not a cricket purist and have never played competitive cricket at any level," he says.
Chris feels one-day internationals are more interesting as co
rajan_77 : mpared to Tests. "Five days for one match is too long," he says. "Tests are boring. I enjoy one-day cricket."
But cricket is nowhere even close to bei

3. 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 plan 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.

Scenario 8: “key word + Speaker id sorted”



Query: cricket

Speaker Id:

Date Filter

	Year	Month	Date	Hour	Minute
From					
To					

Sort By ☐ Date ☑ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard ☐ User

Number of Results Per Page: 03

Display Page: 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 **session000084199527** on [**03/25/2003** Time : **01:59**]

rajan_77 : ry talented side and players are full of potential, but they are disorganised and they must overcome this factor in order to get productive results."
karthik_nt :
Musharraf, who is also the patron of the Pakistan Cricket Board (PCB), said he is a keen follower of cricket and has pinned his hopes on the Pakista
rajan_77 : n team winning the event.

"One has a lot of hope," he said after inaugurating the national cricket academy here.

Musharraf in January promised luc

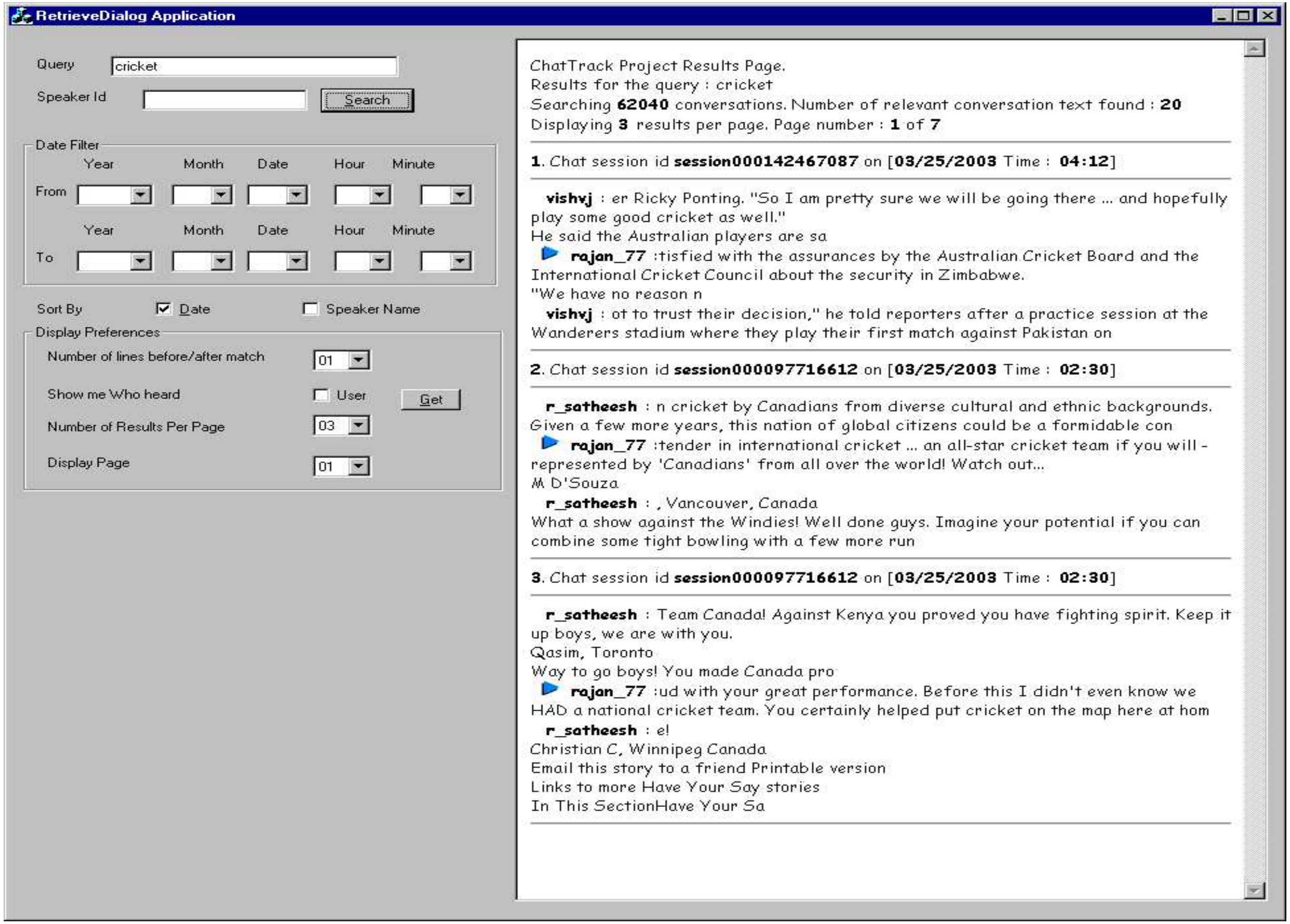
2. Chat session id **session000094205700** on [**03/25/2003** Time : **02:22**]

rajan_77 : e feels that the best part of the job is that he gets to see a lot of matches. And he has watched a lot of matches in his 30 years at the ground.
"I a
meetanush_82 :m not a cricket purist and have never played competitive cricket at any level," he says.
Chris feels one-day internationals are more interesting as co
rajan_77 : mpared to Tests. "Five days for one match is too long," he says. "Tests are boring. I enjoy one-day cricket."
But cricket is nowhere even close to bei

3. 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 plan 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.

Scenario 9: “key word + date sorted”



RetrieveDialog Application

Query: cricket

Speaker Id:

Date Filter

	Year	Month	Date	Hour	Minute
From					
To					

Sort By ☑ Date □ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard □ User

Number of Results Per Page: 03

Display Page: 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 **session000142467087** on [**03/25/2003** Time : **04:12**]

vishvj : er Ricky Ponting. "So I am pretty sure we will be going there ... and hopefully play some good cricket as well."
He said the Australian players are sa
rajan_77 :tisfied with the assurances by the Australian Cricket Board and the International Cricket Council about the security in Zimbabwe.
"We have no reason n
vishvj : ot to trust their decision," he told reporters after a practice session at the Wanderers stadium where they play their first match against Pakistan on

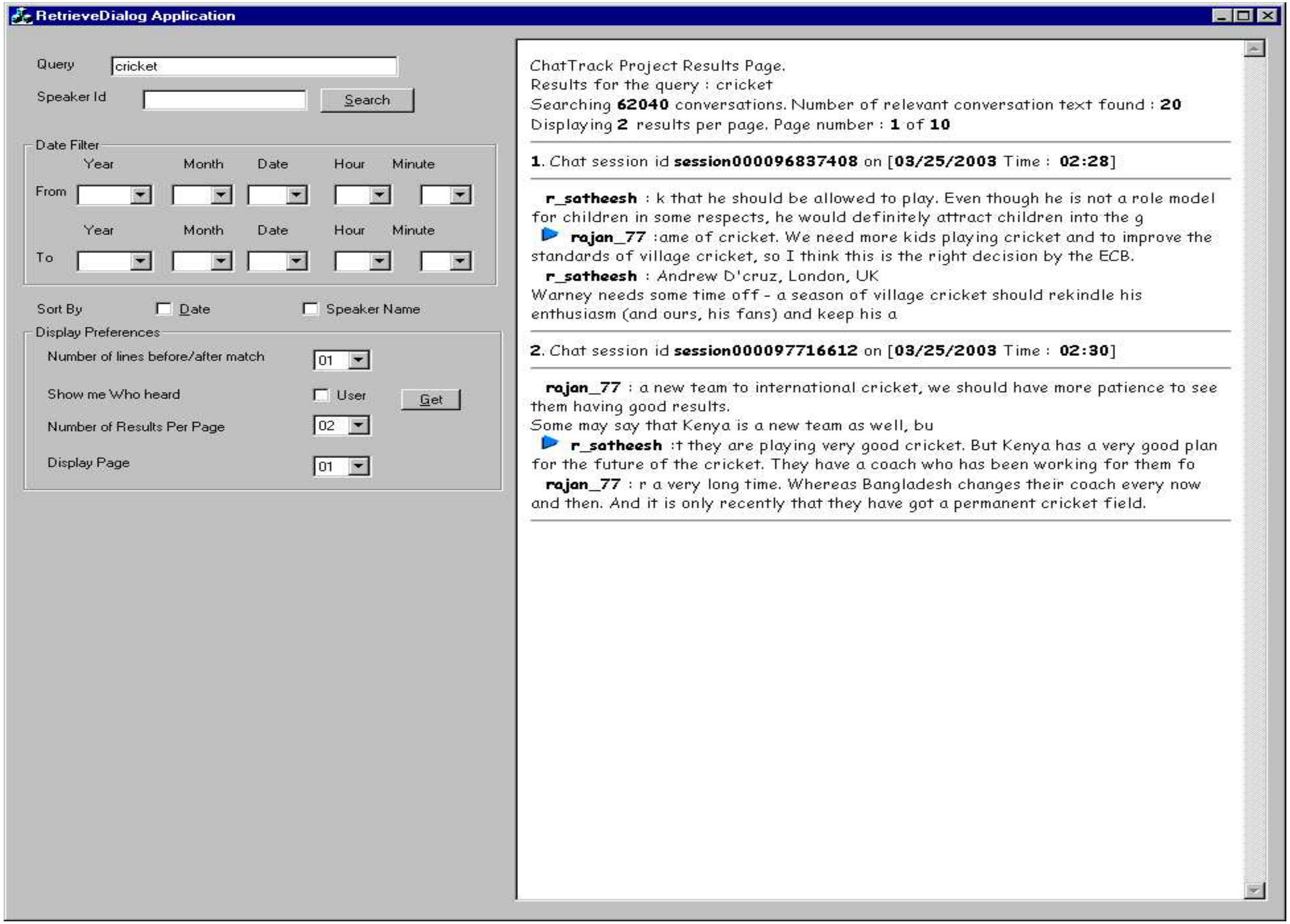
2. Chat session id **session000097716612** on [**03/25/2003** Time : **02:30**]

r_satheesh : n cricket by Canadians from diverse cultural and ethnic backgrounds. Given 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 - represented by 'Canadians' from all over the world! Watch out...
M D'Souza
r_satheesh : , Vancouver, Canada
What a show against the Windies! Well done guys. Imagine your potential if you can combine 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 it up boys, we are with you.
Qasim, Toronto
Way to go boys! You made Canada pro
rajan_77 :ud with your great performance. Before this I didn't even know we HAD a national cricket team. You certainly helped put cricket on the map here at hom
r_satheesh : e!
Christian C, Winnipeg Canada
Email this story to a friend Printable version
Links to more Have Your Say stories
In This SectionHave Your Sa

Scenario 10: “key word + two results per page option”



Query: cricket

Speaker Id:

Date Filter

	Year	Month	Date	Hour	Minute
From					
To					

Sort By ☐ Date ☐ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard ☐ User

Number of Results Per Page: 02

Display Page: 01

ChatTrack Project Results Page.
Results for the query : cricket
Searching **62040** conversations. Number of relevant conversation text found : **20**
Displaying **2** results per page. Page number : **1** of **10**

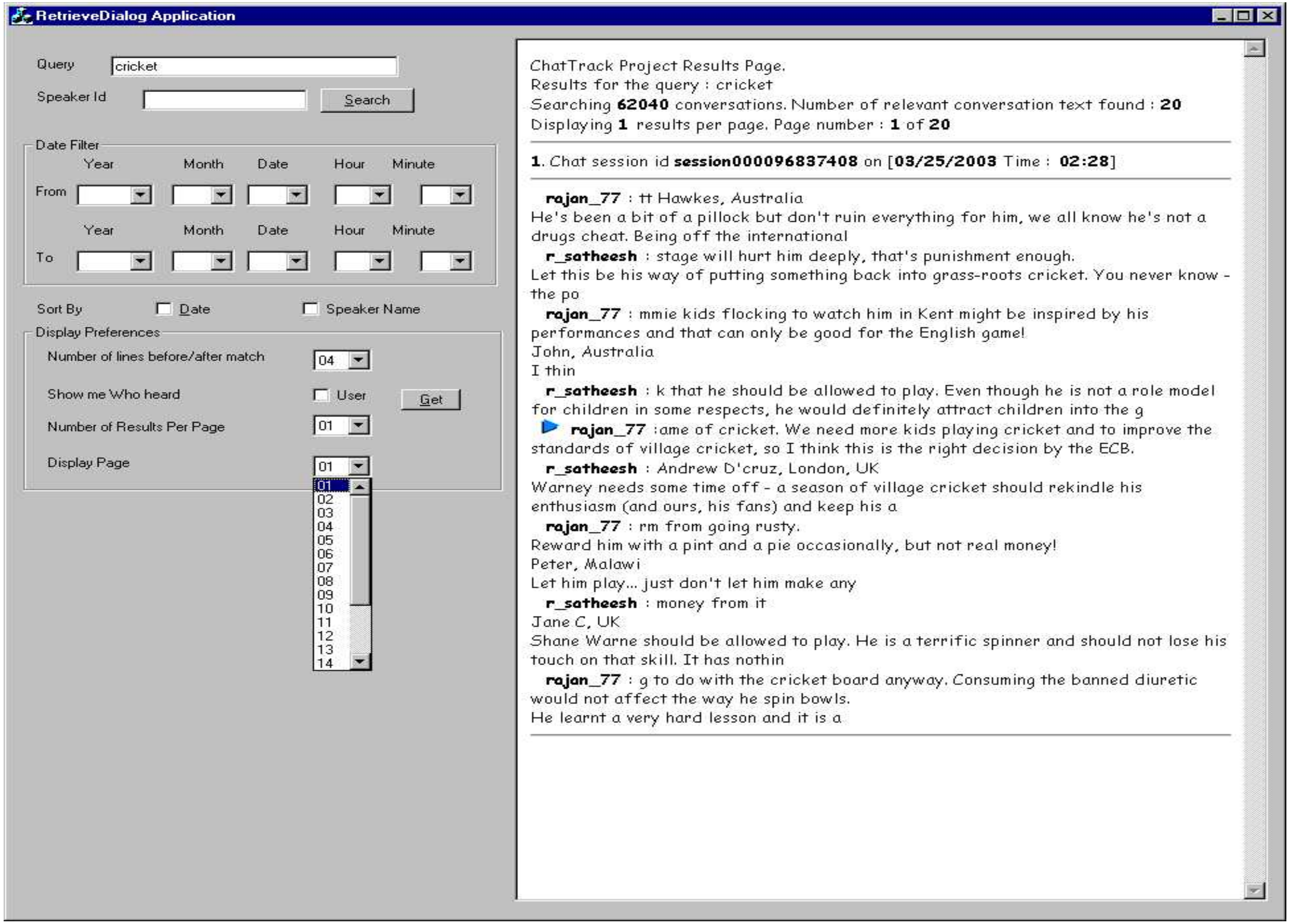
1. Chat session id **session000096837408** on [**03/25/2003** Time : **02:28**]

r_satheesh : k that he should be allowed to play. Even though he is not a role model 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 plan 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.

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



Query: cricket

Speaker Id:

Sort By ☐ Date ☐ Speaker Name

Display Preferences

Number of lines before/after match: 04

Show me Who heard ☐ User

Number of Results Per Page: 01

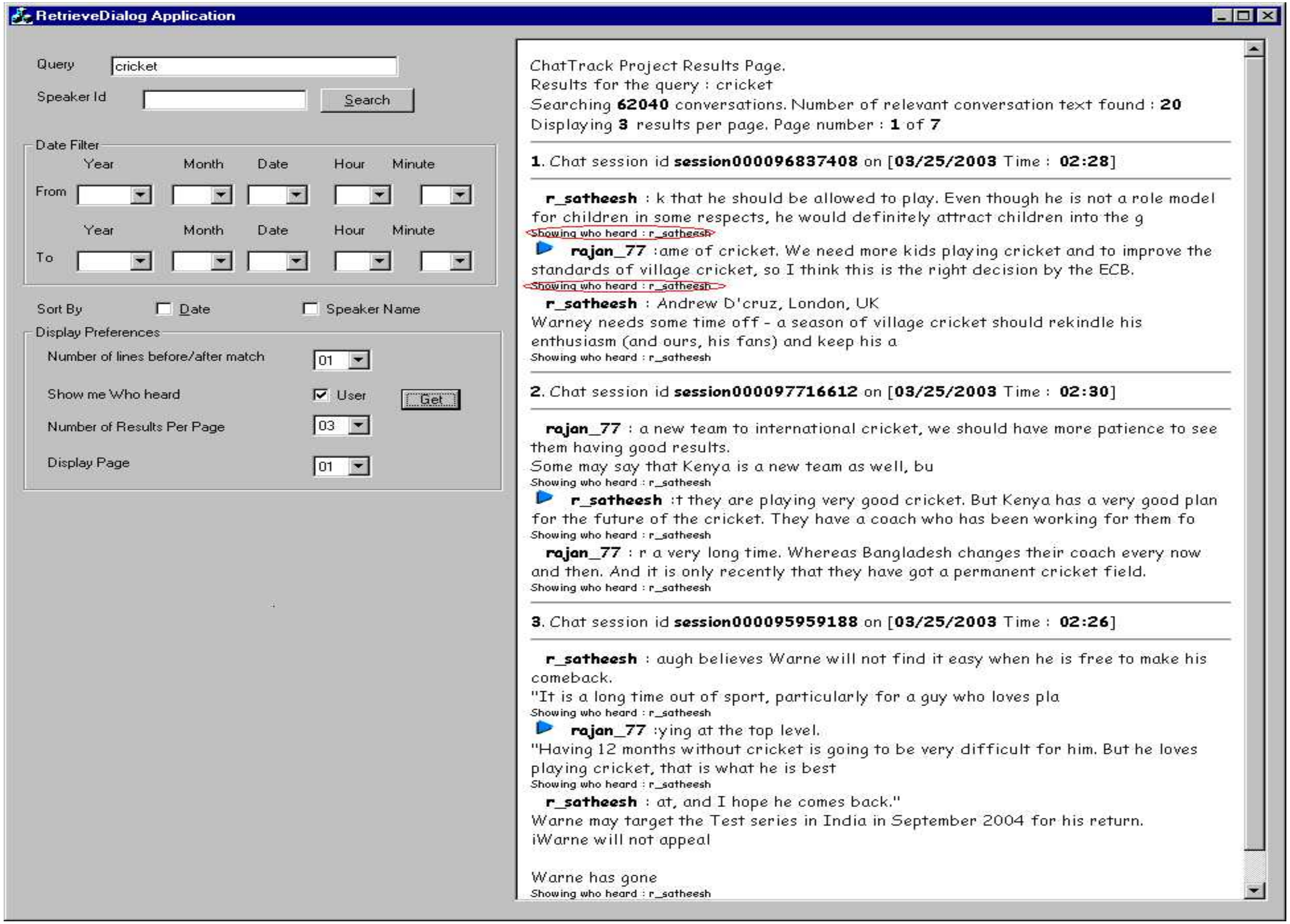
Display Page: 01

ChatTrack Project Results Page.
Results for the query : cricket
Searching **62040** conversations. Number of relevant conversation text found : **20**
Displaying **1** results per page. Page number : **1** of **20**

1. Chat session id **session000096837408** on [**03/25/2003** Time : **02:28**]

rajan_77 : tt Hawkes, Australia
He's been a bit of a pillock but don't ruin everything for him, we all know he's not a drugs cheat. Being off the international
r_satheesh : stage will hurt him deeply, that's punishment enough.
Let this be his way of putting something back into grass-roots cricket. You never know - the po
rajan_77 : mmie kids flocking to watch him in Kent might be inspired by his performances and that can only be good for the English game!
John, Australia
I thin
r_satheesh : k that he should be allowed to play. Even though he is not a role model 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
rajan_77 : rm from going rusty.
Reward him with a pint and a pie occasionally, but not real money!
Peter, Malawi
Let him play... just don't let him make any
r_satheesh : money from it
Jane C, UK
Shane Warne should be allowed to play. He is a terrific spinner and should not lose his touch on that skill. It has nothin
rajan_77 : g to do with the cricket board anyway. Consuming the banned diuretic would not affect the way he spin bowls.
He learnt a very hard lesson and it is a

Scenario 12: “key word + show who listened”



Query: cricket

Speaker Id:

Date Filter

	Year	Month	Date	Hour	Minute
From					
To					

Sort By ☐ Date ☐ Speaker Name

Display Preferences

Number of lines before/after match: 01

Show me Who heard ☑ User

Number of Results Per Page: 03

Display Page: 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 model for children in some respects, he would definitely attract children into the g
Showing who heard : r_satheesh

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.
Showing who heard : r_satheesh

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
Showing who heard : r_satheesh

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
Showing who heard : r_satheesh

r_satheesh :t they are playing very good cricket. But Kenya has a very good plan for the future of the cricket. They have a coach who has been working for them fo
Showing who heard : r_satheesh

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.
Showing who heard : r_satheesh

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 guy who loves pla
Showing who heard : r_satheesh

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
Showing who heard : r_satheesh

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

Warne has gone
Showing who heard : r_satheesh

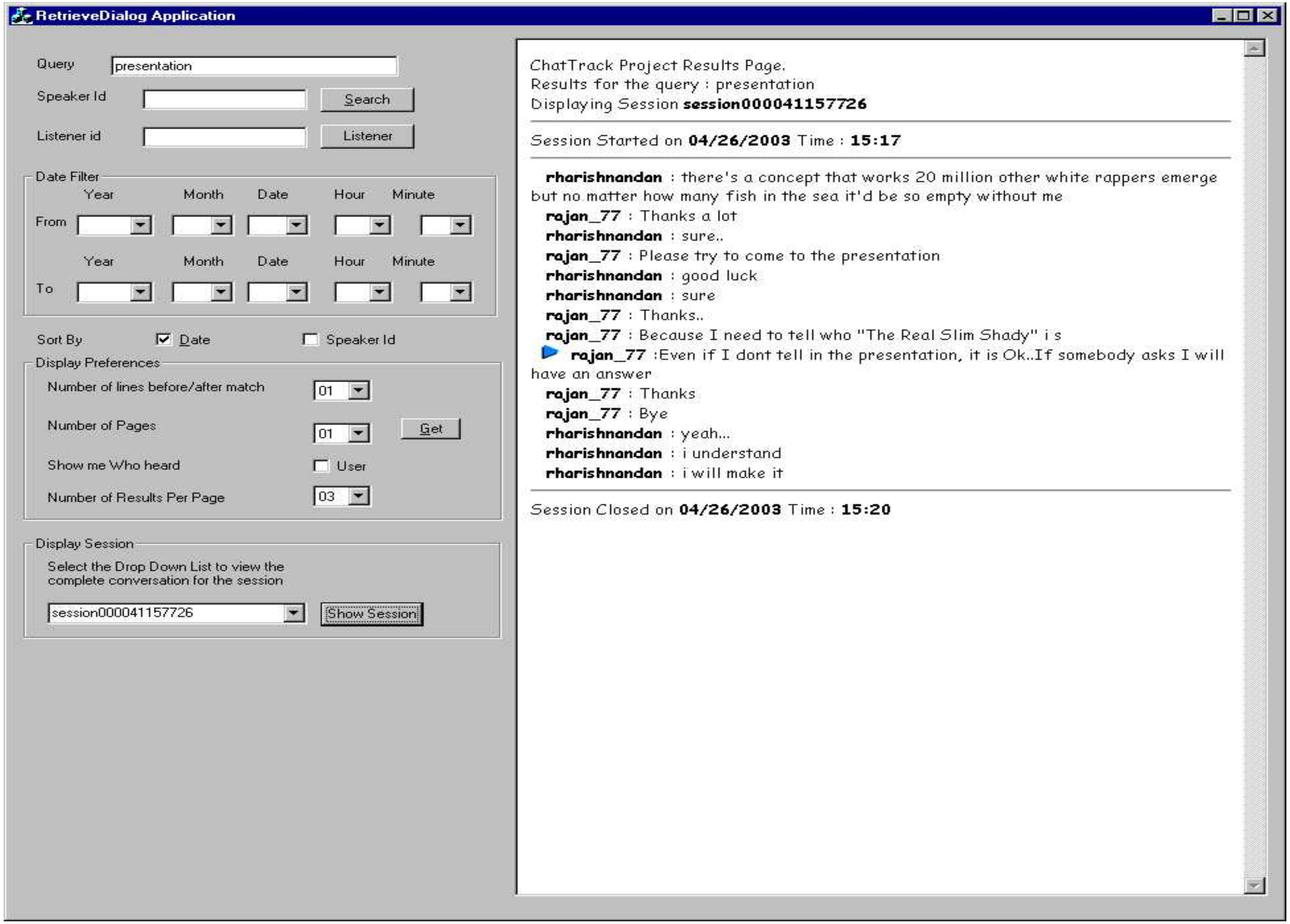
Summary of queries supported

- 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



Query: presentation

Speaker Id:

Listener id:

Date Filter

	Year	Month	Date	Hour	Minute
From					
To					

Sort By ☑ Date ☐ Speaker Id

Display Preferences

Number of lines before/after match: 01

Number of Pages: 01

Show me Who heard ☐ User

Number of Results Per Page: 03

Display Session

Select the Drop Down List to view the complete conversation for the session

session000041157726

ChatTrack Project Results Page.
Results for the query : presentation
Displaying Session **session000041157726**

Session Started on **04/26/2003** Time : **15:17**

rharishnandan : 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
rajan_77 : Thanks a lot
rharishnandan : sure..
rajan_77 : Please try to come to the presentation
rharishnandan : good luck
rharishnandan : sure
rajan_77 : Thanks..
rajan_77 : Because I need to tell who "The Real Slim Shady" i s
rajan_77 :Even if I dont tell in the presentation, it is Ok..If somebody asks I will have an answer
rajan_77 : Thanks
rajan_77 : Bye
rharishnandan : yeah...
rharishnandan : i understand
rharishnandan : i will make it

Session Closed on **04/26/2003** Time : **15:20**

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)

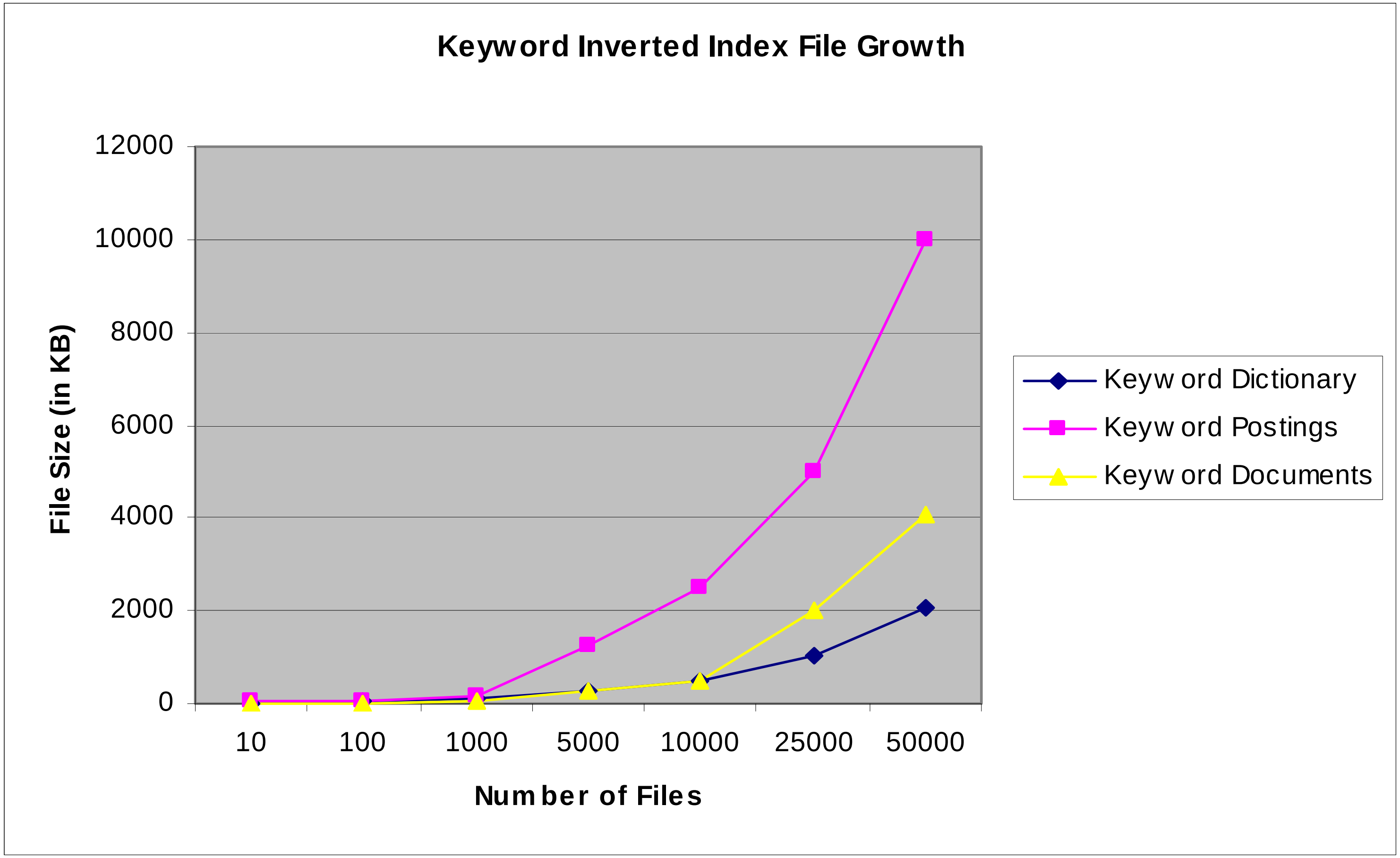
      In a year, the total will be 259,200 files (Total files = 777,600)

      Windows allows up 2^{32}-1 Files = 4,294,967,295 files.
  - Disk usage for files

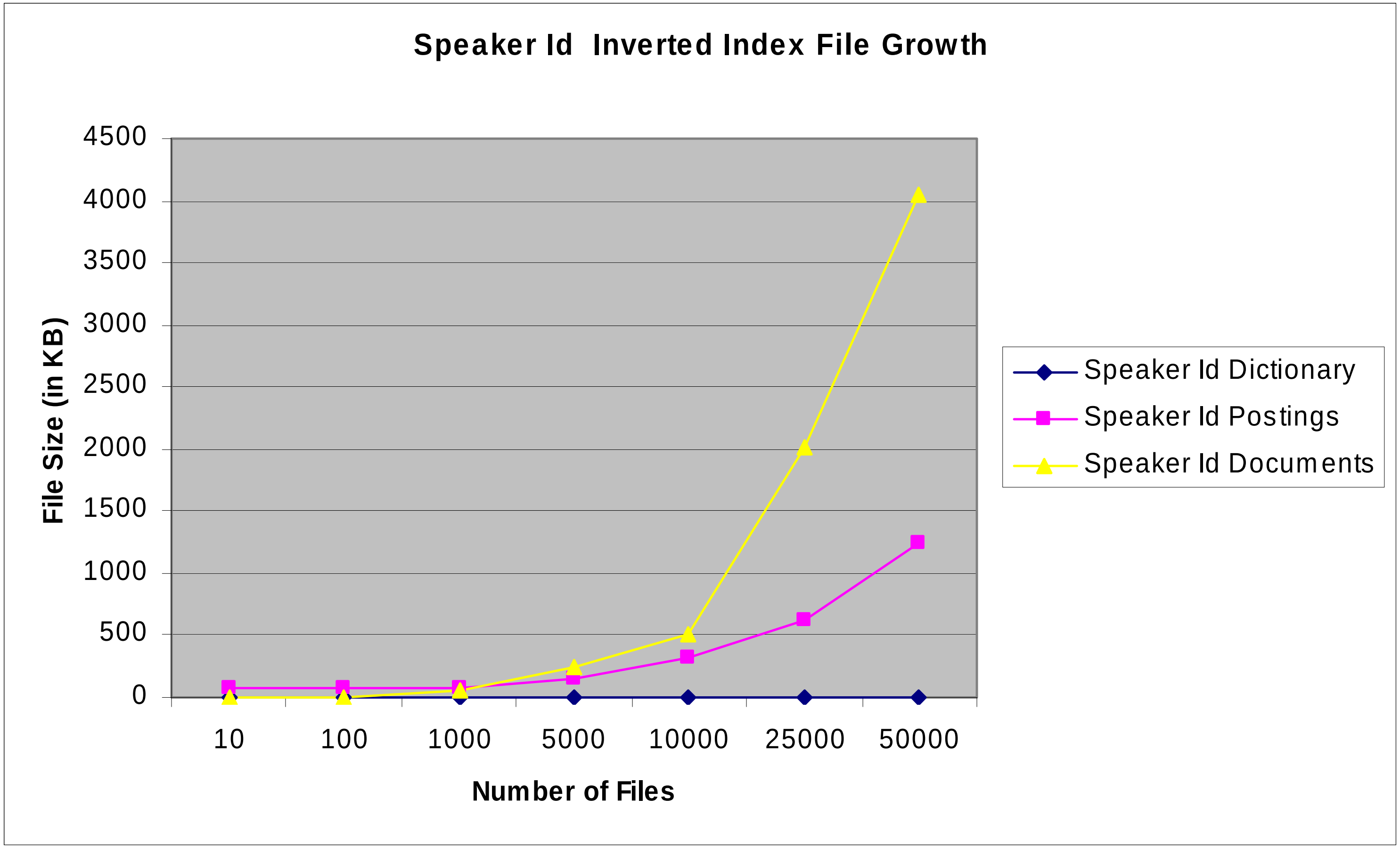
    - Assuming each message file with 150 Bytes of data

      39 MB of Message data + 7 MB for storing user information per year
  - Size of inverted index
  - Time for retrieve

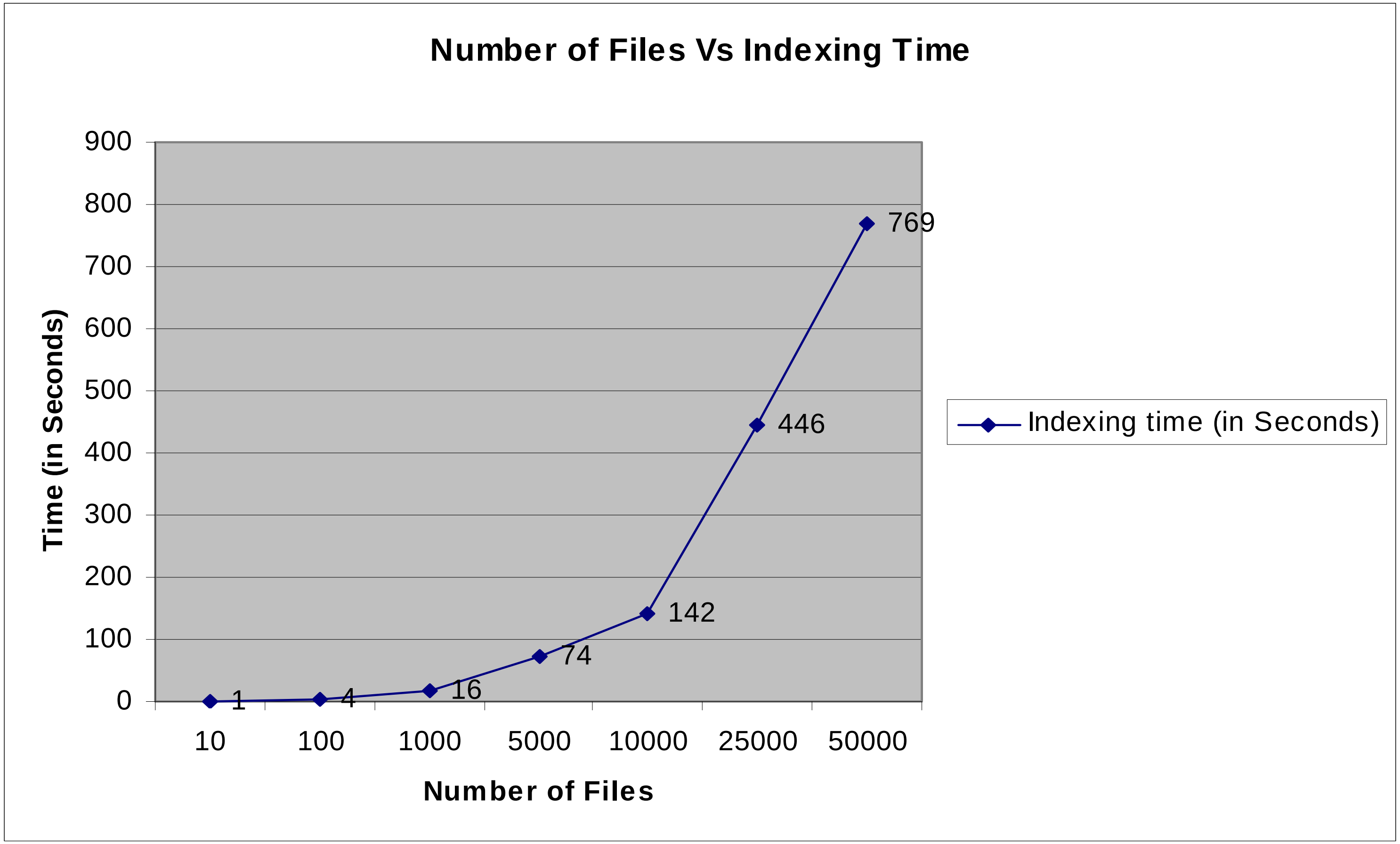
Key Word Inverted Index File Growth



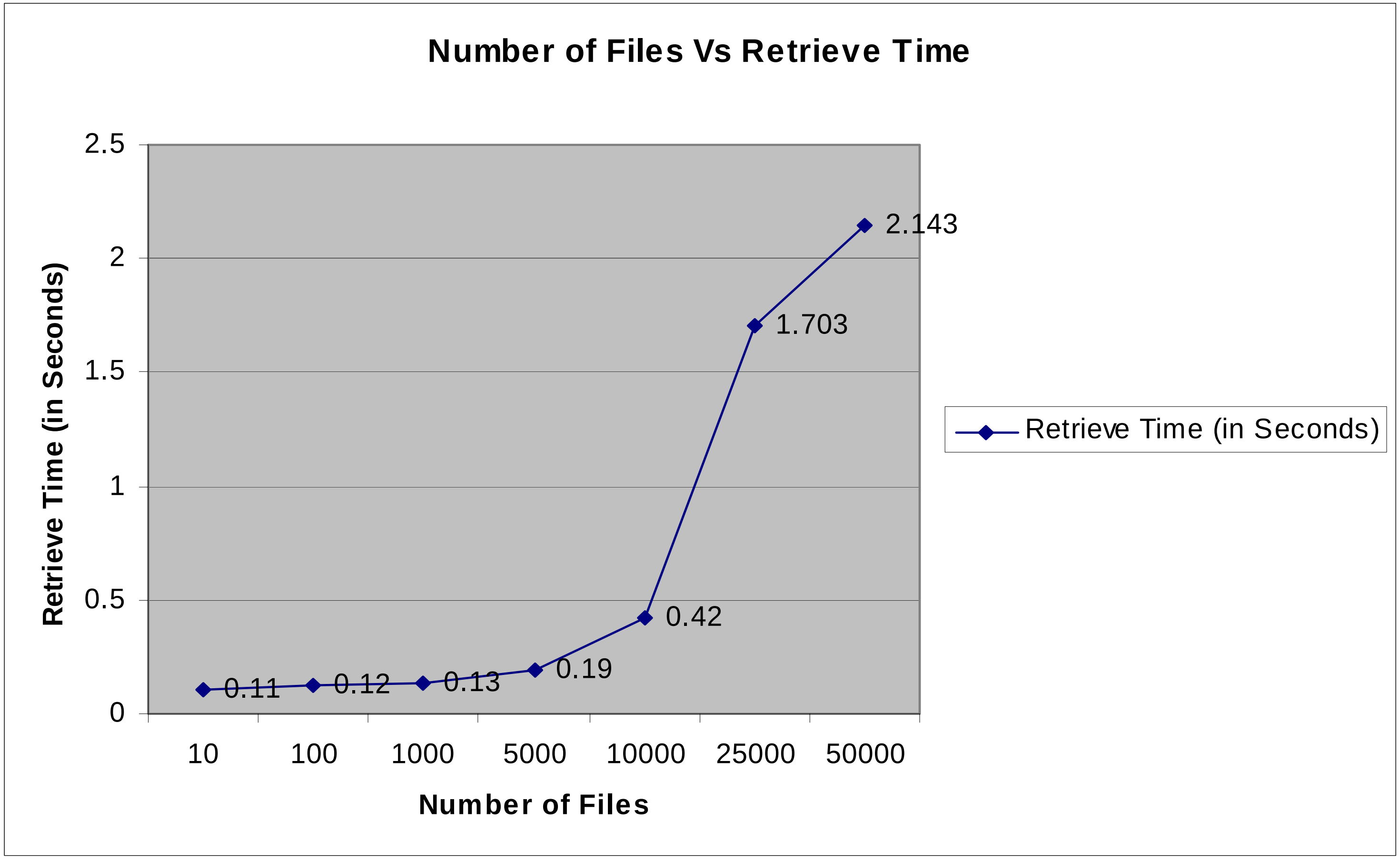
Speaker Id Inverted Index File Growth



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

- The 5 Ss

Questions ??