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
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
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
Example.,

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

The Real Slim Shady says:

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

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
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 = log₂(200/100), pointer to postings }

Postings File { 100 postings record ( 100 documents),
  weight in document j = idf * frequency of ‘prozac’ in document j;
  document id = unique for a document;
  pointer to next postings record; }

14
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

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

This is the ChatTrack project undertaken by br. Susan E. Gough. This system has been intended particularly to archive messages from chat systems. The client side system under usage is Microsoft MSN Messenger but the overall 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

ChatTrack Project Results Page:
Results for the query: cricket
Searching 62040 conversations. Number of relevant conversation text found: 30
Displaying 5 results per page. Page number: 1 of 7

1. Chat session id session000898357408 on [08/25/2003 Time: 02:28]
   _r_sothose: I believe 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 game.
   _rajan_77: I think some 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_sothose: Andrew D’Arcy, London, UK
   Warne needs some time off - a season of village cricket should rekindle his enthusiasm (and ours, his fans) and keep him afloat.

2. Chat session id session000967166412 on [08/25/2003 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, but...
   _r_sothose: If 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 for 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 coach.

3. Chat session id session00095959188 on [08/25/2003 Time: 02:26]
   _r_sothose: I think 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: A top level. Playing 12 months without cricket is going to be very difficult for him. But he loves playing cricket, that is what he is best.
   _r_sothose: Yes, and I hope he comes back.
   Warne may target the Test series in India in September 2004 for his return.
   Warne will not appeal.

Warne has gone
Scenario 2: “key word + user name” query
Scenario 3: “key word + date query” specifying “from” option
Scenario 4: “key word + date query” specifying “to” option

<table>
<thead>
<tr>
<th>Chat session id</th>
<th>session0000096837408</th>
<th>on [08/25/2003 Time : 02:23]</th>
</tr>
</thead>
<tbody>
<tr>
<td>rajan_77</td>
<td>I think 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 game 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.</td>
<td></td>
</tr>
<tr>
<td>r_satheesh</td>
<td>Andrew D'cruz, London, UK</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Waqar Younis has some time off - a season of village cricket should rekindle his enthusiasm (and ours, his fans) and keep his a...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chat session id</th>
<th>session000009716612</th>
<th>on [08/25/2003 Time : 02:30]</th>
</tr>
</thead>
<tbody>
<tr>
<td>rajan_77</td>
<td>I am 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, but...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chat session id</th>
<th>session000095959188</th>
<th>on [08/25/2003 Time : 02:26]</th>
</tr>
</thead>
<tbody>
<tr>
<td>r_satheesh</td>
<td>I agree with 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...</td>
<td></td>
</tr>
<tr>
<td>rajan_77</td>
<td>I am 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 at, and I hope he comes back...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chat session id</th>
<th>session000096837408</th>
<th>on [08/25/2003 Time : 02:23]</th>
</tr>
</thead>
<tbody>
<tr>
<td>r_satheesh</td>
<td>Waqar Younis has some time off - a season of village cricket should rekindle his enthusiasm (and ours, his fans) and keep his a...</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Display Preferences</th>
<th>Number of time before/after switch: 01</th>
<th>Show me who heard: User:</th>
<th>Number of Results Per Page: 01</th>
<th>Display Page: 01</th>
</tr>
</thead>
</table>
Scenario 5: “key word + date query” specifying “from” & “to” options
Scenario 6: “key word + Speaker Id + date sorted”
Scenario 7: “key word + date + Speaker id sorted”
Scenario 8: “key word + Speaker id sorted”
Scenario 9: “key word + date sorted”
Scenario 10: “key word + two results per page option”
Scenario 11: “key word + 4 lines before/after match”
Scenario 12: “key word + show who listened”
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

```
<table>
<thead>
<tr>
<th>Rhetoricon</th>
<th>there's a concept that works 20 million other white rappers emerge</th>
</tr>
</thead>
<tbody>
<tr>
<td>rajan_77</td>
<td>but no matter how many fish in the sea it'd be so empty without me</td>
</tr>
<tr>
<td>rajan_77</td>
<td>Thanks a lot</td>
</tr>
<tr>
<td>rherishmondo</td>
<td>I need to tell who &quot;The Real Slim Shady&quot; is</td>
</tr>
<tr>
<td>rherishmondo</td>
<td>Because I need to tell who &quot;The Real Slim Shady&quot; is</td>
</tr>
<tr>
<td>rajan_77</td>
<td>Thanks</td>
</tr>
<tr>
<td>rajan_77</td>
<td>Bye</td>
</tr>
<tr>
<td>rherishmondo</td>
<td>I will make it</td>
</tr>
</tbody>
</table>
```

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

Session Closed on 04/26/2003 Time : 15:20
Scalability 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 to $2^{32} - 1$ 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

![Graph showing the growth of file size in KB versus the number of files. The graph includes a line for the keyword dictionary, a line for the keyword postings, and a line for the keyword documents.]

Keywords: Inverted Index, Keyword Growth
Speaker Id Inverted Index File Growth

![Graph showing the growth of Speaker Id Index File Size with increasing number of files](image)

- Speaker Id Dictionary
- Speaker Id Postings
- Speaker Id Documents
Number of Files Vs Indexing Time

![Graph showing the relationship between the number of files and indexing time. The x-axis represents the number of files, and the y-axis represents time in seconds. The graph includes data points at 10, 446, 142, and 769.]
Number of Files Vs Retrieval Time

Number of Files Vs Retrieve Time

Retrieve Time (in Seconds)

Number of Files
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 ??