APPENDIX A
IMPLEMENTATION DETAIL

The proposed architecture for incremental parallel web crawler, shown in Figure 4.1 has been implemented. For downloading the web pages the developed software was executed at four systems in parallel. Three were used as client crawlers and one as MT server as shown in snapshot (Figure A.1). Kindly refer to Figure A.2 wherein a seed URL has been provided to initialize the system. The MT server on its own turn has assigned the URL to client crawler as shown in Figure A.5 for downloading the corresponding web page. The client crawler after downloading the web page for seed URL extracts all embedded URLs from it and sends back to MT server. MT server after calculating the ranking of uncrawled URLs sends them back to client crawlers. This process repeats till no more URLs are left to be downloaded. The snapshots of GUI (graphical user interface) for entire workings of MT server and client crawlers are as shown from Figure A.1 to A.5. In this work total 2.5 millions web pages has been downloaded.

A.1 DEVELOPMENT ENVIRONMENT

• **Software**

  *Operating System: Windows XP*
  *Development Platform: Net Beans 6.5*
  *Development Language: Java*
  *Database: SQL Server*

• **Hardware**

  *Processor: 1.5+ GHz*
  *RAM: 512 MB/1 GB*
  *Internet Bandwidth: 80Kbps (will affect run times accordingly)*
  *Hard Disk: Depends on the size of repository (i.e. number of web pages to be downloaded).*
A.2 USER INTERFACE SNAPSHOTS

Figure A.1: MT Server waiting for seed URL from user to start

Figure A.2: MT Server after initiated with seed URL from user to start
**Figure A.3:** MT Server after it is receiving URLs downloaded from Client Crawlers

**Figure A.4:** Client crawler waiting for URL from Server
A.2 CLASS DIAGRAM

The proposed architecture of incremental parallel WebCrawler is based on client server model. The client crawler’s side class diagram is shown in Figure A.1 where as MT server side class diagram is shown in Figure A.2 with the embedded functions. Following are the main classes and their functionality:

ClientCrawler: Contains GUI of the Client Crawler.
WebCrawler: Checks for the robot.txt, downloads the web page, parse the web page for links and extracts links from it.
Runnable: Runnable interface declares a single method: run() used for threads.
Mouselistener: It is used for an applet to detect left, right and double mouse clicks.
Connection: Checks for the IP address of the MT Server and connects to it if MT server has been started on that IP address.
Server: Initializes the connection between the master and client crawlers.
**Floating Array**: This class keeps all the in-built functions pertaining to float type array.

**String Array**: This class keeps all the in-built functions pertaining to String type array.

**Search**: Used to search a particular link saved in the database.

---

**Figure A.6**: Client Side Class Diagram
Figure A.7: Server Side Class Diagram
Figure B.1 and Figure B.2 represent the HTML source code of the web documents shown in Figure 5.5 and Figure 5.7 respectively. These documents were used to find out the structural changes in different versions of a web page. Similarly Figure B.3 to Figure B.9 represent the web pages used to compute the fingerprints for detecting content level changed and the results obtained thereof are shown in Table 5.7 and Table 5.9 respectively.

Figure B.1: HTML code of web page shown in Figure 5.5
Figure B.2: HTML code of web page shown in Figure 5.7
PM to visit Kashmir today, may invite separatists for talks

Security forces were put on alert ahead of Manmohan Singh’s visit to the Valley on Wednesday, during which he is likely to invite Kashmir separatists to join talks and also inaugurate Anantnag-Qazigund rail section.

- Separate inquiries ordered into youth’s killing in J&K
- Farooq blames cops for J&K youth’s death in custody
- Keep off Kashmir, India tells OIC

Dalits enter Tamil Nadu temple first time in 100 years

For the first time in a hundred years, dalits of Chettipulam village stepped inside the local Shiva temple on Tuesday, in a historic moment brought about by government and political intervention.

- Dalit families refused insurance scheme in TN
- Mob blocks Dalits’ entry into temple

Desi-videshi row may ground Air India

Pilots of erstwhile Indian Airlines have threatened to go on strike if expats get paid in full again, before the desis get their flying allowances — that account for 75%-85% of the total pay — of last two months.

- Blog: Satyam way for Air India?

SPORTS

- India look to draw level against Australia
  - Sunny hopes Sachin reaches 17K today
  - Not possible to bat up the order always: Dhoni
  - Brazilian star Robinho to quit Manchester City

INDIA

- Rahul may take on Mulayam in his den
  - Maoists change strategy, leave alone passengers
  - Terrorists not worthy of being called Muslims: Ansari
  - Defence supply: Pvt firms can partner foreign cos
Kabul siege ends, 6 UN staff among 9 killed

AGENCIES 1 hour ago
The United Nations says six of its staff members are dead after an early morning attack on a guest house in the Afghan capital. The attack has ended with the deaths of three Taliban gunmen, police said.

- US losing the Afghan plot?
- Karzai's brother said to receive payments from CIA
- 8 US troops killed in Afghanistan, toll rises to 833

PM to visit Kashmir today, may invite separatists for talks

TNN 3 hours ago
Security forces were put on alert ahead of Manmohan Singh's visit to the Valley on Wednesday, during which he is likely to invite Kashmiri separatists to join talks.

- Separate inquiries ordered into youth's killing in J&K
- Farooq blames cops for J&K youth's death in custody

Dalits enter Tamil Nadu temple first time in 100 years

V MAYILVAGANAN 4 hours ago
For the first time in a hundred years, dalits of Chettipalam village stepped inside the local Shiva temple on Tuesday, in a historic moment brought about by government and political intervention.

SPORTS

- India look to draw level against Australia
  - Sunny hopes Sachin reaches 17K today
  - Not possible to bat up the order always: Dhoni
  - Brazilian star Robinho to quit Manchester City

INDIA

- Ajit Pawar skips NCP portfolio meeting
  - Rahul Gandhi may take on Mulayam in his den
  - Maoists stop Rajdhani train, hold India to ransom
  - Terrorists not worthy of being called Muslims: Ansari

Figure B.4: Web page version 2 (28-Oct-12:00 Noon)
PM, Sonia arrive in Kashmir; separatists call bandh

There are high expectations from the PM’s visit which comes days after the announcement that the Centre would engage every shade of political opinion to find an honourable solution to the problems in J&K.

- Historic Kashmir railway link inaugurated by PM
- PM may invite separatists for talks
- File: Separate inquiries ordered into youth’s killing

6 UN staff among 10 killed in Kabul siege

Gunmen stormed a UN guest house in the Afghan capital early on Wednesday, killing 10 people, including six UN staff. Taliban claimed responsibility, saying it was an assault on the upcoming presidential election.

- US losing the Afghan plot?
- Karzai's brother said to receive payments from CIA

India, UK natural partners: President Patil

Drawing a parallel between the traditions of diversity and democracy embraced by India and Britain, Pratibha Patil on Tuesday said the two ‘natural partners’ are set to shape the course of the 21st century.

- Prince Philip in royal mess over Patel joke

India look to draw level against Australia

- Sunny hopes Sachin reaches 17K today
- Not possible to bat up the order always: Dhoni
- Brazilian star Robinho to quit Manchester City

Ajit Pawar skips NCP portfolio meeting

- Harmohinder Chatha is Haryana assembly speaker
- Rahul Gandhi may take on Mulayam in his den
- Maoists stop Rajdhani train, hold India to ransom
Peshawar blast kills 86 as Hillary Clinton visits Pak

AFP 28 minutes ago
Most of the dead were women and children out shopping in Peshawar’s Meena Bazaar. The blast coincides with Hillary Clinton’s trip, who is in Islamabad, a three-hour drive away.

- Hillary Clinton visits Pakistan
- Pakistanis too scared to leave home, want to flee
- India is funding Taliban fighters, claims Pakistan

PM reaches out to Pakistan with 'curb terror' rider

IANS 3 hours ago
Manmohan Singh reached out to Pakistan today amid persisting tensions between the neighbours over what New Delhi says is Islamabad’s unwillingness to bring to book the Mumbai terror attacks masterminds. [ ]

- PM, Sonia arrive in Kashmir; separatists calls bandh
- Historic Kashmir railway link inaugurated by PM

6 UN staff among 12 killed in Kabul siege

AP 2 hours ago
Gunmen stormed a UN guest house in the Afghan capital early on Wednesday, killing 12 people, including six UN staff. Taliban claimed responsibility, saying it was an assault on the upcoming presidential election.

- US losing the Afghan plot?

SPORTS

- Live: Dhoni slams 33rd ODI fifty
- Somdev Devvarman in Seoul Challenger quarters
- PCB chief Ijaz Butt off to India to revive ties
- Force India hope to sign off with points in Abu Dhabi

INDIA

- India, UK natural partners: President
- Ajit Pawar skips NCP portfolio meeting
- Rahul Gandhi may take on Mulayam in his don
- Maoists stop Rajdhani train, hold India to ransom
**Peshawar blast kills 86 as Hillary Clinton visits Pak**

＊AFP 55 minutes ago＊

Most of the dead were women and children out shopping in Peshawar’s Meena Bazaar. The blast coincides with Hillary Clinton’s trip who is in Islamabad, a three-hour drive away.

- Hillary Clinton visits Pakistan
- Pakistanis too scared to leave home, want to flee
- India is funding Taliban fighters, claims Pakistan

**Chinese participation in PoK projects illegal: India**

＊PTI＊

India has made clear its opposition to China’s participation in projects in Pakistan-occupied Kashmir, saying it treats any such activity as “illegal”.

- China’s claims on Arunachal absurd: CM Khandu
- PM stands up to China on Dalai Lama

**Gandhi did not have answer to environment problem: Amartya Sen**

＊SAIDAL DASGUPTA＊

Nobel laureate Amartya Sen told a conference in China that Gandhi was even against railways, while pointing out that the transportation industry was a crucial aspect of development.

- Found in US: Rare Gandhi recording

**SPORTS**

- Live: Dhoni slams 5th ODI hundred
- Somdev Devwarman in Seoul Challenger quarters
- PCB chief Ijaz Butt off to India to revive ties
- Force India hope to sign off with points in Abu Dhabi

**INDIA**

- CM Hooda wins trust vote
- PM reaches out to Pak with ‘curb terror’ rider
- India, UK natural partners: President
- Ajit Pawar skips NCP portfolio meeting

---

**Figure B.7:** Web page version 5 (28-Oct-6:15PM)
THE TIMES OF INDIA

Rajdhani hostage saga: Motive still a mystery
AJANTA CHAKRABORTY & SUKUMAR MAHATO 3 hours ago
The Bhubaneswar-New Delhi Rajdhani Express taken hostage by armed tribals at Bansiala Halt Station was no mere blockade. The move has laid bare the PCPA’s links with Maoists.
- Railways blames intelligence failure for Rajdhani mess
- Maoists abduct 2 train drivers in Bengal
- 'What Muslims were to BJP, Maoists are to Cong'

Centre plans to pave way for M F Husain’s early return
PTI 11 minutes ago
The 94-year-old painter, who has been living in exile for the last few years, wants to come back to India and had recently made an emotional appeal in this regard.
- Hindu outfit wants museum to remove Husain painting
- Art summit not to display Husain paintings

40 deemed varsities don’t meet norms
AKSHAYA MUKUL 30 minutes ago
HRD ministry and Kapil Sibal have kept the report on deemed universities by its own committee under wraps because the panel has said that over 40 universities do not deserve deemed status.
- No Indian university among world’s top 100

SPORTS
- 2nd ODI: India thrash Oz by 99 runs
- Blog: Team India has a perfect day
- Agassi admits taking drugs and lying to ATP
- Lensmen fight with Virat Kohli at Jamtha

INDIA
- 'No Army in Maoist-affected areas'
- Govt may consider law to deter honour killings
- One week on, still no govt in Maharashtra
- Karnataka infighting adds to BJP’s problems

Figure B.8: Web page version 6 (29-Oct-8:30AM)
Rajdhani hostage saga: Motive still a mystery
AJANTA CHAKRABORTY & SUKUMAR MAHATO 4 hours ago
The Bhubaneswar-New Delhi Rajdhani Express taken hostage by armed tribals at Banstala Halt Station was no mere blockade. The move has laid bare the PCPA’s links with Maoists.
- Probe Mamata role in trainjacking: CPM
- Railways blames intel failure for Rajdhani mess
- ‘What Muslims were to BJP, Maoists are to Cong’

Centre plans to pave way for M F Husain’s early return
PTI 1 hour ago
The 94-year-old painter, who has been living in exile for the last few years, wants to come back to India and had recently made an emotional appeal in this regard.
- Hindu outfit wants museum to remove Husain painting
- Art summit not to display Husain paintings

40 deemed varsities don’t meet norms
AKSHAYA MUKUL 1 hour ago
HRD ministry and Kapil Sibal have kept the report on deemed universities by its own committee under wraps because the panel has said that over 40 universities do not deserve deemed status.
- No Indian university among world’s top 100

SPORTS
- 2nd ODI: India thrash Oz by 99 runs
- Blog: Team India has a perfect day
- ‘I hate tennis’. Andre Agassi’s book adds to his tale
- Lensmen fight with Virat Kohli at Jamtha

INDIA
- Three militants killed in Jammu
- One week on, still no govt in Maharashtra
- Govt may consider law to deter honour killings
- ‘No Army in Maoist-affected areas’

Figure B.9: Web page version 7 (29-Oct-9:50AM)