WEB CRAWLING BASED DATA CONSOLIDATION

Project Report submitted in partial fulfillment of the requirement for the degree of

Bachelor of Technology

in

INFORMATION TECHNOLOGY

Under the Supervision of

Mrs. Sanjana Singh

By

Shubham Bajpai (Roll No: - 111464)

To

Jaypee University of Information and Technology

Waknaghat, Solan – 173234, Himachal Pradesh

CERTIFICATE

This is to certify that project report entitled ”Web Crawling based Data Consolidation”, submitted by Shubham Bajpai in partial fulfillment for the award of degree of Bachelor of Technology in Information Technology to Jaypee University of Information Technology, Waknaghat, Solan has been carried out under my supervision.

This work has not been submitted partially or fully to any other University or Institute for the award of this or any other degree or diploma.

Date: 08th May 2015

Supervisor’s Name: **Mrs. Sanjana Singh**

Designation: **Assistant Professor**

(Department of Computer Science and Engineering)

ACKNOWLEDGEMENT

I am extremely grateful to the Computer Science and Engineering Department of Jaypee University of Information Technology, Waknaghat, Solan for giving me an opportunity to perform the project from August, 2014 to May, 2015.
I take this opportunity to express my profound sense of gratitude & respect to everyone who help me throughout this project.

It is my radiant sentiment to express my sincerest regards to my project supervisor, **Mrs. Sanjana Singh** for her valuable inputs, able guidance, encouragement, whole-hearted cooperation and constructive criticism throughout the duration of my project.

I am indebted to **Prof. Dr. Satya Prakash Ghrera**, Head of Department Computer Science & Engineering for arranging the project program and inspiring us throughout the project work

I express my gratitude towards **Prof. Hemraj Saini,** (Project Coordinator - Computer Science & Engineering Department) for active guidance, project directive and his continuous interaction/monitoring till the completion of the project program.

I take this opportunity to sincerely thank all my lecturers who have directly or indirectly helped my project. I pay my respects and love to my parents,family members and friends for their love and encouragement throughout my career.

DATE: 08th May 2015

SHUBHAM BAJPAI
111464
Information Technology

Jaypee University of Information Technology, Solan

CONTENTS

List of Figures

ABSTRACT

While planning a software project one faces various troubles and need for guidance the obvious choice for such situations is looking up blog sites that can give us some insight into our problem. Today, very large amounts of information are available in online documents. As a part of the effort to better organize this information for users, researchers have been actively investigating the problem of automatic text categorization. Big data is an evolving term that describes any voluminous amount of structured, semi-structured and unstructured data that has the potential to be mined for information. The bulk of such work has focused on topical categorization, attempting to sort documents according to their subject matter (e.g., sports vs. politics).

Contemporary electronic commerce involves everything from ordering "digital" content for immediate online consumption, to ordering conventional goods and services, to "meta" services to facilitate other types of electronic commerce. India has an internet user base of about 250.2 million as of June 2014. The penetration of e-commerce is low compared to markets like the United States and the United Kingdom but is growing at a much faster rate with a large number of new entrants.

The ever-increasing popularity of websites because of e-commerce (online portals) has grown widely in few recent years. The merchants have even emerged with new ideas to provide some new suggestion to the customers. Instead of selling products to the customers, they prefer to suggest them the right websites to buy the best deals.

Moreover, the merchants are trying to work upon their recommender system so as to efficiently touch the user’s choice - the type of search queries and to recommend the user with the variety of items related to it, thus saving the valuable time and attracting interest in the items of products thereby making its user to surf more. The overall propaganda behind the procedure is to provide users with the best in least time. It would be as simple to say convenient shopping amidst the stressful life. Moreover, what is common among every web portals is the agenda of leaving decision in the minds of its customers.

Chapter 1: Introduction

1.1 Problem Statement

Adhering to this idea I aim **To Design a web based platform that consolidates data from various websites at a single location, to suggest users right gifts based on various parameters.**

The parameters considered would be

- Occasion
- Relationship
- Age

Major Parameters

- Personality
- Mood
- Interests
- Price
- Delivery Time

The parameters mean the topical category which would be considered widely in suggesting the gifts to users. As mentioned above the gifts will be recommended on the basis of occasion – the type of event or function of incident and the mood of the function, the relationship with the user- means the relation the user possess with the person to whom he has to give gift. The age of person gifted. The Personality and the area of interests if any of the person gifted. This can be linked with the social networking sites like Facebook and Twitter just to gather sufficient information for the feasibility of user. The price parameter will be again an unimportant one but subdivided into the range of price distribution where the categories of products shall be filtered according to the range of price predetermined.

1.2 A brief Overview

Gift-giving and receiving is something that all of us indulge in regardless of culture. While we may have given and received gifts all our lives, we may not be fully aware of the spiritual implications of giving and receiving gifts. In this article, we explain the various aspects of giving and receiving gifts, from a spiritual perspective.

However, as per the spiritual law of Karma, when we give or accept gifts:

- We are either creating or settling give-and-take accounts with the other person.
- We incur either merits or demerits, depending on the type of gift we are giving and the intention behind.

When we give or receive gifts we are either creating or settling give-and-take accounts with each other
© Spiritual Science Research Foundation

Source: http://www.spiritualresearchfoundation.org

Fig: Importance of Gifts

1.3 Related Area of Work

i. **JUNGLEE – Anand Rajaraman (Cofounder) in 1996**

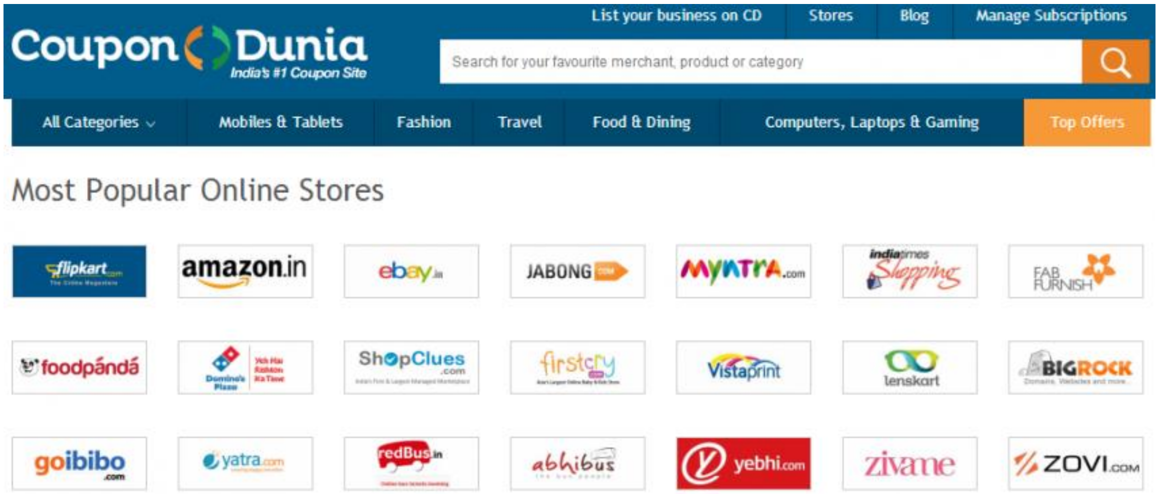
Junglee.com is an online shopping service by Amazon that suggests its customers with the list of all the stores to buy suitable products at the best price by comparing the deals on various parameters as per the decision of the customer.

Source: http://blog.junglee.com

Fig: JUNGLEE Logo

ii. **COUPONDUNIA MEDIA PRIVATE LIMITED - Sameer Parwani (CEO)**

Coupondunia.com is a coupon website which distribute coupons that only can be used for online shopping to also distribute coupons that can be used offline. The combined service's websites, mobile applications, e-mail newsletters alerts and social media presence will enable consumers to search for, discover, and source the best offers from leading retailers and brands. For virtually every e-commerce store in India, the website list coupons and offers that can be availed to save on a purchase. The websites directly contracts with the other companies and provide coupons to consumer so as to save their money. All coupons and offers are free. Everyone loves saving money so the response we've received in the market has been excellent. Coupondunia.com boasts to be the most popular coupon website in India due to its focus on helping people save money.



Source: http://www.coupondunia.com

Fig: COUPONDUNIA Logo

iii. **PINTEREST - Ben Silbermann, Paul Sciarra and Evan Sharp (Founders)**

Pinterest is a web and mobile application company that offers a visual discovery, collection, sharing, and storage tool. Users create and share the collections of visual bookmarks (boards). Boards are created through a user selecting an item, page, website, etc. and pinning it to an existing or newly created board.

Source: http://in.pinterest.com

Fig: PINTEREST Logo

Users save and share pins from multiple resources onto boards based on a plethora of criteria, e.g., similar characteristics, a theme, birthday parties, planning a vacation, writing a book, interior decorating, holidays. Boards can develop projects, organize events, or save pictures and data together. Pinterest acts as a personalized media platform.

iv. TRIPADVISOR MEDIA GROUPS PRIVATE LIMITED - Stephen Kaufer

TripAdvisor is an American travel website providing reviews of travel-related content. It also includes interactive travel forums. TripAdvisor aims to provide the best hotels and places to visit to its customers who make a search to visit in the new places. TripAdvisor also feature user-generated opinions which has led to an abundance of customer reviews that are often too numerous for a user to read. Consequently, there is a growing need for systems that are able to automatically extract, evaluate and present opinions in ways that are both helpful and easy for a user to interpret.

Source: http://tripadvisor-warning.com

Fig: TRIPADVISOR Logo

1.4 My Area of work

After looking at the related area of work, an interests to work in the same domain field concentrating upon the suggesting proper gifts. A web site that can help consumers to redirect to the pages of the various websites that provide the services related to the product. The user has to just enter the parameters details on the homepage. The parameters are the major parameters (Occasion, Age and Relationship) already described in the list of parameters. The query ones send to the crawler crawls the webpages of the various websites and then display the best – top ranked data to the user. The main objective is to let user find the “Simplest way to find right gifts, at the best portal”

Simplest way to find right gifts, at the best portal

We know that this era unlike any, is faced with explosive growth in the size of data generated or captured. Data growth has undergone a renaissance, influenced primarily by ever cheaper computing power and the ubiquity of the internet. This has led to a paradigm shift in the E-commerce sector; as data is no longer seen as the byproduct of their business activities, but as their biggest asset providing: key insights to the needs of their customers, predicting trends in customer’s behavior, democratizing of advertisement to suits consumers varied taste, as well as providing a performance metric to assess the effectiveness in meeting customers’ needs.

With the rapid evolution of e-commerce and increasing number of internet users since the few years the area of my work would be on crawling and consolidating the various gifts website and creating a web based platform to help the users to find the right gifts at the right place. Normally, any e-commerce website focusses upon creating its own new retail store and then managing it only.

In this project, all the concentration will be on the managing the other retail stores and then categorizing them in a fashion that through one login portal or from one application platform the user can access various gifts on the different sites. It is just the same that the customer now prefers to buy all the products at the same place or shop rather than going to different shops for different categories of products. The best example of this is Malls and Multiplexes.

The web based application will provide the same benefit to the users to access all the gifts at the one place. The gifts will not be from one website only but from the different websites which will be considered and discussed in the next heading. The users will also have advantage of finding the best among the available options from different websites. Moreover, the users can also compare the two similar products at different websites based on their pricing, user ratings and reviews.

1.5 Approach to the Problem

There are four different websites that are being worked upon in the project. The websites are as follows:

i. www.indiangiftsportal.com
ii. www.archiesonline.com
iii. www.giftease.com
iv. www.giveter.com

Once the gifts are suggested to the users, the users can compare the gifts at all the websites considered for crawling and consolidation and hence will be redirected to their page for the further purchase. The difference it holds from the other web portals is that it shows all the products from multiple websites at the same page for similar parameters entered by the user.

A Search Engine is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as search engine results pages (SERPs). The information may be a mix of web pages, images, and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain real-time information by running an algorithm.

The overall project uses the concepts of

- **Web Crawling**
- **Data Consolidation**
- **Pruning of Data**
- **Page Rankings**
- **Different types of Login - Login through Facebook (Birthday Calendars and Reminders), Login through email (single login account for the different websites)**

Focusing on the login account. The gift portal will be enabled using **social login**, also known as **social sign-in**, is a form of single sign-on using existing login information from a social networking service such as Facebook to sign into a third party website in lieu of creating a new login account specifically for that website. It is designed to simplify logins for end users as well as provide more and more reliable demographic information to web developers.

The social login platforms would also enable users to get the details like (birthdays of the friends on the friend list on Facebook account) and then let the user suggest gifts on the basis of their personalities (likes and dislikes). The idea of birthday calendar is the main motive behind its implementation. Moreover, the social login users will also lead to enable e-commerce referrals and interaction among users regarding the various products shown on the websites.

The crawling methods discussed from the next page is of the Google Bot that how the Google handles its search query and the bot crawls the webpages and fetches the links according to the different algorithms.

The tools that would be involved in implementing the above would be:

• XAMPP version 5.5.19
• Adobe Dreamweaver CS4
• Technologies used would be:
• PHP
• MYSQL
• JAVASCRIPT (used implicitly by JOOMLA)
• HTML

Features like APACHE and MYSQL are embedded in XAMPP. Apache enables coding in PHP and MYSQL would be used for handling database content. PHP is a server-side scripting language which is used for making dynamic web pages. Moreover, PHP allows database connectivity with no hassles, thus this would make handling huge databases relatively easier. Finally, HTML would be used to design the client side application.

1.5.1 Important Points regarding the implementation

- The interface design and display of the content would be done using the HTML and its looks and formatting would be done using the CSS.

- The image slider on the homepage will be made in the java script is made using the Java Script and the functions in Jscript. Its formatting will be again done through CSS.

- The other backend coding of the crawling technique, pruning of links and their consolidation will be done using the PHP (Preprocessor Hypertext)

i. As we generally see on any website that after the search query is send to the crawler or the search engine

SEARCH STARTS WITH THE WEB. IT'S MADE UP OF OVER

60 TRILLION

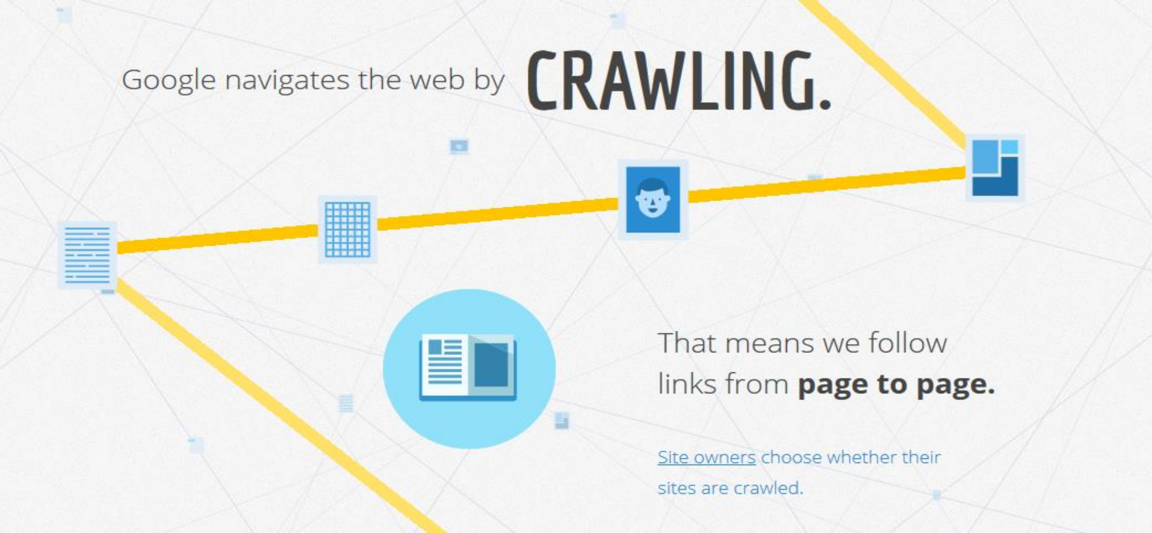
INDIVIDUAL PAGES

AND IT'S CONSTANTLY GROWING.

Source: http://www.google.co.in/insidesearch/howsearchworks

Fig: Big Data on the World Wide Web

ii. Search Engine navigates the webpages using the crawlers.



Google navigates the web by CRAWLING.

That means we follow links from **page to page.**

Site owners choose whether their sites are crawled.

Source: http://www.google.co.in/insidesearch/howsearchworks

Fig: Searching is done by Crawling

iii. Keep track of all the webpages in the index. Perform pruning of the page links. Pruning involves

- Remove the “anchor” or “reference” part of the URL.
- Use heuristics to recognize default web pages.
- Remove “.”, “. //”,”//”,”#” and its parent directory from the URL path.

iv. Programs and Algorithms to deliver the best search results.

- Search Methods
- Autocomplete or Spelling
- Synonym
- Query Understanding

Google
YAHOO!
msn
alltheweb
Your Site
altavista
dmoz
LYCOS
overture

Source: http://www.techiemania.com

Fig: Data Consolidation

v. Based on the results, pull the relevant webpages from the index and rank them.

vi. Fight the spams.

1.5.2 Importance of Gifting

- An expression of love or friendship.
- An expression of gratitude for a gift received.
- An expression of piety, in the form of charity.
- An expression of solidarity, in the form of mutual aid.
- To share wealth.
- To offset misfortune.

1.5.3 Discussion on the basis of Parameters

Major Parameters

- Categorization on the basis of Occasions (often celebrations):
  - Anniversary (the wedding anniversary of couples may receive gifts).
  - Birthday (the person having birthday gives cake, etc. and receives gifts).
  - Wedding of the couple (includes gift for bride, groom and both).
  - Romantic Gifts (just to glorify the romance among the lovers).
  - House warming gifts (signifying the happy occasion of a new house).

- Categorization on the basis of Age:
    - Babies (0-2 years)
    - Kids (2-4 years)
    - Youth
    - Young Married
    - Seniors

- Categories on the basis of Relationships

  MEN

    - Grandfather
    - Father
    - Husband
    - Brother
    - Son
    - Boyfriend
    - Friend: Male

  WOMEN

    - Grandmother
    - Mother
    - Wife
    - Sister
    - Daughter
    - Girlfriend
    - Friend: Female

Some other Parameters to be considered

- Categorization on Price
  - Price Range
  - Page Ranking according to price (Low to High, High to Low)

- Categorization on the Delivery Time
  - Delivery within Today
  - Delivery within Tomorrow
  - Delivery at 00:00 hours

The delivery categorization will be based on the location factors because it would be tremendously difficult to reach the outreached area in the limited area. Only the remote location can be accessed easily.

- Categorization on the basis of some other Occasion
  - Father's Day (the father receives gifts.)
  - Mother's Day (the mother receives gifts.)
  - Retirement Gifts
  - Congratulations Gifts
  - Engagement Gifts
  - Valentine’s Day Gifts
  - Farewell
  - Hindus give Diwali, Holi, Rakhi and Pongal gifts to family and friends.
  - Muslims give gifts to family and friends, known as EId, on Eid al-Fitr (the end of Ramadan) and on Eid al-Adha.

    - Christmas (People give one another gifts, often supposedly receiving them from Santa Claus, the Christ child or Saint Nicholas.)
    - Feast of Saint Nicholas (People give each other gifts, often supposedly receiving them from Saint Nicholas.)
    - Easter baskets with chocolate eggs, jelly beans, and chocolate rabbits are gifts given on Easter.
    - Greek Orthodox Christians in Greece, will give gifts to family and friends on the Feast of Saint Basil.
    - Jews gives Hanukkah gifts to family and friends.

- Categorization on the basis of Personality

    - Area of Interests
    - Size (Height parameters, Figure)
    - Particular Likes (can be linked with the Facebook account)

Like every e-commerce portal's search engines rely on web crawlers to feed content into various indexing and analysis layers, which in turn feed a ranking layer that handles user search queries. Therefore, we discuss the web crawler in the next chapter.

1.6 Motivation

- The ever increasing popularity of websites that feature user generated opinions. (The examples e commerce websites that emerged as a way of providing something new to the people in different areas of the industries are shown above)

- Offering gifts is a tradition followed since years and will continue- sign of donation, love and compliments.

- People who make social commerce click- sellmojo. Sellmojo is a unique application that enables merchants to create their very own storefront, on their Facebook page, where Facebook's 1 billion users can discover products and shop without ever leaving the social network.

- Difficult to decide the gift considering various parameters like the type of occasion, the age of the person who has to be gifted, the relationship with the person.

- More options are accessible taking less time.

- Entrepreneurs should know the opinions and demands of the consumers, to know the needs, wants and the demands of the customers and their area of interest.

- My own area of interests in developing a system that follows recent trends.

Chapter 2: Concepts

2.1 Web Crawling

A *web crawler* (also known as a *spider*) is an Internet bot that systematically browses the World Wide Web, typically for the purpose of Web Indexing. Web crawlers are used for a variety of purposes. Most prominently, they are one of the main components of Web search engines, systems that assemble a corpus of web pages, index them, and allow users to issue queries against the index and find the web pages that match the queries.

The crawling layer has two responsibilities: down loading new pages, and keeping previously downloaded pages fresh. The most well-known crawler is called “Googlebot.”

40% of the Internet traffic is due to the web crawlers.

WWW contains millions of information beneficial for the users, many information seekers usage search engine to initiate their Web activity. Search engines rely on a crawler module to provide the grist for its operation, **Matthew Gray** wrote the first Crawler, the World Wide Web Wanderer, which was used from 1993 to 1996. Web Crawlers describe how the search engines should cope with the evolving Web, in an attempt to provide users with up-to-date results. There exists various studies on crawler policies discussed below which proposes how one can maintain local copies of remote data sources “fresh," when the source data is updated autonomously and independently.

The behavior of a Web crawler is the outcome of a combination of policies:

- A *selection policy* that states which pages to download,
- A *re-visit policy* that states when to check for changes to the pages,
- A *politeness policy* that states how to avoid overloading Web sites, and
- A *parallelization policy* that states how to coordinate distributed web crawlers.

2.1.1 Uses of Web Crawlers

Crawlers look at webpages and follow links on those pages, much like you would if you were browsing content on the web. They go from link to link and bring data about those webpages back to Google's servers.

i. Web Crawlers are used to discover public available webpages.

ii. Web Indexing: - Web indexing (or Internet indexing) refers to various methods for indexing the contents of a website or of the Internet as a whole.

iii. Web archiving: - A service provided by Internet archiving where large sets of web pages are periodically collected and archived for posterity.

iv. Web data mining: In web data mining the different web pages are analyzed for statistical properties, or where data analytics is performed on them.

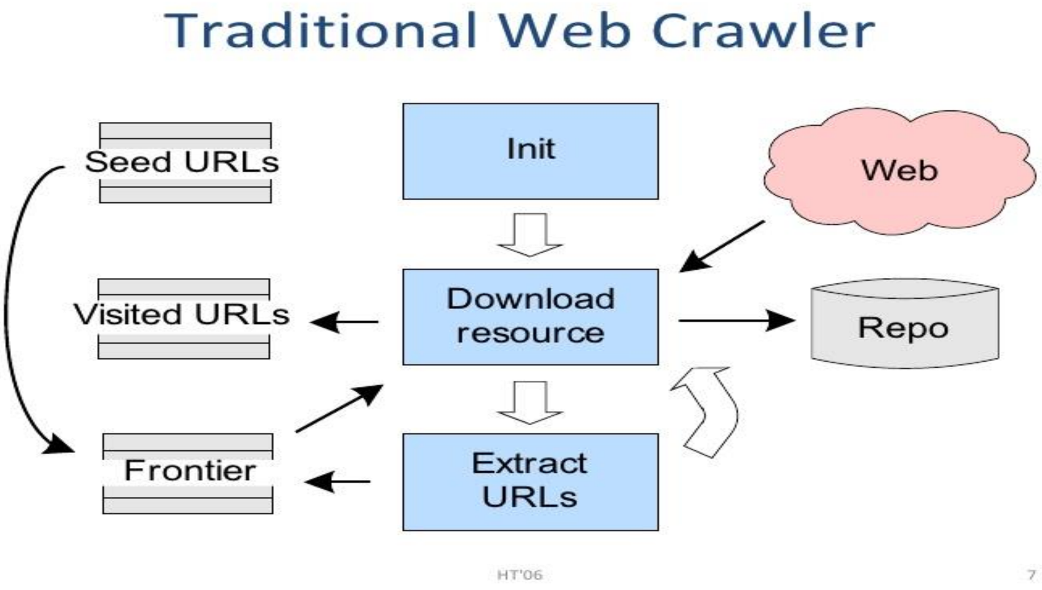
v. Web crawlers can copy all the visited pages by a search engine that indexes the downloaded pages so that the users can search them more quickly.

2.1.2 How do Web Crawler's Work?

i. Crawler starts with a list of URL's to visit – called the seeds.
ii. Identify all the hyperlinks in the page.
iii. Fetch the hyperlinks.
iv. Remove the inappropriate symbols and undesirable content from the hyperlinks.
v. Add the hyperlinks to the list of URLs to visit – called the crawl frontier.
vi. URLs from the frontier are recursively visited.

As the hyperlinks in the seed page are fetched and parsed, the extracted URL's are placed in a data structure (Queue) which I called the crawl frontier. The URL's are revisited from this Queue itself using First in First Out implementation.

- If the crawler is performing archiving of websites it copies and saves the information as it goes.
- Such archives are usually stored such that they can be viewed, read and navigated as they were on the live web, but are preserved as 'snapshots'.
- The large volume implies that the crawler can only download a limited number of the Web pages within a given time from a given web page, so it needs to prioritize its downloads.
- The high rate of change implies that the pages might have already been updated or even deleted.



Source: http://image.slidesharecdn.com/webcrawler

Fig: Traditional Web Crawler

2.1.3 Types of Web Crawlers

i. **Focused Crawler: -** Focused Crawler is also known as a **Topic Crawler** that tries to download pages that are related to each other. It collects documents which are specific and relevant to the given topic. The focused crawler determines the following – Relevancy, Way forward. It determines how far the given page is relevant to the particular topic and how to proceed forward. The benefits of focused web crawler is that it is economically feasible in terms of hardware and network resources, it can reduce the amount of network traffic and downloads.

ii. **Incremental Crawler: -** A traditional crawler, in order to refresh its collection, periodically replaces the old documents with the new documents. On the contrary, an incremental crawler incrementally refreshes the existing collection of pages by visiting them frequently; based upon the estimate as to how often pages change. It also exchanges less important pages by new and more important pages. The benefit of incremental crawler is that only the valuable data is provided to the user, thus network bandwidth is saved and data enrichment is achieved.

iii. **Distributed Crawler: -** In Distributed web crawling many crawlers are working to distribute in the process of web crawling, in order to have the most coverage of the web. A central server manages the communication and synchronization of the nodes, as it is geographically distributed. It basically uses Page rank algorithm for its increased efficiency and quality search. The benefit of distributed web crawler is that it is robust against system crashes and other events, and can be adapted to various crawling applications.

iv. **Parallel Crawler: -** A parallel crawler consists of multiple crawling processes (running in parallel) called as C-procs which can run on network of workstations. The Parallel crawlers depend on Page freshness and Page Selection. A Parallel crawler can be on local network or be distributed at geographically distant locations. Parallelization of crawling system is very vital from the point of view of downloading documents in a reasonable amount of time.

2.1.4 Page Fetching

- Clients need to have timeouts so that unnecessary amount of time is not spent on slow servers or in reading large pages.
- Frontier may be implemented using a queue. A new unvisited URL is added to the tail and the head will point to the URL next to be crawled.
- Error Checking and Exception Handling are important during the fetching process since we need to deal with millions of remote servers using the same code.
- Remote Exclusion Protocol helps server administrator to communicate their file access policies to different user agents.

2.1.5 Parsing

- Convert the protocol and hostname to lowercase.
- Remove the “anchor” or “reference” part of the URL.
- Use heuristics to recognize default web pages.
- Remove “.”, “. //”,”//”,”#” and its parent directory from the URL path.
- Leave the port numbers in the URL unless it is port 80.
- Stop listing and Stemming.

2.1.6 Choice of Website owners

For a certain website, site owners have many choices about how Google crawls and indexes their sites through Webmaster Tools and a file called “**robots.txt**”. With the robots.txt file, site owners can choose not to be crawled by Googlebot, or they can provide more specific instructions about how to process pages on their sites. Site owners have granular choices and can choose how content is indexed on a page-by-page basis. For example, they can opt to have their pages appear without a snippet (the summary of the page shown below the title in search results) or a cached version (an alternate version stored on Google’s servers in case the live page is unavailable).

2.1.7 Crawling in My Area of Work

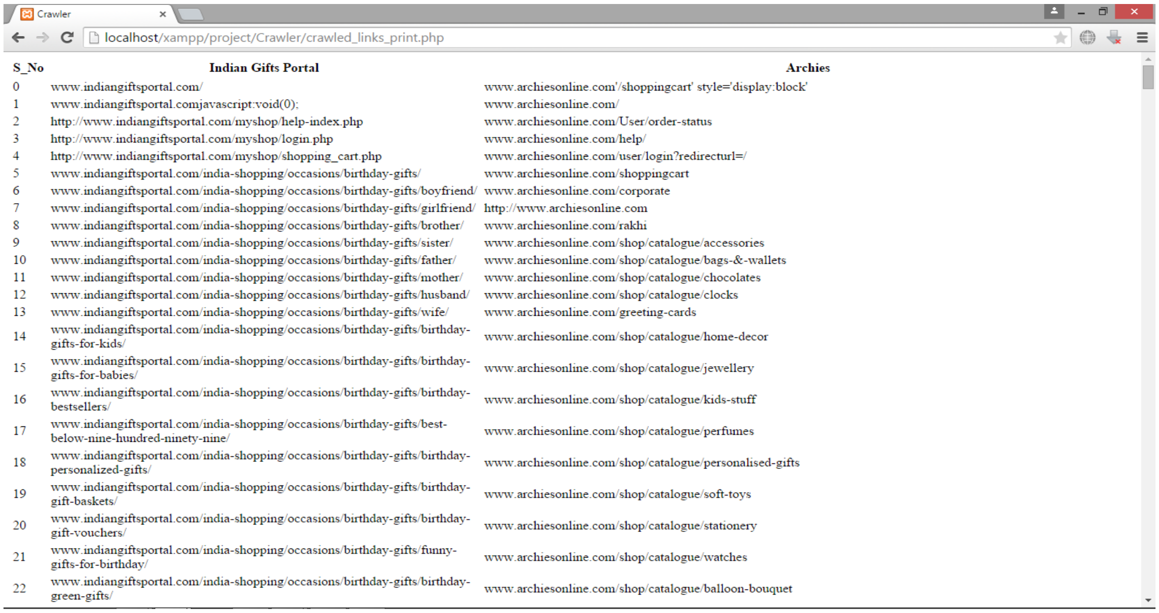
In the project, Focused Crawler is implemented using PHP. The crawler crawls all the four websites considered (www.indiangiftsportal.com, www.archiesonline.com, www.giftease.com, www.giveter.com), fetches and parses all the links as discussed above. The web pages links are stored in the different queues implemented using an array. The parsing of the link has been performed in the crawling function itself. Once the parsing is done the useful links are extracted from the page source and then stored in the frontier (queue). These links will be fetched from the same queue for further usage.

Similarly, a crawler function for extracting image links is also implemented. The image crawler will extract all the images on the webpage under <img src> tag holding extensions like .jpg/.jpeg, .png, .bmp. The image links after parsing are stored in separate arrays for all the four different websites. So, overall there are eight arrays four for the links for each website and the other four arrays for the image links for each website.

No database is created for the storage of links as the crawler performs live crawling. Once the database has an information about any product and then at certain stage the product is not available with the company or the information (price, quality, color, URL) of the product has been changed by the company then the database has to be updated every time. Instead, arrays will fetch and only add those links that are crawled recently. So the server takes less time to load and the array is updated with recent URL's.

The crawler function should be kept in the simplest format. Harder and more complex crawler create troubles in loading when hosted on server. They make the server slow. The images are not loaded properly. The text formatting also goes distorted and deformed contents on the webpages. Moreover, the complex crawler also increases the loading execution time which may lead to execution time out. The default execution time of a localhost server is 30 seconds. If the crawler takes more than 30 seconds to execute there might be a probability that all the links from all the 4 websites are not fetched from the webpage and hence this will create problems for future work of consolidation.

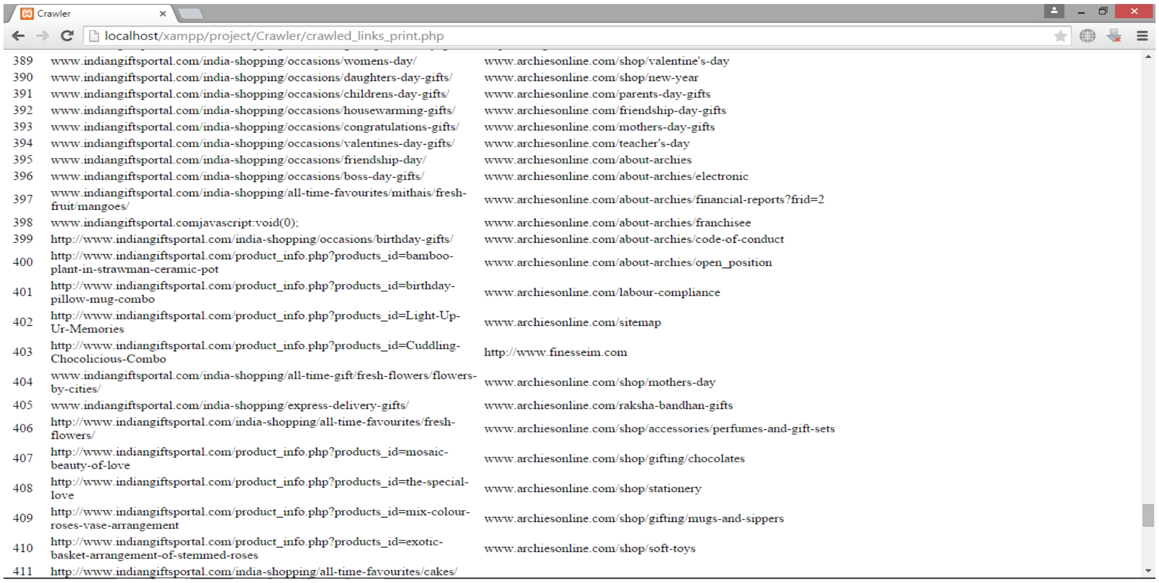
The screenshot of the crawled links from the four different websites (i.e. www.indiangiftsportal.com, www.archiesonline.com, www.giftease.com, www.giveter.com) are shown below:



S_No	Indian Gifts Portal	Archies
0	www.indiangiftsportal.com/	www.archiesonline.com'/shoppingcart' style='display:block'
1	www.indiangiftsportal.comjavascript:void(0);	www.archiesonline.com/
2	http://www.indiangiftsportal.com/myshop/help-index.php	www.archiesonline.com/User/order-status
3	http://www.indiangiftsportal.com/myshop/login.php	www.archiesonline.com/help/
4	http://www.indiangiftsportal.com/myshop/shopping_cart.php	www.archiesonline.com/user/login?redirecturl=/
5	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/	www.archiesonline.com/shoppingcart
6	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/boyfriend/	www.archiesonline.com/corporate
7	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/girlfriend/	http://www.archiesonline.com
8	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/brother/	www.archiesonline.com/rakhi
9	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/sister/	www.archiesonline.com/shop/catalogue/accessories
10	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/father/	www.archiesonline.com/shop/catalogue/bags-&-wallets
11	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/mother/	www.archiesonline.com/shop/catalogue/chocolates
12	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/husband/	www.archiesonline.com/shop/catalogue/clocks
13	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/wife/	www.archiesonline.com/greeting-cards
14	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gifts-for-kids/	www.archiesonline.com/shop/catalogue/home-decor
15	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gifts-for-babies/	www.archiesonline.com/shop/catalogue/jewellery
16	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-bestsellers/	www.archiesonline.com/shop/catalogue/kids-stuff
17	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/best-below-nine-hundred-ninety-nine/	www.archiesonline.com/shop/catalogue/perfumes
18	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-personalized-gifts/	www.archiesonline.com/shop/catalogue/personalised-gifts
19	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gift-baskets/	www.archiesonline.com/shop/catalogue/soft-toys
20	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gift-vouchers/	www.archiesonline.com/shop/catalogue/stationery
21	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/funny-gifts-for-birthday/	www.archiesonline.com/shop/catalogue/watches
22	www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-green-gifts/	www.archiesonline.com/shop/catalogue/balloon-bouquet

Source: http://localhost/xampp/project/Crawler/crawled_links_print.php

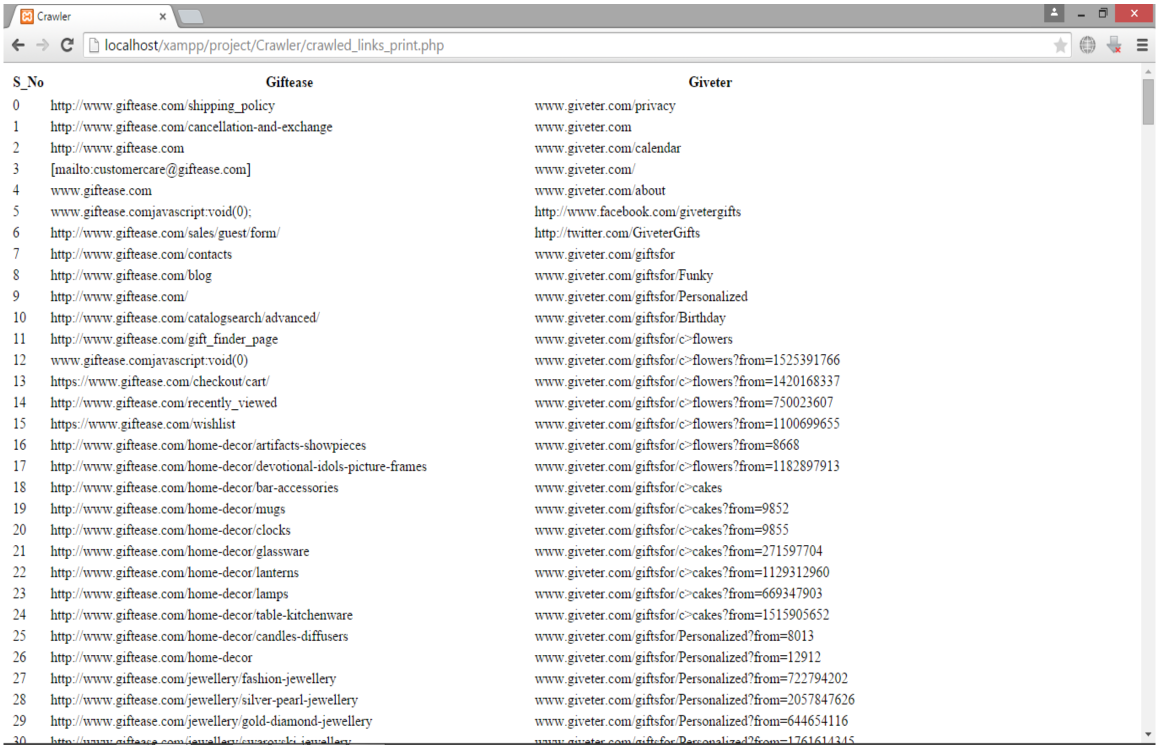
Fig: Screenshot 1 of the Crawled links from Indian gifts portal and Archies websites



389	www.indiangiftsportal.com/india-shopping/occasions/womens-day/	www.archiesonline.com/shop/valentine's-day
390	www.indiangiftsportal.com/india-shopping/occasions/daughters-day-gifts/	www.archiesonline.com/shop/new-year
391	www.indiangiftsportal.com/india-shopping/occasions/childrens-day-gifts/	www.archiesonline.com/parents-day-gifts
392	www.indiangiftsportal.com/india-shopping/occasions/housewarming-gifts/	www.archiesonline.com/friendship-day-gifts
393	www.indiangiftsportal.com/india-shopping/occasions/congratulations-gifts/	www.archiesonline.com/mothers-day-gifts
394	www.indiangiftsportal.com/india-shopping/occasions/valentines-day-gifts/	www.archiesonline.com/teacher's-day
395	www.indiangiftsportal.com/india-shopping/occasions/friendship-day/	www.archiesonline.com/about-archies
396	www.indiangiftsportal.com/india-shopping/occasions/boss-day-gifts/	www.archiesonline.com/about-archies/electronic
397	www.indiangiftsportal.com/india-shopping/all-time-favourites/mithais/fresh-fruit/mangoes/	www.archiesonline.com/about-archies/financial-reports?frid=2
398	www.indiangiftsportal.comjavascript:void(0);	www.archiesonline.com/about-archies/franchisee
399	http://www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/	www.archiesonline.com/about-archies/code-of-conduct
400	http://www.indiangiftsportal.com/product_info.php?products_id=bamboo-plant-in-strawman-ceramic-pot	www.archiesonline.com/about-archies/open_position
401	http://www.indiangiftsportal.com/product_info.php?products_id=birthday-pillow-mug-combo	www.archiesonline.com/labour-compliance
402	http://www.indiangiftsportal.com/product_info.php?products_id=Light-Up-Ur-Memories	www.archiesonline.com/sitemap
403	http://www.indiangiftsportal.com/product_info.php?products_id=Cuddling-Chocolicious-Combo	http://www.finesseim.com
404	www.indiangiftsportal.com/india-shopping/all-time-gift/fresh-flowers/flowers-by-cities/	www.archiesonline.com/shop/mothers-day
405	www.indiangiftsportal.com/india-shopping/express-delivery-gifts/	www.archiesonline.com/raksha-bandhan-gifts
406	http://www.indiangiftsportal.com/india-shopping/all-time-favourites/fresh-flowers/	www.archiesonline.com/shop/accessories/perfumes-and-gift-sets
407	http://www.indiangiftsportal.com/product_info.php?products_id=mosaic-beauty-of-love	www.archiesonline.com/shop/gifting/chocolates
408	http://www.indiangiftsportal.com/product_info.php?products_id=the-special-love	www.archiesonline.com/shop/stationery
409	http://www.indiangiftsportal.com/product_info.php?products_id=mix-colour-roses-vase-arrangement	www.archiesonline.com/shop/gifting/mugs-and-sippers
410	http://www.indiangiftsportal.com/product_info.php?products_id=exotic-basket-arrangement-of-stemmed-roses	www.archiesonline.com/shop/soft-toys
411	http://www.indiangiftsportal.com/india-shopping/all-time-favourites/cakes/	

Source: http://localhost/xampp/project/Crawler/crawled_links_print.php

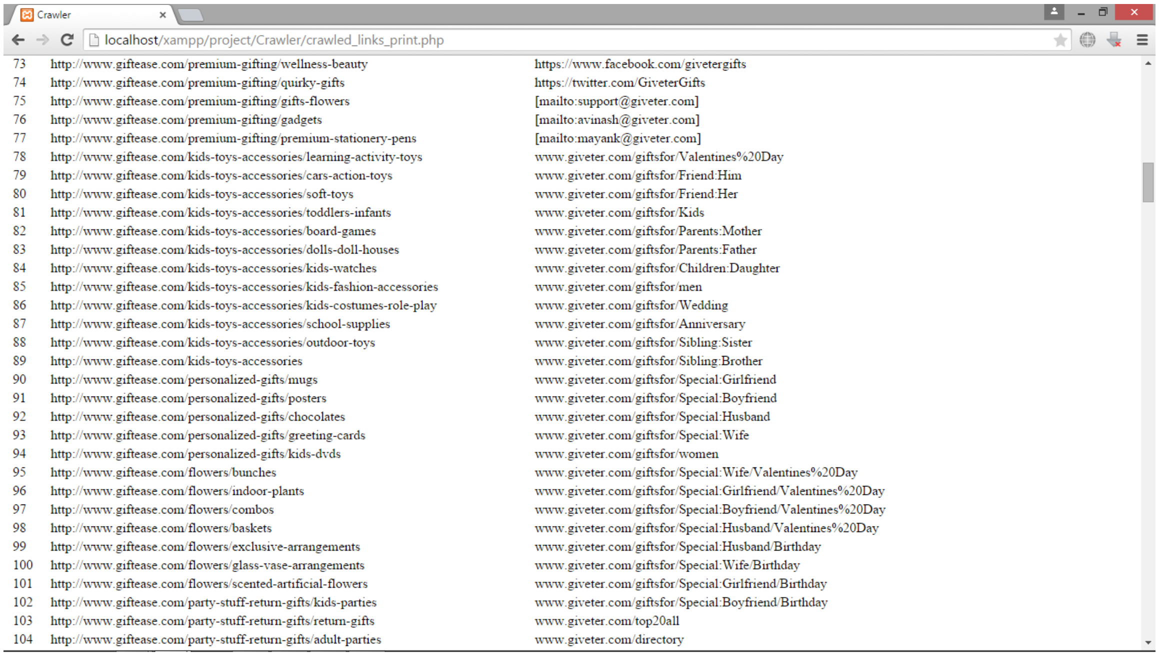
Fig: Screenshot 2 of the Crawled links from Indian gifts portal and Archies websites



S_No	Giftease	Giveter
0	http://www.giftease.com/shipping_policy	www.giveter.com/privacy
1	http://www.giftease.com/cancellation-and-exchange	www.giveter.com
2	http://www.giftease.com	www.giveter.com/calendar
3	[mailto:customercare@giftease.com]	www.giveter.com/
4	www.giftease.com	www.giveter.com/about
5	www.giftease.comjavascript:void(0);	http://www.facebook.com/givetergifts
6	http://www.giftease.com/sales/guest/form/	http://twitter.com/GiveterGifts
7	http://www.giftease.com/contacts	www.giveter.com/giftsfor
8	http://www.giftease.com/blog	www.giveter.com/giftsfor/Funky
9	http://www.giftease.com/	www.giveter.com/giftsfor/Personalized
10	http://www.giftease.com/catalogsearch/advanced/	www.giveter.com/giftsfor/Birthday
11	http://www.giftease.com/gift_finder_page	www.giveter.com/giftsfor/c>flowers
12	www.giftease.comjavascript:void(0)	www.giveter.com/giftsfor/c>flowers?from=1525391766
13	https://www.giftease.com/checkout/cart/	www.giveter.com/giftsfor/c>flowers?from=1420168337
14	http://www.giftease.com/recently_viewed	www.giveter.com/giftsfor/c>flowers?from=750023607
15	https://www.giftease.com/wishlist	www.giveter.com/giftsfor/c>flowers?from=1100699655
16	http://www.giftease.com/home-decor/artifacts-showpieces	www.giveter.com/giftsfor/c>flowers?from=8668
17	http://www.giftease.com/home-decor/devotional-idols-picture-frames	www.giveter.com/giftsfor/c>flowers?from=1182897913
18	http://www.giftease.com/home-decor/bar-accessories	www.giveter.com/giftsfor/c>cakes
19	http://www.giftease.com/home-decor/mugs	www.giveter.com/giftsfor/c>cakes?from=9852
20	http://www.giftease.com/home-decor/clocks	www.giveter.com/giftsfor/c>cakes?from=9855
21	http://www.giftease.com/home-decor/glassware	www.giveter.com/giftsfor/c>cakes?from=271597704
22	http://www.giftease.com/home-decor/lanterns	www.giveter.com/giftsfor/c>cakes?from=1129312960
23	http://www.giftease.com/home-decor/lamps	www.giveter.com/giftsfor/c>cakes?from=669347903
24	http://www.giftease.com/home-decor/table-kitchenware	www.giveter.com/giftsfor/c>cakes?from=1515905652
25	http://www.giftease.com/home-decor/candles-diffusers	www.giveter.com/giftsfor/Personalized?from=8013
26	http://www.giftease.com/home-decor	www.giveter.com/giftsfor/Personalized?from=12912
27	http://www.giftease.com/jewellery/fashion-jewellery	www.giveter.com/giftsfor/Personalized?from=722794202
28	http://www.giftease.com/jewellery/silver-pearl-jewellery	www.giveter.com/giftsfor/Personalized?from=2057847626
29	http://www.giftease.com/jewellery/gold-diamond-jewellery	www.giveter.com/giftsfor/Personalized?from=644654116
30	http://www.giftease.com/jewellery/swarovski-jewellery	www.giveter.com/giftsfor/Personalized?from=1761614345

Source: http://localhost/xampp/project/Crawler/crawled_links_print.php

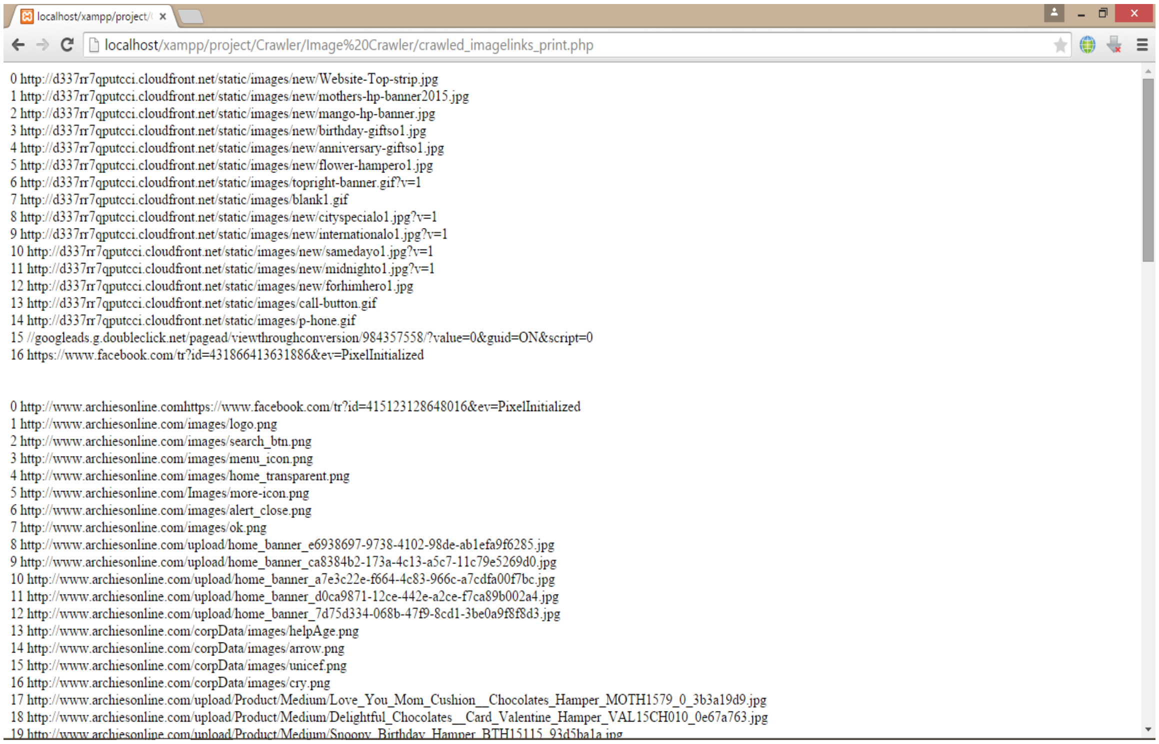
Fig: Screenshot 3 of the Crawled links from Giftease and Giveter websites



73	http://www.giftease.com/premium-gifting/wellness-beauty	https://www.facebook.com/givetergifts
74	http://www.giftease.com/premium-gifting/quirky-gifts	https://twitter.com/GiveterGifts
75	http://www.giftease.com/premium-gifting/gifts-flowers	[mailto:support@giveter.com]
76	http://www.giftease.com/premium-gifting/gadgets	[mailto:avinash@giveter.com]
77	http://www.giftease.com/premium-gifting/premium-stationery-pens	[mailto:mayank@giveter.com]
78	http://www.giftease.com/kids-toys-accessories/learning-activity-toys	www.giveter.com/giftsfor/Valentines%20Day
79	http://www.giftease.com/kids-toys-accessories/cars-action-toys	www.giveter.com/giftsfor/Friend:Him
80	http://www.giftease.com/kids-toys-accessories/soft-toys	www.giveter.com/giftsfor/Friend:Her
81	http://www.giftease.com/kids-toys-accessories/toddlers-infants	www.giveter.com/giftsfor/Kids
82	http://www.giftease.com/kids-toys-accessories/board-games	www.giveter.com/giftsfor/Parents:Mother
83	http://www.giftease.com/kids-toys-accessories/dolls-doll-houses	www.giveter.com/giftsfor/Parents:Father
84	http://www.giftease.com/kids-toys-accessories/kids-watches	www.giveter.com/giftsfor/Children:Daughter
85	http://www.giftease.com/kids-toys-accessories/kids-fashion-accessories	www.giveter.com/giftsfor/men
86	http://www.giftease.com/kids-toys-accessories/kids-costumes-role-play	www.giveter.com/giftsfor/Wedding
87	http://www.giftease.com/kids-toys-accessories/school-supplies	www.giveter.com/giftsfor/Anniversary
88	http://www.giftease.com/kids-toys-accessories/outdoor-toys	www.giveter.com/giftsfor/Sibling:Sister
89	http://www.giftease.com/kids-toys-accessories	www.giveter.com/giftsfor/Sibling:Brother
90	http://www.giftease.com/personalized-gifts/mugs	www.giveter.com/giftsfor/Special:Girlfriend
91	http://www.giftease.com/personalized-gifts/posters	www.giveter.com/giftsfor/Special:Boyfriend
92	http://www.giftease.com/personalized-gifts/chocolates	www.giveter.com/giftsfor/Special:Husband
93	http://www.giftease.com/personalized-gifts/greeting-cards	www.giveter.com/giftsfor/Special:Wife
94	http://www.giftease.com/personalized-gifts/kids-dvds	www.giveter.com/giftsfor/women
95	http://www.giftease.com/flowers/bunches	www.giveter.com/giftsfor/Special:Wife/Valentines%20Day
96	http://www.giftease.com/flowers/indoor-plants	www.giveter.com/giftsfor/Special:Girlfriend/Valentines%20Day
97	http://www.giftease.com/flowers/combos	www.giveter.com/giftsfor/Special:Boyfriend/Valentines%20Day
98	http://www.giftease.com/flowers/baskets	www.giveter.com/giftsfor/Special:Husband/Valentines%20Day
99	http://www.giftease.com/flowers/exclusive-arrangements	www.giveter.com/giftsfor/Special:Husband/Birthday
100	http://www.giftease.com/flowers/glass-vase-arrangements	www.giveter.com/giftsfor/Special:Wife/Birthday
101	http://www.giftease.com/flowers/scented-artificial-flowers	www.giveter.com/giftsfor/Special:Girlfriend/Birthday
102	http://www.giftease.com/party-stuff-return-gifts/kids-parties	www.giveter.com/giftsfor/Special:Boyfriend/Birthday
103	http://www.giftease.com/party-stuff-return-gifts/return-gifts	www.giveter.com/top20all
104	http://www.giftease.com/party-stuff-return-gifts/adult-parties	www.giveter.com/directory

Source: http://localhost/xampp/project/Crawler/crawled_links_print.php

Fig: Screenshot 4 of the Crawled links from Giftease and Giveter websites


localhost/xampp/project/
localhost/xampp/project/Crawler/Image%20Crawler/crawled_imagelinks_print.php

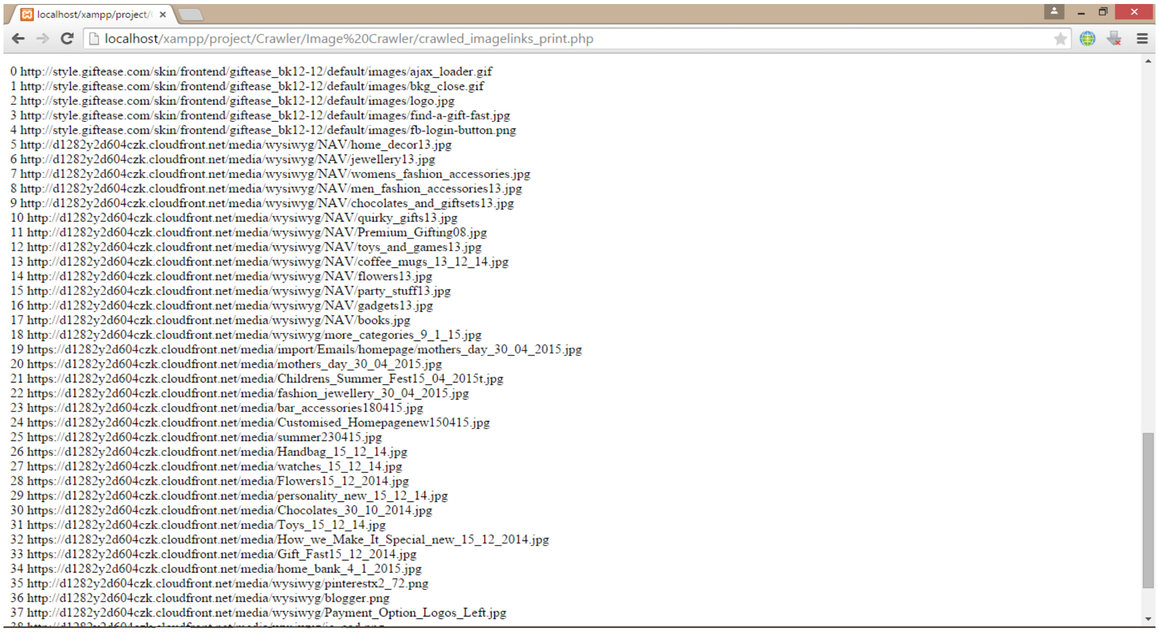
```
0 http://d337rr7qputcci.cloudfront.net/static/images/new/Website-Top-strip.jpg
1 http://d337rr7qputcci.cloudfront.net/static/images/new/mothers-hp-banner2015.jpg
2 http://d337rr7qputcci.cloudfront.net/static/images/new/mango-hp-banner.jpg
3 http://d337rr7qputcci.cloudfront.net/static/images/new/birthday-giftso1.jpg
4 http://d337rr7qputcci.cloudfront.net/static/images/new/anniversary-giftso1.jpg
5 http://d337rr7qputcci.cloudfront.net/static/images/new/flower-hampero1.jpg
6 http://d337rr7qputcci.cloudfront.net/static/images/topright-banner.gif?v=1
7 http://d337rr7qputcci.cloudfront.net/static/images/blank1.gif
8 http://d337rr7qputcci.cloudfront.net/static/images/new/cityspecialo1.jpg?v=1
9 http://d337rr7qputcci.cloudfront.net/static/images/new/internationalo1.jpg?v=1
10 http://d337rr7qputcci.cloudfront.net/static/images/new/samedayo1.jpg?v=1
11 http://d337rr7qputcci.cloudfront.net/static/images/new/midnighto1.jpg?v=1
12 http://d337rr7qputcci.cloudfront.net/static/images/new/forhimhero1.jpg
13 http://d337rr7qputcci.cloudfront.net/static/images/call-button.gif
14 http://d337rr7qputcci.cloudfront.net/static/images/p-hone.gif
15 //googleads.g.doubleclick.net/pagead/viewthroughconversion/984357558/?value=0&guid=ON&script=0
16 https://www.facebook.com/tr?id=431866413631886&ev=PixelInitialized


0 http://www.archiesonline.comhttps://www.facebook.com/tr?id=415123128648016&ev=PixelInitialized
1 http://www.archiesonline.com/images/logo.png
2 http://www.archiesonline.com/images/search_btn.png
3 http://www.archiesonline.com/images/menu_icon.png
4 http://www.archiesonline.com/images/home_transparent.png
5 http://www.archiesonline.com/Images/more-icon.png
6 http://www.archiesonline.com/images/alert_close.png
7 http://www.archiesonline.com/images/ok.png
8 http://www.archiesonline.com/upload/home_banner_e6938697-9738-4102-98de-ab1efa9f6285.jpg
9 http://www.archiesonline.com/upload/home_banner_ca8384b2-173a-4c13-a5c7-11c79e5269d0.jpg
10 http://www.archiesonline.com/upload/home_banner_a7e3c22e-f664-4c83-966c-a7cdfa00f7bc.jpg
11 http://www.archiesonline.com/upload/home_banner_d0ca9871-12ce-442e-a2ce-f7ca89b002a4.jpg
12 http://www.archiesonline.com/upload/home_banner_7d75d334-068b-47f9-8cd1-3be0a9f8f8d3.jpg
13 http://www.archiesonline.com/corpData/images/helpAge.png
14 http://www.archiesonline.com/corpData/images/arrow.png
15 http://www.archiesonline.com/corpData/images/unicef.png
16 http://www.archiesonline.com/corpData/images/cry.png
17 http://www.archiesonline.com/upload/Product/Medium/Love_You_Mom_Cushion__Chocolates_Hamper_MOTH1579_0_3b3a19d9.jpg
18 http://www.archiesonline.com/upload/Product/Medium/Delightful_Chocolates__Card_Valentine_Hamper_VAL15CH010_0e67a763.jpg
19 http://www.archiesonline.com/upload/Product/Medium/Snoopy_Birthday_Hamper_BTH15115_93d5ba1a.jpg
```

Source:

http://localhost/xampp/project/Crawler/Image_Crawler/crawled_imagelinks_print.php

Fig: Screenshot 5 of the Crawled image links from the websites considered


localhost/xampp/project/
localhost/xampp/project/Crawler/Image%20Crawler/crawled_imagelinks_print.php

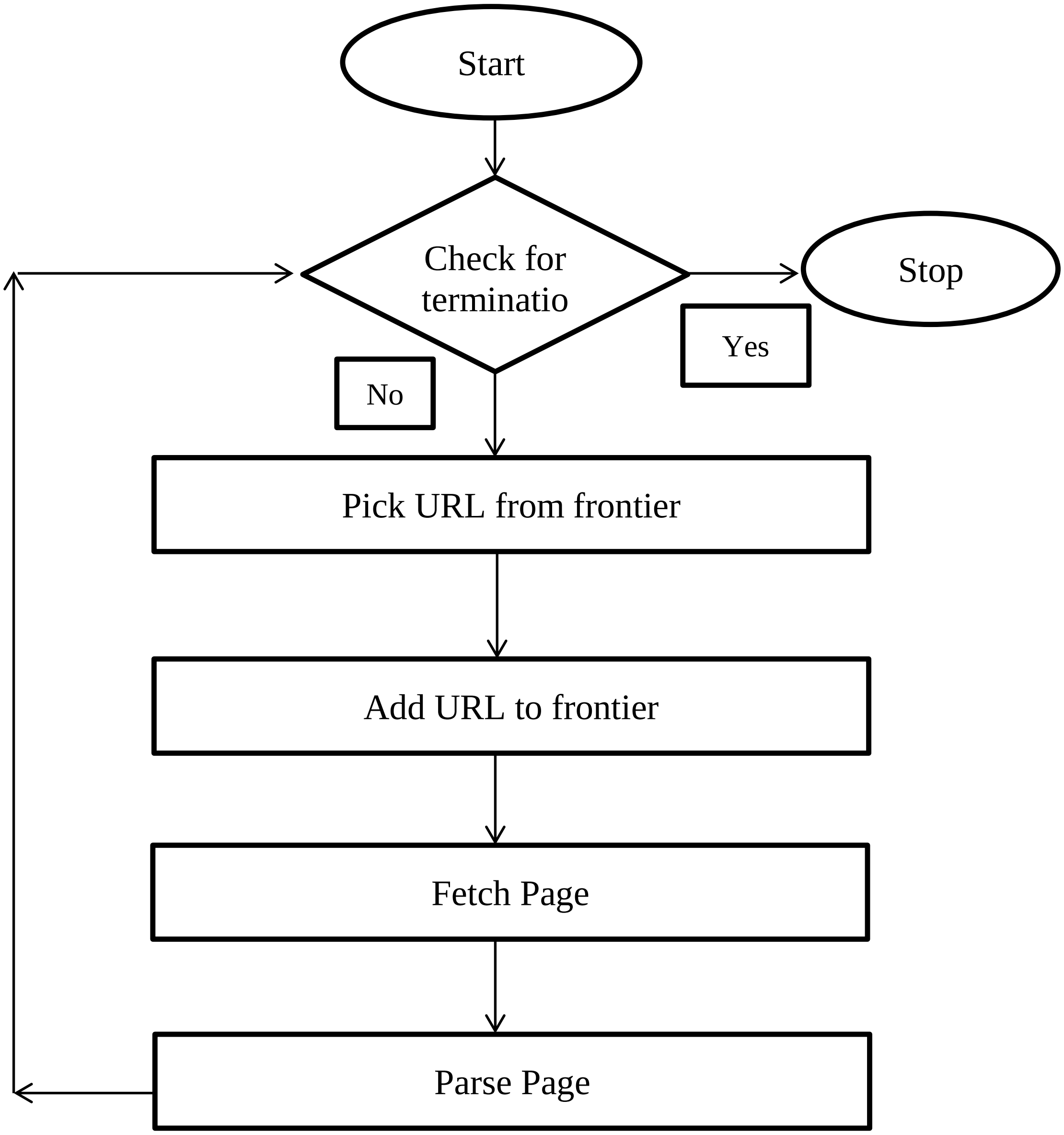
```
0 http://style.giftease.com/skin/frontend/giftease_bk12-12/default/images/ajax_loader.gif
1 http://style.giftease.com/skin/frontend/giftease_bk12-12/default/images/bkg_close.gif
2 http://style.giftease.com/skin/frontend/giftease_bk12-12/default/images/logo.jpg
3 http://style.giftease.com/skin/frontend/giftease_bk12-12/default/images/find-a-gift-fast.jpg
4 http://style.giftease.com/skin/frontend/giftease_bk12-12/default/images/fb-login-button.png
5 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/home_decor13.jpg
6 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/jewellery13.jpg
7 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/womens_fashion_accessories.jpg
8 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/men_fashion_accessories13.jpg
9 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/chocolates_and_giftsets13.jpg
10 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/quirky_gifts13.jpg
11 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/Premium_Gifting08.jpg
12 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/toys_and_games13.jpg
13 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/coffee_mugs_13_12_14.jpg
14 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/flowers13.jpg
15 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/party_stuff13.jpg
16 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/gadgets13.jpg
17 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/NAV/books.jpg
18 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/more_categories_9_1_15.jpg
19 https://d1282y2d604czk.cloudfront.net/media/import/Emails/homepage/mothers_day_30_04_2015.jpg
20 https://d1282y2d604czk.cloudfront.net/media/mothers_day_30_04_2015.jpg
21 https://d1282y2d604czk.cloudfront.net/media/Childrens_Summer_Fest15_04_2015t.jpg
22 https://d1282y2d604czk.cloudfront.net/media/fashion_jewellery_30_04_2015.jpg
23 https://d1282y2d604czk.cloudfront.net/media/bar_accessories180415.jpg
24 https://d1282y2d604czk.cloudfront.net/media/Customised_Homepagenew150415.jpg
25 https://d1282y2d604czk.cloudfront.net/media/summer230415.jpg
26 https://d1282y2d604czk.cloudfront.net/media/Handbag_15_12_14.jpg
27 https://d1282y2d604czk.cloudfront.net/media/watches_15_12_14.jpg
28 https://d1282y2d604czk.cloudfront.net/media/Flowers15_12_2014.jpg
29 https://d1282y2d604czk.cloudfront.net/media/personality_new_15_12_14.jpg
30 https://d1282y2d604czk.cloudfront.net/media/Chocolates_30_10_2014.jpg
31 https://d1282y2d604czk.cloudfront.net/media/Toys_15_12_14.jpg
32 https://d1282y2d604czk.cloudfront.net/media/How_we_Make_It_Special_new_15_12_2014.jpg
33 https://d1282y2d604czk.cloudfront.net/media/Gift_Fast15_12_2014.jpg
34 https://d1282y2d604czk.cloudfront.net/media/home_bank_4_1_2015.jpg
35 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/pinterestx2_72.png
36 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/blogger.png
37 http://d1282y2d604czk.cloudfront.net/media/wysiwyg/Payment_Option_Logos_Left.jpg
```

Source:

http://localhost/xampp/project/Crawler/Image_Crawler/crawled_imagelinks_print.php

Fig: Screenshot 6 of the Crawled image links from the websites considered

2.1.8 Overall Flow of Sequential Crawler



Source: G Pant, P Srinivasan, F. Menczer, “**Crawling the Web,”** in Springer, 2004

Fig: Flow of Sequential Crawler

2.2 Data Consolidation

Data consolidation refers to the collection and integration of data from multiple sources into a single destination. During this process, different data sources are put together, or consolidated, into a single data store. Because data comes from a broad range of sources, consolidation allows organizations to more easily present data, while also facilitating effective data analysis.

For example: In windows Operating system, Microsoft Excel offers a tool that allows users to consolidate data between various worksheets to form a larger, more organized summary of all your sheets. Microsoft Excel's data consolidation tool also allows users to consolidate data from more than one Excel file, allowing the user to summarize data sheets into one easy to read spreadsheet.

2.2.1 Advantages

i. Reduces inefficiencies: Consolidating the content of websites can remove the redundant data using topical categorization.

ii. No data duplication: When the users send a search query on a system, the same webpage/web link appears should not appear more than once on the screen.

iii. Increases costs reliability: Consolidating the data is a costs optimization technique and the money spent in the consolidating methodology turns out to be worth investment.

iv. Improves discoverability: Consolidating the content of websites can improve discoverability (through better search, grouping of content based on audience needs, consistent navigation) and provide more consistent and authoritative content (through better processes and resourcing).

2.2.2 Why consolidate data?

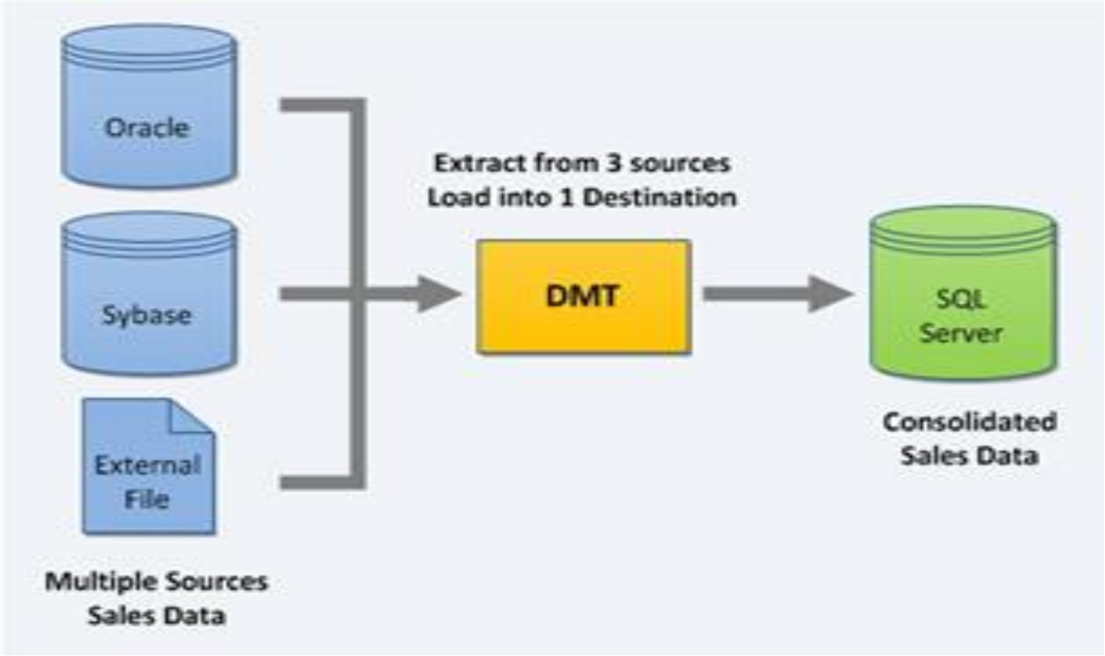
The ultimate goal of consolidating the web content is to help people find and use the information and services they are looking for. Having too many websites can make it difficult for a target audience to know where to look initially and generally complicate the discovery process. Since, the online web portals are increasing continuously day by day, the competition amongst them is also rising to peak. Every company is trying to emerge with new ideas so as to combat their weaknesses, and provide products in a new and efficient way. The categorization of the product is the new methodology to present best to the users in just one click at an instant of time. This leads to consolidating the huge data on the websites into different categories.

2.2.3 How Consolidation is done?

Before consolidating your websites, identify business objectives and user needs for the consolidated website. It is important to know what content one wants to keep and where you will place it in the new location – which means divide the data into several categories according to the business objective. Edit the index page to include the redirect code and customize a message explaining that the old site has been retired and ‘You are now being redirected to (the new URL)’.

The person doing the consolidation should communicate back to the agency governance body/ business administrators and to stakeholders throughout the consolidation process so as to get

- Progress against the consolidation implementation plan
- Issues or obstacles still to be addressed.



In the Fig. we can see the three different databases of the three different businesses where the crawler fetches and parses the webpage of the web link. Once the web links are extracted in the queue the similar type of products will be categorized in the same category so as to provide feasibility to the user

2.2.4 Consolidation in My Area of Work

Once the crawler fetches and parses the webpage and the web links of the specific sites (focusing on the four websites considered particularly) the consolidation of data is done considering the major parameters.

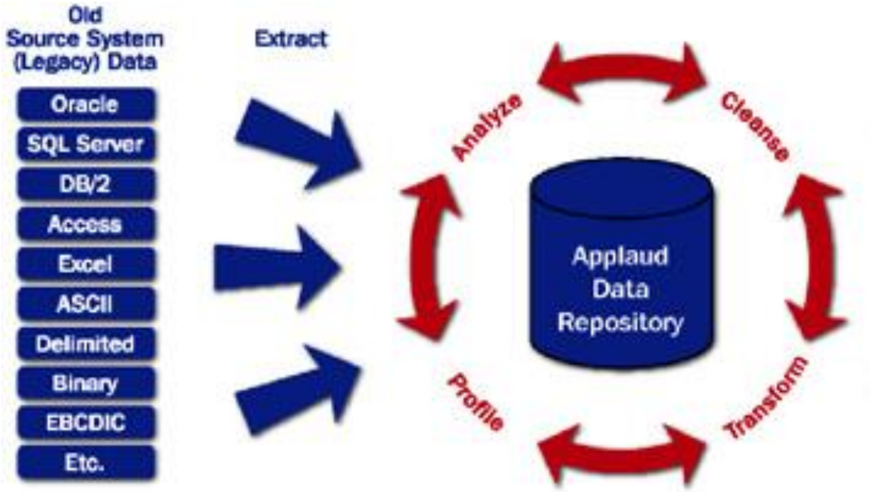
The major parameters include:

- Occasion
- Relationship
- Age
- Personality
- Price
- Delivery Time

Major Parameters

The products in the databases of both the web links will be broadly categorized in the above three parameters so that user gets to see what he wants to see. The topical categorization of the webpages will be done on the major parameters mentioned above. There can be an occurrence that the same product can occur for the two different search queries and also the different parameters clicked. This can happen because the product will be classified in both the segments. In general, a single product can be subcategorized in the different segment of the databases.

The consolidation algorithm has to be such that it has to avoid the loading of the same products twice for the same search query performed by the user.



Source: http://premier-international.com/

Fig: Data Consolidation

The Data warehouses are central repositories of integrated data from one or more disparate sources. They store current and historical data for comparisons. Hence the data warehouse built after analyzing and consolidation would

- free from duplicate links
- Efficient to surf.

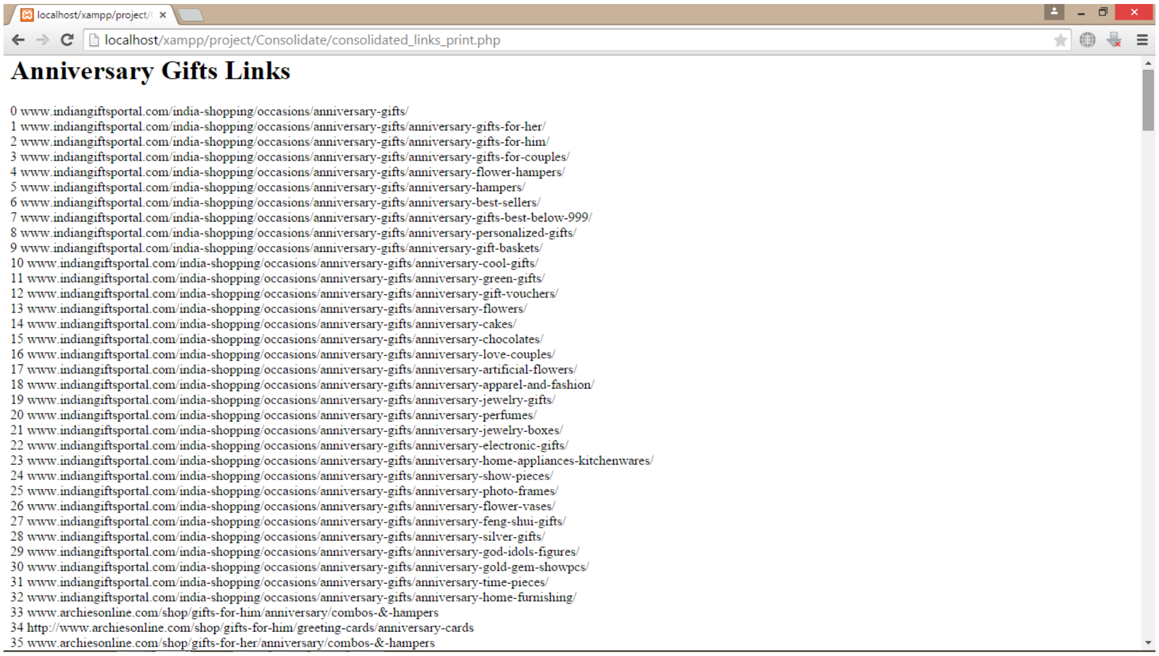
The overall method is costs reliable also as it needs just the efficient algorithm to divide into various segment the categorical data.

As in the previous section [2.1] there was a crawler implementation discussed, the crawler had crawled for both the image links and the web pages links. Similarly, the consolidation will also be done for all the links and images too. There are different queues implemented for the web links consolidation and the image links consolidation. The queues are implemented using arrays in PHP.

There are different arrays on the basis of various Occasional parameters (like anniversary, birthday, wedding, romance, house warming), Age parameters (like teens, babies, kids, youth, young married, seniors) and Relationship Parameters (like Special, Friend, Sibling, Parents, Children). The subdivision of consolidation is also to be done which would include other parameters discussed above in the parameters section.

Different parameters might contain the same links (either of the webpage or of the image links) twice but a single parametric array would not include two same links (neither of the webpage nor of the image links).

The screenshots of the links consolidated are shown for the purpose of reference.

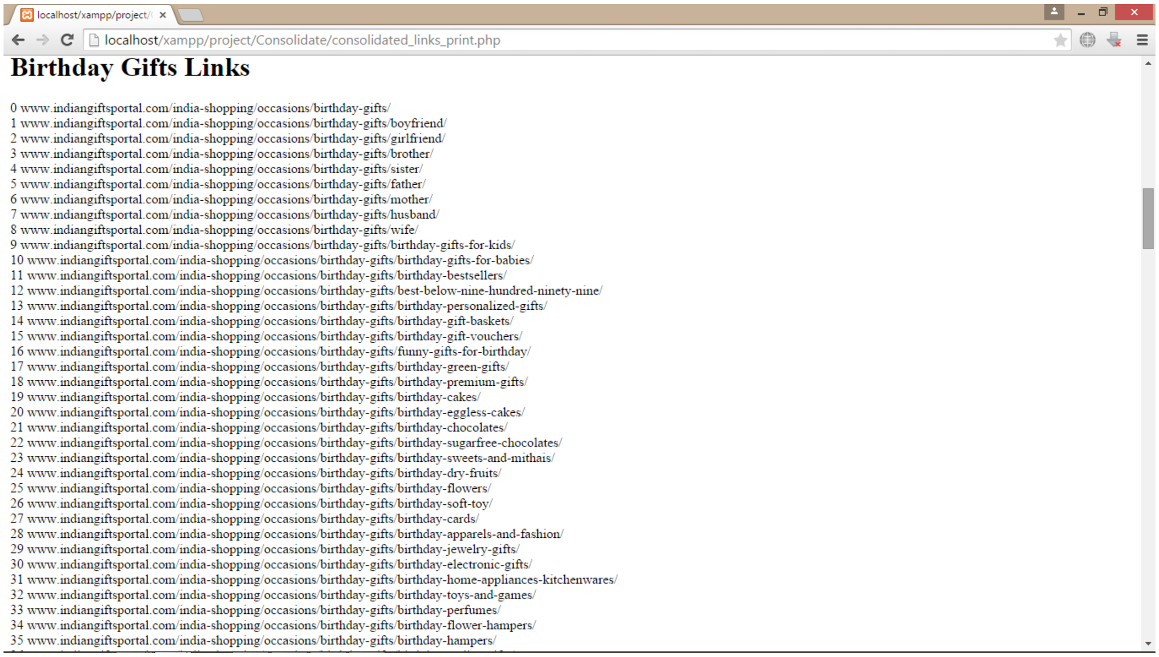

localhost/xampp/project/
localhost/xampp/project/Consolidate/consolidated_links_print.php

Anniversary Gifts Links

```
0 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/
1 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-gifts-for-her/
2 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-gifts-for-him/
3 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-gifts-for-couples/
4 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-flower-hampers/
5 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-hampers/
6 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-best-sellers/
7 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-gifts-best-below-999/
8 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-personalized-gifts/
9 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-gift-baskets/
10 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-cool-gifts/
11 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-green-gifts/
12 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-gift-vouchers/
13 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-flowers/
14 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-cakes/
15 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-chocolates/
16 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-love-couples/
17 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-artificial-flowers/
18 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-apparel-and-fashion/
19 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-jewelry-gifts/
20 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-perfumes/
21 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-jewelry-boxes/
22 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-electronic-gifts/
23 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-home-appliances-kitchenwares/
24 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-show-pieces/
25 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-photo-frames/
26 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-flower-vases/
27 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-feng-shui-gifts/
28 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-silver-gifts/
29 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-god-idols-figures/
30 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-gold-gem-showpcs/
31 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-time-pieces/
32 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-home-furnishing/
33 www.archiesonline.com/shop/gifts-for-him/anniversary/combos-&-hampers
34 http://www.archiesonline.com/shop/gifts-for-him/greeting-cards/anniversary-cards
35 www.archiesonline.com/shop/gifts-for-her/anniversary/combos-&-hampers
```

Source: http://localhost/xampp/project/Consolidate/consolidate_occasion_print.php

Screenshot 7 of the consolidated link on the basis of Anniversary Occasion

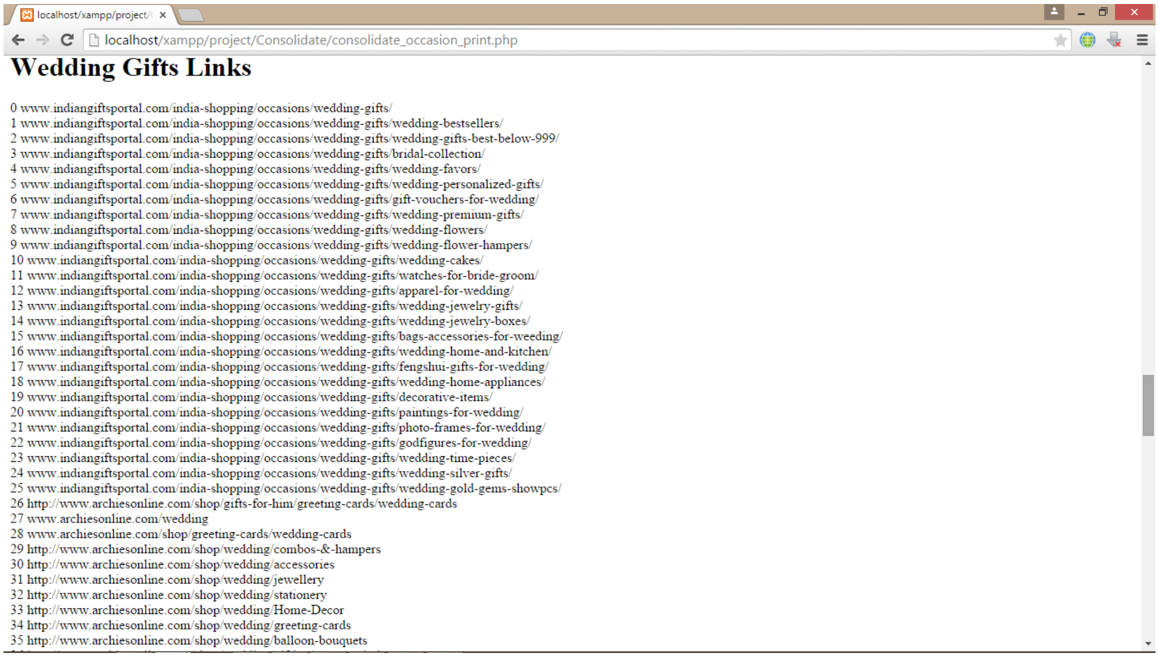

localhost/xampp/project/
localhost/xampp/project/Consolidate/consolidated_links_print.php

Birthday Gifts Links

```
0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/
1 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/boyfriend/
2 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/girlfriend/
3 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/brother/
4 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/sister/
5 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/father/
6 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/mother/
7 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/husband/
8 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/wife/
9 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gifts-for-kids/
10 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gifts-for-babies/
11 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-bestsellers/
12 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/best-below-nine-hundred-ninety-nine/
13 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-personalized-gifts/
14 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gift-baskets/
15 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gift-vouchers/
16 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/funny-gifts-for-birthday/
17 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-green-gifts/
18 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-premium-gifts/
19 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-cakes/
20 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-eggless-cakes/
21 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-chocolates/
22 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-sugarfree-chocolates/
23 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-sweets-and-mithais/
24 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-dry-fruits/
25 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-flowers/
26 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-soft-toy/
27 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-cards/
28 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-apparels-and-fashion/
29 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-jewelry-gifts/
30 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-electronic-gifts/
31 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-home-appliances-kitchenwares/
32 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-toys-and-games/
33 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-perfumes/
34 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-flower-hampers/
35 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-hampers/
```

Source: http://localhost/xampp/project/Consolidate/consolidate_occasion_print.php

Fig: Screenshot 8 of the consolidated link on the basis of Birthday Occasion

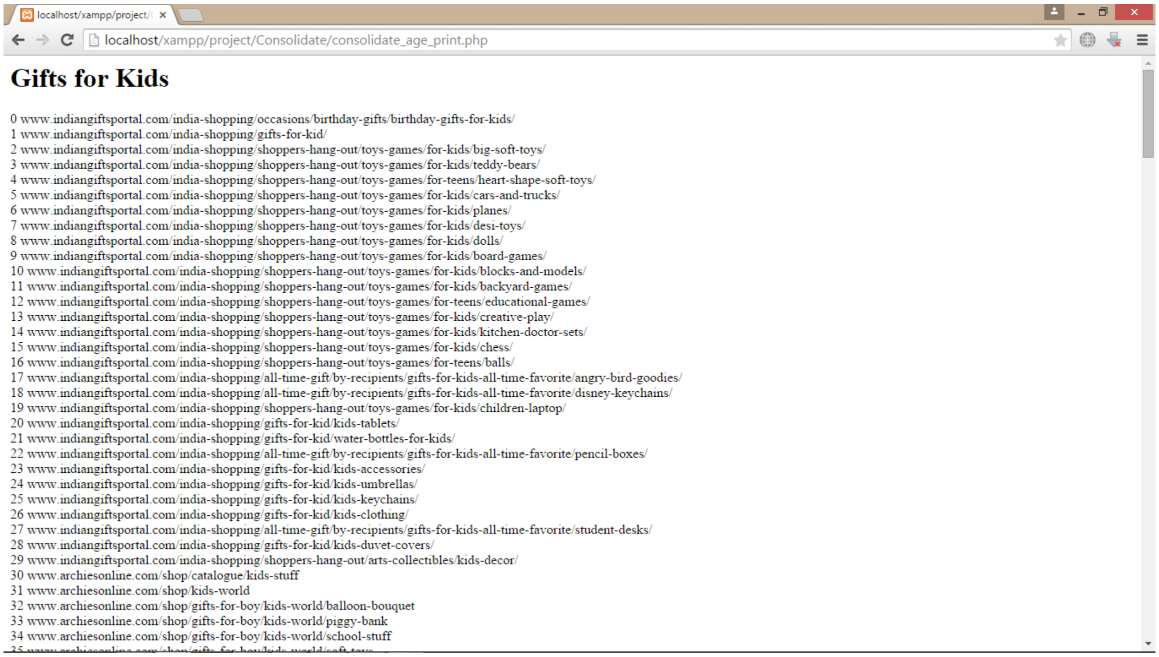

localhost/xampp/project/
localhost/xampp/project/Consolidate/consolidate_occasion_print.php

Wedding Gifts Links

```
0 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/
1 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-bestsellers/
2 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-gifts-best-below-999/
3 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/bridal-collection/
4 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-favors/
5 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-personalized-gifts/
6 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/gift-vouchers-for-wedding/
7 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-premium-gifts/
8 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-flowers/
9 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-flower-hampers/
10 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-cakes/
11 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/watches-for-bride-groom/
12 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/apparel-for-wedding/
13 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-jewelry-gifts/
14 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-jewelry-boxes/
15 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/bags-accessories-for-weeding/
16 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-home-and-kitchen/
17 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/fengshui-gifts-for-wedding/
18 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-home-appliances/
19 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/decorative-items/
20 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/paintings-for-wedding/
21 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/photo-frames-for-wedding/
22 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/godfigures-for-wedding/
23 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-time-pieces/
24 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-silver-gifts/
25 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/wedding-gold-gems-showpcs/
26 http://www.archiesonline.com/shop/gifts-for-him/greeting-cards/wedding-cards
27 www.archiesonline.com/wedding
28 www.archiesonline.com/shop/greeting-cards/wedding-cards
29 http://www.archiesonline.com/shop/wedding/combos-&-hampers
30 http://www.archiesonline.com/shop/wedding/accessories
31 http://www.archiesonline.com/shop/wedding/jewellery
32 http://www.archiesonline.com/shop/wedding/stationery
33 http://www.archiesonline.com/shop/wedding/Home-Decor
34 http://www.archiesonline.com/shop/wedding/greeting-cards
35 http://www.archiesonline.com/shop/wedding/balloon-bouquets
```

Source: http://localhost/xampp/project/Consolidate/consolidate_occasion_print.php

Fig: Screenshot 9 of the consolidated link on the basis of Wedding Occasion

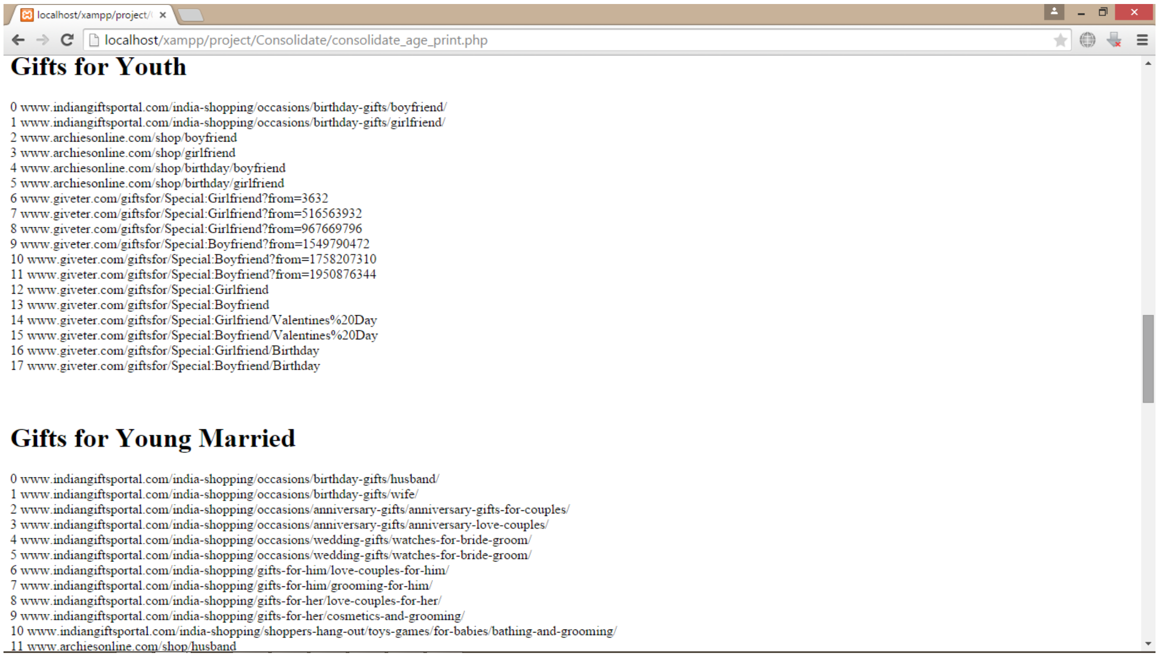

localhost/xampp/project/
localhost/xampp/project/Consolidate/consolidate_age_print.php

Gifts for Kids

0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/birthday-gifts-for-kids/
1 www.indiangiftsportal.com/india-shopping/gifts-for-kid/
2 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/big-soft-toys/
3 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/teddy-bears/
4 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-teens/heart-shape-soft-toys/
5 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/cars-and-trucks/
6 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/planes/
7 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/desi-toys/
8 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/dolls/
9 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/board-games/
10 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/blocks-and-models/
11 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/backyard-games/
12 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-teens/educational-games/
13 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/creative-play/
14 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/kitchen-doctor-sets/
15 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/chess/
16 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-teens/balls/
17 www.indiangiftsportal.com/india-shopping/all-time-gift/by-recipients/gifts-for-kids-all-time-favorite/angry-bird-goodies/
18 www.indiangiftsportal.com/india-shopping/all-time-gift/by-recipients/gifts-for-kids-all-time-favorite/disney-keychains/
19 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-kids/children-laptop/
20 www.indiangiftsportal.com/india-shopping/gifts-for-kid/kids-tablets/
21 www.indiangiftsportal.com/india-shopping/gifts-for-kid/water-bottles-for-kids/
22 www.indiangiftsportal.com/india-shopping/all-time-gift/by-recipients/gifts-for-kids-all-time-favorite/pencil-boxes/
23 www.indiangiftsportal.com/india-shopping/gifts-for-kid/kids-accessories/
24 www.indiangiftsportal.com/india-shopping/gifts-for-kid/kids-umbrellas/
25 www.indiangiftsportal.com/india-shopping/gifts-for-kid/kids-keychains/
26 www.indiangiftsportal.com/india-shopping/gifts-for-kid/kids-clothing/
27 www.indiangiftsportal.com/india-shopping/all-time-gift/by-recipients/gifts-for-kids-all-time-favorite/student-desks/
28 www.indiangiftsportal.com/india-shopping/gifts-for-kid/kids-duvet-covers/
29 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/arts-collectibles/kids-decor/
30 www.archiesonline.com/shop/catalogue/kids-stuff
31 www.archiesonline.com/shop/kids-world
32 www.archiesonline.com/shop/gifts-for-boy/kids-world/balloon-bouquet
33 www.archiesonline.com/shop/gifts-for-boy/kids-world/piggy-bank
34 www.archiesonline.com/shop/gifts-for-boy/kids-world/school-stuff
35 www.archiesonline.com/shop/gifts-for-boy/kids-world/soft-toys

Source: http://localhost/xampp/project/Consolidate/consolidate_age_print.php

Fig: Screenshot 10 of the consolidated link on the basis of different Ages


localhost/xampp/project/
localhost/xampp/project/Consolidate/consolidate_age_print.php

Gifts for Youth

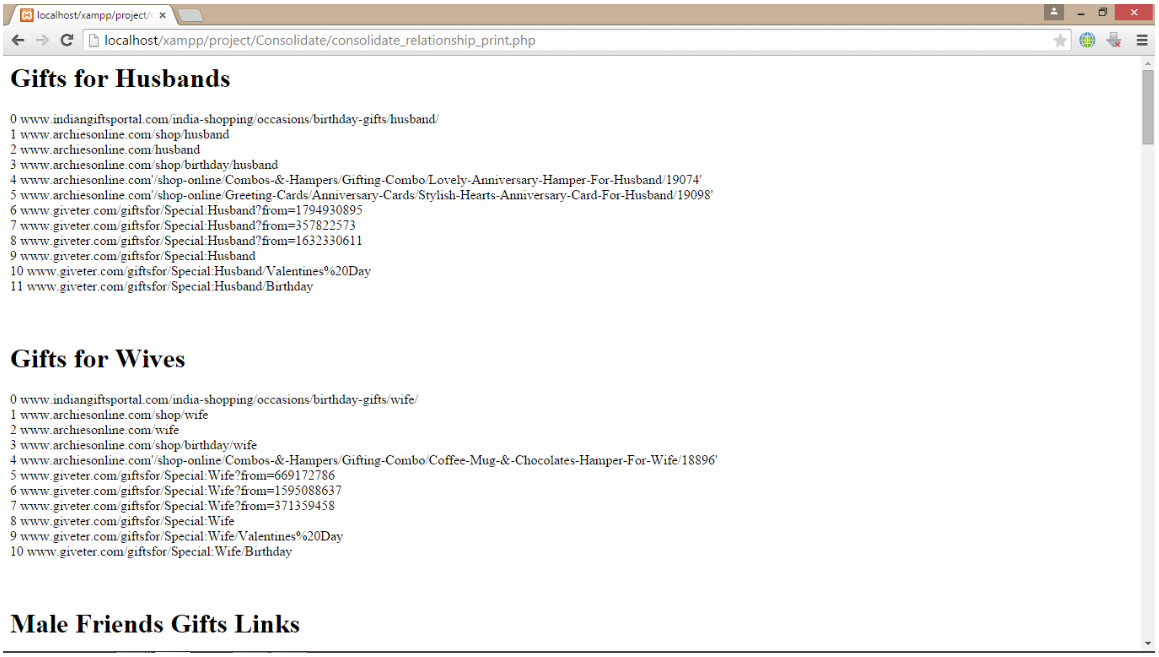
0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/boyfriend/
1 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/girlfriend/
2 www.archiesonline.com/shop/boyfriend
3 www.archiesonline.com/shop/girlfriend
4 www.archiesonline.com/shop/birthday/boyfriend
5 www.archiesonline.com/shop/birthday/girlfriend
6 www.giveter.com/giftsfor/Special:Girlfriend?from=3632
7 www.giveter.com/giftsfor/Special:Girlfriend?from=516563932
8 www.giveter.com/giftsfor/Special:Girlfriend?from=967669796
9 www.giveter.com/giftsfor/Special:Boyfriend?from=1549790472
10 www.giveter.com/giftsfor/Special:Boyfriend?from=1758207310
11 www.giveter.com/giftsfor/Special:Boyfriend?from=1950876344
12 www.giveter.com/giftsfor/Special:Girlfriend
13 www.giveter.com/giftsfor/Special:Boyfriend
14 www.giveter.com/giftsfor/Special:Girlfriend/Valentines%20Day
15 www.giveter.com/giftsfor/Special:Boyfriend/Valentines%20Day
16 www.giveter.com/giftsfor/Special:Girlfriend/Birthday
17 www.giveter.com/giftsfor/Special:Boyfriend/Birthday

Gifts for Young Married

0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/husband/
1 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/wife/
2 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-gifts-for-couples/
3 www.indiangiftsportal.com/india-shopping/occasions/anniversary-gifts/anniversary-love-couples/
4 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/watches-for-bride-groom/
5 www.indiangiftsportal.com/india-shopping/occasions/wedding-gifts/watches-for-bride-groom/
6 www.indiangiftsportal.com/india-shopping/gifts-for-him/love-couples-for-him/
7 www.indiangiftsportal.com/india-shopping/gifts-for-him/grooming-for-him/
8 www.indiangiftsportal.com/india-shopping/gifts-for-her/love-couples-for-her/
9 www.indiangiftsportal.com/india-shopping/gifts-for-her/cosmetics-and-grooming/
10 www.indiangiftsportal.com/india-shopping/shoppers-hang-out/toys-games/for-babies/bathing-and-grooming/
11 www.archiesonline.com/shop/husband

Source: http://localhost/xampp/project/Consolidate/consolidate_age_print.php

Fig: Screenshot 11 of the consolidated link on the basis of different Ages


localhost/xampp/project/
localhost/xampp/project/Consolidate/consolidate_relationship_print.php

Gifts for Husbands

```
0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/husband/
1 www.archiesonline.com/shop/husband
2 www.archiesonline.com/husband
3 www.archiesonline.com/shop/birthday/husband
4 www.archiesonline.com'/shop-online/Combos-&-Hampers/Gifting-Combo/Lovely-Anniversary-Hamper-For-Husband/19074'
5 www.archiesonline.com'/shop-online/Greeting-Cards/Anniversary-Cards/Stylish-Hearts-Anniversary-Card-For-Husband/19098'
6 www.giveter.com/giftsfor/Special:Husband?from=1794930895
7 www.giveter.com/giftsfor/Special:Husband?from=357822573
8 www.giveter.com/giftsfor/Special:Husband?from=1632330611
9 www.giveter.com/giftsfor/Special:Husband
10 www.giveter.com/giftsfor/Special:Husband/Valentines%20Day
11 www.giveter.com/giftsfor/Special:Husband/Birthday
```

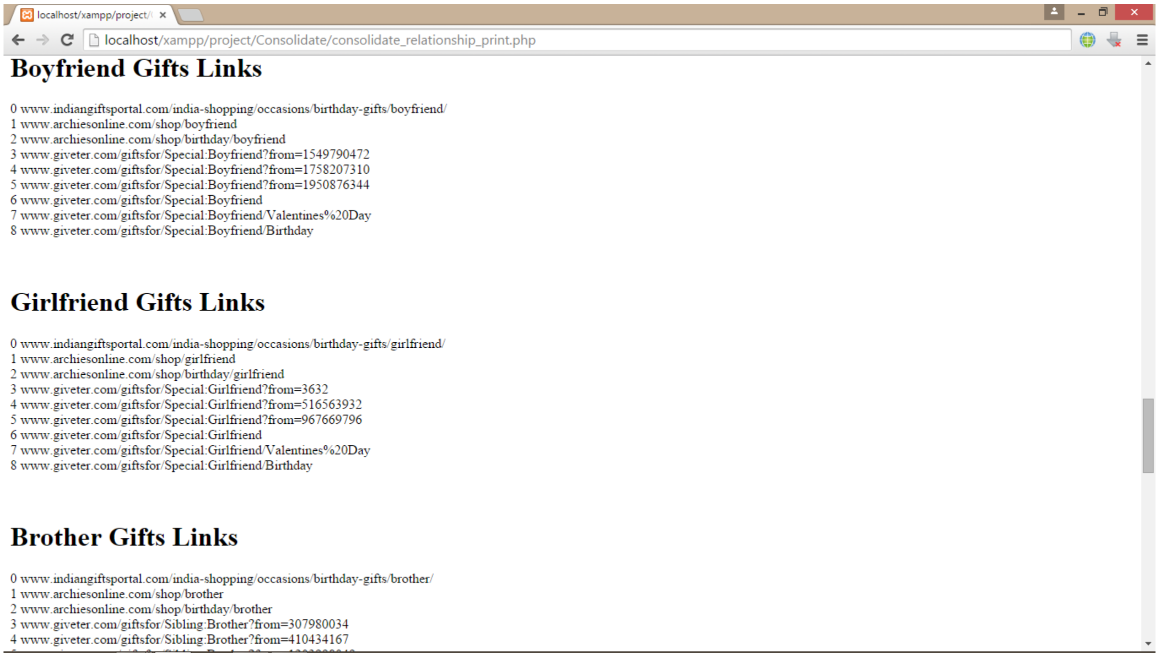
Gifts for Wives

```
0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/wife/
1 www.archiesonline.com/shop/wife
2 www.archiesonline.com/wife
3 www.archiesonline.com/shop/birthday/wife
4 www.archiesonline.com'/shop-online/Combos-&-Hampers/Gifting-Combo/Coffee-Mug-&-Chocolates-Hamper-For-Wife/18896'
5 www.giveter.com/giftsfor/Special:Wife?from=669172786
6 www.giveter.com/giftsfor/Special:Wife?from=1595088637
7 www.giveter.com/giftsfor/Special:Wife?from=371359458
8 www.giveter.com/giftsfor/Special:Wife
9 www.giveter.com/giftsfor/Special:Wife/Valentines%20Day
10 www.giveter.com/giftsfor/Special:Wife/Birthday
```

Male Friends Gifts Links

Source: http://localhost/xampp/project/Consolidate/consolidate_relationship_print.php

Fig: Screenshot 12 of the consolidated link on the basis of different Relationship


localhost/xampp/project/
localhost/xampp/project/Consolidate/consolidate_relationship_print.php

Boyfriend Gifts Links

```
0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/boyfriend/
1 www.archiesonline.com/shop/boyfriend
2 www.archiesonline.com/shop/birthday/boyfriend
3 www.giveter.com/giftsfor/Special:Boyfriend?from=1549790472
4 www.giveter.com/giftsfor/Special:Boyfriend?from=1758207310
5 www.giveter.com/giftsfor/Special:Boyfriend?from=1950876344
6 www.giveter.com/giftsfor/Special:Boyfriend
7 www.giveter.com/giftsfor/Special:Boyfriend/Valentines%20Day
8 www.giveter.com/giftsfor/Special:Boyfriend/Birthday
```

Girlfriend Gifts Links

```
0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/girlfriend/
1 www.archiesonline.com/shop/girlfriend
2 www.archiesonline.com/shop/birthday/girlfriend
3 www.giveter.com/giftsfor/Special:Girlfriend?from=3632
4 www.giveter.com/giftsfor/Special:Girlfriend?from=516563932
5 www.giveter.com/giftsfor/Special:Girlfriend?from=967669796
6 www.giveter.com/giftsfor/Special:Girlfriend
7 www.giveter.com/giftsfor/Special:Girlfriend/Valentines%20Day
8 www.giveter.com/giftsfor/Special:Girlfriend/Birthday
```

Brother Gifts Links

```
0 www.indiangiftsportal.com/india-shopping/occasions/birthday-gifts/brother/
1 www.archiesonline.com/shop/brother
2 www.archiesonline.com/shop/birthday/brother
3 www.giveter.com/giftsfor/Sibling:Brother?from=307980034
4 www.giveter.com/giftsfor/Sibling:Brother?from=410434167
```

Source: http://localhost/xampp/project/Consolidate/consolidate_relationship_print.php

Fig: Screenshot 13 of the consolidated link on the basis of different Relationships

2.3 Pruning of Data

Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. Once the query is made by the user, the results reflect out as the examples of the product. The quality of the examples also matters and that the learning algorithm might be better off when some training examples are discarded. The question is which examples need to be eliminated, so as to improve generalization performance. We propose a general approach, called 'data pruning', to automatically identify and eliminate examples that are troublesome for learning with a given model. The dual goal of pruning is reduced complexity of the final classifier as well as better predictive accuracy by the reduction of over fitting and removal of sections of a classifier that may be based on noisy or erroneous data.

2.3.1 Pruning in My Area of Work

While doing the process of crawling the concept of pruning is required. Once the page links are fetched there are several links which contains the symbols which are encrypted and these create problem to access the page links. It is obvious that to access the webpage the address or the URL of the page has to be correct. So the data pruning is necessary to ignore everything else from the page links and just concentrate upon the page links.

The pruning process includes

- Convert the protocol and hostname to lowercase.
- Remove the “anchor” or “reference” part of the URL.
- Remove “.”, “. //”,”//”,”#” and its parent directory from the URL path.
- Split the links using the ‘#’.
- Adding the website name (domain name) in front of the ‘/ ‘symbol.
- The web link must begin with the domain name and not any symbol.

2.3.2 HTML Tag Tree

```
<html>
<head>          <title> Projects </title> </head>
<body>          <h4> Projects </h4>
<ul>
<li> <a href=”book.html”>BOOKS</a>…..some texts….</li>
<li> <a href=”clothes.html”>CLOTHES</a>…..some texts….</li>
<li> <a href=”Gadges.html”>GADGES</a>…..some texts….</li>
</ul>
</body>
</html>
```

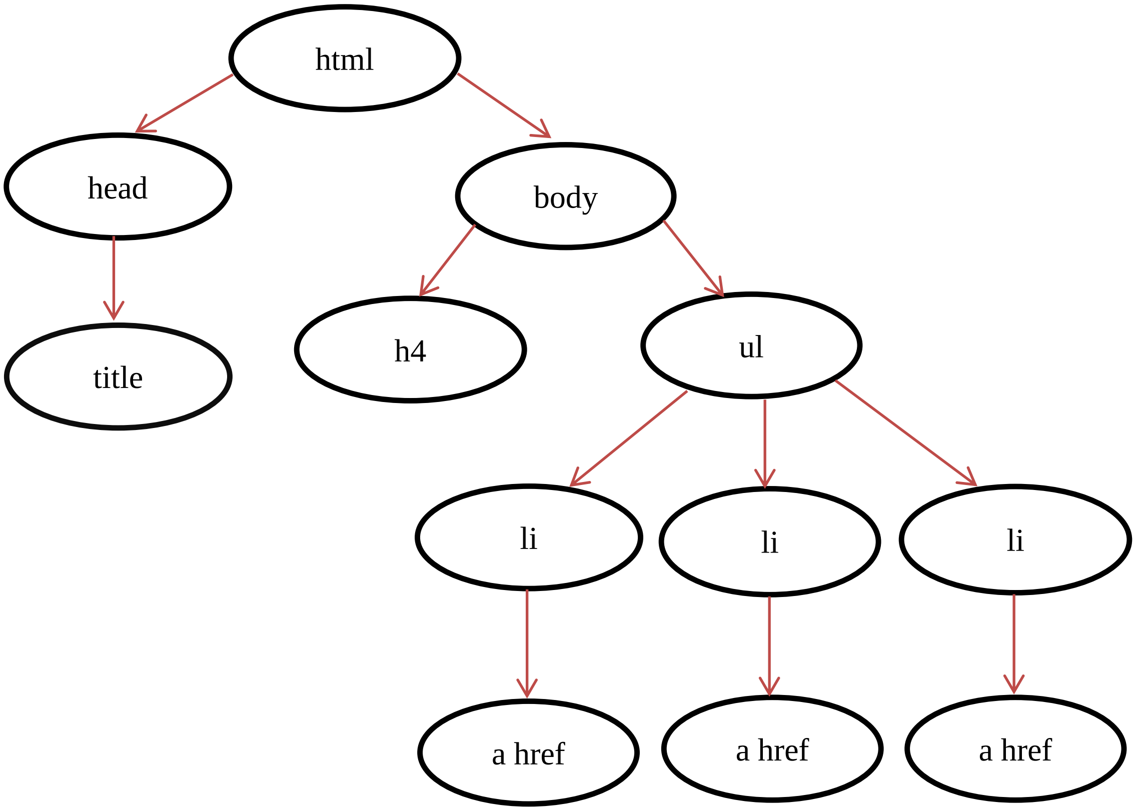
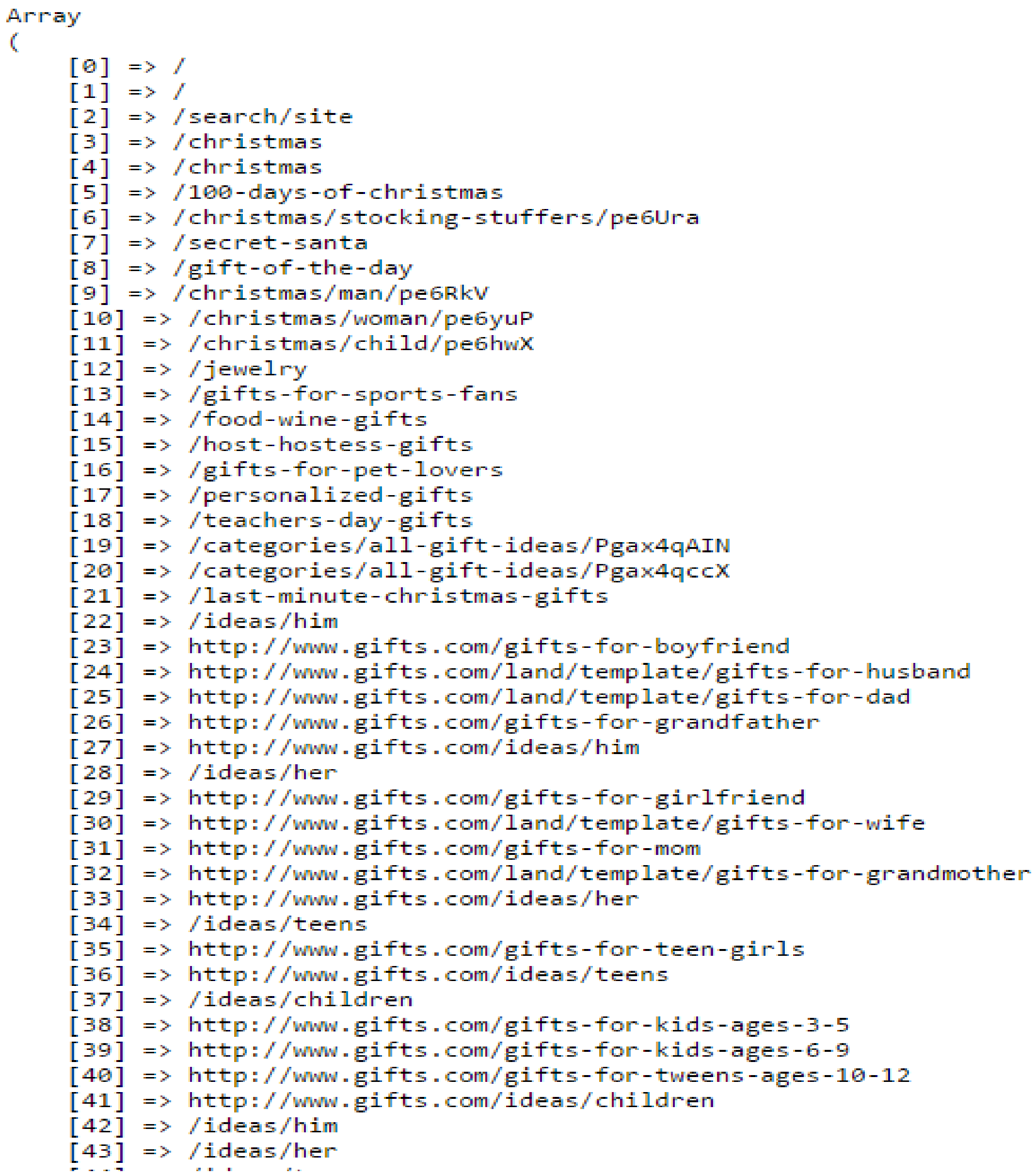


Fig: Diagrammatic representation of the HTML TAG TREE

2.3.3 Effect of Pruning



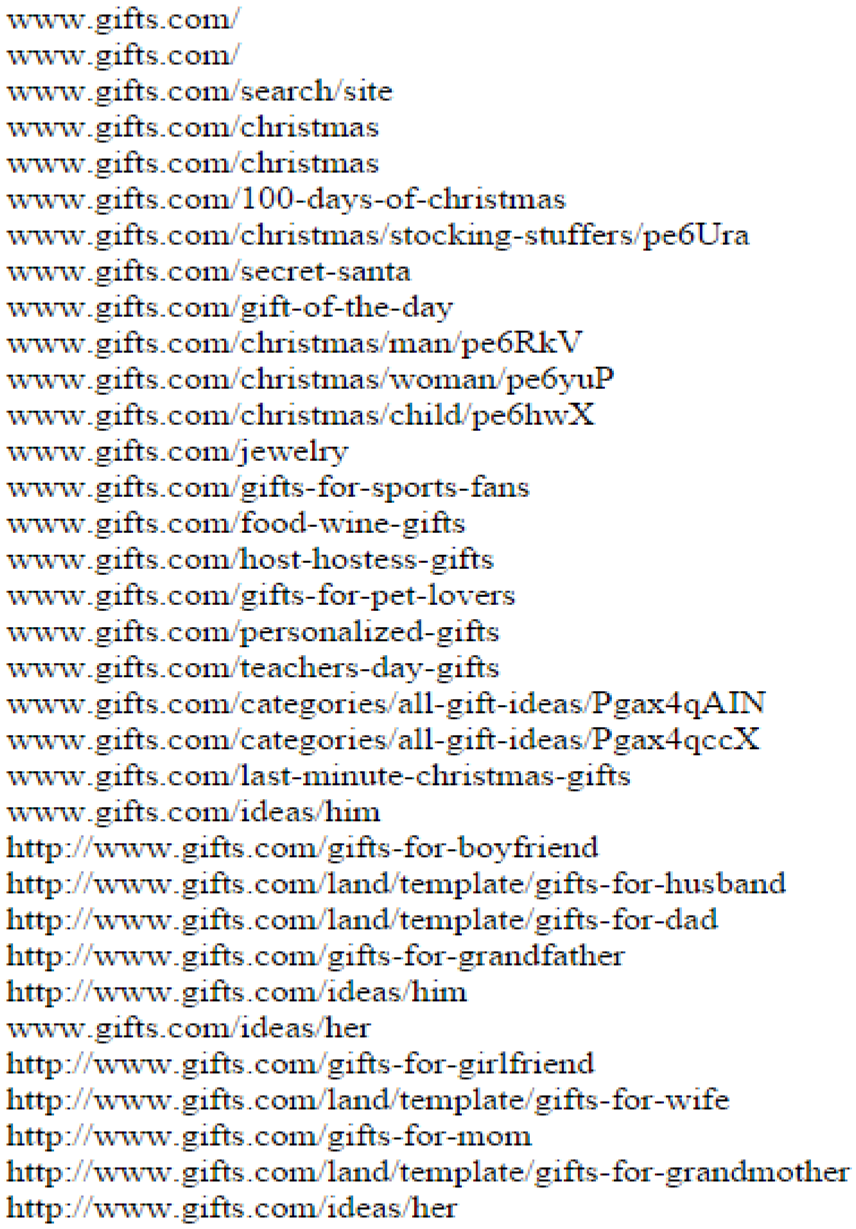
```
Array
(
    [0] => /
    [1] => /
    [2] => /search/site
    [3] => /christmas
    [4] => /christmas
    [5] => /100-days-of-christmas
    [6] => /christmas/stocking-stuffers/pe6Ura
    [7] => /secret-santa
    [8] => /gift-of-the-day
    [9] => /christmas/man/pe6RkV
    [10] => /christmas/woman/pe6yuP
    [11] => /christmas/child/pe6hwX
    [12] => /jewelry
    [13] => /gifts-for-sports-fans
    [14] => /food-wine-gifts
    [15] => /host-hostess-gifts
    [16] => /gifts-for-pet-lovers
    [17] => /personalized-gifts
    [18] => /teachers-day-gifts
    [19] => /categories/all-gift-ideas/Pgax4qAIN
    [20] => /categories/all-gift-ideas/Pgax4qccX
    [21] => /last-minute-christmas-gifts
    [22] => /ideas/him
    [23] => http://www.gifts.com/gifts-for-boyfriend
    [24] => http://www.gifts.com/land/template/gifts-for-husband
    [25] => http://www.gifts.com/land/template/gifts-for-dad
    [26] => http://www.gifts.com/gifts-for-grandfather
    [27] => http://www.gifts.com/ideas/him
    [28] => /ideas/her
    [29] => http://www.gifts.com/gifts-for-girlfriend
    [30] => http://www.gifts.com/land/template/gifts-for-wife
    [31] => http://www.gifts.com/gifts-for-mom
    [32] => http://www.gifts.com/land/template/gifts-for-grandmother
    [33] => http://www.gifts.com/ideas/her
    [34] => /ideas/teens
    [35] => http://www.gifts.com/gifts-for-teen-girls
    [36] => http://www.gifts.com/ideas/teens
    [37] => /ideas/children
    [38] => http://www.gifts.com/gifts-for-kids-ages-3-5
    [39] => http://www.gifts.com/gifts-for-kids-ages-6-9
    [40] => http://www.gifts.com/gifts-for-tweens-ages-10-12
    [41] => http://www.gifts.com/ideas/children
    [42] => /ideas/him
    [43] => /ideas/her
```

Source: http://localhost/xampp/project/Crawler/crawled_links_print.php

Fig: Screenshot 14 for the crawler before pruning of the web pages links fetched.

The figure above represents that the website (www.gifts.com) has been surfed and all the associated URL’s have been fetched from its page source. But the links contain various symbols which would create lots of problem while accessing those webpages. So the data links have to be pruned.

After Pruning:



```
www.gifts.com/
www.gifts.com/
www.gifts.com/search/site
www.gifts.com/christmas
www.gifts.com/christmas
www.gifts.com/100-days-of-christmas
www.gifts.com/christmas/stocking-stuffers/pe6Ura
www.gifts.com/secret-santa
www.gifts.com/gift-of-the-day
www.gifts.com/christmas/man/pe6RkV
www.gifts.com/christmas/woman/pe6yuP
www.gifts.com/christmas/child/pe6hwX
www.gifts.com/jewelry
www.gifts.com/gifts-for-sports-fans
www.gifts.com/food-wine-gifts
www.gifts.com/host-hostess-gifts
www.gifts.com/gifts-for-pet-lovers
www.gifts.com/personalized-gifts
www.gifts.com/teachers-day-gifts
www.gifts.com/categories/all-gift-ideas/Pgax4qAIN
www.gifts.com/categories/all-gift-ideas/Pgax4qccX
www.gifts.com/last-minute-christmas-gifts
www.gifts.com/ideas/him
http://www.gifts.com/gifts-for-boyfriend
http://www.gifts.com/land/template/gifts-for-husband
http://www.gifts.com/land/template/gifts-for-dad
http://www.gifts.com/gifts-for-grandfather
http://www.gifts.com/ideas/him
www.gifts.com/ideas/her
http://www.gifts.com/gifts-for-girlfriend
http://www.gifts.com/land/template/gifts-for-wife
http://www.gifts.com/gifts-for-mom
http://www.gifts.com/land/template/gifts-for-grandmother
http://www.gifts.com/ideas/her
```

Source: http://localhost/xampp/project/Crawler/crawled_links_print.php

Fig: Screenshot 15 for the crawler after pruning of the web pages links fetched.

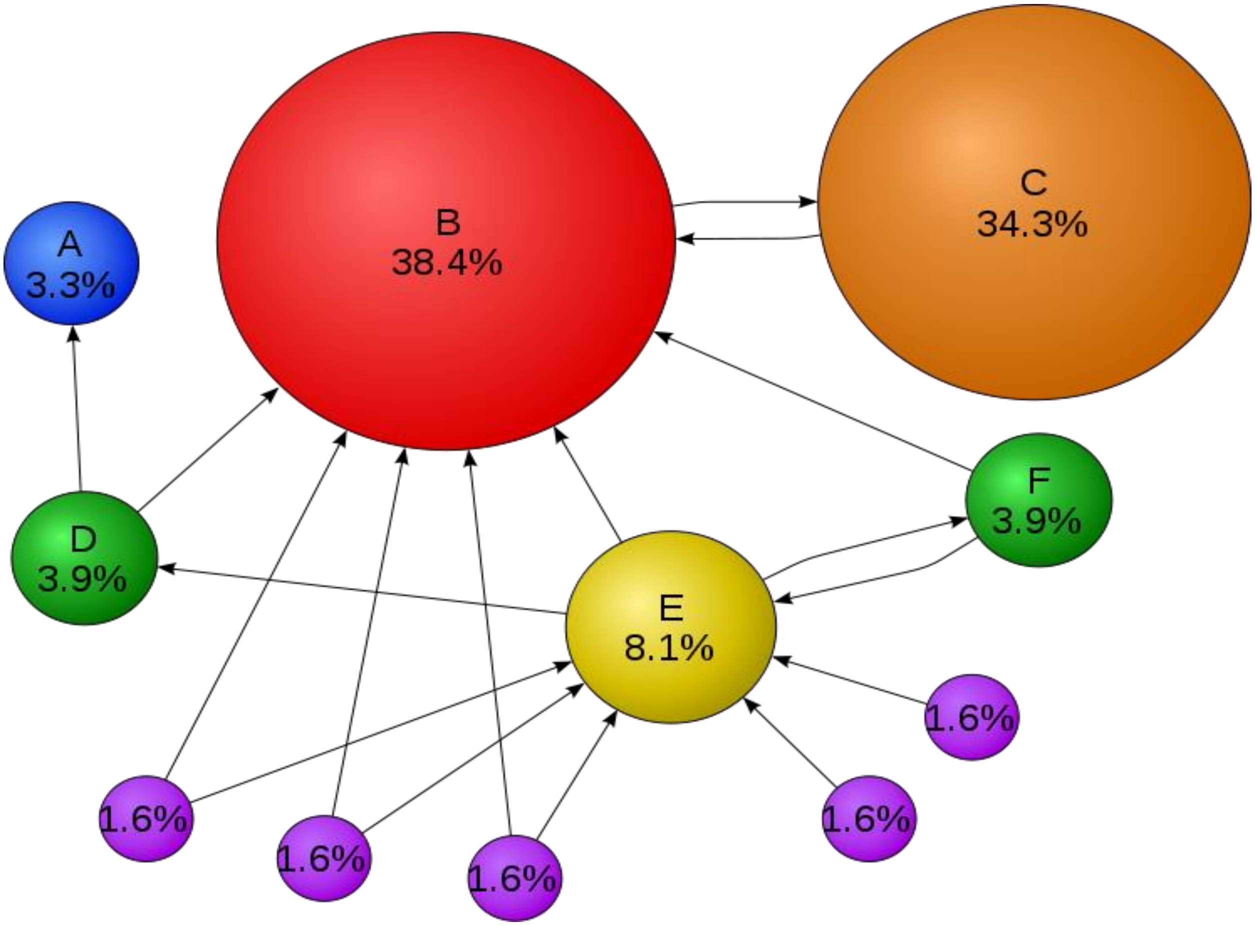
The redundant data from the array has been removed and the proper links are made referring to the previous array and the website (www.gifts.com). This is the effect of pruning on the output. The links are proper now and can be accessed easily.

2.4 Page Ranking

PageRank is an algorithm used by the search engine to rank web pages in their search results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages depending on the frequency it is being visited by Internet users.

PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites.

It is not the only algorithm used by Google to order search engine results, but it is the first algorithm that was used by the company, and it is the best-known.



Source: http://www.techiemania.com

Fig: Showing the frequency of occurrence of different pages on Web

2.4.1 How to increase Page Rankings?

Page Ranking is intended to increase in the following ways:-

- Publish articles (related to the website) that are original and of excellent quality so that the visitors will find it useful and will be attracted to link to it.
- The web developer needs to have a lot of backlinks, or even reciprocal links to the webpage. This can be done by bartering links with other webmasters.
- Register your website with any of the high ranking web directories. When your webpage is linked to a web directory there are better chances to get greater traffic.

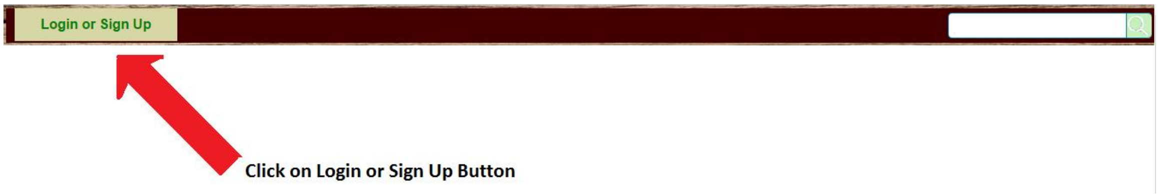
2.4.2 Page Ranking in My Area of Work

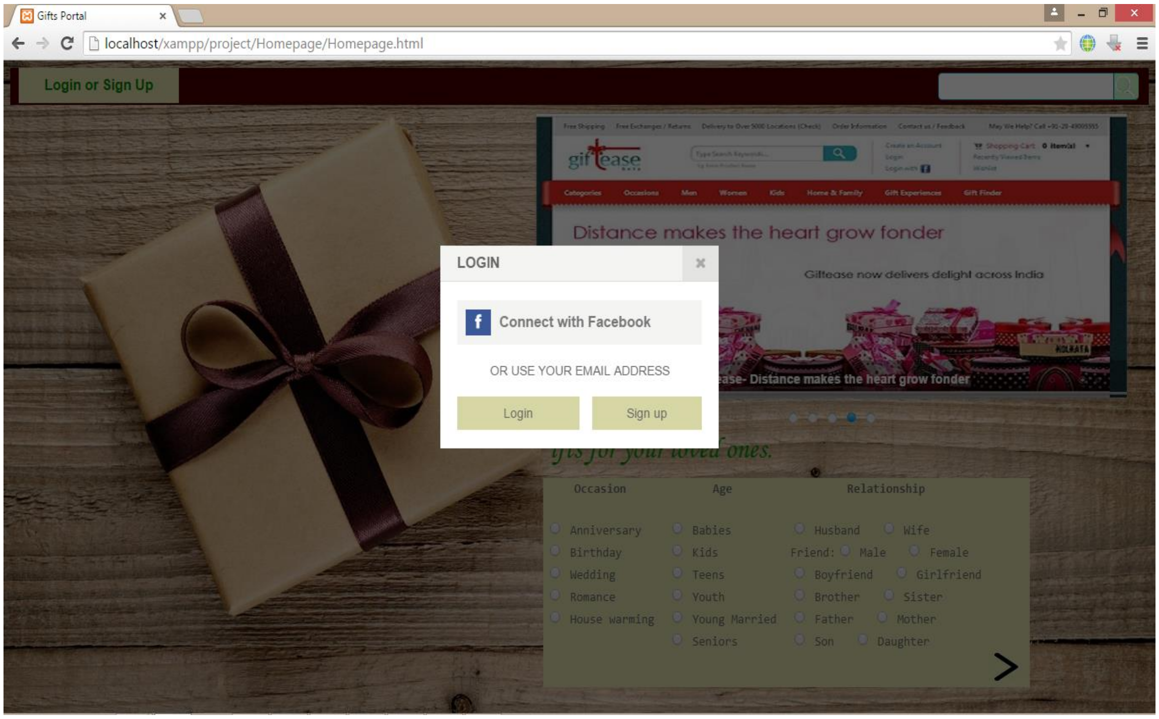
When the crawler fetches the URL's from the page source of the different websites considered, the page links are parsed and pruned. Linked Analysis Algorithm will be used to implement the optimization technique for Page Ranking.

- The webpages links are stored in a data structure (implemented using a queue) and then the consolidation is done over them. After the user decides his/her parameters the web links those are common to all the parameters will be shown first and then the other links are shown under the heading more options.
- The page links are to be stored in the queue after ranking them. More frequently the page link would be visited; greater will be its ranking. There can be a separate parameter of page count to be used with each page links so as to keep count of the frequency of visiting a particular page.
- The highest page rank link product will appear at the top, the second highest page rank link product follows the first one. Thus, the entire search results will display the products in the decreasing order of their page ranks or page counts.

2.5 Log In or Sign Up Pop Up Box

In the project there has been an implementation of Login and Sign up pop up box. The Log In or Sign Up button in the header bar when is clicked there is an image pop up box that is activated and all the other buttons of the webpages are henceforth deactivated.





Source: http://localhost/xampp/project/Homepage/Homepage.html

Fig. Screenshot 16 of the homepage showing the Log in pop up box.

Social login is often considered as a gateway to many of the recent trends in social software and social commerce because it can be used as a mechanism for both authentication and authorization. Social login is created in our web based gifts portal system using the essential configuration and security settings.

The gifts portal system will have its own login page which include Login through different ways.

i.) Login using the Facebook account

ii.) Login using the email account

2.5.1 How Login through Facebook is created?

- Created using Graph API Facebook

- Facebook developers provide a separate App Id and Secret when the administrator takes the permission after sending requests to the Graph API Facebook.

- The App Id acts as a unique Id to provide permission requests to users so as to access the application using Facebook.

- The PHP code has the App Id and the secret key of the application whereas the Facebook will also have the website name and domain name (‘localhost’ in this case) apart from the secret key and the App Id.

- The application is first tested on the test Id before launching on the original website Facebook App Id.

Login through Facebook enable users to login to the website using their Facebook account. There will be no posts made on behalf of the user regarding the Login or purchase made. The security measures are kept into consideration and it is on the hands of the users to grant access to the Facebook features for the website.

2.5.2 What is Graph API in Facebook?

The Graph API is the primary way to get data in and out of Facebook's social graph. It's a low-level HTTP-based API that you can use to query data, post new stories, upload photos and a variety of other tasks that an app might need to do. This guide will teach you how to accomplish all these things in the Graph API.

The **Facebook Platform** is an umbrella term used to describe the set of services, tools, and products provided by the social networking service Facebook for third-party developers to create their own applications and services that access data in Facebook. It was launched in 2010. The platform offers a set of programming interfaces and tools which enable developers to integrate with the open "**social graph**" of personal relations and other things like songs, places, and Facebook pages. Applications on facebook.com, external websites, and devices are all allowed to access the graph.

The Graph API is the core of Facebook Platform, enabling developers to read from and write data into Facebook. The Graph API presents a simple, consistent view of the Facebook social graph, uniformly representing objects in the graph (e.g., people, photos, events, and pages) and the connections between them (e.g., friend relationships, shared content, and photo tags).

Facebook authentication enables developer’s applications to interact with the Graph API on behalf of Facebook users, and it provides a single-sign on mechanism across web, mobile, and desktop apps. The Open Graph protocol enables developers to integrate their pages into the social graph. These pages gain the functionality of other graph objects including profile links and stream updates for connected users

2.5.3 What is Facebook Connect?

Facebook Connect is also known as **Log in with Facebook**, like Open ID. It is a set of authentication APIs from Facebook that developers can use to help their users connect and share with such users' Facebook friends on Facebook and increase engagement for their website or application.

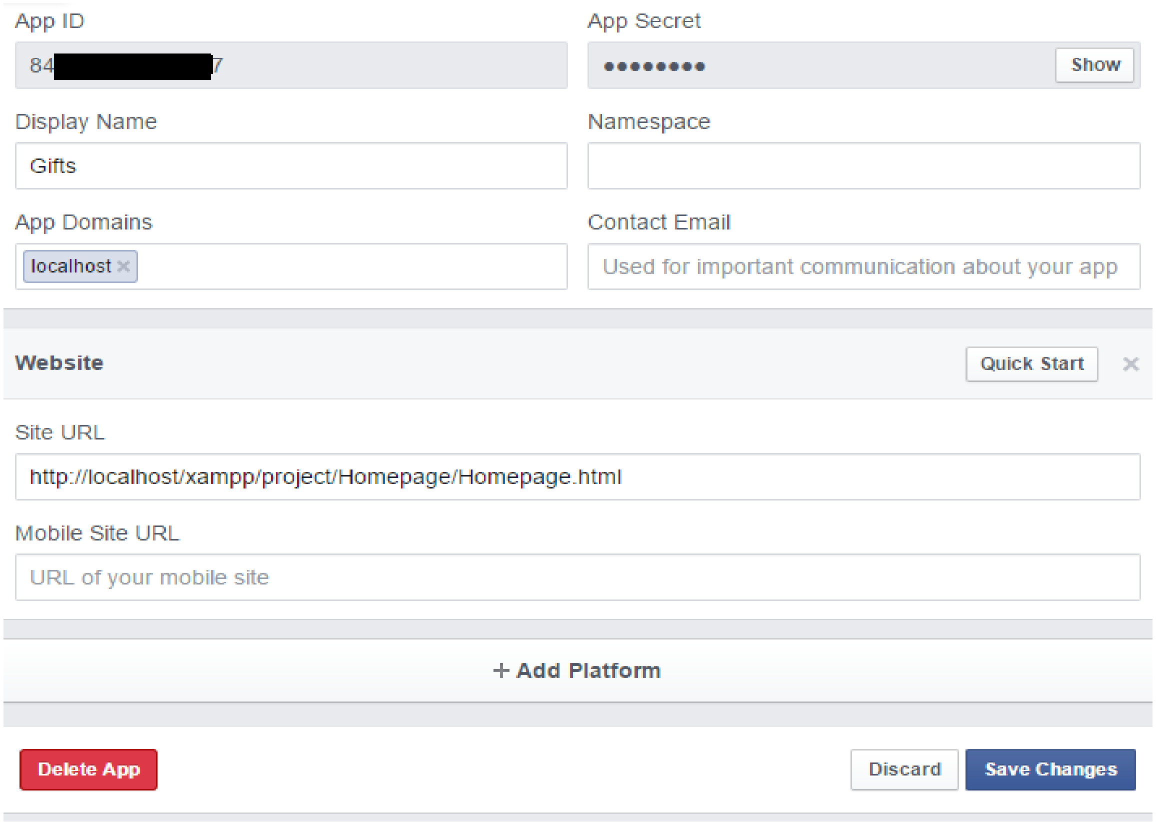
When so used, Facebook members can log on to third-party websites, applications, mobile devices and gaming systems with their Facebook identity and, while logged in, and can connect with friends via these media and post information and updates to their Facebook profile.

Originally unveiled during Facebook’s developer conference, F8, in July 2008, Log in with Facebook became generally available in December 2008. Log in with Facebook cannot be used by users in locations that cannot access Facebook (e.g. China), even if the third-party site is otherwise accessible from that location.

In the project, Login through Facebook where the standard API’s are used with the proper App Id. The App Id is provided from the Facebook developer team on a request of it. The App Id acts as a major key to run the Facebook applications and get the permissions for the users. Apart from the App Id there is also a secret key shared by the Facebook developers. This secret key is to be kept confidential only with the user of Facebook who has requested to Facebook for creating the application on it and use its features. The secret key also helps in fighting the spams. Accurate security of the Facebook account of the users is kept in the minds of the users so that there is no content of the users account is accessible by the websites or the spammers unless the access permission is provided by the users.

Even if the user provides the access to the Facebook account for the particular website there are no posts made from the user accounts unless the user wishes to do so. The user can also block the particular application where the Facebook will restrict the applications features to appear on the user’s account who have blocked it. If many user’s block a website the Facebook blocks the application to run its feature on the Facebook platform. This is done using the Facebook App Id which is also retained by the Facebook.

The user only needs to provide the application domain and the website URL to the Facebook so as to enable connect. The image for the same is shown in the next page.



App ID: 84[redacted]

App Secret: ●●●●●●●●

Display Name: Gifts

Namespace:

App Domains: localhost

Contact Email: Used for important communication about your app

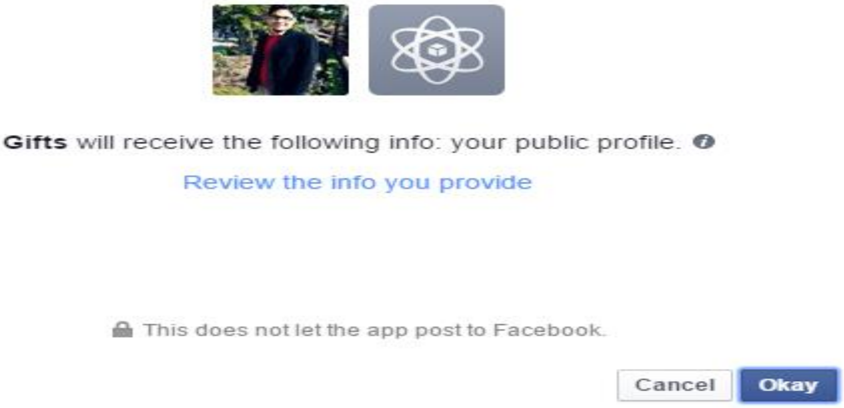
Website

Site URL: http://localhost/xampp/project/Homepage/Homepage.html

Mobile Site URL: URL of your mobile site

Source: http:/developers.facebook.com/

Fig. Screenshot 17 of the Settings done in My Apps on Facebook Account



Gifts will receive the following info: your public profile.

Review the info you provide

This does not let the app post to Facebook.

Source: http://www.facebook.com

Fig. Screenshot 18 of the Gifts App taking permission from the user

Social login links logins to one or more social networking services to a website, typically using either a plug-in or a widget. By selecting the desired social networking service, the user simply uses his or her login for that services to sign on to the web site. This in turn negates the need for the end user to remember login information for multiple electronic commerce and other websites while providing site owners with uniform demographic information as provided by the social networking service. Many sites which offer social login also offer more traditional online registration for those who either desire it or who do not have an account with a compatible social networking service (and therefore would be precluded from creating an account with the website).

2.5.4 Advantages of Social Login

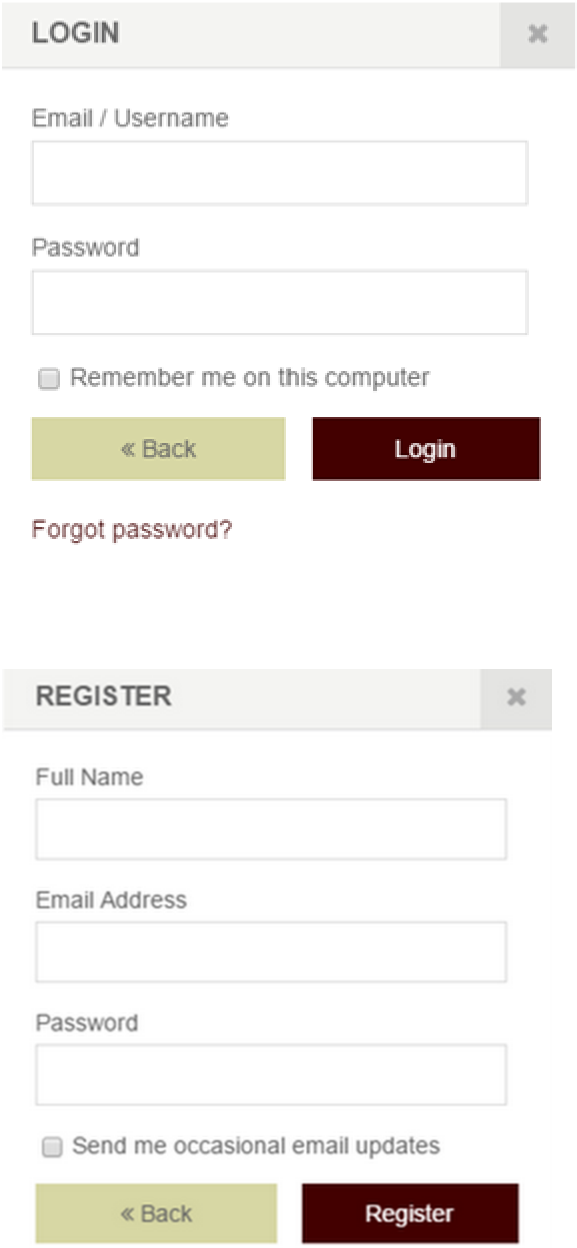
- ***Targeted Content*** - Websites can obtain a profile and social graph data in order to target personalized content to the user. This includes information such as name, email, hometown, interests, activities and friends. However, this can create issues for privacy, resulting in narrowing of the variety of options available on the internet.

- ***Registration Data*** - Many websites use the profile data returned from social login instead of having users manually enter their PII (Personally Identifiable Information) into web forms. This can potentially speed up the registration or sign-up process.

- ***Pre-Validated Email*** - Identity providers who support email such as Google can return the user's email address to the 3rd party website preventing the user from supplying a fabricated email address during the registration process.

- ***Account linking*** – Social login is used for authentication, many sites allow users to link pre-existing site account with their social login account without re-registration.

Log in through Email would help users to sign in through their registered email id. If the email id is not registered the user needs to register using the specific email id. In this case there will be an email send to the users regarding the specific deals and discounts of the day. The emails will act like newsletters to attract user's attention.



Source: http://localhost/xampp/project/Homepage/Homepage.html

Fig. Screenshot 19 of the Login and Sign up form in the Login through Email pop up

2.5.5 Advantages of creating different Log in Methods

- Separate Log in methods would reduce the number of entries in the particular database created backend, hence faster response to queries.

- User need not create a separate login for every website he is redirected to.
  A single login would help him surf for any website through the same login account. The product details will be added to the same login account no matter from which website it is brought. A unique id will be allotted which will differentiate the user login and their purchase or cart history.

- Social Log in would provide users benefit of providing birthday reminder for the friends added in his friend list.

- Social Log in would also encourage users to analyze the areas of interests of users in buying the product.

Chapter 3 Design & Modeling

3.1 Design Diagrams

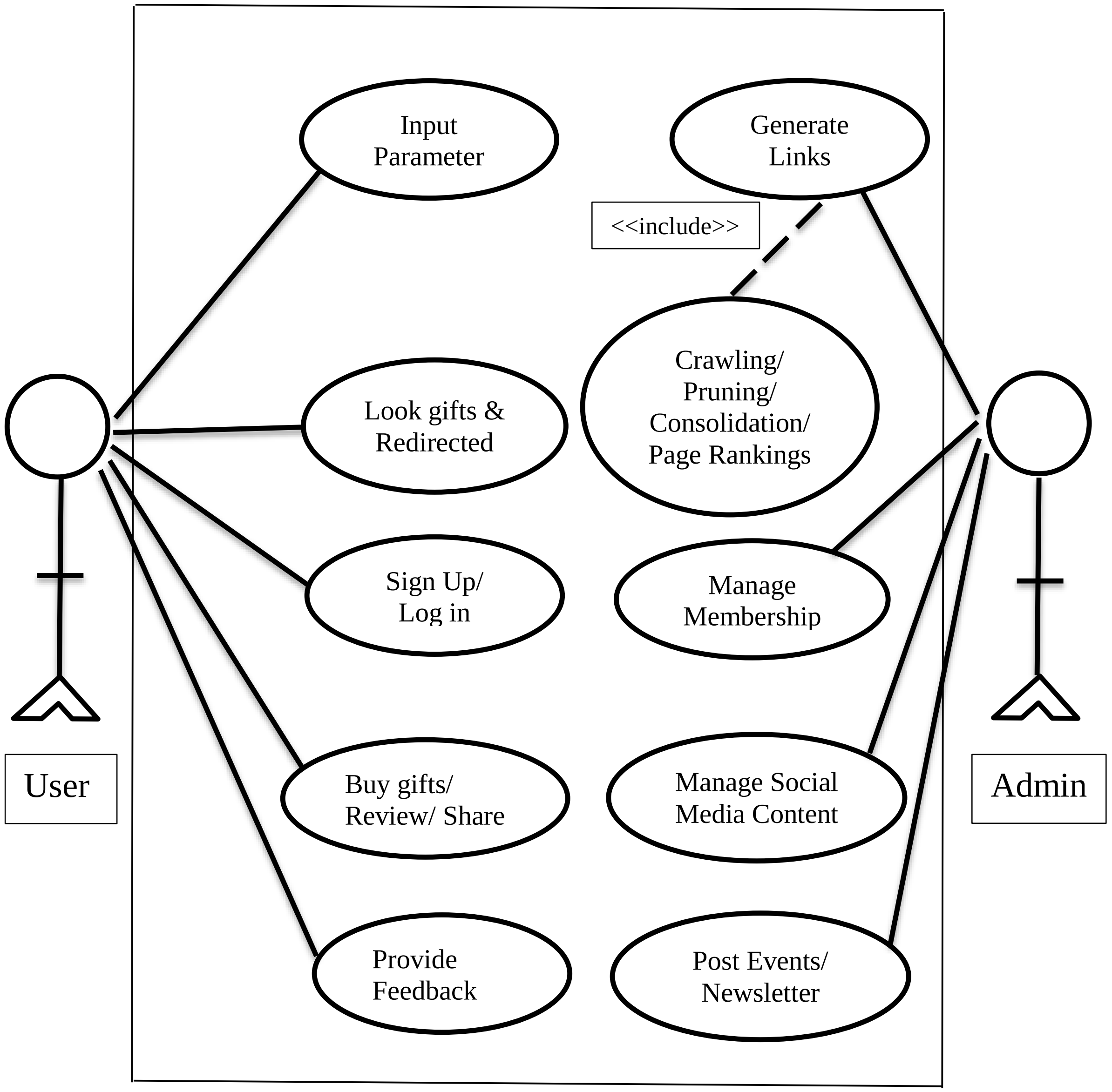


Fig. Use case Diagram

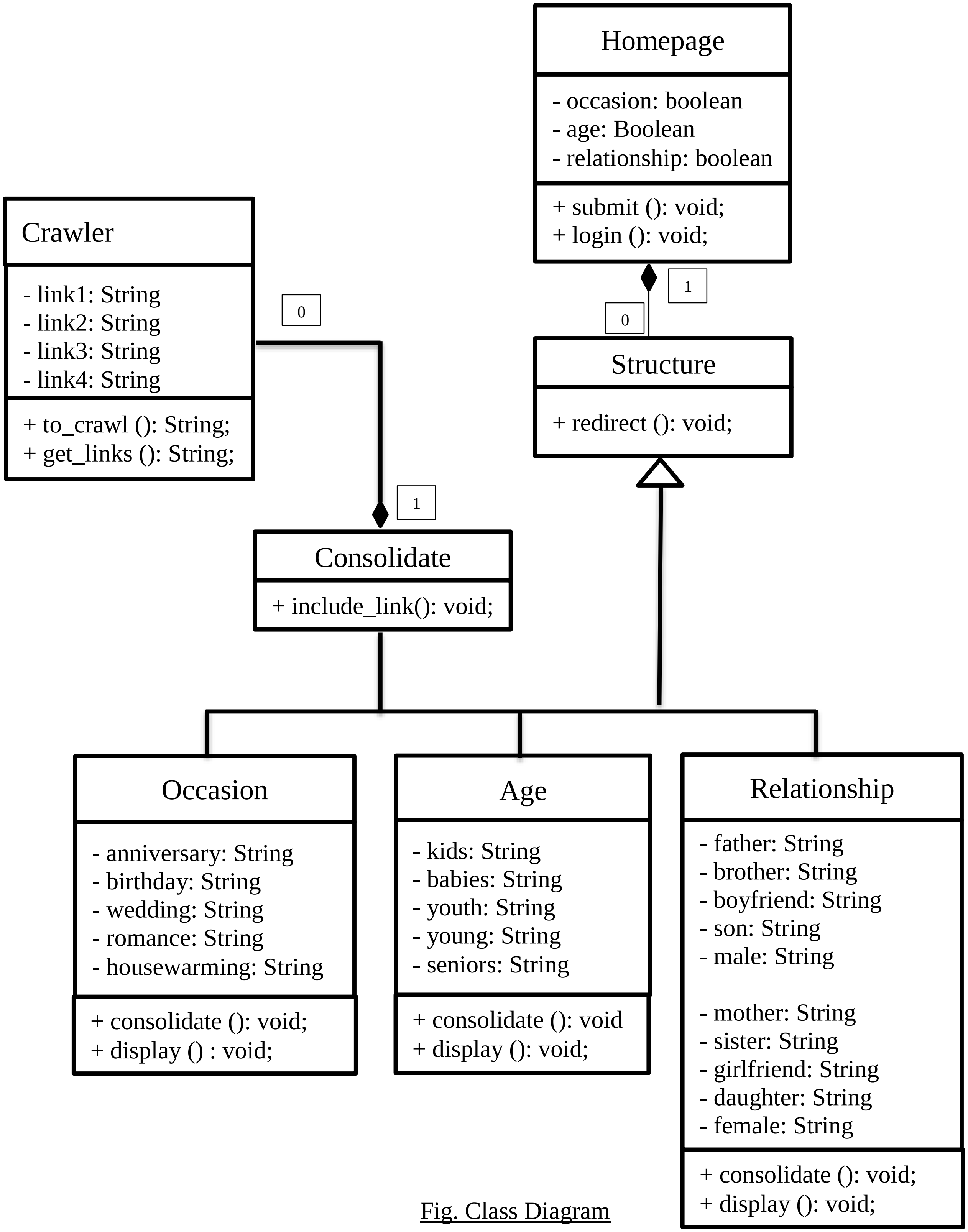


Fig. Class Diagram

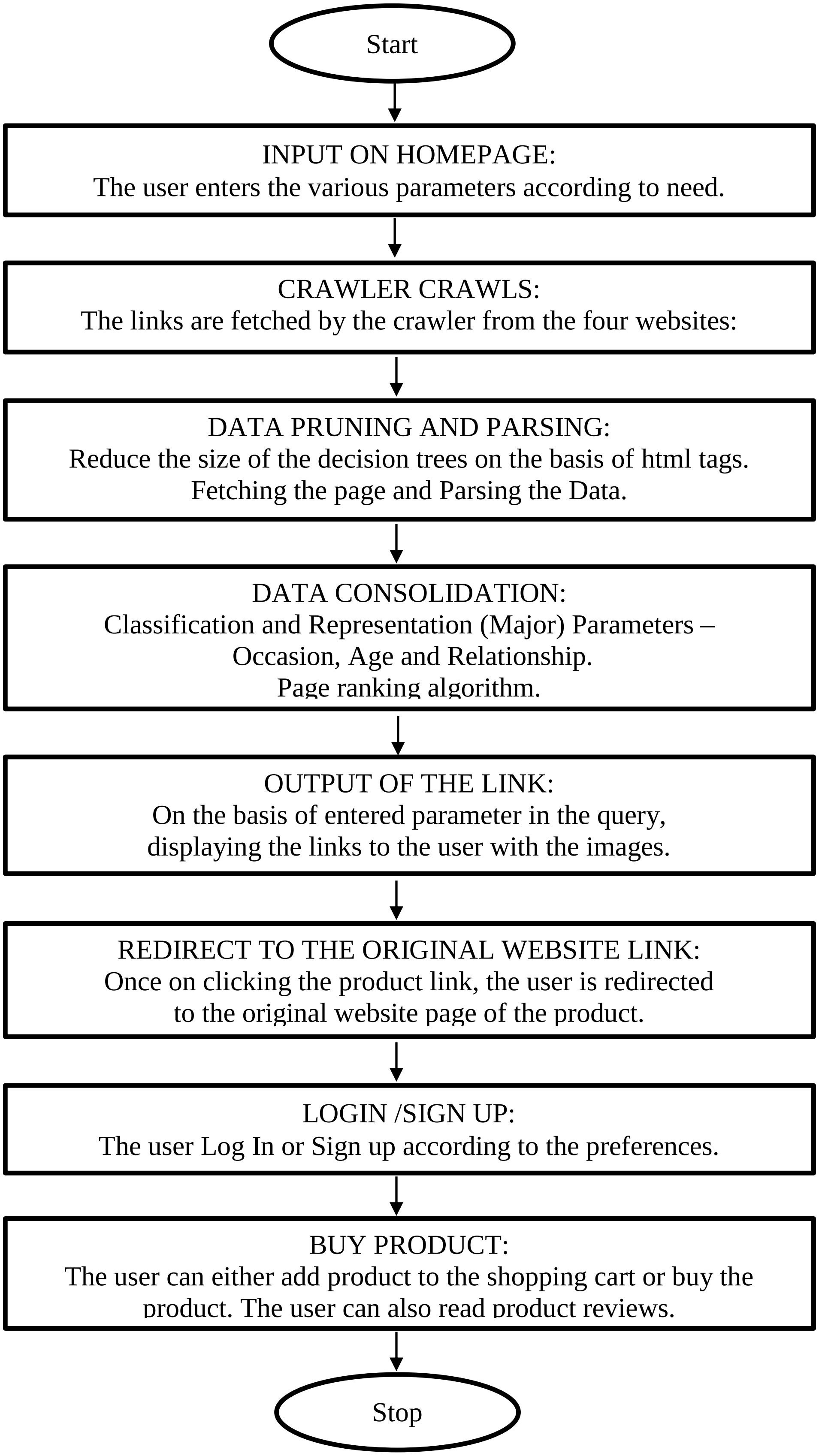

Start
INPUT ON HOMEPAGE:
The user enters the various parameters according to need.
CRAWLER CRAWLS:
The links are fetched by the crawler from the four websites:
DATA PRUNING AND PARSING:
Reduce the size of the decision trees on the basis of html tags.
Fetching the page and Parsing the Data.
DATA CONSOLIDATION:
Classification and Representation (Major) Parameters –
Occasion, Age and Relationship.
Page ranking algorithm.
OUTPUT OF THE LINK:
On the basis of entered parameter in the query,
displaying the links to the user with the images.
REDIRECT TO THE ORIGINAL WEBSITE LINK:
Once on clicking the product link, the user is redirected
to the original website page of the product.
LOGIN /SIGN UP:
The user Log In or Sign up according to the preferences.
BUY PRODUCT:
The user can either add product to the shopping cart or buy the
product. The user can also read product reviews.
Stop

Fig. Flowchart representing overall procedure for the execution of the Application

3.2 Overall Description of Project

MODULE 1: *Input Key*

Input	Parameters filled by the user.
Functionality	To take input from the user interface and provide to the system as user query.

MODULE 2: *Crawling*

Input	The query posted by the user.
Output	To fetch the pages for the particular search query made by the user.
Functionality	To crawl and search the webpages associated with the search query.

MODULE 3: *Pruning and Parsing*

Input	The URL links of the pages fetched
Output	Proper links removing the redundant data from the page links.
Functionality	Perform parsing of the links fetched from the page source of the website.

MODULE 4: ***Match***

Input	The URL of the links parsed.
Output	Display the webpages link along with its images.
Functionality	1) To match the page URL's in the queue and display the webpages link from the arrays if found. 2) Otherwise redirect to proper module to perform live crawling on the website.

MODULE 5: ***Pop up Login***

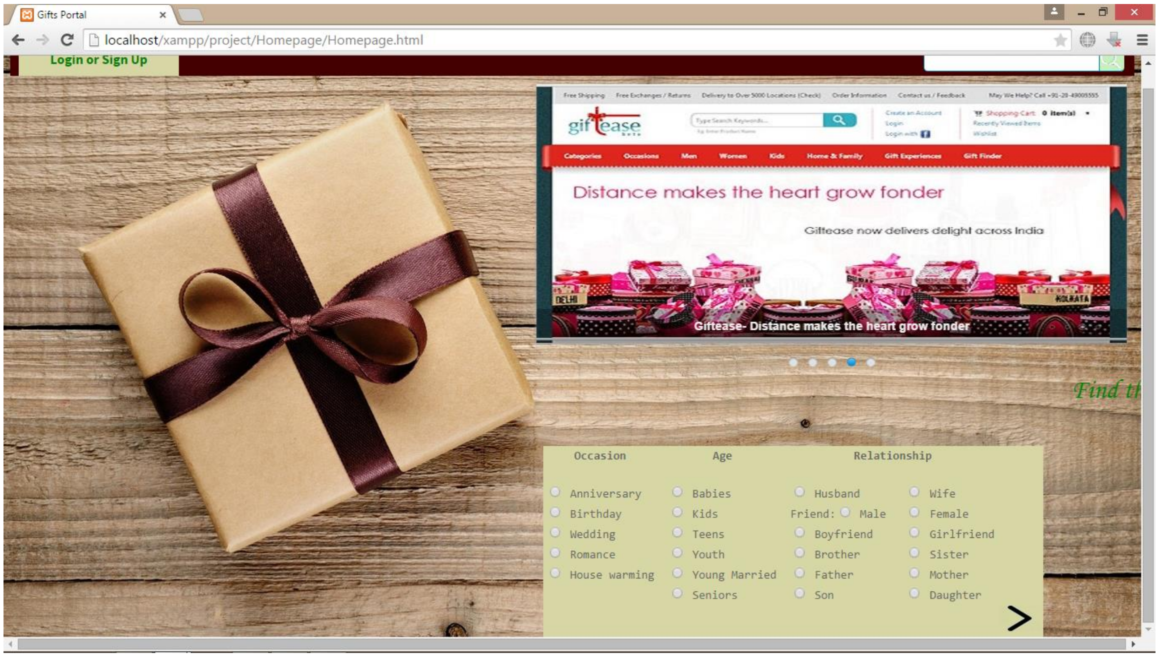
Input	User Sign in or Sign up (if account does not exist)
Output	Login/Sign Up success
Functionality	1) Identify the username and password match. 2) Proper Sign Up for the new user. 3) Login through Social Networking site. 4) Managing Shopping cart.

.

MODULE 6: ***Compare and Buy Products***

Input	User selects the product
Output	Redirected to the original websites
Functionality	Once the user selects particular product the link of the website is activated and redirected and the user can buy the product through it.

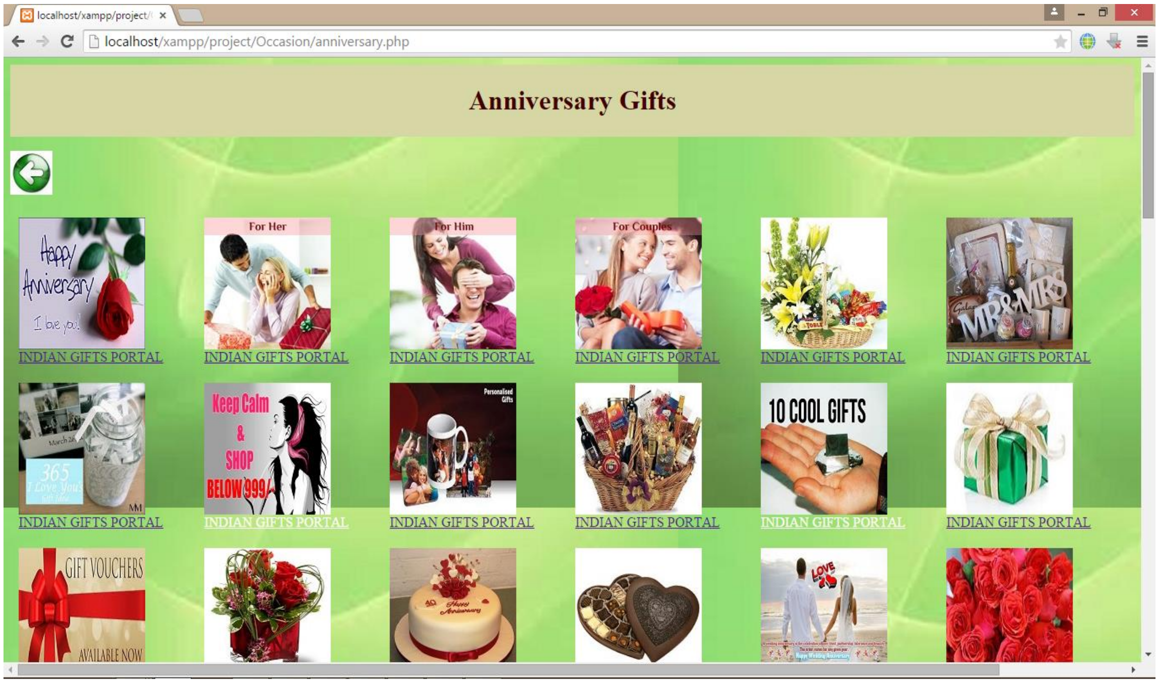
3.3 Screenshots of the Implementation in Project


Gifts Portal
localhost/xampp/project/Homepage/Homepage.html
Login or Sign Up
Free Shipping Free Exchanges / Returns Delivery to Over 5000 Locations (Check) Order Information Contact us / Feedback May We Help? Call +91-20-49005555
giftease
Type Search Keywords...
Create an Account
Login
Login with
Shopping Cart 0 item(s)
Recently Viewed Items
Wishlist
Categories Occasions Men Women Kids Home & Family Gift Experiences Gift Finder
Distance makes the heart grow fonder
Giftease now delivers delight across India
DELHI
KOLKATA
Giftease- Distance makes the heart grow fonder
Find tl

Occasion	Age	Relationship	
○ Anniversary	○ Babies	○ Husband	○ Wife
○ Birthday	○ Kids	Friend: ○ Male	○ Female
○ Wedding	○ Teens	○ Boyfriend	○ Girlfriend
○ Romance	○ Youth	○ Brother	○ Sister
○ House warming	○ Young Married	○ Father	○ Mother
	○ Seniors	○ Son	○ Daughter

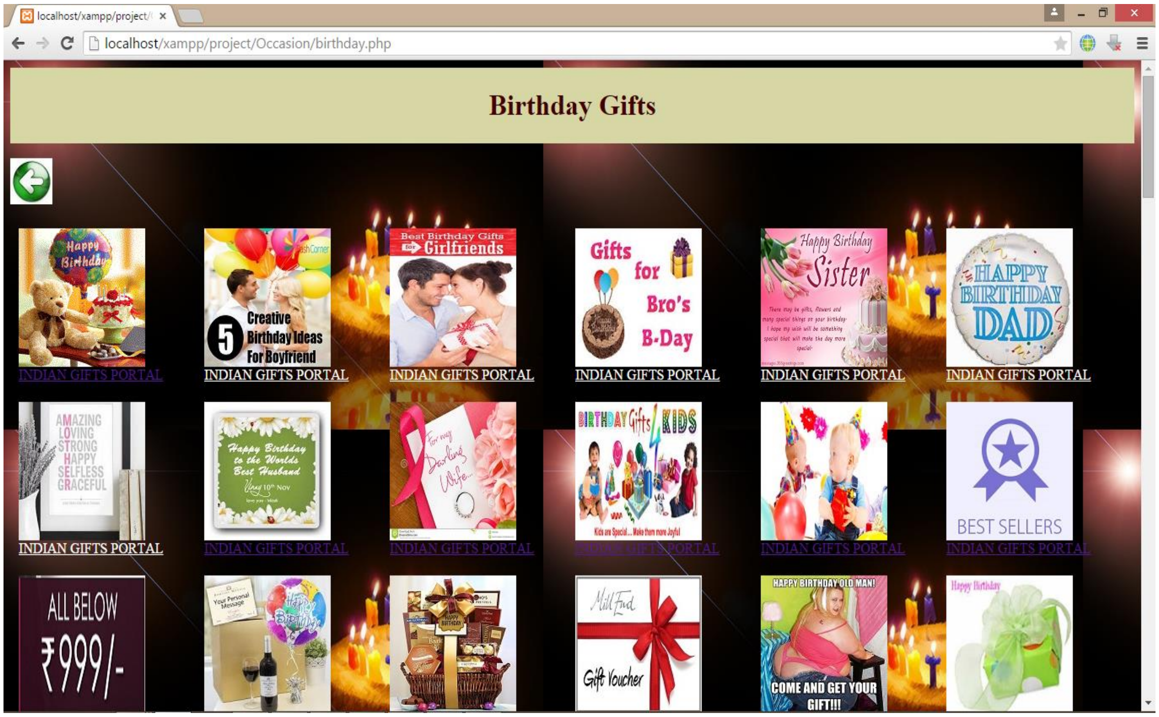
Source: http://localhost/xampp/project/Homepage/Homepage.html

Fig. Screenshot 20 reflecting the Homepage for the end users



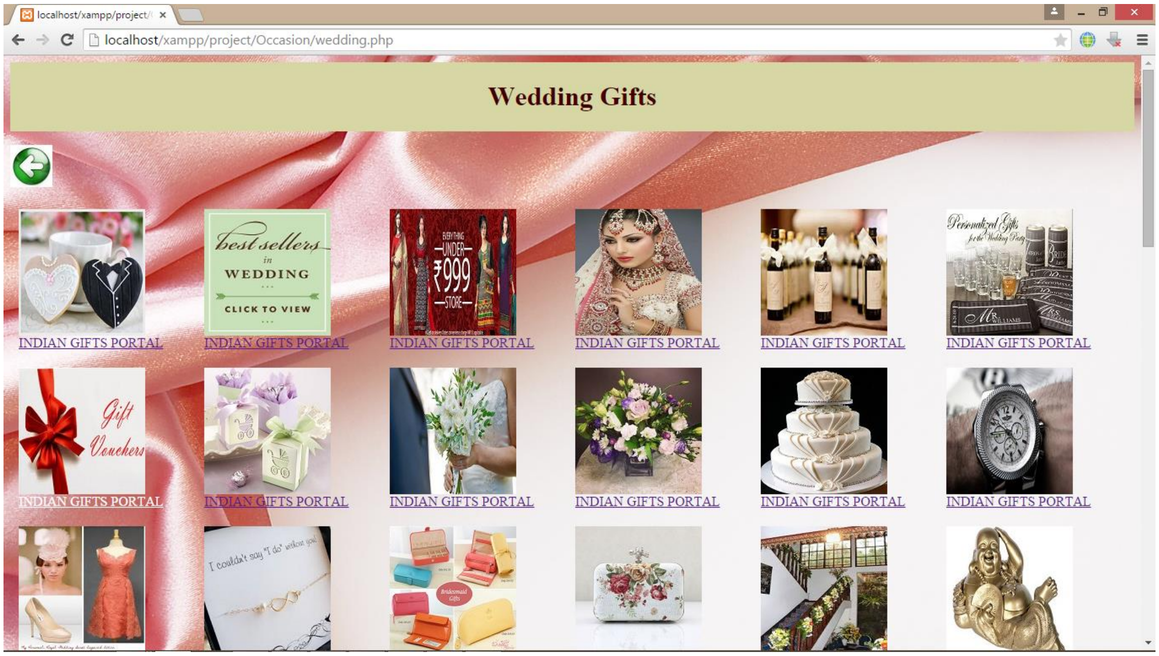
Source: http://localhost/xampp/project/Occasion/anniversary.php

Fig. Screenshot 21 showing the Anniversary Gifts page



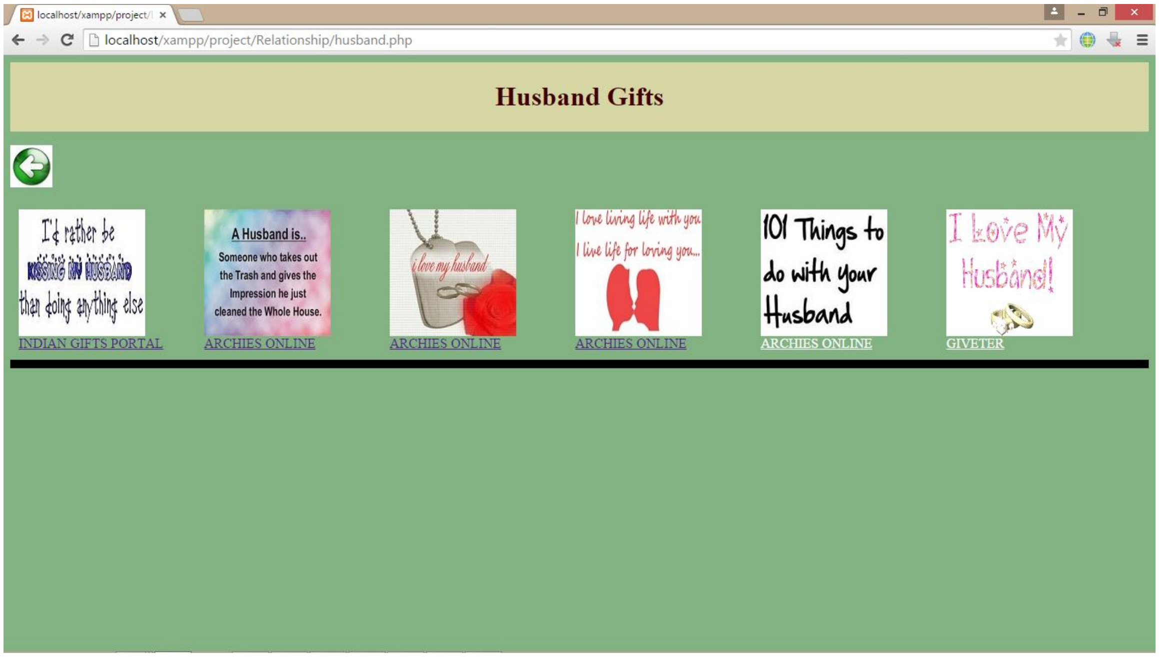
Source: http://localhost/xampp/project/Occasion/birthday.php

Fig. Screenshot 22 showing the Birthday Gifts page


localhost/xampp/project/
localhost/xampp/project/Occasion/wedding.php
Wedding Gifts
INDIAN GIFTS PORTAL
best sellers in WEDDING
CLICK TO VIEW
EVERYTHING UNDER ₹999 STORE
Personalized Gifts for the Wedding Party
Gift Vouchers
I couldn't say "I do" without you

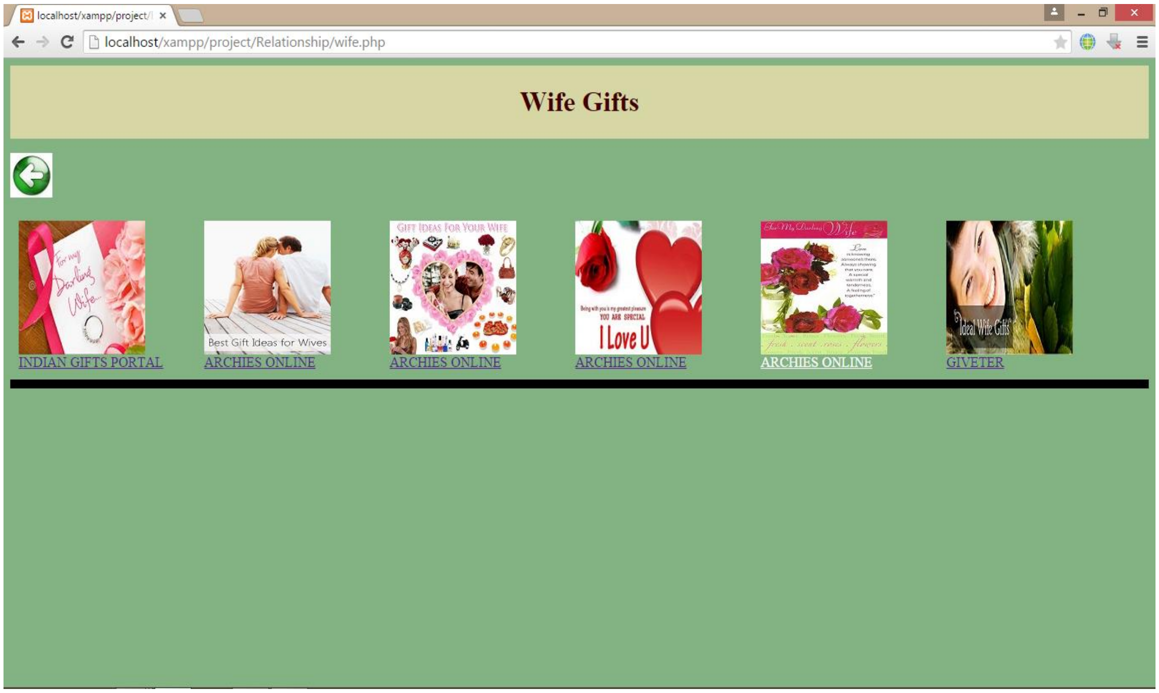
Source: http://localhost/xampp/project/Occasion/wedding.php

Fig. Screenshot 23 showing the Wedding Gifts page



Source: http://localhost/xampp/project/Relationship/husband.php

Fig. Screenshot 24 showing the Husband Gifts page



Source: http://localhost/xampp/project/Relationship/wife.php

Fig. Screenshot 25 showing the Wife Gifts page

3.4 Specific Requirements

3.4.1 External Interfaces

User Interfaces

i. Name of Item:
   - Web Browser (Internet Explorer, Google Chrome, Mozilla Firefox)
   - Xampp Server (3.2.1)
   - PHP 5 or above
   - Adobe Flash Player
   - Image Editor Software (Microsoft Paint or Adobe Photoshop 5+)
   - Adobe Dreamweaver (CS4)
   - Java Script (used implicitly by JOOMLA)

ii. Description of Purpose: To provide a platform for taking the input from user and displaying the output to the user.

iii. Source of input or destination of output:
   a. Input source: Query (by the end user)
   b. Output Destination: Web page on the browser

iv. Data Formats: Test

3.4.2 Functions

i. The system shall ask the user to fill some parameters.
ii. Do live crawling and consolidation on PHP.
iii. Display the relevant products on the webpage to the user
iv. Login display
v. Buy the product if interested.

3.4.3 Performance Requirements

Performance requirements deals with both the static and the dynamic numerical requirements placed on the software or on human interaction with the software or the application as a whole.
Static numerical requirements involve:
The number of terminals to be supported: There shall be a single terminal when the application would be tested on local host and when on web it would have no limit as it would be available to an undefined number of users.

3.4.4 Software Attributes

- Reliability: A reliable PHP and MYSQL platform, embedded in XAMPP is been used for the development of the web application. A jar application of it might be made in the end stages to make sure reliability at the time of delivery.

- Availability: Several checkpoints shall be made in the development to ensure defined availability level of the system. These checkpoints shall occur after every main module. Our system insures check pointing as it follows a step wise approach.

- Security:

  i. For security purposes, specific history data sets in the form of session handling shall be implemented.
  ii. Restrict communications between some areas of the program such as restricting the database access from the user and allow him to perform only end user functions.
  iii. A User is supposed to register so as to be able to buy the gifts from the web sites to the developers.

- Maintainability: This specifies to the attributes of the software that relate to the ease of maintenance of the software. In the project, modularity is taken into consideration and separate modules like search key, match and crawl shall be implemented up gradation and revising the current version is an easy task with the modular structure defined.

- Portability: As the system is an online web application, thus there would be no issues of portability as it would be available on the internet. Thus, making it OS independent and host independent.

Chapter 4: Conclusion and Future Works

4.1 Conclusion

To conclude with, the gifts website is a site that is aimed to suggest gifts to its user on the basis of different parameters entered by the user. The overall procedure of suggestion is discussed in the other chapters. The gifts from the four websites are consolidated and put forward for the user. The main objective was to show all the gifts under the same parameters from different websites in the same page and it is achieved.

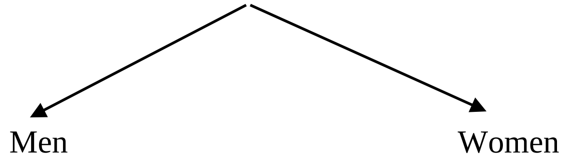
This will keep the website a little different from the rest other gifts websites and will also enable users to access it because gifting is a segment where users are much confused regarding what to gift and what not to gift. Consolidating different website products and showing them on one single page would help users find the alternative of the product at different sites. Moreover, the users can also compare price difference, delivery option, product reviews for the similar products at one go. The site also gives new innovating ideas in the market in the segment which helps users to be updated with the recent trends being followed.

The project was designed to create a website that crawls all the links on the page source of all the four websites considered. The crawler fetches and parses the links crawled and prunes the unwanted links. Once the pruned links are stored in the array then comes the implementation of the Consolidation.

Consolidation is the technique of categorizing the data on the basis of various categories. In the project the pruned links fetched are categorized on the basis of major parameters (Occasion, Age and Relationship).

The parameters are discussed in details below:

- Occasions which include anniversary, birthday, wedding, romance, and house warming.
- Age which include babies, kids, youth, young married, and seniors.
- Relationships which include two categories



Men

Women

(Grandfather, Father, Husband, Brother, Son, Boyfriend, Friend: Male)

(Grandmother, Mother, Wife, Sister, Daughter, Girlfriend, Friend: Female)

The consolidation was the most difficult part to be done. The arrays involved certain links which were getting repeated so the array of unique links was created. Websites like www.giftease.com and www.giveter.com had the products categorized on the basis of their names and product id respectively so the consolidation involved the manual study of the links and then implementing its consolidating technique.

Once the consolidation was over then came the part of putting the consolidated links in the interface where the end user can see the links along with their image and description. The best part in the implementation of the project is that there is no sort of the database created for the techniques of both crawling and the consolidation. Every time the crawler bot performs live crawling and consolidation. Hence, with this the benefit is that whenever there are any changes made by the developer teams to the original websites that are considered they will automatically be reflected in my web application. Therefore, only the database of user accounts and activities are to be taken aware of. The application also uses the Social Login which we have already discussed in the above chapters.

4.2 Advantages from the Project

- A List of option available at one place which provides easiness to the customers in buying the product.
- A good product comes into limelight and hence increases buyers.
- Entrepreneurs gets to know the demands among the users.
- No need to create a database for the product links and images.
- Data will always be updated using the live crawling.
- Applications that are able to automatically extract, evaluate and present opinions are both helpful and easy for a user to interpret.

4.3 Limitations of Solutions

- www.giftease.com website have page URL's categorized on the basis of product names rather than occasion, age and relationship parameters. So there was a problem in consolidating those links .Hence, consolidation is done manually.

- www.giveter.com websites have page URL's categorized on the basis of product Id therefore there were several links crawled and fetched under one category but all redirecting to the same webpage on the original website. Hence fetching only one link from the www.giveter.com website.

- The Log In database is not implemented because it requires permissions from the administrators of the four websites considered.

- The search bar in the header is not implemented using the internal search operation and rather is done using the external search. Search engine used is of Google (www.google.com).

4.4 Difficulties faced while implementing the Project

- The difficulty aroused in creating the pruning strategies within the crawler function. The links were to be analyzed well before applying the pruning techniques for it.

- The consolidation part had the difficulty when the predefined function took several other categories of data in the array.
  For example the strops (“male”) took the links having both “male” and “female” in the respective male array because “female” = “fe” + “male” and so the strops function extracted the substring “male” from “female”.
  Similarly there was a problem for the word “son” to categorize the links according to the links. “son” fetched the links having “personal” or ”personalized” gifts.

- Facebook connect is creating problems may be because the website URL is blocked by the university network manager.

- The image links crawled were not appearing with the proper links of the gifts. The array indexes were mismatching when the end interface was implemented.

- The description tag only had the description of the original website from where the link was fetched and not the description of the product the link was redirecting to at the end.

- The product images were saved in different dimensions so they were not getting loaded properly on the end user interface page. The images were taking time to be fetched from the respective URL’s and getting loaded. So there were certain links the images were not getting loaded properly.

4.5 Future Works

- Topical Categorization of the Parameters and their consolidation.
- Several other gifts websites should be considered.
- User friendly access to the website.
- Better display of the products with their description like an official website.
- Log In should be made after consulting the different websites companies.
- For the security purpose, the webpages should be made more secured according to the privacy policy of the Internet. The parent directory should be kept separated from the other webpages location.
- Email sender mechanisms should be made for the newsletter.
- Caching mechanisms should be implemented so as to fetch the links from the application server before going to the database server.
- Sentiment Analysis and Opinion Mining of the reviews and the product ratings on the basis of Opinion Mining Summarization.

Chapter 5: References

[1] Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik. (2014).
“Study of Web Crawler and its different types” e-ISSN: 2278-0661,
p- ISSN: 2278-8727Volume 16, Issue 1, Ver. VI
Available: http://www.iosrjournals.org

[2] Margaret Rouse. (2013). “An architect guide: How to use Big Data?”
Available: http://www.searchcloudcomputing.techtarget.com/definition/big-data-Big-Data

[3] Matt Cutts. (2010). “Google inside Search”
Available: http://www.google.co.in/insidesearch/howsearchworks/

[4] Cory Janssen. (2010). “Data Consolidation”
Available: http://www.techopedia.com/definition/28034/data-consolidation

[5] Marc Najork. (2009). “Web Crawler Architecture” in Encyclopedia of Database Systems, Springer.

[6] Jonathan Strickland. (2008). “How Google Docs works?”
Available: http://computer.howstuffworks.com/howgooglestuff works

[7] Casey Helmick. (2008). “What is Data Consolidation?” in eHow Contributor
Available: http://www.ehow.com/about_6617667_data-consolidation.html

[8] Ian Rogers. “The Google Pagerank Algorithm and How It Works”
Available: http://www.cs.princeton.edu/

[9] Gautam Pant, Padmini Srinivasan, Filippo Menczer. (2004).
“Crawling the Web” in Springer.

Appendix

Web References

[A] The websites considered for crawling and consolidation

Indian Gifts Portal	http://www.indiangiftsportal.com
Archies	http://www.archiesonline.com
Giftease	http://www.giftease.com
Giveter	http://www.giveter.com

[B] The website considered for showing the effect of pruning

Gifts	http://www.gifts.com

[C] Related work in the same domain

Sellmojo	http://www.sellmojo.com
Junglee	http://www.junglee.com
Couponduinia	http://www.coupondunia.com
Pinterest	http://www.pinterest.com
TripAdvisor	http://www.tripadvisor.com

[D] Related to the Social Login implementation

Facebook developer	http://developers.facebook.com/apps