1. **INTRODUCTION**

The deep (or hidden) web refers to the contents lie behind searchable web interfaces that cannot be indexed by searching engines. Based on extrapolations from a study done at University of California, Berkeley, it is estimated that the deep web contains approximately 91,850 terabytes and the surface web is only about 167 terabytes in 2003. More recent studies estimated that 1.9 zettabytes were reached and 0.3 zettabytes were consumed worldwide in 2007. An IDC report estimates that the total of all digital data created, replicated, and consumed will reach 6 zettabytes in 2014. A significant portion of this huge amount of data is estimated to be stored as structured or relational data in web databases — deep web makes up about 96% of all the content on the Internet, which is 500-550 times larger than the surface web. These data contain a vast amount of valuable information and entities such as Infomine, Clusty, BooksInPrint may be interested in building an index of the deep web sources in a given domain (such as book). Because these entities cannot access the proprietary web indices of search engines (e.g., Google and Baidu), there is a need for anefficient crawler that is able to accurately and quickly explore the deep web databases.

It is challenging to locate the deep web databases because they are not registered with any search engines, are usually sparsely distributed, and keep constantly changing. To address this problem, previous work has proposed two types of crawlers, *generic* *crawlers* and *focused crawlers*. Generic crawlers, fetch all searchable forms and cannot focus on a specific topic. Focused crawlers such as Form-Focused Crawler (FFC) and Adaptive Crawler for Hidden-web Entries (ACHE) can automatically search online databases on a specific topic. FFC is designed with link, page, and form classifiers for focused crawling of web forms, andis extended by ACHE with additional components for form filtering and adaptive link learner. The link classifiers in these crawlers play a pivotal role in achieving higher crawling efficiency than the best-first crawler. However, these link classifiers are usedto predict the distance to the page containing searchable forms, which is difficult to estimate, especially for the delayed benefit links (links eventually lead to pages with forms). As a result, the crawler can be inefficiently led to pages without targeted forms.

Besides efficiency, quality and coverage on relevant deep web sources are also challenging. Crawler must produce a large quantity of high-quality results from the most relevant content sources. For assessing source quality, SourceRank ranks the results from the selected sources by computing the agreement between them. When selecting a relevant subset from the available content sources, FFC and ACHE prioritize links that bring immediate return (links directly point to pages con- taining searchable forms) and delayed benefit links. But the set of retrieved forms is very heterogeneous. For example, from a set of representative domains, on average only 16% of forms retrieved by FFC are relevant. Furthermore, little work has been done on the source selection problem when crawling more content sources. Thus it is crucial to develop smart crawling strategies that are able to quickly discover relevant content sources from the deep web as much as possible.

In this paper, we propose an effective deep web harvesting framework, namely *SmartCrawler*, for achieving both wide coverage and high efficiency for a focused crawler. Based on the observation that deep websites usually contain a few searchable forms and most of them are within a depth of three, our crawler is divided into two stages: *site locating* and *in-site exploring*. The site locating stage helps achieve wide coverage of sites for a focused crawler, and the in-site exploring stage can efficiently perform searches for web forms within a site. Our main contributions are:

* We propose a novel two-stage framework to address the problem of searching for hidden-web resources. Our site locating technique employs a *reverse searching* technique (e.g., using Google’s”link:” facility to get pages pointing to a given link) and incremental two-level site prioritizing technique for unearthing relevant sites, achieving more data sources. During the in-site exploring stage, we design a link tree for balanced link prioritizing, eliminating bias toward webpages in popular directories.
* We propose an adaptive learning algorithm that performs online feature selection and uses these features to automatically construct link rankers. In the site locating stage, high relevant sites are prioritized and the crawling is focused on a topic using the contents of the root page of sites, achieving more accurate results. During the insite exploring stage, relevant links are prioritized for fast in-site searching.

We have performed an extensive performance evaluation of *SmartCrawler* over real web data in 12 representative domains and compared with ACHE and a site-based crawler. Our evaluation shows that our crawling framework is very effective, achieving substantially higher harvest rates than the state-of-the-art ACHE crawler. The results also show the effectiveness of the reverse searching and adaptive learning.

**Existing System:**

The existing system is a manual or semi-automated system, i.e. The Textile Management System is the system that can directly sent to the shop and will purchase clothes whatever you wanted.

The users are purchase dresses for festivals or by their need. They can spend time to purchase this by their choice like color, size, and designs, rate and so on. They But now in the world everyone is busy. They don’t need time to spend for this. Because they can spend whole the day to purchase for their whole family. So we proposed the new system for web crawling.

**Disadvantages:**

1. Consuming large amount of data’s.

2. Time wasting while crawl in the web.

**Proposed System:**

We propose a two-stage framework, namely Smart Crawler, for efficient harvesting deep web interfaces. In the first stage, Smart Crawler performs site-based searching for center pages with the help of search engines, avoiding visiting a large number of pages. To achieve more accurate results for a focused crawl, SmartCrawlerranks websites to prioritize highly relevant ones for a given topic. In the second stage, Smart Crawler achieves fast in-site searching by excavating most relevant links with an adaptive link-ranking. To eliminate bias on visiting some highly relevant links in hidden web directories, we design a link tree data structure to achieve wider coverage for a website. Our experimental results on a set of representative domains show the agility and accuracy of our proposed crawler framework, which efficiently retrieves deep-web interfaces from large-scale sites and achieves higher harvest rates than other crawlers. Propose an effective harvesting framework for deep-web interfaces, namely Smart-Crawler. We have shown that our approach achieves both wide coverage for deep web interfaces and maintains highly efficient crawling. Smart Crawler is focused crawler consisting of two stages: efficient site locating and balanced in-site exploring. SmartCrawlerperforms site-based locating by reversely searching the known deep web sites for center pages, which can effectively find many data sources for sparse domains. By ranking collected sites and by focusing the crawling on a topic, Smart Crawlerachieves more accurate results

**MODULE DESCRIPTION:**

# Number of Modules:

After careful analysis the system has been identified to have the following modules:

1. **two-stage crawler.**

2**. Site Ranker**

**3. Adaptive learning**

**1 Two-stage crawler.**

It is challenging to locate the deep web databases, because they are not registered with any search engines, are usually sparsely distributed, and keep constantly changing. To address this problem, previous work has proposed two types of crawlers, generic crawlers and focused crawlers. Generic crawlers fetch all searchable forms and cannot focus on a specific topic. Focused crawlers such as Form-Focused Crawler (FFC) and Adaptive Crawler for Hidden-web Entries (ACHE) can automatically search online databases on a specific topic. FFC is designed with link, page, and form classifiers for focused crawling of web forms, and is extended by ACHE with additional components for form filtering and adaptive link learner. The link classifiers in these crawlers play a pivotal role in achieving higher crawling efficiency than the best-first crawler However, these link classifiers are used to predict the distance to the page containing searchable forms, which is difficult to estimate, especially for the delayed benefit links (links eventually lead to pages with forms). As a result, the crawler can be inefficiently led to pages without targeted forms.

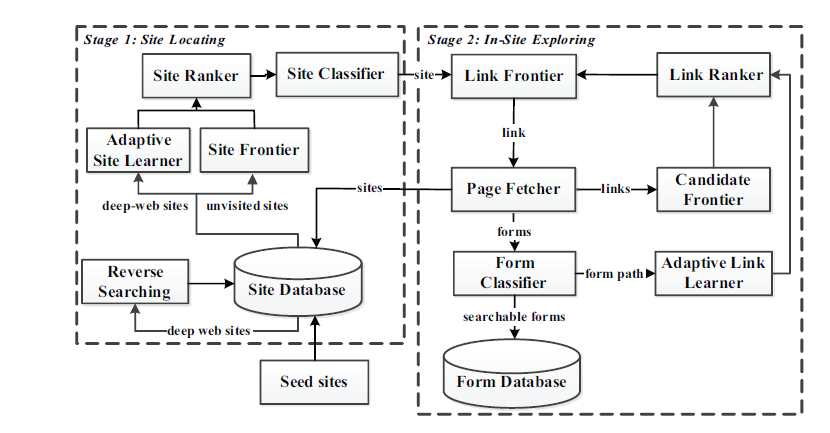
**2. Site Ranker:**

When combined with above stop-early policy. We solve this problem by prioritizing highly relevant links with link ranking. However, link ranking may introduce bias for highly relevant links in certain directories. Our solution is to build a linktreefor a balanced link prioritizing. Figure 2 illustrates an example of a link tree constructed from the homepage of http://www.abebooks.com. Internal nodes of the tree represent directory paths. In this example, servlet directory is for dynamic request; books directory is for displaying different catalogs of books; Amdocs directory is for showing help information. Generally each directory usually represents one type of files on web servers and it is advantageous to visit links in different directories. For links that only differ in the query string part, we consider them as the same URL.Because links are often distributed unevenly in server directories, prioritizing links by the relevance can potentially bias toward some directories. For instance, the links under books might be assigned a high priority, because “book” is an important feature word in the URL. Together with the fact that most links appear in the books directory, it is quite possible that links in other directories will not be chosen due to low relevance score. As a result, the crawler may miss searchable forms in those directories.

**3. Adaptive learning**

Adaptive learning algorithm that performs online feature selection and uses these features to automatically construct link rankers. In the site locating stage, high relevant sites are prioritized and the crawling is focused on atopic using the contents of the root page of sites, achieving more accurate results. During the insiteexploring stage, relevant links are prioritized for fast in-site searching. We have performed an extensive performance evaluation ofSmart Crawler over real web data in 1representativedomains and compared with ACHE and site-based crawler. Our evaluation shows that our crawling framework is very effective, achieving substantially higher harvest rates than the state-of-the-art ACHE crawler. The results also show the effectiveness of the reverse searching and adaptive learning.

**Two-Stage Architecture**



**1.2 ORGANIZATION PROFILE**

**About Cloud Technologies:**

**Cloud technologies** are an internationally established software development company with offices In India. The Indian office and development centre is located in a state-of-art facility in the heart of Hyderabad; India. Cloud Technologies is into software product development and enterprise consulting services. Cloud Technologies service offerings cater to multiple industry domains and verticals, following full or partial SDLC, optimally customized to cater to specific client needs. Cloud Technologies also provide human resources for the different areas of a software development life cycle.

**Cloud Technologies Vision:**

**Cloud technologies** are an interactive and technology solution Provider. We help clients to plan their online strategy, budget their technology investments, integrate critical applications, and implement projects that achieve business objectives and improve performance. Since our inception we have focused on building dynamic, user-focused web sites, intranets and extranets, supported by interactive marketing and e-communication campaigns.

**Cloud Technologies College Training:**

**CLOUD TECHNOLOGIES** give confidence and assist students to study and perform in their project reports by our College Training. Our devoted team helps students to understand project ideas and execute it as per the current technology. We support throughout the growth of the project and also assist to deliver the abstract and project report for all Projects. Selection of the project is a very important aspect of the Final Year Program. We have a very detailed discussion with all of our students in Colleges about the project and finalization of the project would be done based on their interest in the subject, skill assessment and feasibility of the project in our College Training. We would always encourage students to come out with innovative concepts so as to exploit their inbuilt creativity and if the concept is really strong we would also partner with the students in marketing the project in our College Training

**College Training By Cloud technologies**

**CLOUD TECHNOLOGIES** give confidence and assist students to study and perform in their project reports by our College Training. Our devoted team helps students to understand project ideas and execute it as per the current technology. We support throughout the growth of the project and also assist to deliver the abstract and project report for all Projects. Selection of the project is a very important aspect of the Final Year Program. We have a very detailed discussion with all of our students in Colleges about the project and finalization of the project would be done based on their interest in the subject, skill assessment and feasibility of the project in our College Training. We would always encourage students to come out with innovative concepts so as to exploit their inbuilt creativity and if the concept is really strong we would also partner with the students in marketing the project in our College Training

**Cloud Technologies Software Development**

Cloud technologies not only providing academic projects but also have our own software development center to develop real time projects based on the student requirements and feasibility. We keep trying to serve our level best services in providing Academic and Student projects for various branches of engineering, MCA and Match students on multiple platforms. We are proud to announce that, Cloud Technologies is about to start a complete online training program in various courses by well experienced trainers. Today, we are supporting several real time projects that help various individuals add value to their resumes and annual sheets. Academic Projects works towards transforming, cultivating and nurturing the spirit of project development, thus playing a vital role in creating intellectual corporate executives.

1. **LITERATURE SURVEY**
2. A Survey on **“An Adaptive Crawler for Locating Hidden-Web Entry Points”**

**Abstract**

In this paper we describe new adaptive crawling strategies to efficiently locate the entry points to hidden-Web sources. The fact that hidden-Web sources are very sparsely distributed makes the problem of locating them especially challenging. We deal with this problem by using the contents of pages to focus the crawl on a topic; by prioritizing promising links within the topic; and by also following links that may not lead to immediate benefit. We propose a new framework whereby crawlers automatically learn patterns of promising links and adapt their focus as the crawl progresses, thus greatly reducing the amount of required manual setup and tuning. Our experiments over real Web pages in a representative set of domains indicate that online learning leads to significant gains in harvest rates—the adaptive crawlers retrieve up to three times as many forms as crawlers that use a fixed focus strategy.

**Conclusion**

We have presented a new adaptive focused crawling strategy for efficiently locating hidden-Web entry points. This strategy effectively balances the exploitation of acquired knowledge with the exploration of links with previously unknown patterns, making it robust and able to correct biases introduced in the learning process. We have shown, through a detailed experimental evaluation, that substantial increases in harvest rates are obtained as crawlers learn from new experiences. Since crawlers that learn from scratch are able to obtain harvest rates that are comparable to, and sometimes higher than manually configured crawlers, this framework can greatly reduce the effort to configure a crawler. In addition, by using the form classifier, ACHE produces highquality results that are crucial for a number information integration tasks.

There are several important directions we intend to pursue in future work. As discussed in Section 5, we would like to integrate the apprentice of into the ACHE framework. To accelerate the learning process and better handle very sparse domains, we will investigate the effectiveness and trade-offs involved in using back-crawling during the learning iterations to increase the number of sample paths. Finally, to further reduce the effort of crawler configuration, we are currently exploring strategies to simplify the creation of the domain-specific form classifiers. In particular, the use of form clusters obtained by the online-database clustering technique described in as the training set for the classi- fier.

1. A Survey on **“Understanding the Deep Web”**

**Abstract**

The most coveted commodity of the information age is indeed information. Information has become a basic need after food, shelter, and clothing. Due to technological advancements, a large amount of information is available on the Web, which has become a complex entity containing information from a variety of sources. Information is found using search engines. A searcher has access to a large amount of information, but it still far from the huge treasury of information lying beneath the Web, a vast store of information beyond the reach of conventional search engines: the “Deep Web” or “Invisible Web.”

The contents of the Deep Web are not included up in the search results of conventional search engines. The crawlers of conventional search engines identify only static pages and cannot access the dynamic Web pages of Deep Web databases. Hence, the Deep Web is alternatively termed the “Hidden” or “Invisible Web.” The term Invisible Web was coined by Dr. Jill Ellsworth to refer to information inaccessible to conventional search engines. But using the term Invisible Web to describe recorded information that is available but not easily accessible, is not accurate.

**Conclusion**

The advent of Internet and access to global information was a great benefit, even though information managers had the difficult task of organizing, retrieving, and providing access to precise information. Users depend on the popular search engines and portals, which cannot provide access to the hidden store of valuable information available in the Deep Web. To access the information available on these databases, users will have to become familiar with the structure of the Deep Web. Any information created should be shared and used, since that alone leads to the creation of more information. When a specific database is created, information regarding its existence should published so that users will be aware and make maximum use of available information.

1. A Survey on **“Relevance and Trust Assessment for Deep Web Sources Based on Inter-Source Agreement”**

**Abstract**

One immediate challenge in searching the deep web databases is source selection—i.e. selecting the most relevant web databases for answering a given query. The existing database selection methods (both text and relational) assess the source quality based on the query-similarity-based relevance assessment. When applied to the deep web these methods have two deficiencies. First is that the methods are agnostic to the correctness (trustworthiness) of the sources. Secondly, the query based relevance does not consider the importance of the results. These two considerations are essential for the open collections like the deep web. Since a number of sources provide answers to any query, we conjuncture that the agreements between these answers are likely to be helpful in assessing the importance and the trustworthiness of the sources. We compute the agreement between the sources as the agreement of the answers returned. While computing the agreement, we also measure and compensate for possible collusion between the sources. This adjusted agreement is modeled as a graph with sources at the vertices. On this agreement graph, a quality score of a source that we call SourceRank, is calculated as the stationary visit probability of a random walk. We evaluate SourceRank in multiple domains, including sources in Google Base, with sizes up to 675 sources. We demonstrate that the SourceRank tracks source corruption. Further, our relevance evaluations show that SourceRank improves precision by 22-60% over the Google Base and other baseline methods. SourceRank has been implemented in a system called Factal.

**Conclusion**

A compelling holy grail for the information retrieval research is to integrate and search the structured deep web sources. An immediate problem posed by this quest is source selection, i.e. selecting relevant and trustworthy sources to answer a query. Past approaches to this problem depended on purely query based measures to assess the relevance of a source. The relevance assessment based solely on query similarity is easily tampered by the content owner, as the measure is insensitive to the popularity and trustworthiness of the results. The sheer number and uncontrolled nature of the sources in the deep web leads to significant variability among the sources, and necessitates a more robust measure of relevance sensitive to source popularity and trustworthiness. To this end, we proposed SourceRank, a global measure derived solely from the degree of agreement between the results returned by individual sources. SourceRank plays a role akin to PageRank but for data sources. Unlike PageRank however, it is derived from implicit endorsement (measured in terms of agreement) rather than from explicit hyperlinks. For added robustness of the ranking, we assess and compensate for the source collusion while computing the agreements. Our comprehensive empirical evaluation shows that SourceRank improves relevance sources selected compared to existing methods and effectively removes corrupted sources. We also demonstrated that combining SourceRank with Google Product search ranking significantly improves the quality of the results.

1. A Survey on **“MODEL-BASED RICH INTERNET APPLICATIONS CRAWLING: “MENU” AND “PROBABILITY” MODELS”**

**Abstract**

Strategies for “crawling” Web sites efficiently have been described more than a decade ago. Since then, Web applications have come a long way both in terms of adoption to provide information and services and in terms of technologies to develop them. With the emergence of richer and more advanced technologies such as AJAX, “Rich Internet Applications” (RIAs) have become more interactive, more responsive and generally more user friendly. Unfortunately, we have also lost our ability to crawl them.

Building models of applications automatically is important not only for indexing content, but also to do automated testing, automated security assessments, automated accessibility assessment and in general to use software engineering tools. We must regain our ability to efficiently construct models for these RIAs. In this paper, we present two methods, based on “Model-Based Crawling” (MBC) first introduced in [1]: the “menu” model and the “probability” model. These two methods are shown to be more effective at extracting models than previously published methods, and are much simpler to implement than previous models for MBC. A distributed implementation of the probability model is also discussed. We compare these methods and others against a set of experimental and “real” RIAs, showing that in our experiments, these methods find the set of client states faster than other approaches, and often finish the crawl faster as well.

**Conclusion**

Two new model-based crawling algorithms were introduced: the Menu model and the Probability model. We models RIAs as a graph, based on the JavaScript events in each state of the DOM. The approach consists of making assumptions about the category of events in order to derive a strategy, then learn, and adapt these categories as the crawling proceeds. In the Menu model, the events are assumed to either always lead to the same state (menu events), or to always lead to new states. In the probability strategy, an actual probability of leading to a new state is calculated for each event, based on how successful it was in the past at finding new states. A concurrent implementation of the probability model was also described. A prototype implementation for these strategies was built and the performance of these strategies is compared with several other algorithms. We have shown empirically that our new strategies are better than other known strategies when it comes to finding all the states of the application being modeled. We have also evidences showing that the concurrent implementation of the probability strategy provides an good additional speed up, opening the door to fast crawling of RIA using concurrent crawlers.

Some directions for future studies are:

* Better definition of states: Learning about independent widgets, as suggested by [24], can substantially improve performance of both Menu and Probabilistic model by reducing the number of events to execute.
  + Adaptive Crawling: starting from a generic meta-model and recognizing the meta model of a web-application on-the-fly is another area of improvement.
  + Relaxing assumptions: In future work, it will be important to relax or waive some of the assumptions made in regard to this work. In particular the assumptions that the applications is deterministic from the viewpoint of the model builder.

1. A Survey on **“Optimal Algorithms for Crawling a Hidden Database in the Web”**

**Abstract**

A hidden database refers to a dataset that an organization makes accessible on the web by allowing users to issue queries through a search interface. In other words, data acquisition from such a source is not by following static hyper-links. Instead, data are obtained by querying the interface, and reading the result page dynamically generated. This, with other facts such as the interface may answer a query only partially, has prevented hidden databases from being crawled effectively by existing search engines.

This paper remedies the problem by giving algorithms to extract all the tuples from a hidden database. Our algorithms are provably efficient, namely, they accomplish the task by performing only a small number of queries, even in the worst case. We also establish theoretical results indicating that these algorithms are asymptotically optimal – i.e., it is impossible to improve their efficiency by more than a constant factor. The derivation of our upper and lower bound results reveals significant insight into the characteristics of the underlying problem. Extensive experiments confirm the proposed techniques work very well on all the real datasets examined.

**Conclusion**

Currently, search engines cannot effectively index hidden databases, and are thus unable to direct queries to the relevant data in those repositories. With the rapid growth in the amount of such hidden data, this problem has severely limited the scope of information accessible to ordinary Internet users. In this paper, we attacked an issue that lies at the heart of the problem, namely, how to crawl a hidden database in its entirety with the smallest cost. We have developed algorithms for solving the problem when the underlying dataset has only numeric attributes, only categorical attributes, or both. All our algorithms are asymptotically optimal, i.e., none of them can be improved by more than constant times in the worst case. Our theoretical analysis has also revealed the factors that determine the hardness of the problem, as well as how much influence each of those factors has on the hardness.

1. **SYSTEM ANALYSIS**

**3.1. EXISTING SYSTEM**

The existing system is a manual or semi-automated system, i.e. The Textile Management System is the system that can directly sent to the shop and will purchase clothes whatever you wanted.

The users are purchase dresses for festivals or by their need. They can spend time to purchase this by their choice like color, size, and designs, rate and so on.They But now in the world everyone is busy. They don’t need time to spend for this. Because they can spend whole the day to purchase for their whole family. So we proposed the new system for web crawling.

**Disadvantages:**

1. Consuming large amount of data’s.

2. Time wasting while crawl in the web.

**3.2. PROPOSED SYSTEM**

We propose a two-stage framework, namely Smart Crawler, for efficient harvesting deep web interfaces. In the first stage, Smart Crawler performs site-based searching for center pages with the help of search engines, avoiding visiting a large number of pages. To achieve more accurate results for a focused crawl, SmartCrawlerranks websites to prioritize highly relevant ones for a given topic. In the second stage, Smart Crawler achieves fast in-site searching by excavating most relevant links with an adaptive link-ranking. To eliminate bias on visiting some highly relevant links in hidden web directories, we design a link tree data structure to achieve wider coverage for a website. Our experimental results on a set of representative domains show the agility and accuracy of our proposed crawler framework, which efficiently retrieves deep-web interfaces from large-scale sites and achieves higher harvest rates than other crawlers. Propose an effective harvesting framework for deep-web interfaces, namely Smart-Crawler. We have shown that our approach achieves both wide coverage for deep web interfaces and maintains highly efficient crawling. Smart Crawler is focused crawler consisting of two stages: efficient site locating and balanced in-site exploring. SmartCrawlerperforms site-based locating by reversely searching the known deep web sites for center pages, which can effectively find many data sources for sparse domains. By ranking collected sites and by focusing the crawling on a topic, Smart Crawlerachieves more accurate results

**3.3. THE STUDY OF THE SYSTEM**

# Number of Modules:

After careful analysis the system has been identified to have the following modules:

1. **two-stage crawler.**

2**. Site Ranker**

**3. Adaptive learning**

**1 Two-stage crawler.**

It is challenging to locate the deep web databases, because they are not registered with any search engines, are usually sparsely distributed, and keep constantly changing. To address this problem, previous work has proposed two types of crawlers, generic crawlers and focused crawlers. Generic crawlers fetch all searchable forms and cannot focus on a specific topic. Focused crawlers such as Form-Focused Crawler (FFC) and Adaptive Crawler for Hidden-web Entries (ACHE) can automatically search online databases on a specific topic. FFC is designed with link, page, and form classifiers for focused crawling of web forms, and is extended by ACHE with additional components for form filtering and adaptive link learner. The link classifiers in these crawlers play a pivotal role in achieving higher crawling efficiency than the best-first crawler However, these link classifiers are used to predict the distance to the page containing searchable forms, which is difficult to estimate, especially for the delayed benefit links (links eventually lead to pages with forms). As a result, the crawler can be inefficiently led to pages without targeted forms.

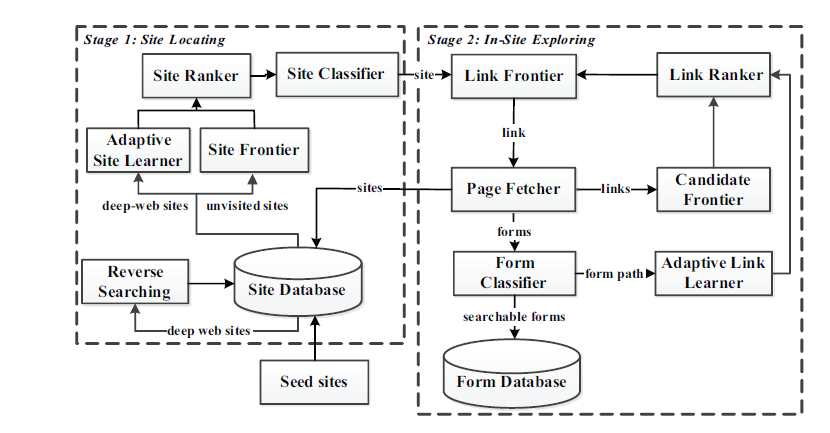
**2. Site Ranker:**

When combined with above stop-early policy. We solve this problem by prioritizing highly relevant links with link ranking. However, link ranking may introduce bias for highly relevant links in certain directories. Our solution is to build a linktreefor a balanced link prioritizing. Figure 2 illustrates an example of a link tree constructed from the homepage of http://www.abebooks.com. Internal nodes of the tree represent directory paths. In this example, servlet directory is for dynamic request; books directory is for displaying different catalogs of books; Amdocs directory is for showing help information. Generally each directory usually represents one type of files on web servers and it is advantageous to visit links in different directories. For links that only differ in the query string part, we consider them as the same URL.Because links are often distributed unevenly in server directories, prioritizing links by the relevance can potentially bias toward some directories. For instance, the links under books might be assigned a high priority, because “book” is an important feature word in the URL. Together with the fact that most links appear in the books directory, it is quite possible that links in other directories will not be chosen due to low relevance score. As a result, the crawler may miss searchable forms in those directories.

**3. Adaptive learning**

Adaptive learning algorithm that performs online feature selection and uses these features to automatically construct link rankers. In the site locating stage, high relevant sites are prioritized and the crawling is focused on atopic using the contents of the root page of sites, achieving more accurate results. During the insiteexploring stage, relevant links are prioritized for fast in-site searching. We have performed an extensive performance evaluation ofSmart Crawler over real web data in 1representativedomains and compared with ACHE and site-based crawler. Our evaluation shows that our crawling framework is very effective, achieving substantially higher harvest rates than the state-of-the-art ACHE crawler. The results also show the effectiveness of the reverse searching and adaptive learning.

**Two-Stage Architecture**



**3.4. INPUT & OUTPUT REPRESENTATION**

**3.4.1. Input Design**

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

* What data should be given as input?
* How the data should be arranged or coded?
* The dialog to guide the operating personnel in providing input.
* Methods for preparing input validations and steps to follow when error occur.

**Objectives**

Input Design is the process of converting a user-oriented description of the input into a computer-based system. This design is important to avoid errors in the data input process and show the correct direction to the management for getting correct information from the computerized system.

It is achieved by creating user-friendly screens for the data entry to handle large volume of data. The goal of designing input is to make data entry easier and to be free from errors. The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

When the data is entered it will check for its validity. Data can be entered with the help of screens. Appropriate messages are provided as when needed so that the user will not be in maize of instant. Thus the objective of input design is to create an input layout that is easy to follow

**3.4.2. Output Design**

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system’s relationship to help user decision-making.

* 1. Designing computer output should proceed in an organized, well thought out manner; the right output must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.
  2. Select methods for presenting information.
  3. Create document, report, or other formats that contain information produced by the system.

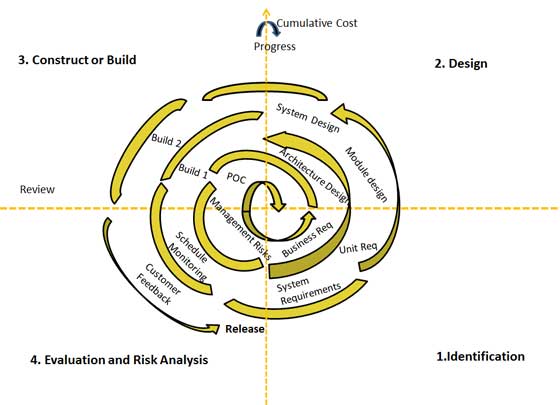
The output form of an information system should accomplish one or more of the following objectives.

* Convey information about past activities, current status or projections of the future.
* Signal important events, opportunities, problems, or warnings.
* Trigger an action.
* Confirm an action.

**3.5. PROCESS MODEL USED WITH JUSTIFICATION**

SDLC is nothing but Software Development Life Cycle. It is a standard which is used by software industry to develop good software.

**SDLC (Spiral Model):**

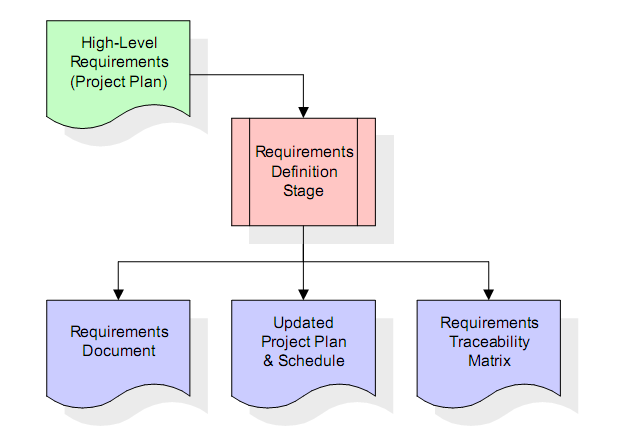
****

**Stages of SDLC:**

* Requirement Gathering and Analysis
* Designing
* Coding
* Testing
* Deployment

**3.5.1. Requirements Definition Stage and Analysis:**

The requirements gathering process takes as its input the goals identified in the high-level requirements section of the project plan. Each goal will be refined into a set of one or more requirements. These requirements define the major functions of the intended application, define operational data areas and reference data areas, and define the initial data entities. Major functions include critical processes to be managed, as well as mission critical inputs, outputs and reports. A user class hierarchy is developed and associated with these major functions, data areas, and data entities. Each of these definitions is termed a Requirement. Requirements are identified by unique requirement identifiers and, at minimum, contain a requirement title and textual description.



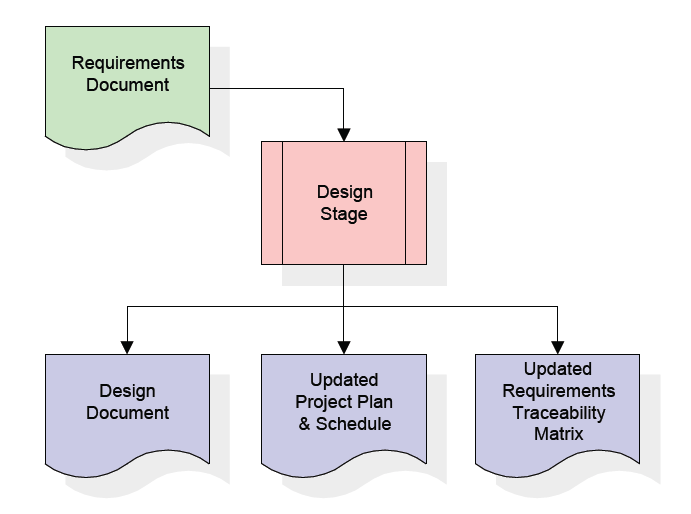
These requirements are fully described in the primary deliverables for this stage: the Requirements Document and the Requirements Traceability Matrix (RTM). the requirements document contains complete descriptions of each requirement, including diagrams and references to external documents as necessary. Note that detailed listings of database tables and fields are *not* included in the requirements document. The title of each requirement is also placed into the first version of the RTM, along with the title of each goal from the project plan. The purpose of the RTM is to show that the product components developed during each stage of the software development lifecycle are formally connected to the components developed in prior stages.

In the requirements stage, the RTM consists of a list of high-level requirements, or goals, by title, with a listing of associated requirements for each goal, listed by requirement title. In this hierarchical listing, the RTM shows that each requirement developed during this stage is formally linked to a specific product goal. In this format, each requirement can be traced to a specific product goal, hence the term *requirements traceability*. The outputs of the requirements definition stage include the requirements

document, the RTM, and an updated project plan.

**3.5.2. Design Stage:**

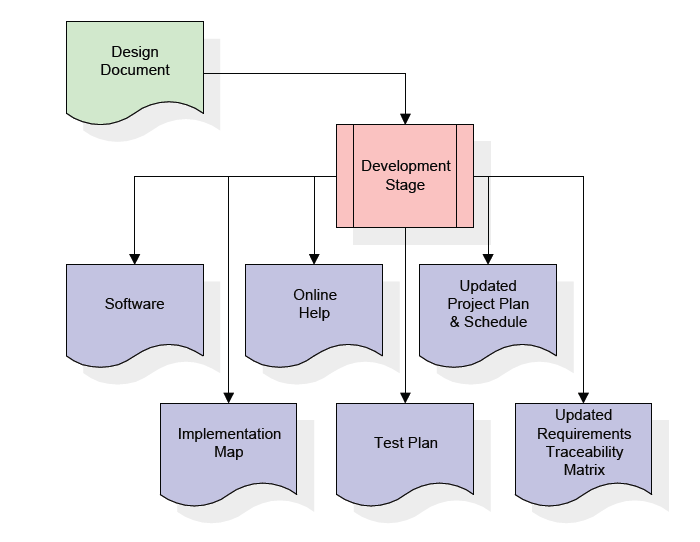
The design stage takes as its initial input the requirements identified in the approved requirements document. For each requirement, a set of one or more design elements will be produced as a result of interviews, workshops, and/or prototype efforts. Design elements describe the desired software features in detail, and generally include functional hierarchy diagrams, screen layout diagrams, tables of business rules, business process diagrams, pseudo code, and a complete entity-relationship diagram with a full data dictionary. These design elements are intended to describe the software in sufficient detail that skilled programmers may develop the software with minimal additional input.



When the design document is finalized and accepted, the RTM is updated to show that each design element is formally associated with a specific requirement. The outputs of the design stage are the design document, an updated RTM, and an updated project plan.

**Development Stage:**

The development stage takes as its primary input the design elements described in the approved design document. For each design element, a set of one or more software artifacts will be produced. Software artifacts include but are not limited to menus, dialogs, data management forms, data reporting formats, and specialized procedures and functions. Appropriate test cases will be developed for each set of functionally related software artifacts, and an online help system will be developed to guide users in their interactions with the software.



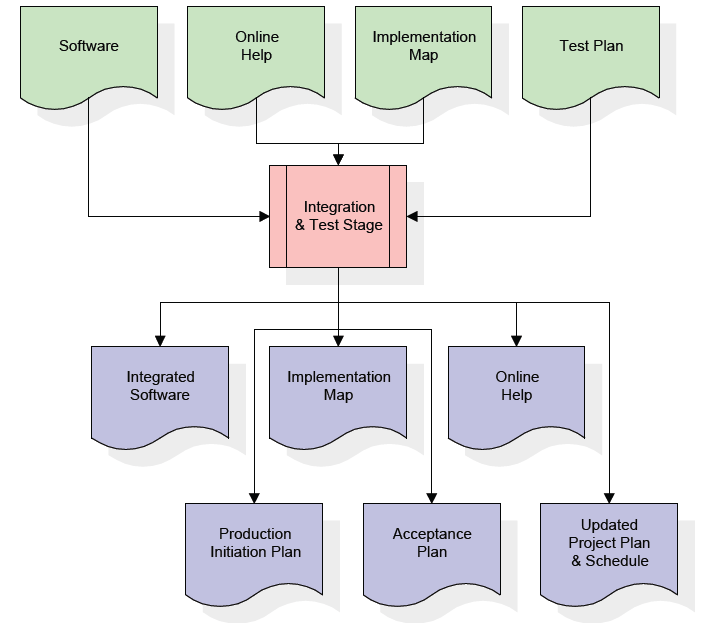
The RTM will be updated to show that each developed artifact is linked to a specific design element, and that each developed artifact has one or more corresponding test case items. At this point, the RTM is in its final configuration. The outputs of the development stage include a fully functional set of software that satisfies the requirements and design elements previously documented, an online help system that describes the operation of the software, an implementation map that identifies the primary code entry points for all major system functions, a test plan that describes the test cases to be used to validate the correctness and completeness of the software, an updated RTM, and an updated project plan.

**Integration Test Phage:**

During the integration and test stage, the software artifacts, online help, and test data are migrated from the development environment to a separate test environment. At this point, all test cases are run to verify the correctness and completeness of the software. Successful execution of the test suite confirms a robust and complete migration capability.

During this stage, reference data is finalized for production use and production users are identified and linked to their appropriate roles. The final reference data (or links to reference data source files) and production user list are compiled into

the Production Initiation Plan.

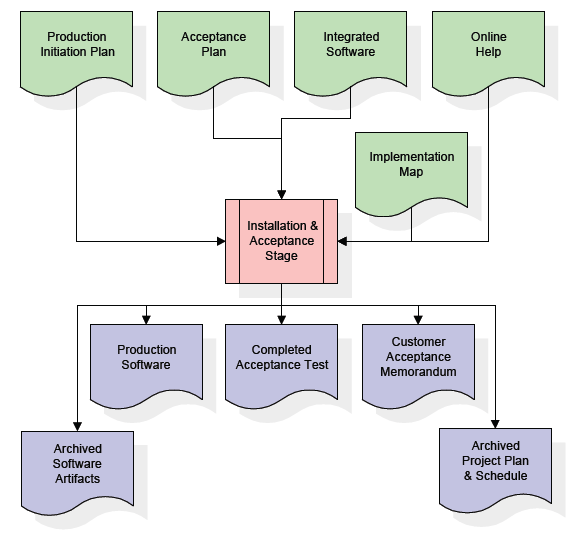


The outputs of the integration and test stage include an integrated set of software, an online help system, an implementation map, a production initiation plan that describes reference data and production users, an acceptance plan which contains the final suite of test cases, and an updated project plan.

**Installation & Acceptance Stage**

During the installation and acceptance stage, the software artifacts, online help, and initial production data are loaded onto the production server. At this point, all test cases are run to verify the correctness and completeness of the software. Successful execution of the test suite is a prerequisite to acceptance of the software by the customer.

After customer personnel have verified that the initial production data load is correct and the test suite has been executed with satisfactory results, the customer formally accepts the delivery of the software.



The primary outputs of the installation and acceptance stage include a production application, a completed acceptance test suite, and a memorandum of customer acceptance of the software. Finally, the PDR enters the last of the actual labor data into the project schedule and locks the project as a permanent project record. At this point the PDR "locks" the project by archiving all software items, the implementation map, the source code, and the documentation for future reference.

* 1. **SYSTEM ARCHITECTURE**

**Architecture Flow:**

Below architecture diagram represents mainly flow of request from the users to database through servers. In this scenario overall system is designed in three tiers separately using three layers called presentation layer, business layer, data link layer. This project was developed using 3-tier architecture.

**3-Tier Architecture:**

The three-tier software architecture (a three layer architecture) emerged in the 1990s to overcome the limitations of the two-tier architecture. The third tier (middle tier server) is between the user interface (client) and the data management (server) components. This middle tier provides process management where business logic and rules are executed and can accommodate hundreds of users (as compared to only 100 users with the two tier architecture) by providing functions such as queuing, application execution, and database staging.

The three tier architecture is used when an effective distributed client/server design is needed that provides (when compared to the two tier) increased performance, flexibility, maintainability, reusability, and scalability, while hiding the complexity of distributed processing from the user. These characteristics have made three layer architectures a popular choice for Internet applications and net-centric information systems

**Advantages of Three-Tier:**

* Separates functionality from presentation.
* Clear separation - better understanding.
* Changes limited to well define components.
* Can be running on WWW.
* Effective network performance**.**
  1. **FEASIBILITY STUDY**

Feasibility study is conducted once the problem is clearly understood. The feasibility study which is a high level capsule version of the entire system analysis and design process. The objective is to determine whether the proposed system is feasible or not and it helps us to the minimum expense of how to solve the problem and to determine, if the Problem is worth solving. The following are the three important tests that have been carried out for feasibility study.

* + 1. **Technical feasibility**

In the technical feasibility study, one has to test whether the proposed system can be developed using existing technology or not. It is planned to implement the proposed system in JSP. The project entitled is technically feasible because of the following reasons.

* + All necessary technology exists to develop the system.
  + The existing system is so flexible that it can be developed further.
    1. **Economic feasibility**

As a part of this, the costs and benefits associated with the proposed systems are to be compared. The project is economically feasible only if tangible and intangible benefits outweigh the cost. We can say the proposed system is feasible based on the following grounds.

* + The cost of developing the full system is reasonable.
  + The cost of hardware and software for the application is less.
    1. **Operational feasibility**

The project is operationally feasible because there is sufficient support from the project management and the users of the proposed system .Proposed system definitely does not harm and will not produce the bad results and no problem will arise after implementation of the system.

**User-friendly**

Customer will use the forms for their various transaction i.e. for adding new routes, viewing the routes details. Also the Customer wants the reports to view the various transactions based on the constraints. Theses forms and reports are generated as user-friendly to the Client.

**Reliability**

The package wills pick-up current transactions on line. Regarding the old transactions, User will enter them in to the system.

**Security**

The web server and database server should be protected from hacking, virus etc.

**Availability**

This software will be available always.

1. **SYSTEM REQUIREMENTS SPECIFICATION**
   1. **SOFTWARE SYSTEM SPECIFICATION**

Operating System : Windows Family

Technology : Java and J2EE

Web Technologies : Html, JavaScript, CSS

Web Server : Tomcat

Database : My SQL

Java Version : J2SDK1.5

* 1. **Hardware System Specification**

Hardware - Pentium

Speed - 1.1 GHz

RAM - 1GB

Hard Disk - 20 GB

Key Board - Standard Windows Keyboard

Mouse - Two or Three Button Mouse

Monitor - SVGA

* 1. **. PLATFORM USED:**

**4.3.1 THE JAVA PROGRAMMING LANGUAGE**

**History**

Java is a programming language created by James Gosling from Sun Microsystems in 1991. The first publicly available version of Java (Java 1.0) was released in 1995. Over time new enhanced versions of Java have been released. The current version of Java is Java 1.7 which is also known as *Java 7*.

From the Java programming language the Java platform evolved. The Java platform allows that the program code is written in other languages than the Java programming language and still runs on the Java virtual machine.

**Java Virtual machine**

The Java virtual machine (JVM) is a software implementation of a computer that executes programs like a real machine. The Java virtual machine is written specifically for a specific operating system, e.g. for Linux a special implementation is required as well as for Windows.

Java programs are compiled by the Java compiler into so-called *byte code*. The Java virtual machine interprets this *byte code* and executes the Java program.

**Java Runtime Environment vs. Java Development Kit**

Java comes in two flavors, the *Java Runtime Environment* (JRE) and the *Java Development Kit* (JDK). The Java runtime environment (JRE) consists of the JVM and the Java class libraries and contains the necessary functionality to start Java programs. The JDK contains in addition the development tools necessary to create Java programs. The JDK consists therefore of a Java compiler, the Java virtual machine, and the Java class libraries.

The Java Development Kit (JDK) is provided by Sun Microsystems as a basic development environment for Java. The JDK provides similar facilities to the cc compiler for C programs, plus a JVM simulator and some additional facilities such as a debugger. To use the JDK, programs are constructed as ascii text files (by using an editor, for example). The program files are compiled, which translates the Java code to JVM bytecode in .class files.

Each public class must be in a file having the class name (case sensitive on Unix) followed by a .java suffix. There may be any number of classes defined in a .java file, but the compiler produces a separate .class file for each class. A file is compiled with the **javac** command, which is similar to the cc (or gcc) command. A class is executed (or more precisely, the method **main** in a class is executed) by the command **java** with the class name (not the .class file) as the parameter. Thus, for example, to compile the program in file Hi.java, we would use the command

***javac Hi.java***

and then to execute the program we would use the command

***java Hi***

Both compile-time and execution-time (exceptions) error messages include the file name and line where the error occurred. No .class file is produced if there is a compile-time error.

**Characteristics of Java**

The target of Java is to write a program once and then run this program on multiple operating systems.

Java has the following properties:

* Platform independent: Java programs use the Java virtual machine as abstraction and do not access the operating system directly. This makes Java programs highly portable. A Java program which is standard complaint and follows certain rules can run unmodified on all supported platforms, e.g. Windows or Linux.
* Object-orientated programming language: Except the primitive data types, all elements in Java are objects.
* Strongly-typed programming language: Java is strongly-typed, e.g. the types of the used variables must be pre-defined and conversion to other objects is relatively strict, e.g. must be done in most cases by the programmer.
* Interpreted and compiled language: Java source code is transferred into the byte code format which does not depend on the target platform. These byte code instructions will be interpreted by the Java Virtual machine (JVM). The JVM contains a so called Hotspot-Compiler which translates performance critical byte code instructions into native code instructions.
* Automatic memory management: Java manages the memory allocation and de-allocation for creating new objects. The program does not have direct access to the memory. The so-called garbage collector deletes automatically objects to which no active pointer exists.

The Java syntax is similar to C++. Java is case sensitive, e.g. the variables my Value and my value will be treated as different variables.

**4.3.2 JAVA PLATFORM**

The programmer writes Java source code in a text editor which supports plain text. Normally the programmer uses an *Integrated Development Environment* (IDE) for programming. An IDE supports the programmer in the task of writing code, e.g. it provides auto-formatting of the source code, highlighting of the important keywords, etc.

At some point the programmer (or the IDE) calls the Java compiler (javac). The Java compiler creates the *byte code* instructions. . These instructions are stored in .class files and can be executed by the Java Virtual Machine.

**J2SE**

J2SE is a collection of Java Programming Language API (Application programming interface) that is very useful to many Java platform programs. It is derived from one of the most dynamic programming language known as "JAVA"

J2SE is a collection of Java Programming Language API (Application programming interface) that is very useful to many Java platform programs. It is derived from one of the most dynamic programming language known as "JAVA" and one of its three basic editions of Java known as Java standard edition being used for writing Applets and other web based applications.

J2SE platform has been developed under the Java umbrella and primarily used for writing applets and other Java-based applications. It is mostly used for individual computers. Applet is a type of fast-working subroutine of Java that is platform-independent but  work within other frameworks. It is a mini application that perform a variety of functions, large and small, ordinary and dynamic, within the framework of larger applications.

J2SE provide the facility to users to see Flash movies or hear audio files by clicking on a Web page link. As the user clicks, page goes into the browser environment and begin the process of launching application-within-an-application to play the requested video or sound application. So many online games are being developed on J2SE. JavaBeans can also be developed by using J2SE.

**J2EE**

Java ***2*** Platform ***E***nterprise ***E***dition. J2EE is a [platform](http://www.webopedia.com/TERM/P/platform.html)-independent, [Java](http://www.webopedia.com/TERM/J/Java.html)-centric environment from Sun for developing, building and deploying Web-based enterprise applications online. The J2EE platform consists of a set of services, [APIs](http://www.webopedia.com/TERM/A/API.html), and [protocols](http://www.webopedia.com/TERM/P/protocol.html) that provide the functionality for developing multitier, Web-based applications.

**Key features and services of J2EE:**

* At the [client](http://www.webopedia.com/TERM/C/client.html) tier, J2EE supports pure [HTML](http://www.webopedia.com/TERM/H/HTML.html), as well as Java [applets](http://www.webopedia.com/TERM/A/applet.html) or applications. It relies on [Java Server Pages](http://www.webopedia.com/TERM/J/JSP.html) and [servlet](http://www.webopedia.com/TERM/S/servlet.html) code to create HTML or other formatted data for the client.
* [Enterprise JavaBeans](http://www.webopedia.com/TERM/E/Enterprise_JavaBeans.html) (EJBs) provide another layer where the platform's logic is stored. An EJB server provides functions such as threading, concurrency, security and memory management. These services are transparent to the author.
* [Java Database Connectivity](http://www.webopedia.com/TERM/J/JDBC.html) (JDBC), which is the Java equivalent to [ODBC](http://www.webopedia.com/TERM/O/ODBC.html), is the standard interface for Java databases.
* The Java servlet API enhances consistency for developers without requiring a [graphical user interface](http://www.webopedia.com/TERM/G/Graphical_User_Interface_GUI.html).

* 1. **JAVA DATA BASE CONNECTION**

JDBC stands for **J**ava **D**ata**b**ase **C**onnectivity, which is a standard Java API for database-independent connectivity between the Java programming language and a wide range of databases.

The JDBC library includes APIs for each of the tasks commonly associated with database usage:

* Making a connection to a database
* Creating SQL or MySQL statements
* Executing that SQL or MySQL queries in the database
* Viewing & Modifying the resulting records

Fundamentally, JDBC is a specification that provides a complete set of interfaces that allows for portable access to an underlying database. Java can be used to write different types of executables, such as:

* Java Applications
* Java Applets
* Java Servlets
* Java ServerPages (JSPs)
* Enterprise JavaBeans (EJBs)

All of these different executables are able to use a JDBC driver to access a database and take advantage of the stored data.

JDBC provides the same capabilities as ODBC, allowing Java programs to contain database-independent code.

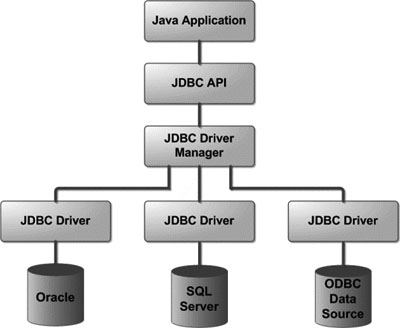
**JDBC Architecture:**

The JDBC API supports both two-tier and three-tier processing models for database access but in general JDBC Architecture consists of two layers:

* **JDBC API:** This provides the application-to-JDBC Manager connection.
* **JDBC Driver API:** This supports the JDBC Manager-to-Driver Connection.

The JDBC API uses a driver manager and database-specific drivers to provide transparent connectivity to heterogeneous databases.The JDBC driver manager ensures that the correct driver is used to access each data source. The driver manager is capable of supporting multiple concurrent drivers connected to multiple heterogeneous databases.

Following is the architectural diagram, which shows the location of the driver manager with respect to the JDBC drivers and the Java application.



**Common JDBC Components:**

The JDBC API provides the following interfaces and classes:

* **DriverManager:** This class manages a list of database drivers. Matches connection requests from the java application with the proper database driver using communication subprotocol. The first driver that recognizes a certain subprotocol under JDBC will be used to establish a database Connection.
* **Driver:** This interface handles the communications with the database server. You will interact directly with Driver objects very rarely. Instead, you use DriverManager objects, which manages objects of this type. It also abstracts the details associated with working with Driver objects
* **Connection :** This interface with all methods for contacting a database. The connection object represents communication context, i.e., all communication with database is through connection object only.
* **Statement :** You use objects created from this interface to submit the SQL statements to the database. Some derived interfaces accept parameters in addition to executing stored procedures.
* **ResultSet:** These objects hold data retrieved from a database after you execute an SQL query using Statement objects. It acts as an iterator to allow you to move through its data.
* **SQLException:** This class handles any errors that occur in a database application.

**What is JDBC Driver?**

JDBC drivers implement the defined interfaces in the JDBC API for interacting with your database server.For example, using JDBC drivers enable you to open database connections and to interact with it by sending SQL or database commands then receiving results with Java.

The *Java.sql* package that ships with JDK contains various classes with their behaviours defined and their actual implementaions are done in third-party drivers. A third party vendor implements the *java.sql.Driver* interface in their database driver.

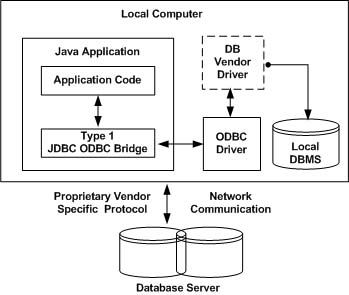
**JDBC Drivers Types:**

JDBC driver implementations vary because of the wide variety of operating systems and hardware platforms in which Java operates. Sun has divided the implementation types into four categories, Types 1, 2, 3, and 4, which is explained below:

**Type 1: JDBC-ODBC Bridge Driver:**

In a Type 1 driver, a JDBC bridge is used to access ODBC drivers installed on each client machine. Using ODBC requires configuring on your system a Data Source Name (DSN) that represents the target database.

When Java first came out, this was a useful driver because most databases only supported ODBC access but now this type of driver is recommended only for experimental use or when no other alternative is available.

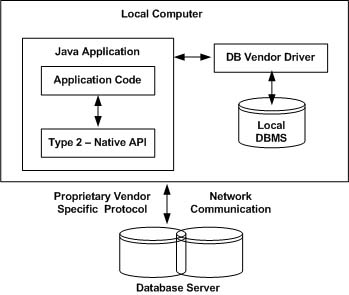


The JDBC-ODBC bridge that comes with JDK 1.2 is a good example of this kind of driver.

**Type 2: JDBC-Native API:**

In a Type 2 driver, JDBC API calls are converted into native C/C++ API calls which are unique to the database. These drivers typically provided by the database vendors and used in the same manner as the JDBC-ODBC Bridge, the vendor-specific driver must be installed on each client machine.

If we change the Database we have to change the native API as it is specific to a database and they are mostly obsolete now but you may realize some speed increase with a Type 2 driver, because it eliminates ODBC's overhead.

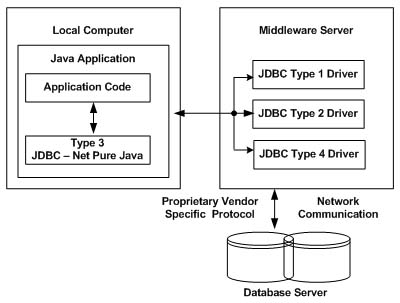


The Oracle Call Interface (OCI) driver is an example of a Type 2 driver.

**Type 3: JDBC-Net pure Java:**

In a Type 3 driver, a three-tier approach is used to accessing databases. The JDBC clients use standard network sockets to communicate with an middleware application server. The socket information is then translated by the middleware application server into the call format required by the DBMS, and forwarded to the database server.

This kind of driver is extremely flexible, since it requires no code installed on the client and a single driver can actually provide access to multiple databases.

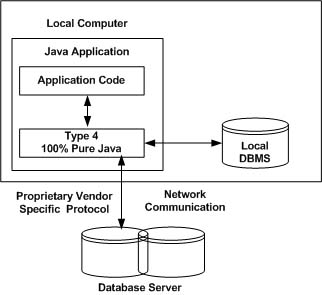


You can think of the application server as a JDBC "proxy," meaning that it makes calls for the client application. As a result, you need some knowledge of the application server's configuration in order to effectively use this driver type.

Your application server might use a Type 1, 2, or 4 driver to communicate with the database, understanding the nuances will prove helpful.

**Type 4: 100% pure Java:**

In a Type 4 driver, a pure Java-based driver that communicates directly with vendor's database through socket connection. This is the highest performance driver available for the database and is usually provided by the vendor itself. This kind of driver is extremely flexible, you don't need to install special software on the client or server. Further, these drivers can be downloaded dynamically.



MySQL's Connector/J driver is a Type 4 driver. Because of the proprietary nature of their network protocols, database vendors usually supply type 4 drivers.

**JDBC Database Connections**

After you've installed the appropriate driver, it's time to establish a database connection using JDBC.

The programming involved to establish a JDBC connection is fairly simple. Here are these simple four steps:

* **Import JDBC Packages:** Add **import** statements to your Java program to import required classes in your Java code.
* **Register JDBC Driver:** This step causes the JVM to load the desired driver implementation into memory so it can fulfill your JDBC requests.
* **Database URL Formulation:** This is to create a properly formatted address that points to the database to which you wish to connect.
* **Create Connection Object:** Finally, code a call to the *DriverManager* object's *getConnection( )* method to establish actual database connection.

**Import JDBC Packages:**

The **Import** statements tell the Java compiler where to find the classes you reference in your code and are placed at the very beginning of your source code.

To use the standard JDBC package, which allows you to select, insert, update, and delete data in SQL tables, add the following *imports* to your source code:

import java.sql.\* ; // for standard JDBC programs

import java.math.\* ; // for BigDecimal and BigInteger support

**Register JDBC Driver:**

You must register the your driver in your program before you use it. Registering the driver is the process by which the Oracle driver's class file is loaded into memory so it can be utilized as an implementation of the JDBC interfaces.

You need to do this registration only once in your program. You can register a driver in one of two ways.

**Approach (I) - Class.forName():**

The most common approach to register a driver is to use Java's **Class.forName()** method to dynamically load the driver's class file into memory, which automatically registers it. This method is preferable because it allows you to make the driver registration configurable and portable.

The following example uses Class.forName( ) to register the Oracle driver:

try {

Class.forName("oracle.jdbc.driver.OracleDriver");

}

catch(ClassNotFoundException ex) {

System.out.println("Error: unable to load driver class!");

System.exit(1);

}

You can use **getInstance()** method to work around noncompliant JVMs, but then you'll have to code for two extra Exceptions as follows:

try {

Class.forName("oracle.jdbc.driver.OracleDriver").newInstance();

}

catch(ClassNotFoundException ex) {

System.out.println("Error: unable to load driver class!");

System.exit(1);

catch(IllegalAccessException ex) {

System.out.println("Error: access problem while loading!");

System.exit(2);

catch(InstantiationException ex) {

System.out.println("Error: unable to instantiate driver!");

System.exit(3);

}

**Approach (II) - DriverManager.registerDriver():**

The second approach you can use to register a driver is to use the static **DriverManager.registerDriver()** method.

You should use the *registerDriver()* method if you are using a non-JDK compliant JVM, such as the one provided by Microsoft.

The following example uses register Driver() to register the Oracle driver:

try {

Driver myDriver = new oracle.jdbc.driver.OracleDriver();

DriverManager.registerDriver( myDriver );

}

catch(ClassNotFoundException ex) {

System.out.println("Error: unable to load driver class!");

System.exit(1);

}

**Database URL Formulation:**

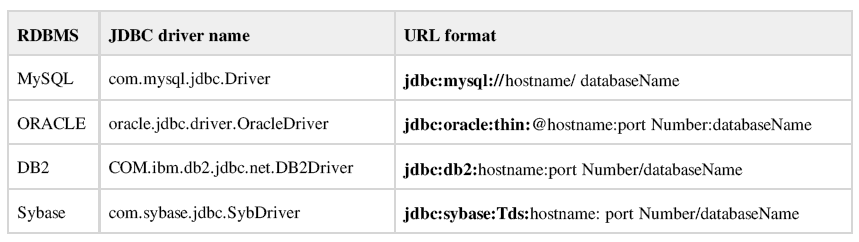
After you've loaded the driver, you can establish a connection using the **DriverManager.getConnection()** method. For easy reference, let me list the three overloaded DriverManager.getConnection() methods:

* getConnection(String url)
* getConnection(String url, Properties prop)
* getConnection(String url, String user, String password)

Here each form requires a database **URL**. A database URL is an address that points to your database.

Formulating a database URL is where most of the problems associated with establishing a connection occur.

Following table lists down popular JDBC driver names and database URL.



All the highlighted part in URL format is static and you need to change only remaining part as per your database setup.

**Create Connection Object:**

**Using a database URL with a username and password:**

I listed down three forms of **DriverManager.getConnection()** method to create a connection object. The most commonly used form of getConnection() requires you to pass a database URL, a *username*, and a *password*:

Assuming you are using Oracle's **thin** driver, you'll specify a host:port:databaseName value for the database portion of the URL.

If you have a host at TCP/IP address 192.0.0.1 with a host name of amrood, and your Oracle listener is configured to listen on port 1521, and your database name is EMP, then complete database URL would then be:

jdbc:oracle:thin:@amrood:1521:EMP

Now you have to call getConnection() method with appropriate username and password to get a **Connection** object as follows:

String URL = "jdbc:oracle:thin:@amrood:1521:EMP";

String USER = "username";

String PASS = "password"

Connection conn = DriverManager.getConnection(URL, USER, PASS);

**Using only a database URL:**

A second form of the DriverManager.getConnection( ) method requires only a database

**URL:**

DriverManager.getConnection(String url);

However, in this case, the database URL includes the username and password and has the following general form:

jdbc:oracle:driver:username/password@database

So the above connection can be created as follows:

String URL = "jdbc:oracle:thin:username/password@amrood:1521:EMP";

Connection conn = DriverManager.getConnection(URL);

**Using a database URL and a Properties object:**

A third form of the DriverManager.getConnection( ) method requires a database URL and a Properties object:

DriverManager.getConnection(String url, Properties info);

A Properties object holds a set of keyword-value pairs. It's used to pass driver properties to the driver during a call to the getConnection() method.

To make the same connection made by the previous examples, use the following code:

import java.util.\*;

String URL = "jdbc:oracle:thin:@amrood:1521:EMP";

Properties info = new Properties( );

info.put( "user", "username" );

info.put( "password","password");

Connection conn = DriverManager.getConnection(URL, info);

**Closing JDBC connections:**

At the end of your JDBC program, it is required explicitly close all the connections to the database to end each database session. However, if you forget, Java's garbage collector will close the connection when it cleans up stale objects.

Relying on garbage collection, especially in database programming, is very poor programming practice. You should make a habit of always closing the connection with the close() method associated with connection object.

To ensure that a connection is closed, you could provide a finally block in your code. A *finally* block always executes, regardless if an exception occurs or not.

To close above opened connection you should call close() method as follows:

conn.close();

Explicitly closing a connection conserves DBMS resources, which will make your database administrator happy.

**HYPER TEXT MARKUP LANGUAGE**

HTML stands for "Hypertext Markup Language". HTML is a SGML (Standard Generalized Markup Language) application widely used to create web pages. It is basically a formatting language and not a programming language.HTML is a language that is easy to write, easy to understand and highly portable. HTML is not a compiled language and is directly interpreted by a browser.HTML is the set of instructions. Each instruction is called as an element or Markup. It is used to structure and format documents for presentation on the web. HTML enhances ASCII files with markup tags that permit the display of a variety of fonts, images, and highlighting options. It also designates structural elements such as headers, lists, and paragraphs, and provides hypertext links to other documents on the Internet.

In plain English, for those of us who are not programming wizards, it is the language our web browser understands, so that it can display what the author of the page wanted us to see... and we can do some really neat stuff with it.

**Interactive HTML:**

**HTML Tag:-**

The first and last tags in a document should always be the HTML tags. These are the tags that tell a Web browser where the HTML in your document begins and ends. The absolute most basic of all possible Web documents is:

<Html >

</Html>

That's it. If we were to load such a page into a Web browser, it wouldn't do anything except give us a blank screen, but it is technically a valid Web page.

**HEAD Tag:-**

The HEAD tags contain all of the document's header information. When I say "header," I don't mean what appears at the top of the browser window, but things like the document title and so on. Speaking of which...

**BODY Tag:-**

BODY comes after the HEAD structure. Between the BODY tags, all of the stuff that gets displayed in the browser window is found. All of the text, the graphics, and links, and so on these things occur between the BODY tags.

**FORMS Tag: -**

Forms provide a unique feature to HTML. Forms allow you to collect data from the end user and return that data to an executable code. The <FORM> element is used to start a form. The <FORM> element specifies what program to run when the form is submitted and how the data is to be transferred.

**ATTRIBUTES Tag:-**

ACTION-The ACTION attribute specifies what program or HTML file is to be called when the submit button is pressed. The ACTION is specified as a URL.

METHOD-The METHOD attribute specifies the protocol to be used when the client sends data to the server. There are two methods. GET (the default) or POST. Using GET method data is attached to the URL mentioned in ACTION attribute. The data that can be sent here is limited here. Where as in POST the data is posted through environment variables and unlimited amounts of data can be sent.

NAME- the NAME attribute specifies the name of the form with which the elements of the form can be referred to later.

**The <INPUT> Element:-**

The <INPUT> tag provides some type of data entry in the form depending on the value of its type attribute.

**ATTRIBUTES:**

NAME- the NAME attribute is used to specify the name of the field.

VALUE- the VALUE attribute is used to specify a default value for the field.

TYPE- the TYPE attribute is used to specify what type of input mechanism is to be drawn on the screen. The permissible values are as follows:

1. Text - The TEXT value is one line-text field.
2. Password - The PASSWORD value is used for entering a password. The text typed in the box is not displayed.
3. Checkbox - The CHECKBOX value provides the user with two choices, selected or not selected.
4. Radio - The RADIO value is used to select one of many from a set of choices.
5. Hidden - The HIDDEN value is used to hide information on the form. This data is not displayed but is returned to the server when submit button is selected.
6. Submit - The SUBMIT value is used to draw a button on the screen that the user can select once the form is completed.
7. Reset - The RESET value is used to draw a button on the screen that a user can select to reset the form to its original default values.
8. Select - The SELECT value is used to select any one of the available options. It is a drop down list and the values are placed in SELECT using OPTION element.
9. Text area - The TEXTAREA allows entering multiple lines of text.

**The <A> Element:**

The <A>tag is used to insert a hyperlink. Hyperlink is a feature of HTML, which allows to link to any document or external files from the current page.

**ATTRIBUTES:**

HREF - HREF stands for "Hypertext Reference," which is another way of saying, "The location of the file I want to load." Most anchors are in the form <A HREF="URL">, where URL is the location of the resource to which you want the link to point. So the HREF attribute of the Anchor element specifies a URL.

**The <TABLE> Element:**

Table allows to create boundaries that make tables position easier in the creation of web pages. The <TABLE> element has four sub elements. They are Table Row<TR>, Table Header<TH>, Table Data<TD> and Caption<CAPTION>.

**ATTRIBUTES:-**

BORDER – The <BORDER> attribute sets the border of the table.

Cell Spacing- It represents the space between the cells.

Cell Padding – It represents the space between the cell border and the cell contents.

WIDTH – The <WIDTH> attribute sets the width of the table and its cells.

**IMAGE tag:**

Images are placed in Web documents using the IMG tag. This tag is empty, and therefore has no closing tag. The basic form of the image tag is <IMG>, but just like <A>, <IMG> by itself is pointless-- it will do nothing.

Visually speaking, images are part of a Web document, but in reality an HTML file and any graphics it refers to are actually all separate files. In other words, one HTML file which has five graphics within it makes a total of six files required to make the page look right. These files are all stored on a Web server, but don't have to all be in the same exact place. (Often, server administrators will set up separate directories for pictures.)

In order to make the IMG tag work, you need to use an SRC attribute. SRC stands for "source," as in, "the source of this graphic." (One way to read a typical image tag is "image source equals...".) The value of SRC is the URL of the graphic you want to have displayed on your Web page. Thus, a typical image tag will take the form:

<IMG SRC="URL of graphic">

The URL of the graphic is just like the URLs used in the anchor tag except in this case the location used is that of the graphic file. A graphic named "blat.gif" located in the directory "pics" on the server "www.site.edu" would have the URL

http://www.site.edu/pics/blat.gif. You can use either relative or full URLs to refer to the graphic file.

**JAVA SCRIPT**

Java Script is Netscape’s cross–platform, object-based scripting language for client server application. JavaScript is mainly used as a client side scripting language. This means that JavaScript code is written into an HTML page. When a user requests an HTML page with JavaScript in it, the script is sent to the browser and it's up to the browser to do something with it. JavaScript can be used in other contexts than a Web browser. Netscape created server-side JavaScript as a CGI-language that can do roughly the same as Perl or ASP.

Fortunately most browsers can handle JavaScript nowadays, but of course some browsers do not support some bits of script.

**Types of Java Script:**

a. Navigator Java Script also called client-side Java Script.

b. Live Wire Java Script also called server-side Java Script.

Using Java Script, dynamic HTML pages can be created that process user input and maintain persistent data using special objects, files and relational databases. Browser interprets JavaScript statements embedded in an HTML page. Netscape Navigator 2.0 and Internet Explorer 3.0 versions and later recognize Java Script. Through JavaScript Live Connect functionally, application can access Java and CORBA distributed-object applications. Navigator 3.0 and later versions supports Live Connect.

**Features of JavaScript (JS):**

1. Browser interprets JavaScript.
2. JavaScript is object based and uses built-in, extensible objects and have no classes or inheritance
3. JavaScript is loosely typed language
4. In JavaScript object reference are checked at runtime
5. JavaScript is designed to supplement the capabilities of HTML with script that are capable of responding to web pages events. JSP has access to some extent of aspects of the web browser window.
6. JavaScript control browser and content but cannot draw graphics or perform networking.

**Client side JavaScript features:**

Client–side JavaScript has expressly been developed for use in a web browser in conjunction with HTML pages. This has certain consequences for security.

* JavaScript cannot read files from or write them to the file system on the computer. This would be a clear security hazard
* JavaScript cannot execute any other programs. This would also be unacceptable.
* JavaScript cannot establish any connection to whatever computer, except to download a new HTML page or to send mail. This, too, would create unacceptable hazards.

The Client-Side JavaScript also has the following features:

* Controls Document’s appearance and content
* Control the browser
* Interact with the HTML forms
* Interact with the user
* Read and write client state with cookies

**Server- Side JavaScript Features:**

1. Embedded in HTML page
2. Executed at the server
3. Pre-complied for faster response
4. Access to Server-side objects
5. Encapsulation of the request

**4.5 TOMCAT 7.0 WEB SERVERS**

Apache Tomcat is a web container developed at the Apache Software Foundation (ASF). Tomcat implements the servlet and the Java Server Pages (JSP) specifications from Sun Microsystems, providing an environment for Java code to run in cooperation with a web server. It adds tools for configuration and management but can also be configured by editing configuration files that are normally XML-formatted. Tomcat includes its own HTTP server internally.

**Environment:**

Tomcat is a web server that supports servlets and JSPs. The accompanying Tomcat Jasper compiler compiles JSPs into servlets. The Tomcat servlet engine is often used in combination with an Apache HTTP Server or other web servers. Tomcat can also function as an independent web server. Earlier in its development, the perception existed that standalone Tomcat was only suitable for development environments and other environments with minimal requirements for speed and transaction handling. However, that perception no longer exists; Tomcat is increasingly used as a standalone web server in high-traffic, high-availability environments.

Tomcat is cross-platform, running on any operating system that has a Java Runtime Environment. The following properties can be specified, either as system properties, or by using a deployer. Properties file located in the root folder of the deployer package:

 Build: The build folder used will be, by default, ${build}/webapp${path}. After the end of the execution of the compile target, the web application WAR will be located at ${build}/webapp${path}.war.

 Webapp: Folder containing the expanded web application which will be compiled and validated. By default, the folder is myapp.

 Path: Deployed context path of the web application, by default /myapp.

 URL: Absolute URL to the manager web application of a running Tomcat server, which will be used to deploy and undeploy the web application. By default, the deployer will attempt to access a Tomcat instance running on local host, at http://localhost:8080/manager.

 Username: Username to be used to connect to the Tomcat manager.

**** Password: Password to be used to connect to the Tomcat manager

A web application which is programmatically developed by the developer is stored in the webapps folder and WEB-INF folder also saved in that location which consists of folder named classes which supports to run application automatically.

The deployer package includes a ready to use Ant script, with the following targets:

 compile (default): Compile and validate the web application. This can be used standalone, and does not need a running Tomcat server. The compiled application will only run on the associated Tomcat 5.0.x server release, and is not guaranteed to work on another Tomcat release, as the code generated by Jasper depends on its runtime component. It should also be noted that this target will also compile automatically any Java source file located in the /WEB-INF/classes folder of the web application.

 deploy: Deploy a web application (compiled or not) to a Tomcat server

 undeploy: Undeploy a web application

 Start: Start web application

 reload: Reload web application

 Stop: Stop web application

**4.6 JAVA SERVER PAGES**

**Introduction:**

Jsp technology enables you to mix regular static html with dynamically generated content from servlets. Separating the static html from the dynamic content provides a number of benefits over servlets alone.

**Why use JSP:**

Jsp is easy to learn and allows developers to quickly produce wed sites and application in an open and standard way. Jsp is based on java, an object-oriented language. Jsp offers a robust platform for web development.

Main reasons to Jsp:

1. Multi platform
2. Component reuse by using java beans and Ejb
3. Advantages if java

We can take one Jsp file and move it to another platform, web server or Jsp servlet engine.

**JSP compared to ASP:**

Jsp and Asp are fairly similar in the functionality that they provide. Jsp may have slightly higher learning curve. Both allow embedded code in an html page, session variables Platform i.e., NT, JSP can operate on any platform that conforms to the J2EE specification. Jsp allow component reuse by using JavaBeans and Ejbs. Asp provides the use of Com/activeX controls.

**JSP compared to servlets:**

A servlet is java class that provides special server side service. It is hard to write HTML code in servlets. In servlets you need to have lots of println statement to generate HTML.

**Description:**

JSP looks like html, but they get compiled into java servlets the first time they are invoked. The resulting servlet is a combination of the html from the jsp file and embedded dynamic content specified by the new tags. That is not to say that jsp must contain html. Some of them contain only java code; this is particularly useful when the jsp is responsible for a particular task like maintaining application flow.

Everything in Jsp can be broken into 2 categories:

1. Elements that are processed on the server.
2. Template data or everything other than elements that the engine processing the Jsp ignores.

A Jsp page is executed by a Jsp engine or container, which is installed on a web server, or on an application server. When the client asks for a jsp resource the engine wraps up that request and response object to incorporate the communication with the client. The container then wraps up the underlying layer for a jsp that of a servlet implementation. The abstractions of the request and response are the same as javax.servlet.ServletRequest and javax.servlet.ServletReesponse respectively.

**JSP Architecture:**

Jsp are built on top of sun’s servlet technology. Jsp is essentially an html page with special jsp tags embedded. These jsp tags can contain java code. The jsp file extension is .jsp rather than .htm or .html. The jsp engine parses the .jsp and creates a java servlet source file. It then compiles the source file into a class file; this is done the first time and this why the jsp is probably slower the first time it is accessed. Any time after this, the special compiled servlet is executed and is therefore returns faster.

**Steps required to JSP request:**

1. The user goes to a web site made using jsp. The user goes to a jsp page. The web browser makes the request via the internet.
2. The jsp request gets sent to the web server
3. The web server recognizes that the file required is special (.jsp), therefore passes the jsp file to the jsp servlet engine.
4. If the jsp file has been called the first time, the jsp file is parsed, otherwise go to step 7
5. The next step is to generate a special servlet form the jsp file. The entire html required is converted to println statements.
6. The servlet source code is compiled into a class
7. The servlet is instantiated, calling the init and service methods
8. Html from the servlet output is sent via the internet
9. Html results are displayed on the user’s web browser.

Servlets are server-side java programs that can be deployed on a web server.

The servlet interface provides the basic frame work for coding servlets.

Java Server Pages, SQL, HTML Forms and Databases

This section examines how to communicate with a database from Java. We have already seen how to interface an HTML Form and a JSP. Now e has to see how that JSP can talk to a database. The objectives of this section are to understand how to:

* Administratively register databases.
* Connect a JSP to an Access database.
* Insert records in a database using JSP.
* Inserting data from HTML Form in a database using JSP.
* Delete Records from Database based on Criteria from HTML Form.
* Retrieve data from a database using JSP – result sets.
* Apply SQL operations like sort, create table, remove table, delete, and Access-based arithmetic functions.

**JSP Declarations**:

Used to define page level variables and methods are placed with in the <%!and %> tags and always end with a semicolon.

Example:

<%!

Int I=0;

Int j=0;

Int z=0 ;%>

**JSP Scriptlets**:

Consists of valid code snippets enclosed with in the <% and %> JSP tags.

Example:

To accept the user name and display the name 10 times:

<%@ page import=”java.util.\*” %>

<%@ page import=”java” %>

<HTML><BODY>

<%out.println (“<HTML>”);

out.println (“<BODY>”);

out.println (“<BODY>”);

out.println (“<HTML>”);%>

</BODY></HTML>

**JSP Expressions:**

Used to directly insert values into the output

Example :< %=msg %>

**JSP Implicit Objects:**

Are predefined variables that can be included in JSP expressions and Scriplets can be Created.

1. Implicitly by using directives.
2. Explicitly by using standard actions.
3. Directly by declaring objects with in scriptlets.

Include variables such as

PAGE: To represent the current instance of the jsp page.

REQUEST: To represent an object of HTTP Servlet request used to retrieve the request data.

RESPONSE: To represent an object of HTTP servlet response used to write the HTML response output.

**JSP Actions**:

<jsp: getProperty>: To retrieve the property of the specified bean and direct it as

output.

Attributes used are: Name and Property.

<jsp: setProperty>: To set the property of specified bean.

Attributes used are: Name, Property, Value and Param.

<jsp: forward>: To forward a request to a different page.

Attribute used is Page.

<jsp: param>: Used as a sub attribute with jsp: include and jsp: forward to passadditional

request parameters.

Attributes used are Name and Value.

<jsp: include> : To insert a file into a particular jsp page.

Attributes used are Page and Flush.

**4.7. SERVLETS**

**Introduction:**

The Java web server is Java Soft own web Server. The Java web server is just a part of a larger framework, intended to provide you not just with a web server, but also with tools. To build customized network servers for any Internet or Intranet client/server system. Servlets are to a web server, how applets are to the browser.

**About SERVLETS:**

Servlets provide a Java-based solution used to address the problems currently associated with doing server-side programming, including inextensible scripting solutions, platform-specific APIs, and incomplete interfaces. Servlets are objects that conform to a specific interface that can be plugged into a Java-based server.

Servlets are to the server-side what applets are to the client-side - object byte codes that can be dynamically loaded off the net. They differ from applets in that they are faceless objects (without graphics or a GUI component). They serve as platform independent, dynamically loadable, plug able helper byte code objects on the server side that can be used to dynamically extend server-side functionality. For example, an HTTP Servlet can be used to generate dynamic HTML content. When you use Servlets to do dynamic content you get the following advantages:

1. They are faster and cleaner than CGI scripts
2. They use a standard API (the Servlet API)
3. They provide all the advantages of Java (run on a variety of servers without needing to be rewritten).

**Attractiveness of SERVLETS:**

There are many features of Servlets that make them easy and attractive to use. These include:

1. Easily configured using the GUI-based Admin tool
2. Can be loaded and invoked from a local disk or remotely across the network.
3. Can be linked together, or chained, so that one Servlet can call another Servlet or several Servlets in sequence.
4. Can be called dynamically from within HTML pages, using server-side include tags.
5. Are secure - even when downloading across the network, the Servlet security model and Servlet sandbox protect your system from unfriendly behavior.

**Advantages of the SERVLET API:**

One of the great advantages of the Servlet API is protocol independence. It assumes nothing about:

1. The protocol being used to transmit on the net
2. How it is loaded
3. The server environment it will be running in

These qualities are important, because it allows the Servlet API to be embedded in many different kinds of servers. There are other advantages to the Servlet API as well. These include:

1. It's extensible - you can inherit all your functionality from the base classes made available to you.
2. It's simple, small, and easy to use.

**Features of SERVLETS:**

Servlets are persistent. Servlet are loaded only by the web server and can maintain services between requests.

1. Servlets are fast. Since Servlets only need to be loaded once, they offer much better performance over their CGI counterparts.
2. Servlets are platform independent.
3. Servlets are extensible. Java is a robust, object-oriented programming language, which easily can be extended to suit your needs
4. Servlets are secure.

**Loading SERVLETS:**

Servlets can be loaded from three places:

From a directory that is on the CLASSPATH. The CLASSPATH of the Java Web Server includes service root/classes/which is where the system classes reside. From the <SERVICE\_ROOT/Servlets/directory> this is \*not\* in the server’s class path. A class loader is used to create Servlets from this directory. New Servlets can be added - existing Servlets can be recompiled and the server will notice these changes from a remote location. For this a code base like http://nine.eng/classes/foo/ is required in addition to the Servlet class name. Refer to the admin GUI docs on Servlet section to see how to set this up.

**Loading Remote SERVLETS:**

Remote Servlets can be loaded by:

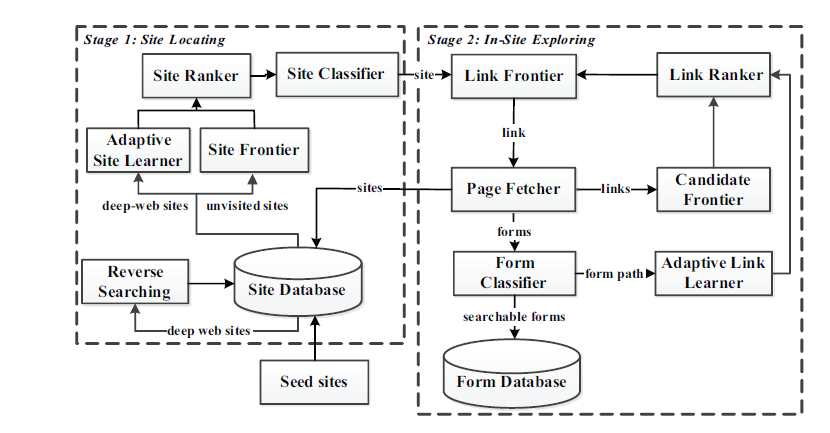
1. Configuring the Admin Tool to setup automatic loading of remote Servlets
2. Setting up server side include tags in .html files
3. Defining a filter chain configuration

**Invoking Servlets**:

A Servlet invoker is a Servlet that invokes the "service" method on a named Servlet. If the Servlet is not loaded in the server, then the invoker first loads the Servlet (either from local disk or from the network) and the then invokes the "service" method. Also like applets, local Servlets in the server can be identified by just the class name. In other words, if a Servlet name is not absolute, it is treated as local.

A client can invoke Servlets in the following ways:

1. The client can ask for a document that is served by the Servlet.
2. The client (browser) can invoke the Servlet directly using a URL, once it has been mapped using the Servlet aliases section of the admin GUI.
3. The Servlet can be invoked through server side include tags.
4. The Servlet can be invoked by placing it in the Servlets / directory.
5. The Servlet can be invoked by using it in a filter chain.
6. **SYSTEM DESIGN**
   1. **ARCHITECTURE DIAGRAM**

****

Fig**:** The two-stage architecture of *SmartCrawler*.

**5.2. UML DIAGRAMS**

The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

**Global Use Case Diagrams:**

Identification of actors:

**Actor:** Actor represents the role a user plays with respect to the system. An actor interacts with, but has no control over the use cases.

Graphical representation:



<<Actor name>>

An actor is someone or something that:

Interacts with or uses the system.

* Provides input to and receives information from the system.
* Is external to the system and has no control over the use cases.

Actors are discovered by examining:

* Who directly uses the system?
* Who is responsible for maintaining the system?
* External hardware used by the system.
* Other systems that need to interact with the system.

Questions to identify actors:

* + Who is using the system? Or, who is affected by the system? Or, which groups need help from the system to perform a task?
  + Who affects the system? Or, which user groups are needed by the system to perform its functions? These functions can be both main functions and secondary functions such as administration.
  + Which external hardware or systems (if any) use the system to perform tasks?
  + What problems does this application solve (that is, for whom)?
  + And, finally, how do users use the system (use case)? What are they doing with the system?

The actors identified in this system are:

1. **System Administrator**
2. **Customer**
3. **Customer Care**

Identification of usecases:

**Usecase:** A use case can be described as a specific way of using the system from a user’s (actor’s) perspective.

**Graphical representation:**



A more detailed description might characterize a use case as:

* Pattern of behavior the system exhibits
* A sequence of related transactions performed by an actor and the system
* Delivering something of value to the actor

Use cases provide a means to:

* capture system requirements
* communicate with the end users and domain experts
* test the system

Use cases are best discovered by examining the actors and defining what the actor will be able to do with the system.

Guide lines for identifying use cases:

* For each actor, find the tasks and functions that the actor should be able to perform or that the system needs the actor to perform. The use case should represent a course of events that leads to clear goal
* Name the use cases.
* Describe the use cases briefly by applying terms with which the user is familiar.

This makes the description less ambiguous

Questions to identify use cases:

* What are the tasks of each actor?
* Will any actor create, store, change, remove or read information in the system?
* What use case will store, change, remove or read this information?
* Will any actor need to inform the system about sudden external changes?
* Does any actor need to inform about certain occurrences in the system?
* What usecases will support and maintains the system?

**Flow of Events**

A flow of events is a sequence of transactions (or events) performed by the system. They typically contain very detailed information, written in terms of what the system should do, not how the system accomplishes the task. Flow of events are created as separate files or documents in your favorite text editor and then attached or linked to a use case using the Files tab of a model element.

A flow of events should include:

* When and how the use case starts and ends
* Use case/actor interactions
* Data needed by the use case
* Normal sequence of events for the use case
* Alternate or exceptional flows

Construction of Usecase diagrams:

Use-case diagrams graphically depict system behavior (use cases). These diagrams present a high level view of how the system is used as viewed from an outsider’s (actor’s) perspective. A use-case diagram may depict all or some of the use cases of a system.

A use-case diagram can contain:

* actors ("things" outside the system)
* use cases (system boundaries identifying what the system should do)
* Interactions or relationships between actors and use cases in the system including the associations, dependencies, and generalizations.

Relationships in use cases:

**1. Communication:**

The communication relationship of an actor in a usecase is shown by connecting the actor symbol to the usecase symbol with a solid path. The actor is said to communicate with the usecase.

**2. Uses:**

A Uses relationship between the usecases is shown by generalization arrow from the usecase.

**3. Extends:**

The extend relationship is used when we have one usecase that is similar to another usecase but does a bit more. In essence it is like subclass.

**SEQUENCE DIAGRAMS**

A sequence diagram is a graphical view of a scenario that shows object interaction in a time-based sequence what happens first, what happens next. Sequence diagrams establish the roles of objects and help provide essential information to determine class responsibilities and interfaces.

There are two main differences between sequence and collaboration diagrams: sequence diagrams show time-based object interaction while collaboration diagrams show how objects associate with each other. A sequence diagram has two dimensions: typically, vertical placement represents time and horizontal placement represents different objects.

**Object:**

An object has state, behavior, and identity. The structure and behavior of similar objects are defined in their common class. Each object in a diagram indicates some instance of a class. An object that is not named is referred to as a class instance.

The object icon is similar to a class icon except that the name is underlined:

An object's concurrency is defined by the concurrency of its class.

**Message:**

A message is the communication carried between two objects that trigger an event. A message carries information from the source focus of control to the destination focus of control. The synchronization of a message can be modified through the message specification. Synchronization means a message where the sending object pauses to wait for results.

**Link:**

A link should exist between two objects, including class utilities, only if there is a relationship between their corresponding classes. The existence of a relationship between two classes symbolizes a path of communication between instances of the classes: one object may send messages to another. The link is depicted as a straight line between objects or objects and class instances in a collaboration diagram. If an object links to itself, use the loop version of the icon.

**CLASS DIAGRAM:**

Identification of analysis classes:

A class is a set of objects that share a common structure and common behavior (the same attributes, operations, relationships and semantics). A class is an abstraction of real-world items.

There are 4 approaches for identifying classes:

1. Noun phrase approach:
2. Common class pattern approach.
3. Use case Driven Sequence or Collaboration approach.
4. Classes , Responsibilities and collaborators Approach
5. **Noun Phrase Approach:**

The guidelines for identifying the classes:

* + Look for nouns and noun phrases in the usecases.
  + Some classes are implicit or taken from general knowledge.
  + All classes must make sense in the application domain; Avoid computer

implementation classes – defer them to the design stage.

* + Carefully choose and define the class names After identifying the classes we have to eliminate the following types of classes:
* Adjective classes.

1. **Common class pattern approach:**

The following are the patterns for finding the candidate classes:

* + Concept class.
  + Events class.
  + Organization class
  + Peoples class
  + Places class
  + Tangible things and devices class.

1. **Use case driven approach:**

We have to draw the sequence diagram or collaboration diagram. If there is need for some classes to represent some functionality then add new classes which perform those functionalities.

1. **CRC approach:**

The process consists of the following steps:

* + Identify classes’ responsibilities ( and identify the classes )
  + Assign the responsibilities
  + Identify the collaborators.

Identification of responsibilities of each class:

The questions that should be answered to identify the attributes and methods of a class respectively are:

1. What information about an object should we keep track of?
2. What services must a class provide?

Identification of relationships among the classes:

Three types of relationships among the objects are:

Association: How objects are associated?

Super-sub structure: How are objects organized into super classes and sub classes?

Aggregation: What is the composition of the complex classes?

Association:

The **questions** that will help us to identify the associations are:

1. Is the class capable of fulfilling the required task by itself?
2. If not, what does it need?
3. From what other classes can it acquire what it needs?

Guidelines for identifying the tentative associations:

* A dependency between two or more classes may be an association. Association often corresponds to a verb or prepositional phrase.
* A reference from one class to another is an association. Some associations are implicit or taken from general knowledge.

Some common association patterns are:

Location association like part of, next to, contained in…..

Communication association like talk to, order to ……

We have to eliminate the unnecessary association like implementation associations, ternary or n-ary associations and derived associations.

Super-sub class relationships:

Super-sub class hierarchy is a relationship between classes where one class is the parent class of another class (derived class).This is based on inheritance.

Guidelines for identifying the super-sub relationship, a generalization are

1***.* Top-down*:***

Look for noun phrases composed of various adjectives in a class name. Avoid excessive refinement. Specialize only when the sub classes have significant behavior.

2.**Bottom-up*:***

Look for classes with similar attributes or methods. Group them by moving the common attributes and methods to an abstract class. You may have to alter the definitions a bit.

3.**Reusability*:***

Move the attributes and methods as high as possible in the hierarchy.

4. **Multiple inheritances*:***

Avoid excessive use of multiple inheritances. One way of getting benefits of multiple inheritances is to inherit from the most appropriate class and add an object of another class as an attribute.

**Aggregation or a-part-of relationship:**

It represents the situation where a class consists of several component classes. A class that is composed of other classes doesn’t behave like its parts. It behaves very difficultly. The major properties of this relationship are transitivity and anti symmetry.

The **questions** whose answers will determine the distinction between the part and whole relationships are:

* Does the part class belong to the problem domain?
* Is the part class within the system’s responsibilities?
* Does the part class capture more than a single value?( If not then simply include it as an attribute of the whole class)
* Does it provide a useful abstraction in dealing with the problem domain?

There are three types of aggregation relationships. They are:

**Assembly:**

It is constructed from its parts and an assembly-partsituation physically exists.

**Container:**

A physical whole encompasses but is not constructed from physical parts.

**Collection member:**

A conceptual whole encompasses parts that may be physical or conceptual. The container and collection are represented by hollow diamonds but composition is represented by solid diamond.

**UML Diagrams**

**Class Diagram:**

****

**Use Case Diagram:**

****

**User Sequence Diagram:**

****

**Admin Sequence Diagram:**

****

**User Activity Diagram:**

****

**Admin Activity Diagram:**

****

**User State Chart Diagram:**



**Admin State Chart Diagram:**



**User Collaboration Diagram**



**Admin Collaboration Diagram**

****

**Component Diagram**

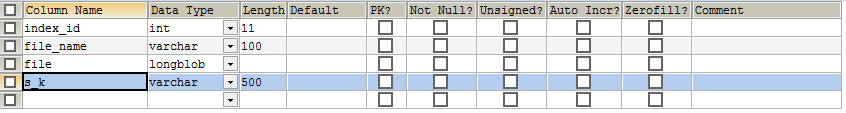
****

**Deployment Diagram**

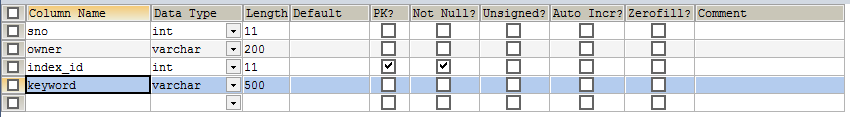
****

**5.3 Data Base Tables**

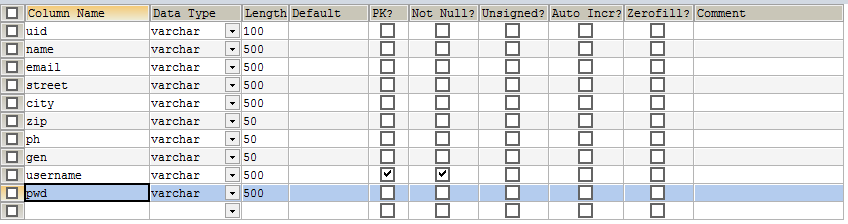
**Data Table:**

****

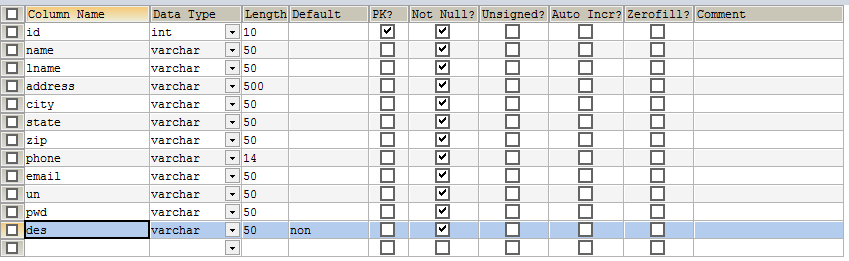
**Meta\_data Table:**

****

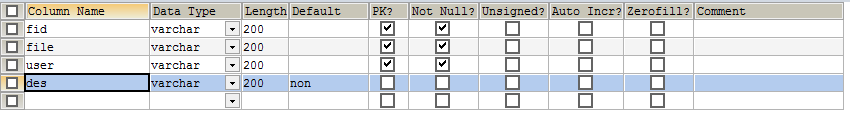
**Owner Table:**

****

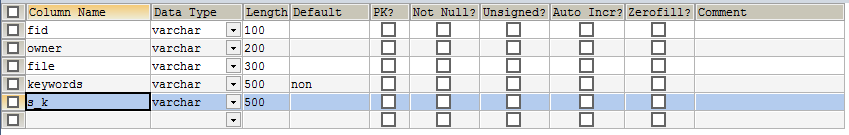
**Register Table:**

****

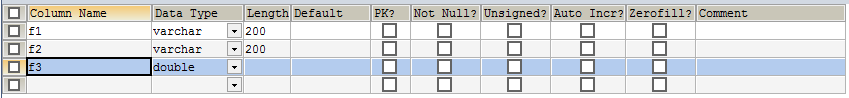
**Request Table**

****

**Temp Table**

****

**tfidf Table**

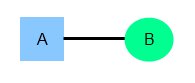
****

**Entity-relationship model**

A relationship is how the data is shared between entities. There are three types of relationships between entities:

**1. One-to-One**

One instance of an entity (A) is associated with one other instance of another entity (B). For example, in a database of employees, each employee name (A) is associated with only one social security number (B).

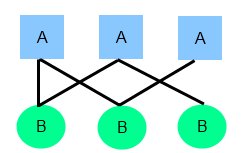


**2. One-to-Many**

One instance of an entity (A) is associated with zero, one or many instances of another entity (B), but for one instance of entity B there is only one instance of entity A. For example, for a company with all employees working in one building, the building name (A) is associated with many different employees (B), but those employees all share the same singular association with entity A.

**3. Many-to-Many**

One instance of an entity (A) is associated with one, zero or many instances of another entity (B), and one instance of entity B is associated with one, zero or many instances of entity A. For example, for a company in which all of its employees work on multiple projects, each instance of an employee (A) is associated with many instances of a project (B), and at the same time, each instance of a project (B) has multiple employees (A) associated with it.



**ER-Diagram  
**

1. **SYSTEM IMPLEMENTATION**

# Number of Modules:

After careful analysis the system has been identified to have the following modules:

1. **two-stage crawler.**

2**. Site Ranker**

**3. Adaptive learning**

**1 Two-stage crawler.**

It is challenging to locate the deep web databases, because they are not registered with any search engines, are usually sparsely distributed, and keep constantly changing. To address this problem, previous work has proposed two types of crawlers, generic crawlers and focused crawlers. Generic crawlers fetch all searchable forms and cannot focus on a specific topic. Focused crawlers such as Form-Focused Crawler (FFC) and Adaptive Crawler for Hidden-web Entries (ACHE) can automatically search online databases on a specific topic. FFC is designed with link, page, and form classifiers for focused crawling of web forms, and is extended by ACHE with additional components for form filtering and adaptive link learner. The link classifiers in these crawlers play a pivotal role in achieving higher crawling efficiency than the best-first crawler However, these link classifiers are used to predict the distance to the page containing searchable forms, which is difficult to estimate, especially for the delayed benefit links (links eventually lead to pages with forms). As a result, the crawler can be inefficiently led to pages without targeted forms.

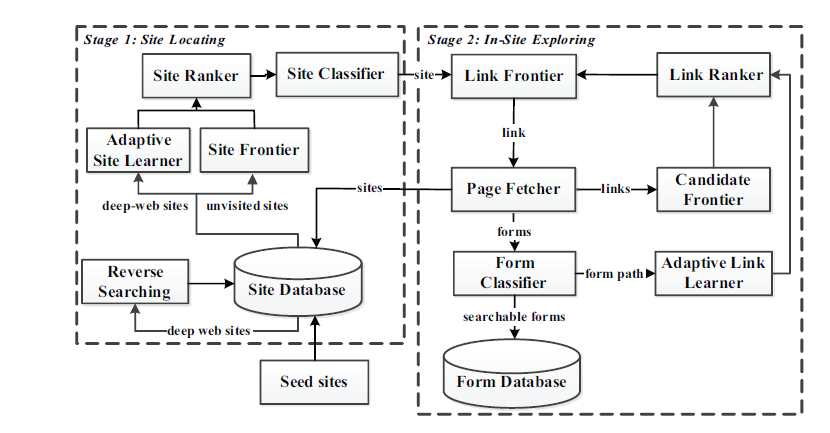
**2. Site Ranker:**

When combined with above stop-early policy. We solve this problem by prioritizing highly relevant links with link ranking. However, link ranking may introduce bias for highly relevant links in certain directories. Our solution is to build a linktreefor a balanced link prioritizing. Figure 2 illustrates an example of a link tree constructed from the homepage of http://www.abebooks.com. Internal nodes of the tree represent directory paths. In this example, servlet directory is for dynamic request; books directory is for displaying different catalogs of books; Amdocs directory is for showing help information. Generally each directory usually represents one type of files on web servers and it is advantageous to visit links in different directories. For links that only differ in the query string part, we consider them as the same URL.Because links are often distributed unevenly in server directories, prioritizing links by the relevance can potentially bias toward some directories. For instance, the links under books might be assigned a high priority, because “book” is an important feature word in the URL. Together with the fact that most links appear in the books directory, it is quite possible that links in other directories will not be chosen due to low relevance score. As a result, the crawler may miss searchable forms in those directories.

**3. Adaptive learning**

Adaptive learning algorithm that performs online feature selection and uses these features to automatically construct link rankers. In the site locating stage, high relevant sites are prioritized and the crawling is focused on atopic using the contents of the root page of sites, achieving more accurate results. During the insiteexploring stage, relevant links are prioritized for fast in-site searching. We have performed an extensive performance evaluation ofSmart Crawler over real web data in 1representativedomains and compared with ACHE and site-based crawler. Our evaluation shows that our crawling framework is very effective, achieving substantially higher harvest rates than the state-of-the-art ACHE crawler. The results also show the effectiveness of the reverse searching and adaptive learning.

**Two-Stage Architecture**



1. **TESTING**

Testing is the debugging program is one of the most critical aspects of the computer programming triggers, without programming that works, the system would never produce an output of which it was designed. Testing is best performed when user development is asked to assist in identifying all errors and bugs. The sample data are used for testing. It is not quantity but quality of the data used the matters of testing. Testing is aimed at ensuring that the system was accurately an efficiently before live operation commands.

**Testing objectives:**

The main objective of testing is to uncover a host of errors, systematically and with minimum effort and time. Stating formally, we can say, testing is a process of executing a program with intent of finding an error.

1. A successful test is one that uncovers an as yet undiscovered error.
2. A good test case is one that has probability of finding an error, if it exists.
3. The test is inadequate to detect possibly present errors.
4. The software more or less confirms to the quality and reliable standards.

**Levels of Testing:**

In order to uncover present in different phases we have the concept of levels of testing.

**The basic levels of Testing:**

Client needs acceptance testing

Requirements system testing

Design integration testing

Code unit testing

Figure 7.1: Levels of Testing

**Code testing:**

This examines the logic of the program. For example, the logic for updating various sample data and with the sample files and directories were tested and verified.

**Specification Testing:**

Executing this specification starting what the program should do and how it should performed under various conditions. Test cases for various situation and combination of conditions in all the modules are tested.

**Unit testing:**

In the unit testing we test each module individually and integrate with the overall system. Unit testing focuses verification efforts on the smallest unit of software design in the module. This is also known as module testing. The module of the system is tested separately. This testing is carried out during programming stage itself. In the testing step each module is found to work satisfactorily as regard to expected output from the module. There are some validation checks for fields also. For example the validation check is done for varying the user input given by the user which validity of the data entered. It is very easy to find error debut the system.

Each Module can be tested using the following two Strategies:

1. Black Box Testing
2. White Box Testing

**7.1 BLACK BOX TESTING**

**What is Black Box Testing?**

Black box testing is a software testing techniques in which **functionality of the software under test (SUT) is tested without looking at the internal code structure**, implementation details and knowledge of internal paths of the software. This type of testing is based entirely on the software requirements and specifications.

**In Black Box Testing we just focus on inputs and output of the software system** without bothering about internal knowledge of the software program.



The above Black Box can be any software system you want to test. For example : an operating system like Windows, a website like Google ,a database like Oracle or even your own custom application. Under Black Box Testing , you can test these applications by just focusing on the inputs and outputs without knowing their internal code implementation.

**Black box testing - Steps**

Here are the generic steps followed to carry out any type of Black Box Testing.

* Initially requirements and specifications of the system are examined.
* Tester chooses valid inputs (positive test scenario) to check whether SUT processes them correctly. Also some invalid inputs (negative test scenario) are chosen to verify that the SUT is able to detect them.
* Tester determines expected outputs for all those inputs.
* Software tester constructs test cases with the selected inputs.
* The test cases are executed.
* Software tester compares the actual outputs with the expected outputs.
* Defects if any are fixed and re-tested.

**Types of Black Box Testing**

There are many types of Black Box Testing but following are the prominent ones -

* **Functional testing** – This black box testing type is related to functional requirements of a system; it is done by software testers.
* **Non-functional testing** – This type of black box testing is not related to testing of a specific functionality, but non-functional requirements  such as performance, scalability, usability.
* **Regression testing** – Regression testing is done  after code fixes , upgrades or any other system maintenance to check the new code has not affected the existing code.

**7.2 WHITE BOX TESTING**

White Box Testing is the testing of a software solution's internal coding and infrastructure.It focuses primarily on strengthening security, the flow of inputs and outputs through the application, and improving design and usability.White box testing is also known as **clear, open, structural, and glass box testing**.

It is one of two parts of the **"box testing" approach** of software testing. Its counter-part, blackbox testing, involves testing from an external or end-user type perspective. On the other hand, Whitebox testing is based on the inner workings of an application and revolves around internal testing. The term "whitebox" was used because of the see-through box concept. The clear box or whitebox name symbolizes the ability to see through the software's outer shell (or "box") into its inner workings. Likewise, the "black box" in "black box testing" symbolizes not being able to see the inner workings of the software so that only the end-user experience can be tested

## What do you verify in White Box Testing ?

White box testing involves the testing of the software code for the following:

* Internal security holes
* Broken or poorly structured paths in the coding processes
* The flow of specific inputs through the code
* Expected output
* The functionality of conditional loops
* Testing of each statement, object and function on an individual basis

The testing can be done at system, integration and unit levels of software development. One of the basic goals of whitebox testing is to verify a working flow for an application. It involves testing a series of predefined inputs against expected or desired outputs so that when a specific input does not result in the expected output, you have encountered a bug.

## How do you perform White Box Testing?

  To give you a simplified explanation of white box testing, we have divided it into **two basic steps**. This is what testers do when testing an application using the white box testing technique:

**STEP 1) UNDERSTAND THE SOURCE CODE**

The first thing a tester will often do is learn and understand the source code of the application. Since white box testing involves the testing of the inner workings of an application, the tester must be very knowledgeable in the programming languages used in the applications they are testing. Also, the testing person must be highly aware of secure coding practices. Security is often one of the primary objectives of testing software. The tester should be able to find security issues and prevent attacks from hackers and naive users who might inject malicious code into the application either knowingly or unknowingly.

**Step 2) CREATE TEST CASES AND EXECUTE**

The second basic step to white box testing involves testing the application’s source code for proper flow and structure. One way is by writing more code to test the application’s source code. The tester will develop little tests for each process or series of processes in the application. This  method requires that the tester must have intimate knowledge of the code and is often done by the developer. Other methods include manual testing, trial and error testing and the use of testing tools as we will explain further on in this article.

**System testing:**

Once the individual module testing is completed, modules are assembled and integrated to perform as a system. The top down testing, which began from upper level to lower level module, was carried out to check whether the entire system is performing satisfactorily.

**There are three main kinds of System testing:**

Alpha Testing

Beta Testing

Acceptance Testing

**Alpha Testing:**

This refers to the system testing that is carried out by the test team with the Organization.

**Beta Testing**:

This refers to the system testing that is performed by a selected group of friendly customers

**Acceptance Testing:**

This refers to the system testing that is performed by the customer to determine whether or not to accept the delivery of the system.

**Acceptance Testing:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Test Condition ID** | **Description of coverage** | **Expected results** | **Covered by script** |
| 1. | Verification of a particular record | If a particular record already exists it displays a message | This type of test in {verify} procedure in every Jsp file where a record is inserted via an interface |
| 2. | Updating of a particular record | All the details should not be updated. | This type of test is covered in all the Asp files where updations are made. |
| 3. | Validity of login | Only the authorized persons must access system. | This is covered in the login procedure for the validity of a user |

**Integration Testing:**

Data can be lost across an interface, one module can have an adverse effort on the other sub functions, when combined, may not produce the desired major functions. Integrated testing is the systematic testing for constructing the uncover errors within the interface. The testing was done with sample data. The developed system has run successfully for this sample data. The need for integrated test is to find the overall system performance.

**Output testing:**

After performance of the validation testing, the next step is output testing. The output displayed or generated by the system under consideration is tested by asking the user about the format required by system. The output format on the screen is found to be correct as format was designed in the system phase according to the user needs. Hence the output testing does not result in any correction in the system.

**Test plan:**

The test-plan is basically a list of testcases that need to be run on the system. Some of the testcases can be run independently for some components (report generation from the database, for example, can be tested independently) and some of the testcases require the whole system to be ready for their execution. It is better to test each component as and when it is ready before integrating the components. It is important to note that the testcases cover all the aspects of the system (ie, all the requirements stated in the RS document).

Figure.7.3: Test plan

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sno** | **Test case Title** | **Description** | **Expected Outcome** | **The requirement in RS that is being tested** | **Result** |
| 1 | Successful User Verification | The login to the system should be tried with the login assigned by the admin and the correct password | Login should be successful and the user should enter in to the system | RS1 | Passed |
| 2 | Unsuccessful User Verification due to wrong password | Login to the system with a wrong password | Login should fail with an error ‘Invalid Password’ | RS1 | Passed |
| 3 | Unsuccessful User Verification due to invalid login id | Login to the system with a invalid login id | Login should fail with an error ‘Invalid user id’ | RS1 | Passed |

1. **SCREEN SHOTS**

**Home page:**

****

1. **CONCLUSION**

In this paper, we propose an effective harvesting framework for deep-web interfaces, namely *Smart-* *Crawler*. We have shown that our approach achieves both wide coverage for deep web interfaces and maintains highly efficient crawling. *SmartCrawler* is a focused crawler consisting of two stages: efficient site locating and balanced in-site exploring. *SmartCrawler* performs site-based locating by reversely searching the known deep web sites for center pages, which can effectively find many data sources for sparse domains. By ranking collected sites and by focusing the crawling on a topic, *SmartCrawler* achieves more accurate results. The in-site exploring stage uses adaptive link-ranking to search within a site; and we design a link tree for eliminating bias toward certain directories of a website for wider coverage of web directories. Our experimental results on a representative set of domains show the effectiveness of the proposed two-stage crawler, which achieves higher harvest rates than other crawlers. In future work, we plan to combine pre-query and post-query approaches for classifying deep-web forms to further improve the accuracy of the form classifier.

**10. BIBLOGRAPHY**

[1] Peter Lyman and Hal R. Varian. How much information? 2003. Technical report, UC Berkeley, 2003.

[2] Roger E. Bohn and James E. Short. How much information? 2009 report on american consumers. Technical report, University of California, San Diego, 2009.

[3] Martin Hilbert. How much information is there in the ”information society”? *Significance*, 9(4):8–12, 2012.

[4] Idc worldwide predictions 2014: Battles for dominance – and survival – on the 3rd platform. <http://www.idc.com/> research/Predictions14/index.jsp, 2014.

[5] Michael K. Bergman. White paper: The deep web: Surfacing hidden value. *Journal of electronic publishing*, 7(1), 2001.

[6] Yeye He, Dong Xin, Venkatesh Ganti, Sriram Rajaraman, and Nirav Shah. Crawling deep web entity pages. In *Proceedingsof the sixth ACM international conference on Web search and data* *mining*, pages 355–364. ACM, 2013.

[7] Infomine. UC Riverside library. <http://lib-www.ucr.edu/>, 2014.

[8] Clusty’s searchable database dirctory. <http://www.clusty>. com/, 2009.

[9] Booksinprint. Books in print and global books in print access.http://booksinprint.com/, 2015.

[10] Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. Toward large scale integration: Building a metaquerier over databases on the web. In *CIDR*, pages 44–55, 2005.

[11] Denis Shestakov. Databases on the web: national web domainsurvey. In *Proceedings of the 15th Symposium on InternationalDatabase Engineering & Applications*, pages 179–184. ACM, 2011.

[12] Denis Shestakov and Tapio Salakoski. Host-ip clustering technique for deep web characterization. In *Proceedings of the* *12th International Asia-Pacific Web Conference (APWEB)*, pages 378–380. IEEE, 2010.

[13] Denis Shestakov and Tapio Salakoski. On estimating the scale of national deep web. In *Database and Expert Systems* *Applications*, pages 780–789. Springer, 2007.

[14] Shestakov Denis. On building a search interface discoverysystem. In *Proceedings of the 2nd international conference on* *Resource discovery*, pages 81–93, Lyon France, 2010. Springer.

[15] Luciano Barbosa and Juliana Freire. Searching for hidden-web databases. In *WebDB*, pages 1–6, 2005.

[16] Luciano Barbosa and Juliana Freire. An adaptive crawler for locating hidden-web entry points. In *Proceedings of the* *16th international conference on World Wide Web*, pages 441–450. ACM, 2007.

[17] Soumen Chakrabarti, Martin Van den Berg, and Byron Dom. Focused crawling: a new approach to topic-specific web resource discovery. *Computer Networks*, 31(11):1623–1640, 1999.

[18] Jayant Madhavan, David Ko, Łucja Kot, Vignesh Ganapathy, Alex Rasmussen, and Alon Halevy. Google’s deep web crawl. *Proceedings of the VLDB Endowment*, 1(2):1241–1252, 2008.

[19] Olston Christopher and Najork Marc. Web crawling. *Foundations and Trends in Information Retrieval*, 4(3):175–246, 2010.

[20] Balakrishnan Raju and Kambhampati Subbarao. Sourcerank: Relevance and trust assessment for deep web sources based on inter-source agreement. In *Proceedings of the 20th international* *conference on World Wide Web*, pages 227–236, 2011.

[21] Balakrishnan Raju, Kambhampati Subbarao, and Jha Manishkumar. Assessing relevance and trust of the deep web sources and results based on inter-source agreement. *ACM* *Transactions on the Web*, 7(2):Article 11, 1–32, 2013.

[22] Mustafa Emmre Dincturk, Guy vincent Jourdan, Gregor V.

**Websites Referred:**

http://java.sun.com

http://www.sourcefordgde.com

http://www.networkcomputing.com/

<http://www.roseindia.com/>