

Crawling the Hidden Web



*Application/task specific approach to hidden web
crawling*

Sriram Raghavan & Hector Garcia-Molina
Stanford University

How can we access Hidden Web?

- Presented by **Kalyan Boggavarapu**, graduate student at Lehigh University
- Brief features of Hidden Web
- How and why Hidden web is interesting
- Challenges in Hidden web
- Our approach of the crawler

Outline

- Problem categories
- Crawler form interaction
- HiWE (Hidden Web Exposure)
- Populating LVS
- LITE (Layout based Information Extraction Technique)
- Experimental results

Definitions

- Hidden comprises of:

Publicly Non Indexable Web

- Indexable Web = pages reachable by **hypertext links**
- Task-specificity:
 - those pages which are **relevant** to the given topic

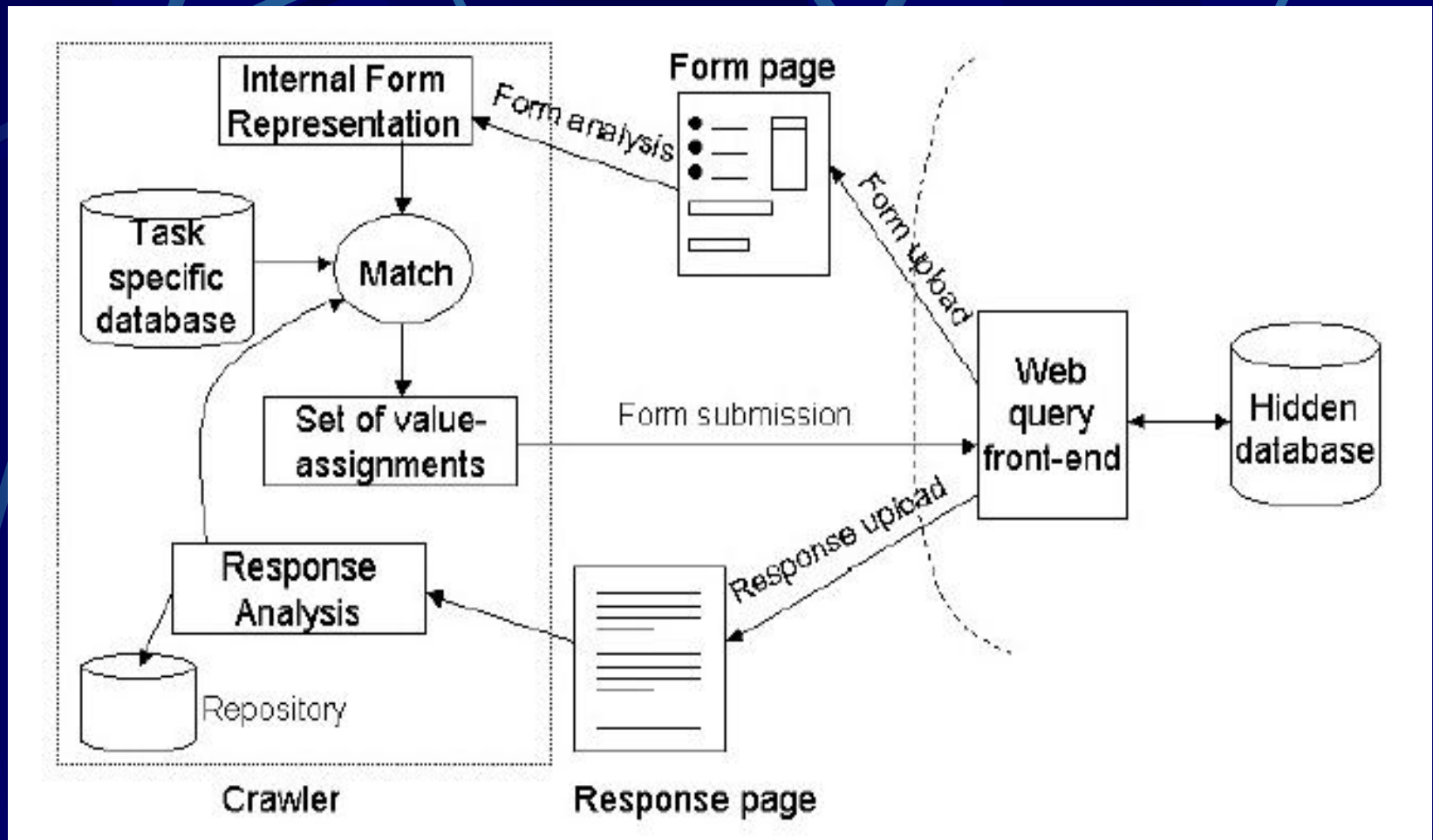
Challenging because ..

- Estimated Size
 - 500 * (size of(Indexable Web));
source: www.brightplanet.com
- Need for crawlers that can handle interfaces **designed for humans**
- Extract the semantic content
 - Eg: label of the text box
- **Filling out forms**

Problem Categories

- Implemented **task specific** crawling
 - We have two problems to address
 - Resource Discovery:
 - Identifying **relevant** databases and sites
 - Content Extraction:
 - Read the form, **prepare the query**, submit the form, extract the hidden web page

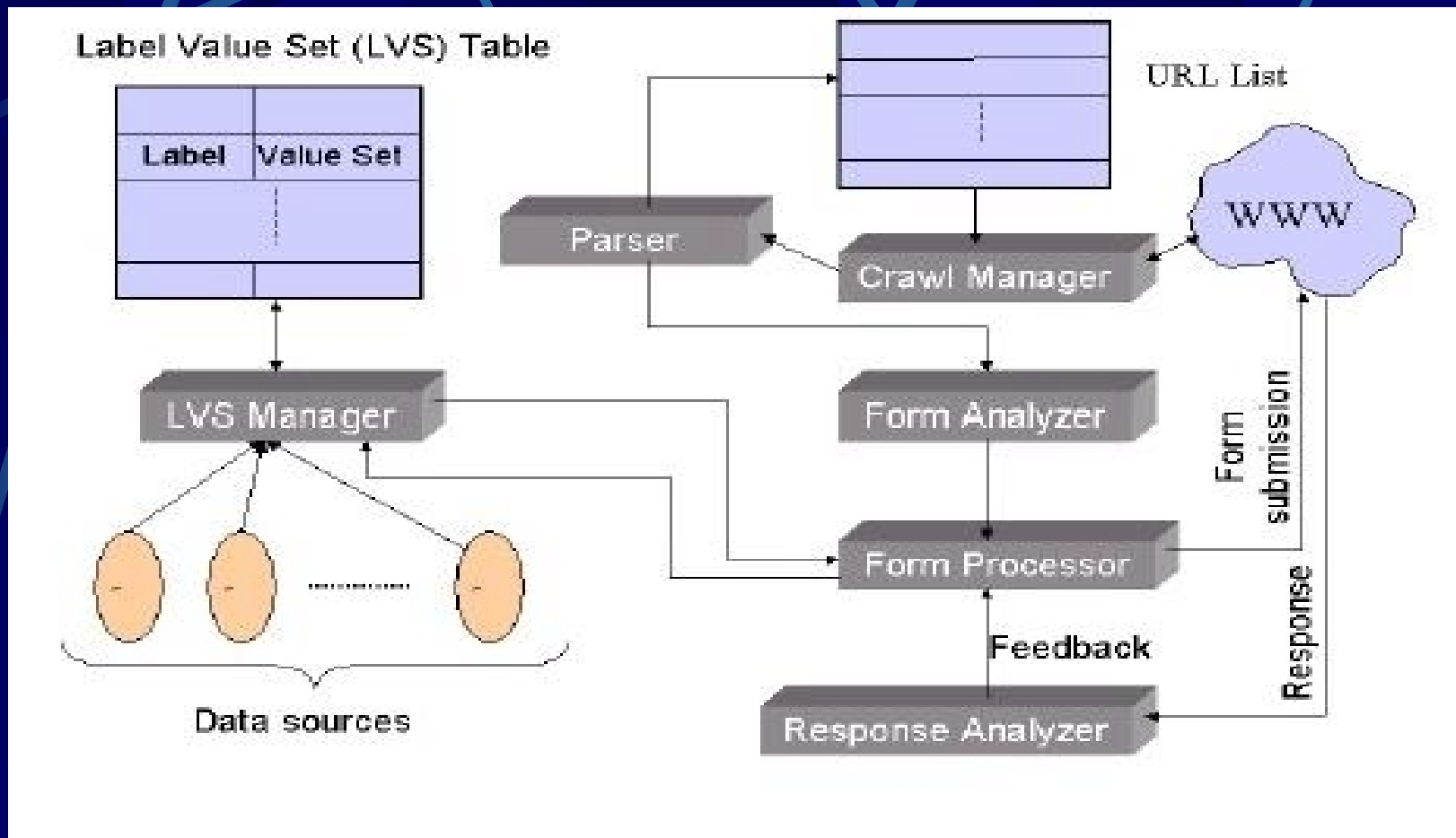
Crawler form Interaction



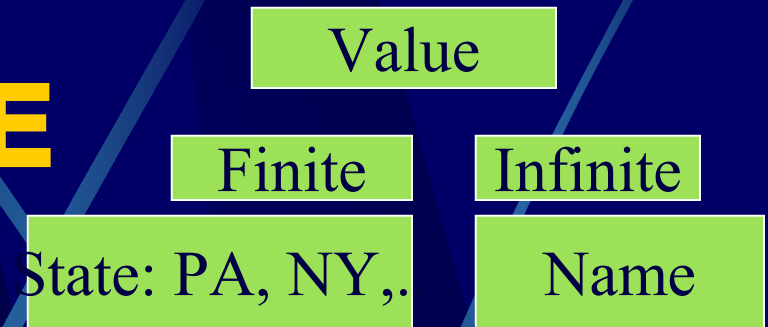
Human Assistance

- Eg: A Market Analyst is interested in gathering information regarding Semiconductor industry
- He supplies list of keywords, companies, products to the crawler
- Crawler uses these values while filling out the forms initially

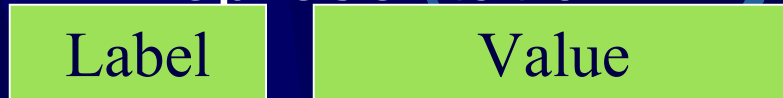
HiWE architecture



HIWE



- Form Representation:



- LVS (Label Value Set):

- is a database of all the label value pairs

- Label Matching

- Minimum edit distances:

- To rectify the typing errors, for Eg: 'state' similar to 'stata'
- Word re-ordering: Eg: company type = type company

Populating LVS

- **Explicit Initialization:**

- ask from the user

- **Built-in entries:**

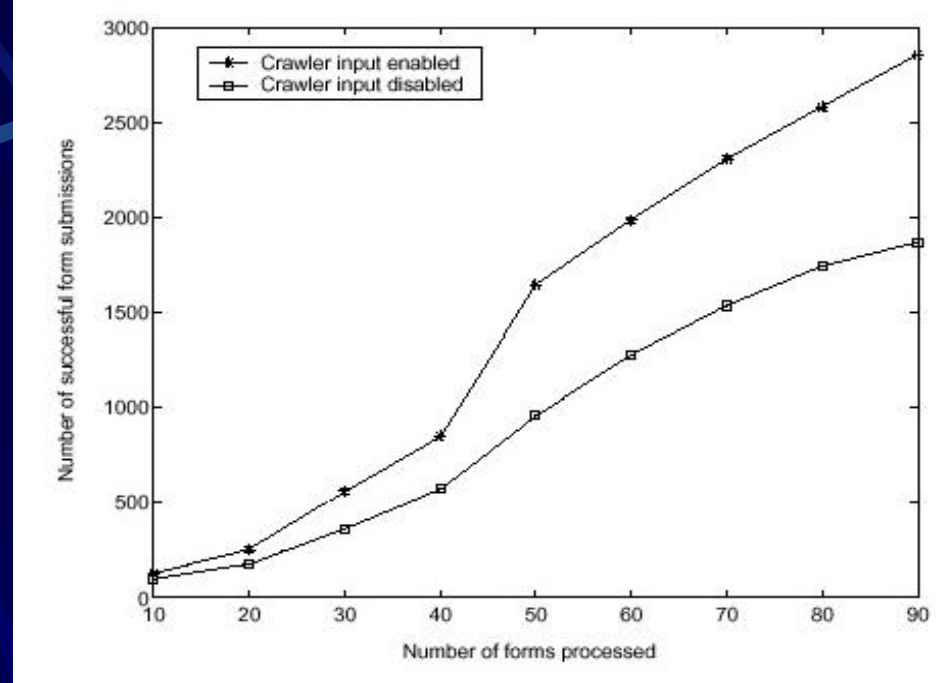
- Eg: dates, cities, states, countries

- **Wrapped data sources:**

- Well defined interface to query good data sources, like the yahoo and extract the information from the results

- **Using Crawler:**

- Crawls the forms on the web and populates LVS



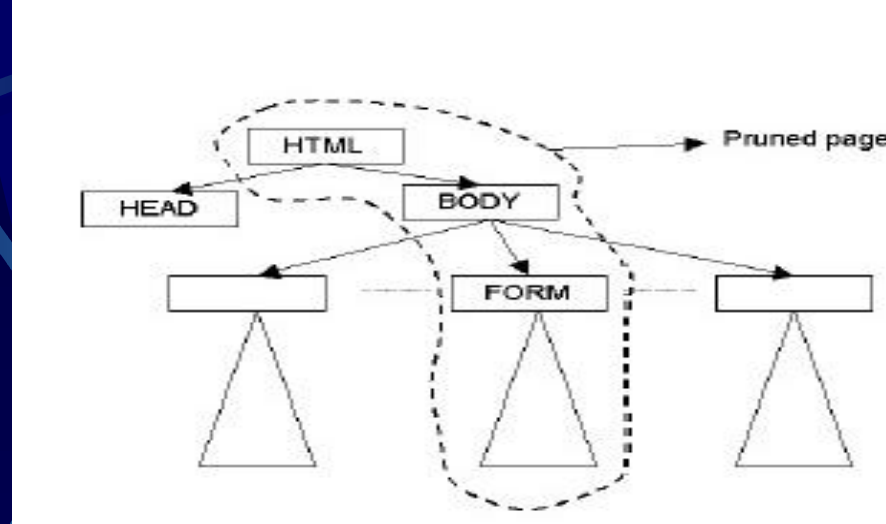
Layout based Information Extraction

- Problem:

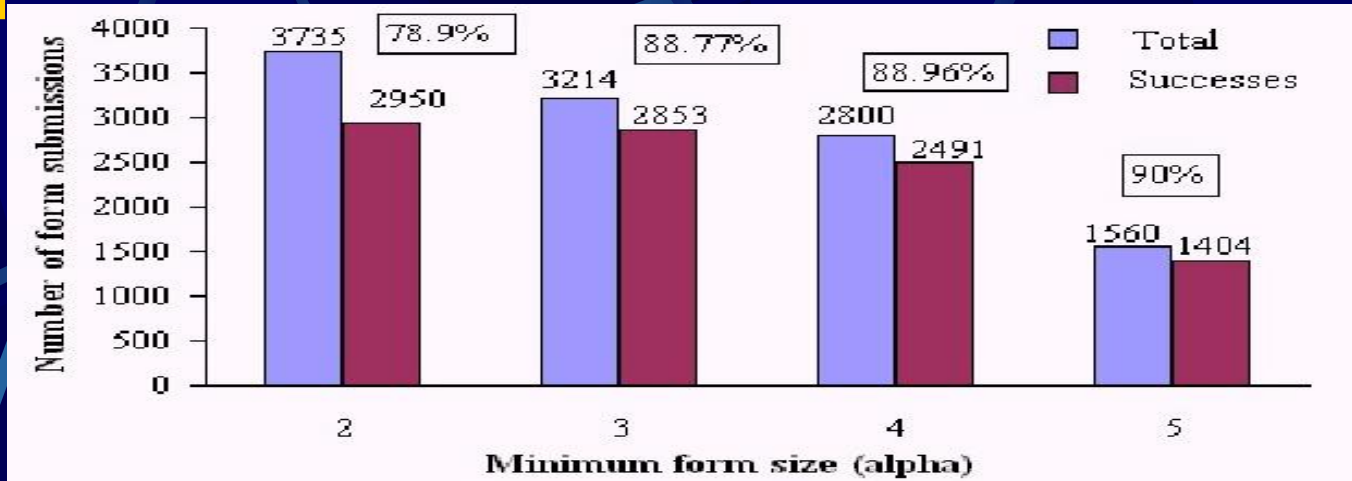
- Label elements are embedded inside a table
- Alignment through explicit spaces and line breaks

- Sol:

- Construct a Pruned tree which has:
 - subtree of the form
 - Nodes on the path from form to root
- Use **Layout Engine** to calculate the physical distance between text elements
- Rank the text elements by position, font size, etc....



Experiment



- Minimum number of form elements set to 3 ?

To eliminate Local site searches

- Using LITE, the highest performance observed:

- 90% of the forms where correctly filled

Do you think Minimum Number of form elements set to 3 is justified?

What we discussed

- Hidden web crawling is different because it involves
 - Data extraction from databases
 - Form extraction and analysis
 - Fill out forms
- We use HiWE crawler to query databases
- Label extraction is another major challenge
We use the LITE for extracting information which are visibly close to each other