2004

A Genetic Algorithm Assisted Hybrid Approach to Web Information Integration

Jia-Lang Seng
National Chengchi University

Ming-Hsiung Ying
Chin Min Institute of Technology

Follow this and additional works at: https://scholarworks.lib.csusb.edu/jiim
Part of the Management Information Systems Commons

Recommended Citation

This Article is brought to you for free and open access by CSUSB ScholarWorks. It has been accepted for inclusion in Journal of International Information Management by an authorized editor of CSUSB ScholarWorks. For more information, please contact scholarworks@csusb.edu.
A Genetic Algorithm Assisted Hybrid Approach to Web Information Integration

Jia-Lang Seng
National Chengchi University, Taiwan
Ming-Hsiung Ying
Chin Min Institute of Technology, Taiwan

ABSTRACT

Heterogeneity and interoperability of Web data sources represent the current key issue in Web information extraction and integration. Warehouse approach and virtual approach are the common approaches adopted to integrate heterogeneous Web data sources. However, few analytic model and cost model were developed to measure and assess the efficiency and effectiveness of either approach or a combination. Hence, a contingency model cannot be produced to assist the search engine to select and mix the warehouse method and the virtual method. In this study, we present a genetic algorithm assisted hybrid approach to aid the search engine to evaluate the cost and performance factors. We apply genetic algorithm technique to formulate a cost optimization model and compute and compare the cost of extraction and integration. The cost model is based on a collection and compilation of the property data of the query analysis and path expression of the involved Web data sources. Six property analyses are conducted and six evolution steps are created to formulate the genetic algorithm of optimization. Further, we conduct a preliminary experiment using 15 local and global Web bookstores to install and test the method. Our experimental results show that the cost optimization can be achieved with the genetic algorithm and factor analysis.

INTRODUCTION

World Wide Web (WWW or Web) technology significantly changes the business and personal information computing environment. It alters the way the business and individuals search and secure data. It changes the way the data is accessed and presented. With the wireless technology emerging, Web information extraction and integration over the mobile computing platform gives the compelling reason to tackle the heterogeneity and interoperability issue with quantitative measure of cost and performance. Further, business and individuals spend ever more time and efforts searching and surfing the Web network. People place ever more attention and resource on the Web media to communicate and disseminate data and information. Heterogeneity and interoperability, as in the distributed computing, becomes one of the key issues in Web information extraction and integration.

In this paper, we present a genetic algorithm assisted method to tackle the research issue. We develop a genetic algorithm assisted optimal cost model and contingency model. We formulate a cost optimization with factor analysis to suggest a quantitative measure to select or combine the warehouse approach and the virtual approach. Six property analyses and six evolution steps are generated. Continuous query mechanism is applied to segment query request and assemble query response. We conduct a preliminary experiment using a standard set of local and global Internet bookstores to install and test the method. Experimental results show that the optimal cost can be achieved on a dynamic process. The exclusive use and hybrid use of either approach can be determined. This paper is organized into five sections. Section one introduces this research. Section two surveys the warehouse approach and the virtual approach. Section three reviews the genetic algorithm. Section four presents the cost optimization model and factor analysis. Section five describes the preliminary experiments. Section six concludes this paper with a discussion and a brief summary.

WEB INFORMATION EXTRACTION AND INTEGRATION

Web information extraction and integration is to process the Web query and operation from single or multiple Web data sources. Various researches including (Etzioni et al., 1994; Woelk et al., 1995; Arens et al., 1996; Levy et al., 1996; Garcia-Molina et al., 1997; Duschka et al., 1997; Friedman et al., 1997; Ambite et al., 1998; Beeri...
et al., 1998; Cohen, 1998) have proposed different resolutions to tackle the heterogeneity and interoperability of Web sources. Most of the resolutions centers on the techniques of wrapping and mediation. Wrappers, mediators, and containers are the proposed solution. However, these techniques are based on a set of assumptions that we have (1) an ever evolving Web, (2) few poor meta-data available, and (3) the ever greater autonomy of Web sources. Hence, prior researches have created the warehouse approach and the virtual approach to work with the wrapper and mediator. These assumptions have been challenged that more and more coherent and compatible Web standards have been proposed and accepted such as XML, SOAP, and Rosetta Net. The warehouse approach and the virtual approach do not have to be mutually exclusive. The reason that the current practice is an exclusive application of independent approaches is due to an assumption that a uniform access to data obtainable from different sources available through the Web. Hence, either all necessary data is collected in a central repository before a user query is issued or the data is collected from the integrated sources dynamically during query evaluation (Vdovajk et al., 2001) (Voida et al., 2001). In this study, there can be a contingency strategy to select either independent approach or hybrid approach. The continuous query mechanism to segment and assemble query group can be applied to work with and enable the hybrid approach. Either way a cost and performance analysis model is needed to justify the decision.

Information Extraction

WWW documents are saved in the HTML/XML format. These are the semi-structured and heterogeneous data documents stored at different Web sites. Web information extraction means a process of three technical tasks to be conducted to access and retrieve the desired information including the data extraction, the data analysis, and the data presentation.

1. Data extraction: Information extraction must first extract data and deal with the problems of the data source location, the data variation, and the un-structured relationship of data objects. It needs the user intervention to decide the data source and format.

2. Data analysis: Information extraction must store and represent the hypermedia data in tables and objects. It must process user query and depend on the metadata analysis of data sources. Current methods such as the Object-Based Logical Models (OBLM) and Record-Based Logical Models (RBLM) describe the data sources by object relationships and by data fields. However, overhead of object storage and evolution is high.

3. Data presentation: Information extraction must present the query result in different data formats simultaneously. The proper amount and stream of Web pages must be determined based on the time and cost calculated.

Further, information extraction involves the process of searching and surfing over the Web. Speed, accuracy, and response time is vital. Information extraction needs an efficient and effective search strategy. There are three basic search methods including the index robot, the agent, and the wrapper/mediator.

1. Index robot: Traditional search engine utilizes the index robot to extract Web data sources. Users first enter the Web URLs. Index robot searches and builds its Web searching list. Web pages are treated as graph structures of hypertexts and hyperlinks. It adopts the depth-first method and the breadth-first method to traverse every node and link. WWW architecture is neither a tree nor a directed cyclic graph structure. Circles result must be dealt with.

2. Agent: With the above-mentioned index robot method, each searched Web page must be sent to the search engine and be differentiated. This process causes a large amount of data to be transferred and managed. More, users are required to (1) know the difference between these Web sources, (2) spend time connecting multiple Web sources and know whether the query is processed, (3) know the different query interfaces, and (4) re-process data when conflicts and incompatible formats occur. These cause problems and inconveniences for users. Agent is an alternative. Agent’s work on the collection and management of distributed data by themselves. Each agent concentrates on one type of task. Each collaborates with one another. Tasks including mapping, matching, routing, dispatching, searching, and securing. (Okada et al. 1996) proposed an architecture commonly referred to and called CAS-IG (Cooperative Agent Society for Information Gathering). The architecture has three parts: the user agent, the manager, and the machine agent.

3. Wrapper/Mediator: Wrappers are translators. A wrapper is a procedure tailored to one single Web data
source. It translates a query request and response to a relational or object format. Wrappers convert the query into local source-specific format. It can be a schema builder to compile the outside data sources into one single global schema. Mediators collaborate with wrappers. A mediator examines and decides which wrapper works on which query or Web source. It dispatches works, combines results, and directs response to users.

**Information Integration**

As described above, two main approaches, the warehouse approach and the virtual approach, are used to integrate the heterogeneous Web data sources. The common practice is to install one approach over another, that is, an exclusive use of one approach on one Web site. Below is a brief introduction to both approaches and the reason why the current practice is an exclusive application. We add an analysis of the conditions fit to each approach.

1. **Warehouse Approach:** Data from multiple Web data sources is loaded into a local Web warehouse. Queries are applied with the warehoused data. Adequate performance is assured with the query response time. The cost includes the time to download all multiple Web sources and to maintain them on the local site. However, it cannot assure the extracted warehouse data is current and consistent.

2. **Virtual Approach:** Web data sources are not warehoused. Web queries are processed through the real time data mediation and wrapping. Data is guaranteed to be current and consistent. The cost includes the time to query, translate, and integrate these heterogeneous data sources at the run time. The disadvantage is that if the connection is broken, real time alternative is hard to generate.

Table 1 provides a set of suggested conditions when each approach should be used. We explain each condition and its justification following the presentation. These suggested conditions represent the core of the property analysis we develop. There six factors consisting of principle elements studied and discussed in standard heterogeneous information integration (HII) literature where query optimization and cost computation is presented (Tzitzikas et al., 2002) (Voida et al., 2001) (Vdovajk et al., 2001) (Vrajitoru 1998).

**Table 1. Suggested Conditions of Warehouse Approach and Virtual Approach**

<table>
<thead>
<tr>
<th>Source Feature</th>
<th>Warehouse Approach</th>
<th>Virtual Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of retrieval and response time</td>
<td>Long</td>
<td>Fast</td>
</tr>
<tr>
<td>Average of update frequency</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Average of retrieval failure frequency</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Average of data structure homogeneity</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Average of retrieval contribution degree</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Average of information integration complexity</td>
<td>High</td>
<td>Low</td>
</tr>
</tbody>
</table>

1. The average of retrieval and response time: The average of retrieval and response time means the average time to connect to a specific Web data source and to secure the query data. If the average time is long, the portion of the Web data source should be warehoused, otherwise, the virtual approach is suggested.

2. The average of update frequency: The average of update frequency means the average of change frequency for a standard period of time. When the query requires the real time data accuracy, Web data source should be in the virtual approach. If the query inquires less frequently updated data such as the published dissertation, the warehouse approach should be used.
3. The average of retrieval failure frequency: The average of retrieval failure frequency means the average of the times that a Web data source cannot be connected over a standard period of time. If the average is higher, the warehouse approach should be used, otherwise the virtual approach is suggested.

4. The average of data structure homogeneity: The consistent degree of data scope means the degree the homogeneity of the contents of one particular Web data sources. For instance, if we query the price of computer books, the Web site queried has both the book information on computers and other un-related information such as the travel, politics, law, and economics book information. We say this Web data source has a low consistent degree and suggest a virtual approach, otherwise we suggest a warehouse approach.

5. The average of retrieval contribution degree: The average of retrieval contribution degree means that the degree that one particular Web data source can fulfill most of the query request. If the degree is low, a warehouse approach is a better choice because multiple Web data sources are un-avoidable. Poor and broken Web connections are what we try to avoid. On the other hand, if the degree is high, we recommend the virtual approach.

6. The average of information integration complexity: The complicated degree means the degree of the heterogeneity of one Web data source in terms of its Web structure, source semantics, and retrieval paths. If the source is less complicated, less time is required to integrate different data sources. The virtual approach is fine. Data currency is increased. If the source is more complicated, the warehouse approach is better than the virtual approach.

**TECHNIQUE OF GENETIC ALGORITHM**

Genetic algorithm (GA) is a set of probabilistic search techniques, which mimic the process of evolution. The fundamental principles of genetics lead to the development of GA. A genetic algorithm consists of five components, as described in Davis's (Davis, 1987), are:

1. A method for encoding potential solutions into chromosomes.
2. A means of creating the initial population.
3. An evaluation function that can evaluate the fitness of chromosomes.
4. Genetic operators that can create the next generation population.
5. A way to set up control parameters, e.g., the population size, the probability of applying a genetic operator, and so on.

In fact, there are two operations in the traditional genetic algorithm that we adopt in this research:

2. Evolution operation: selection.

John Holland was inspired by the theory of evolution to create a computer algorithm (Holland, 1992). First, Holland used a list of zeros and ones - a bit string to encode chromosome. Computers can represent everything they do with bit string. In Holland’s genetic algorithm, all evaluation functions returned positive numbers. The higher the number is, the better the chromosome is. Holland used a technique, now called the roulette wheel selection, to determine which population members are chosen for reproduction event. As shown in Figure 1, crossover operator swaps the left potions of two chromosomes. Mutation operator replaces bits on a chromosome with randomly generated bits.

**Figure 1. Genetic Operators of Holland’s Genetic Algorithm**

```
11000101 11001111
10001111 10001100
11000101 10001100
10001111 11001111
```

(a) Crossover operation
The following brief outline of GA illustrates the functioning of GA, where the notation $S(t)$ is the population in the $t$th generation; $s_i(t)$ is the $i$th member in $S(t)$; $f(s_i(t))$ is the fitness value of $s_i(t)$, and $\text{TOTFIT}(t)$ is the sum of $f(s_i(t))$ for all $s_i(t)$ and $S(t)$ that we will adopt and adapt in our new method.

**Step 1**

Generate the initial population, $S(t)$, where $t=0$. Determine the size of the population, $\text{POPSIZE}$, and the number of generations, $\text{GENER}$.

**Step 2**

Calculate the fitness value of each member, $f(s_i(t))$, for population, $S(t)$.

**Step 3**

Calculate the selection probability for each $s_i(t)$, where the selection probability is defined as

$$P(s_i(t)) = \frac{f(s_i(t))}{\text{TOTFIT}}$$

where

$$\text{TOTFIT} = \frac{\text{POPSIZE}}{\sum_{i=1}^{\text{POPSIZE}} f(s_i(t))}$$

In our case, we have revised the formula, the $P(s_i(t)) = \frac{[f(s_i(t)) - \text{Min}(f(s_i(t)))]}{\text{TOTFIT}}$, and $\text{TOTFIT} = \frac{\text{POPSIZE}}{\sum_{i=1}^{\text{POPSIZE}} [f(s_i(t)) - \text{Min}(f(s_i(t)))]}$

The purpose that revises the formula is to make the higher fitness value have a higher probability of being selected with more significance.

**Step 4**

Select a pair of members (parents) that will be used for reproduction via the selection probability.

**Step 5**

Apply genetic operators (crossover, mutation, inversion) to the parents. Replace the parents with the resulting offspring to form a new population, $S(t+1)$, for generation $t+1$. If the size of the new population is equal to the $\text{POPSIZE}$, then go to step 6, else go to step 4.

**Step 6**

If current generation, $t+1$, is equal to $\text{GENER}$, then stop, else go to step 2.
A GENETIC ALGORITHM ASSISTED HYBRID METHOD

In this section, we present a genetic algorithm assisted hybrid method to formulate a cost optimization model with property analysis to assess and measure the time and cost of the warehouse approach or the virtual approach or the hybrid approach. The cost optimization model works with the search engine and the continuous query mechanism to evaluate and determine if an effective information extraction and integration strategy is chosen.

Architecture

Figure 2 shows an experimental architecture of the genetic algorithm assisted hybrid method. Both the warehouse approach and the virtual approach are built. Wrapper and mediator techniques are used. An evaluation and adjustment mechanism is applied to adjust the integration approach. In specific, we first classify Web data sources into the “virtual group” and the “warehouse group”. We formulate continuous queries. We compile and compute the property data of Web data sources and their status. We apply and execute the genetic algorithm through the property analyses and the evolution steps. We generate and produce the cost optimization table of results. The selection is made based on the results and contingent suggestions. A two-phase data presentation is developed to pipeline and assemble query results from the warehouse approach and the virtual approach.

The implementation of the cost optimization model implies an optimal degree of the query satisfaction reached under the user-given constraints. When the warehouse approach is chosen, it implies: (1) the Web data sources have the average of retrieval and response time longer, (2) the Web data sources have the average of update frequency lower, (3) the Web data sources have the average of retrieval failure frequency higher, (4) the Web data sources have the consistent degree of data scope lower, (5) the Web data sources have the average of retrieval contribution degree higher, and (6) the Web data sources have the complicated degree of information integration higher. When the virtual approach, it implies: (1) the Web data sources have the average of retrieval and response time faster, (2) the Web data sources have the average of update frequency higher, (3) the Web data sources have the average of retrieval failure frequency lower, (4) the Web data sources have the consistent degree of data scope higher, (5) the Web data sources have the average of retrieval contribution degree lower, and (6) the Web data sources have the complicated degree of information integration lower.

Property Analysis

Property analysis analyzes the Web source factors in information integration. As described in literature review, we generate a generic property set including (1) the average of retrieval and response time, (2) the average of update frequency, (3) the average of retrieval failure frequency, (4) the average of data structure homogeneity, (5) the average of retrieval contribution degree, and (6) the average of information integration complexity. Table 2 is an example of our preliminary experiment that we execute a property analysis. We analyze each Web data source property using the discrete distribution of weights. We derive the fitness values of each property using the translation formulas.
Figure 2. Experimental Architecture of Hybrid Approach

Internet

Query Application

Global Mediator

GA Calculator

Property Database

Adjuster

Hybrid Approach

Virtual Approach

Query/Schema

Mediator

Wrapper 1

Wrapper 2

......

Wrapper n

Warehouse Approach

Mediator

Wrapper 1

Wrapper 2

......

Wrapper n

Internet

Web
Bookstore 1

Web
Bookstore 2

Web
Bookstore n
Table 2. Property Data for Each Web Data Source

<table>
<thead>
<tr>
<th></th>
<th>BS₁</th>
<th>BS₂</th>
<th>BS₃</th>
<th>BS₄</th>
<th>BS₅</th>
<th>BS₆</th>
<th>BS₇</th>
<th>BS₈</th>
<th>BS₉</th>
<th>BS₁₀</th>
<th>BS₁₁</th>
<th>BS₁₂</th>
<th>BS₁₃</th>
<th>BS₁₄</th>
<th>BS₁₅</th>
</tr>
</thead>
<tbody>
<tr>
<td>C₁</td>
<td>7.33</td>
<td>47.67</td>
<td>13.67</td>
<td>3.67</td>
<td>2</td>
<td>12.33</td>
<td>6.67</td>
<td>1.33</td>
<td>46.33</td>
<td>8</td>
<td>27.67</td>
<td>1</td>
<td>30</td>
<td>5</td>
<td>249</td>
</tr>
<tr>
<td>W₁</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C₂</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>15</td>
<td>25</td>
<td>20</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>30</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>W₂</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C₃</td>
<td>0</td>
<td>0.14</td>
<td>0</td>
<td>0</td>
<td>0.003</td>
<td>0.001</td>
<td>0.003</td>
<td>0</td>
<td>0</td>
<td>0.021</td>
<td>0</td>
<td>0.012</td>
<td>0</td>
<td>0.002</td>
<td>0.01</td>
</tr>
<tr>
<td>W₃</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C₄</td>
<td>0.15</td>
<td>0.75</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>W₄</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C₅</td>
<td>0.669</td>
<td>0.460</td>
<td>0.451</td>
<td>0.089</td>
<td>0.058</td>
<td>0.326</td>
<td>0.048</td>
<td>0.05</td>
<td>0.36</td>
<td>0.019</td>
<td>0.575</td>
<td>0.031</td>
<td>0.089</td>
<td>0.074</td>
<td>1</td>
</tr>
<tr>
<td>W₅</td>
<td>0</td>
<td>0.677</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
<td>0.667</td>
</tr>
<tr>
<td>C₆</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>1</td>
<td>8</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>5</td>
<td>8</td>
<td>10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>W₆</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: *Gray background color is the default value, white color is the source data value
BS₁ - BS₁₅: notate the 15 Internet bookstores, i = 1, ..., 15
C₁ - C₆: notate the six evaluation properties
C₁: average of retrieval and response time
C₂: average of update frequency
C₃: average of retrieval failure frequency
C₄: consistent degree of data scope
C₅: average of retrieval contribution degree
C₆: complicated degree of information integration
W_ij: parameter of adopting warehouse approach
V_ij: parameter of adopting virtual approach

Table 3. Weight of Each Property

<table>
<thead>
<tr>
<th></th>
<th>C₁</th>
<th>C₂</th>
<th>C₃</th>
<th>C₄</th>
<th>C₅</th>
<th>C₆</th>
</tr>
</thead>
<tbody>
<tr>
<td>WTᵢ</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

Each property has a different unit of measurement. We convert the property data value to the standard unit. Standard unit notates the satisfaction degree of a Web query. There are six translation formulas used to undertake these property analyses and generate the fitness values. We present each of the formulas with an explanation as follows.

Property 1(C₁):
If \( V_{ij} > 90 \) sec, then \( V_{ij} = 90 \)
\[
W_{ij} = (90 - W_{ij}) \times 100, \quad V_{ij} = (90 - V_{ij}) \times 100
\]
Property 2(C2):
If \( V_{i2} > 30 \text{ day} \), then \( V_{ii} = 30 \)
\[
W\_{i2} = (30 - W\_{i2})/30*100, \quad V\_{i2} = (30 - V\_{i2})/30*100
\]

Property 3(C3):
\[
W\_{i3} = \sqrt{1 - W\_{i3}} \times 100, \quad V\_{i3} = \sqrt{1 - V\_{i3}} \times 100
\]

Property 4(C4):
\[
W\_{i4} = W\_{i4} \times 100, \quad V\_{i4} = V\_{i4} \times 100
\]

Property 5(C5):
\[
W\_{i5} = W\_{i5} \times 100, \quad V\_{i5} = V\_{i5} \times 100
\]

Property 6(C6):
\[
W\_{i6} = (10 - W\_{i6})/10*100, \quad V\_{i6} = (10 - V\_{i6})/10*100
\]

Property 1 notates the average of retrieval and response time and must be no greater than a user-defined level of 90 seconds. If the average of retrieval and response time has exceeded 90 seconds, the satisfaction degree equals 0. Otherwise, when the average of retrieval and response time is smaller, and the satisfaction degree is higher.

Property 2 notates the average of update frequency once and must be no greater than per 30 days (or one month) as user requests. If the average of update frequency once has exceeded 30 seconds, the satisfaction degree equals 0. Otherwise, when the average of update frequency is smaller, and the satisfaction degree is higher.

Property 3 is the average of retrieval failure frequency. If the average of retrieval failure frequency is 100% (very high), the satisfaction degree equals 0.

Property 4 is the homogeneity degree of data structure. If the homogeneity degree is higher, the satisfaction degree is higher too.

Property 5 is the average of retrieval contribution degree. If the average of retrieval contribution degree is higher, the satisfaction degree is higher too (the range of original value is from 0 to 100%).

Property 6 is the complication degree of information integration. If the complication degree of information integration is higher, the satisfaction degree is lower (the range of original value is from 0 to 10).

These properties have to be transformed to a satisfaction degree using one of these formulas. We calculate the satisfaction degree values using the fitness function described as follows.

\[
\text{Fitness Function} = \frac{\sum (1-a_i) \cdot \Sigma V_{ij} \cdot WT_{ij} + a_i \cdot \Sigma W_{ij} \cdot WT_{ij} \cdot 100}{\sum \text{ bits}}, \quad a_i = 0 \text{ or } 1
\]

The \( a_i = 0 \) means \( i^{th} \) Web data source that can be classified into the virtual approach group. The \( a_i = 1 \) means \( i^{th} \) Web data source that can be classified into the warehouse approach group. There are 15 bits to compose one chromosome. Each bit has to be encoded by 0 or 1 and to indicate one Web data source.

In our example experiment, each data source is notated by one bit and each bit is 0 or 1 to indicate whether it is the virtual group or the warehouse group. The experiment consists of 15 Web data sources. We have \( 2^{15} = 1024 \times 32 = 32,768 \) combinations of variety. Though it is a large simulation, we can easily solve the problem using the genetic algorithm technique.

If a chromosome structures \( \{a_1, a_2, \ldots, a_{15}\} \) has been written as \( \{101001011000101\} \), it means that the 1\(^{st}\), 3\(^{rd}\), 6\(^{th}\), 8\(^{th}\), 9\(^{th}\), 13\(^{th}\), 15\(^{th}\) Web data source can be classified into the warehouse group, and use the warehouse approach to integrate information; whereas, the 2\(^{nd}\), 4\(^{th}\), 5\(^{th}\), 7\(^{th}\), 10\(^{th}\), 11\(^{th}\), 12\(^{th}\), 14\(^{th}\) Web source can be classified into the virtual group, and use the virtual approach to integrate information. Each chromosome gets the fitness value by mapping every bit to a particular source property that is calculated.
IMPLEMENTATION OF GENETIC ALGORITHM

We need to implement the developed set of genetic algorithm into a process of evolution steps as discussed in literature review. The implementation is described as follows.

**Step 1**
Generate the initial population $S(t)$ in Table 4, where $t=0$. GA parameter is GA (30, 0.975, 0.015, 1.0, 1, E), the number of generations, GENER=300.

<table>
<thead>
<tr>
<th>I</th>
<th>$s_i(t)$</th>
<th>$f(s_i(t))$</th>
<th>$P(s_i(t))$</th>
<th>Accumulative probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>00110010001100</td>
<td>74.30109</td>
<td>0.03804</td>
<td>0.03804</td>
</tr>
<tr>
<td>002</td>
<td>00011001000111</td>
<td>73.63920</td>
<td>0.02839</td>
<td>0.06642</td>
</tr>
<tr>
<td>003</td>
<td>100110110000111</td>
<td>72.05098</td>
<td>0.00522</td>
<td>0.07165</td>
</tr>
<tr>
<td>004</td>
<td>101111010101011</td>
<td>72.94028</td>
<td>0.01819</td>
<td>0.08984</td>
</tr>
<tr>
<td>005</td>
<td>011100100011100</td>
<td>74.93188</td>
<td>0.04724</td>
<td>0.13708</td>
</tr>
<tr>
<td>006</td>
<td>110100110000011</td>
<td>71.69271</td>
<td>0.00000</td>
<td>0.13708</td>
</tr>
<tr>
<td>007</td>
<td>0111010010001011</td>
<td>74.81274</td>
<td>0.04550</td>
<td>0.18257</td>
</tr>
<tr>
<td>008</td>
<td>0110011000110111</td>
<td>75.64617</td>
<td>0.05765</td>
<td>0.24023</td>
</tr>
<tr>
<td>009</td>
<td>0001000101100111</td>
<td>72.82082</td>
<td>0.01645</td>
<td>0.25668</td>
</tr>
<tr>
<td>010</td>
<td>1010101001110101</td>
<td>72.15601</td>
<td>0.00676</td>
<td>0.26343</td>
</tr>
<tr>
<td>011</td>
<td>1000010101111111</td>
<td>72.76434</td>
<td>0.01563</td>
<td>0.27906</td>
</tr>
<tr>
<td>012</td>
<td>1000110000000101</td>
<td>75.28389</td>
<td>0.05237</td>
<td>0.33143</td>
</tr>
<tr>
<td>013</td>
<td>1011001001100001</td>
<td>73.79508</td>
<td>0.03066</td>
<td>0.36209</td>
</tr>
<tr>
<td>014</td>
<td>0111101000110111</td>
<td>74.44418</td>
<td>0.04012</td>
<td>0.40221</td>
</tr>
<tr>
<td>015</td>
<td>1111000111010101</td>
<td>73.95589</td>
<td>0.03300</td>
<td>0.43521</td>
</tr>
<tr>
<td>016</td>
<td>1111000001111011</td>
<td>74.51656</td>
<td>0.04118</td>
<td>0.47639</td>
</tr>
<tr>
<td>017</td>
<td>0001111001000011</td>
<td>75.53661</td>
<td>0.05605</td>
<td>0.53245</td>
</tr>
<tr>
<td>018</td>
<td>1001010000000111</td>
<td>74.22493</td>
<td>0.03693</td>
<td>0.56938</td>
</tr>
<tr>
<td>019</td>
<td>1010011110111111</td>
<td>73.68130</td>
<td>0.02900</td>
<td>0.59837</td>
</tr>
<tr>
<td>020</td>
<td>1001001011101111</td>
<td>72.82786</td>
<td>0.01655</td>
<td>0.61493</td>
</tr>
<tr>
<td>021</td>
<td>1101010111000001</td>
<td>74.91785</td>
<td>0.04703</td>
<td>0.66196</td>
</tr>
<tr>
<td>022</td>
<td>0000010100000101</td>
<td>75.12695</td>
<td>0.05008</td>
<td>0.71204</td>
</tr>
<tr>
<td>023</td>
<td>0110001000000111</td>
<td>74.39539</td>
<td>0.03941</td>
<td>0.75145</td>
</tr>
<tr>
<td>024</td>
<td>1100111100111101</td>
<td>74.00224</td>
<td>0.03368</td>
<td>0.78513</td>
</tr>
<tr>
<td>025</td>
<td>1011001001111001</td>
<td>73.54758</td>
<td>0.02705</td>
<td>0.81218</td>
</tr>
<tr>
<td>026</td>
<td>0011010011111110</td>
<td>76.17761</td>
<td>0.06540</td>
<td>0.87758</td>
</tr>
<tr>
<td>027</td>
<td>1011101011110110</td>
<td>71.76176</td>
<td>0.00101</td>
<td>0.87859</td>
</tr>
<tr>
<td>028</td>
<td>1000110100011111</td>
<td>74.01786</td>
<td>0.03391</td>
<td>0.91250</td>
</tr>
</tbody>
</table>
Step 2
Calculate the fitness value of each member, $f(s_i(t))$. The fitness value of each number (see Table 4), e.g. the twenty-sixth chromosome has the optimal fitness value, and the satisfaction degree for query is 76.17761%.

Step 3
Calculate the selection probability for each $s_i(t)$, in the thesis. The selection probability is defined as $P(s_i(t)) = \frac{f(s_i(t)) - \min(f(s_i(0)))}{\text{TOTFIT}}$, and

\[
\text{TOTFIT} = \sum_{i=1}^{\text{POPSIZE}} \left[ f(s_i(t)) - \min(f(s_i(0))) \right]
\]

In Table 4, the twenty-sixth chromosome has the maximal fitness value, and the chromosome has the maximal probability to be selected as a parent, and the probability value is 0.0654. The sixth chromosome has the minimal fitness value, and the selected probability is 0.

Step 4
Select a pair of members (parents) that compare the accumulation probability with a random number (ranging from 0 to 1), and reproduce the chromosome into the new population.

Step 5
Apply the genetic operators (crossover, mutation, and inversion) to the parents. Replace the parents with the resulting offspring to form a new population, $S(t+1)$, for the generation $t+1$. If the size of the new population is equal to 30, then go to step 6, else go to step 4. In our case, every generation including 30 chromosomes, we use the higher crossover rate of 0.975 that can generate more newly structure, and prevent dropping into the local optimal solution. Our mutation rate is 0.015. If the result falls in the optimal local area, we will consider it in the computation of the global optimal area to be a possible final number.

Step 6
If the current generation, $t + 1$, is equal to 300, then stop, else go to step 2.

Two-Phase Presentation

Traditional query result presentation has to wait for complete query results and show them in one single window. Our method provides a two-phases presentation of query results in increments. The first phase presents the query results from the warehouse approach and from the virtual approach. The second phase pipelines both results and feeds them into the continuous query mechanism for assembly. The increments at the first phase retrieve query data from the local warehoused data sources and from the virtual data sources. Two-phase presentation allows the flexibility and efficiency of continuous heterogamous query parsing and processing.

A PRELIMINARY EXPERIMENT

We conducted a preliminary experiment to test and execute our method. The virtual approach and the warehouse approach were programmed into a set of wrappers and mediators. A Web bookstore information integration query engine was built to run with the genetic algorithm. Adjuster was created to perform the factor analysis. In this example experiment, we asked users to issue queries on book catalog and price information. Multiple heterogeneous Web sources needed to be sought and integrated. For a standard set of Internet bookstores, we got the optimal fitness value 78.63039%. Its chromosome structure was {011001001001101}. The 1st, 4th, 5th, 7th, 8th, 10th, 11th, 14th Web data sources were classified into the virtual group. We decided to use the virtual approach to test the optimization. The 2nd, 3rd, 6th, 9th, 12th, 13th, 15th Web data sources were classified into the warehouse group.
We decided to use the warehouse approach to test the optimization.

In the case where all Web bookstore data sources have chosen the warehouse approach (the chromosome structure was \{1111111111\}), the fitness function value was 72.70293%. When all Web bookstore data sources have used the virtual approach (the chromosome structure was \{0000000000000\}), the fitness function value was 75.70533%. In this situation, we selected the combination model to extract and integrate information. The hybrid approach gave us the optimal satisfaction degree of query. Comparing with Table 2, we have found that 2^th, 3^th, 6^th, 9^th, 13^th, 15^th Web data sources had a longer average of retrieval and response time. We classified them into the warehouse group. But the 12^th Web data source’s average of retrieval and response time was faster. We classified it into the virtual group. This meant that we could not classify all Web sources at one time. We could only group them one at a time based on the property analysis. Table 5 shows the variation of the fitness function values. The fitness has changed six times. The sixth time was the final optimal solution. The 9^th to 139^th generations had the same fitness function value of 78.58757, and the fitness became a local optimal solution. But the 140^th generation left the local optimal solution and was moved to another better optimal solution. The mutation moved it.

Table 5. Variation of Fitness Function Value

<table>
<thead>
<tr>
<th>Generation</th>
<th>Fitness Function Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>76.84378</td>
</tr>
<tr>
<td>2</td>
<td>77.31331</td>
</tr>
<tr>
<td>3–5</td>
<td>78.03445</td>
</tr>
<tr>
<td>6–8</td>
<td>78.1096</td>
</tr>
<tr>
<td>9–139</td>
<td>78.58757</td>
</tr>
<tr>
<td>140–300</td>
<td>78.63039</td>
</tr>
</tbody>
</table>

Figure 3 shows the stable maximal fitness value after the 140^th generation. The average fitness value rose very fast after the initial stage reaches the steady state. The minimal fitness descended gracefully with the chromosome mutation. The mutant chromosome had the minimal fitness value. Therefore, we were sure the optimal cost was found and the hybrid approach was chosen.

Figure 3. Fitness Value of Each Generation
SUMMARY

Discussion and Limitation

In this section, we discuss the implication and limitation of this research. The genetic algorithm assisted hybrid approach is innovative and incremental. However, there is the scale and scope limitation in terms of the factors compiled and applied in the property analysis in the computation and decision making. We perceive Web information integration as a process of cost optimization and property analysis. Measurement and evolution has to be conducted. The current practice is inadequate because it always chooses one approach over the other. No cost and performance is considered. No combined approach is suggested. No continuous query mechanism is applied. This study perceives the search engine in need of a cost optimization model and a contingency mechanism. Genetic algorithm is a viable approach to incorporate the heterogeneity and integration costs with generation of value evolution.

Concluding Remarks

In this paper, we have presented a genetic algorithm assisted hybrid method. The method allows the information integration to select either warehouse or virtual approach or combine both. Basic continuous query technique is used to assist the query group and assembly. Property analysis of condition and cost factors is processed in a dynamic manner. Prior researches have not been able to formulate the genetic algorithm to give the cost optimization and translation formula. The development is traceable and provable. In implementation, six evolution steps are created. We have conducted a preliminary experiment using 15 Internet bookstore Web sites to run and test the genetic algorithm assisted method. Property analyses and evolution steps are applied and measured. Experimental results show that the optimal cost was achieved and a hybrid approach was chosen. However, this is only an initial effort before any undertaking of a larger scale of experimentation and modification. In the near future, we expect to incorporate more factors of heterogeneous Web sources into the genetic algorithm method and incrementally revise the cost model. We expect to experiment with more heterogeneous Web sources to further test the robustness and generality of the method and model.

REFERENCES


Codd, E.F. A relation model for large shared data banks. *CACM* 13(6), 377-387


