01 Jan 1999

Cost-benefit Analysis of Web Bag in a Web Warehouse

Sanjay Kumar Madria
Missouri University of Science and Technology, madrias@mst.edu

Wee Keong Ng

Ee-Peng Lim

Sourav S. Bhowmick

Recommended Citation

S. K. Madria et al., "Cost-benefit Analysis of Web Bag in a Web Warehouse," *Proceedings of the International Symposium on Database Engineering and Applications, 1999*, Institute of Electrical and Electronics Engineers (IEEE), Jan 1999.
The definitive version is available at https://doi.org/10.1109/IDEAS.1999.787249

Cost-Benefit Analysis of Web Bag in a Web Warehouse [*]

Sourav S Bhowmick[1] Sanjay Madria[2] Wee-Keong Ng[1] Ee-Peng Lim[1]

Centre for Advanced Information Systems[1]
Nanyang Technological University
Singapore 639798
{p517026, awkng, aseplim}@ntu.edu.sg

Department of Computer Science[2]
Purdue University
West Lafayette, IN 47907
skm@cs.purdue.edu

Abstract

Sets and bags are closely related structures and have been studied in relational databases. A bag is different from a set in that it is sensitive to the number of times an element occurs while a set is not. In this paper, we introduce the concept of web bag *in the context of a* web warehouse *called* WHOWEDA *(*W*are*ho*use* O*f* We*b* Da*ta) which we are currently building. Informally, a web bag is a* web table *which allows multiple occurrences of* identical web tuples*. Web bag helps to discover useful knowledge from a web table such as* visible *documents (or web sites),* luminous *documents and* luminous paths*. In this paper, we provide a cost-benefit analysis of materializing web bags as compared to web tables with distinct web tuples.*

1 Introduction

Given the high rate of growth of the volume of data available on the WWW, locating information of interest in such an anarchic setting becomes a more difficult process everyday. Thus, there is the recognition of the undeferring need for effective and efficient tools for information consumers, who must be able to easily locate information in the Web. The current approach for locating information of interest mostly depends on browsing or sending a keyword or a combination of keywords to search engines such as Alta Vista and Yahoo. These approaches of locating information have the following shortcomings. Note that these shortcomings are not meant to be exhaustive. Our intention is to highlight only those shortcomings which are addressed in this paper. Other limitations of the search engines are discussed in [5, 8].

1. From the query's result returned by search engines, a user may wish to locate *visible* Web sites [4] or documents for reference. That is, sites or documents which can be reached by many paths (high fan in). The significance of visible web documents or sites is that it enables us to identify popular web documents or sites for a given query. Visible documents for a query are those documents which can be reached by many different paths. Presently, one may only do so manually by visiting the documents in the query result, follow each links in the web documents and then download the visible documents as files on user's hard disk for future reference. Nevertheless, this method is tedious due to the large volume of results returned by search engines.
2. Reversing the concept of visibility, a user may wish to locate *luminous* Web sites [4] or documents for reference. That is, web sites or documents which have many number of outgoing links. Luminous documents or web sites define a document's or a site's exposure to other related web documents or sites. Thus, *luminosity* is a measure of a web site's or web document's connectivity to different web documents or web sites.

[*] This work was supported in part by the Nanyang Technological University, Ministry of Education (Singapore) under Academic Research Fund #4-12034-5060, #4-12034-3012, #4-12034-6022. Any opinions, findings, and recommendations in this paper are those of the authors and do not reflect the views of the funding agencies.

Currently, one may locate this information by manually visiting each Web documents.

3. Current search engines fail to measure efficiently the *inter-site connectivity* of web documents or sites. By web connectivity, we mean how richly connected is a web document or web site from/to other off-site servers. We may determine the richness by measuring the inter-site connectivity of visible or luminous documents. Inter-site connectivity helps us to determine if the visibility or luminosity of these documents are due to links from local servers or from off-site URLs. The importance of inter-site connectivity is that it enables us to quantify the popularity of a web document with respect to other sites.
4. Furthermore, a user may wish to find out the most traversed path for a particular query result. This is important since it helps the user to identify the set of most popular interlinked Web documents which traversed frequently to obtain the query result. Presently, one may only do so by visiting each documents in the search result and comparing their link information. This method is time consuming due to the quantity of results returned by search engines.

Researchers in the area of the WWW have emphasized the importance of resolving the limitations of present search engines [2, 3, 10, 12, 13, 14]. However, existing web query processing systems [11] do not address the issues raised above with respect to discovering useful knowledge from query results. If we consider the general problem of identifying visible and luminous web sites, the authors in [4, 16] have ranked various web sites based on the number of links to or from these web sites. However, they do not address the issue of determining visibility or luminosity of web document and luminous paths with respect to user's query result. This is important because one may only be interested in popular web documents relevant to his query. Thus, identification of a set of visible or luminous web sites may not be useful to him.

1.1 Overview

We have introduced *web bag* in a *web warehouse* as a part of our *Web Information Coupling Systems* (WICS) in [5]. WICS is one of the capabilities of our web warehousing system, called Whoweda[1] (*W*arehouse *O*f *We*b *Da*ta) [1, 7] which we are currently building. It is a system for managing and manipulating coupled information extracted from the Web. WICS is based on a collection of methods for organizing web information centered on the notion of *web table*. A web table is a set of *web tuples*. A web tuple is a set of inter-linked documents retrieved from the WWW which satisfies a *query graph* or *web schema*. In WICS, a user specifies a query using a query graph. A query graph is a directed connected graph consisting of *node* and *link variables* and *constraints* over some or all of these variables. Figure 1 is an example of a query graph specified by an user. We have also defined a set of web algebraic operators such as *web select*, *web project*, *web join* etc. with web semantics to manipulate web tables and correlate additional, useful, related web information residing in the web tables. For more details about WICS, the reader is referred to [8, 15].

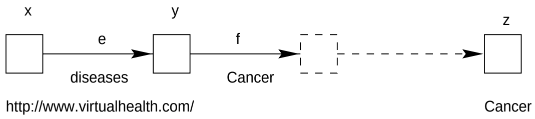

Figure 1. Web schema (query graph) of 'Cancer' web table.

Informally, a web bag is a web table containing multiple occurrences of *identical* web tuples. We are interested in the three components of web bag in the context of Web data: (1) resolving the above limitations of search-engines and existing web query systems. Specifically, how web bag can help us to discover knowledge related to query traversed path, visible documents or web sites, luminous documents or web sites; (2) analyzing the computational efficiency of different web operations with respect to web bags; and (3) performing cost-benefit analysis of materialization of web bags. We have studied (1) and (2) in [5] and [9] respectively. This paper addresses the component (3). We perform a cost-benefit analysis with respect to storage, transmission and operational cost of web bags and discussed issues and implication of materializing web bags as opposed to web tables containing distinct web tuples.

A web bag may be created by eliminating some of the nodes from web tuples of a web table using the *web project* operator. A web project operator is used to isolate data of interest, allowing subsequent queries to run over a smaller, perhaps more structured web data. Unlike its relational counterpart, a web project operator does not eliminate *identical* web tuples autonomously. Thus, the projected web table may contain identical web tuples (web bag). The duplicate elimination process is initiated explicitly by a user. The justification for not eliminating duplicate web tuples autonomously is three fold. First, existence of identical web tuples (web bag) enables us to discover useful knowledge (visible documents, luminous documents and luminous paths) from a web table [5]. Second, existence of duplicate web tuples in a web table eliminates the cost of duplicate removal from that web table. Third, the computa-

[1] Pronounced as 'hoo-eh-da'.

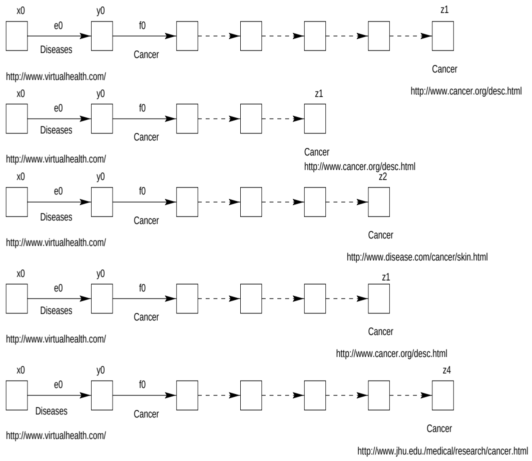
x0
y0
e0
f0
z1
Diseases
Cancer
http://www.virtualhealth.com/
Cancer
http://www.cancer.org/desc.html
x0
y0
e0
f0
z1
Diseases
Cancer
http://www.virtualhealth.com/
Cancer
http://www.cancer.org/desc.html
x0
y0
e0
f0
z2
Diseases
Cancer
http://www.virtualhealth.com/
Cancer
http://www.disease.com/cancer/skin.html
x0
y0
e0
f0
z1
Diseases
Cancer
http://www.virtualhealth.com/
Cancer
http://www.cancer.org/desc.html
x0
y0
e0
f0
z4
Diseases
Cancer
http://www.virtualhealth.com/
Cancer
http://www.jhu.edu./medical/research/cancer.html

Figure 2. Web table 'Cancer'.

tional efficiency of some of the web operation may increase if a web table contains identical tuples [9]. In this paper, we have analyzed the conditions when the materialization of web bag may be cost effective as opposed to web tables with distinct web tuples. The following example briefly illustrates the notion of web project and web bag.

Example 1 Assume that the web site at `http://www.virtualhealth.com/` integrates disease related information from various web sites. Suppose a user wish to integrate cancer related information from this web site and stores the set of related documents in a web table labeled Cancer. The user specifies the query to couple cancer-related information by providing a query-graph as shown in Figure 1. The coupling of related information is performed by the *global web coupling* operator and the result (a set of web tuples) is stored in the web table Cancer as shown in Figure 2(a). [2] The global web coupling operation retrieves or couples those portions of the Web that matches the query graph. Once the query result is materialized in Cancer, the query graph is assigned as the web schema of the web table Cancer.

Suppose the user now wish to eliminate all instances of *node variables* or nodes between y and z from each web tuple in the web table Cancer. This is performed by the web project operator and the resultant web table is shown in Figure 4. The first, second and fourth web tuples in the figure are now identical (URL and connectivity of instances of node variables in each web tuple are *identical* to that of other web tuple) and they form a web bag.

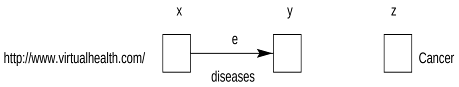

Figure 3. Web schema after projecting node variables between y and z.

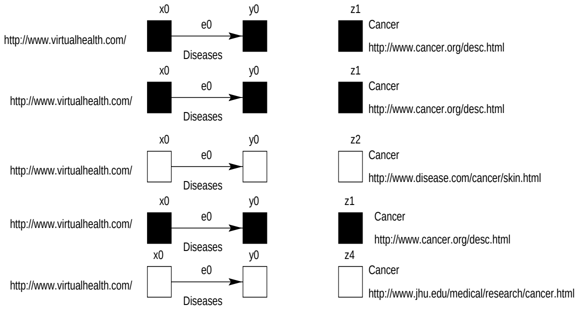
x0
y0
z1
http://www.virtualhealth.com/
e0
Diseases
Cancer
http://www.cancer.org/desc.html
x0
y0
z1
http://www.virtualhealth.com/
e0
Diseases
Cancer
http://www.cancer.org/desc.html
x0
y0
z2
http://www.virtualhealth.com/
e0
Diseases
Cancer
http://www.disease.com/cancer/skin.html
x0
y0
z1
http://www.virtualhealth.com/
e0
Diseases
Cancer
http://www.cancer.org/desc.html
x0
y0
z4
http://www.virtualhealth.com/
e0
Diseases
Cancer
http://www.jhu.edu/medical/research/cancer.html

Figure 4. Web table after projecting node variables between y and z.

The significance of the web bag indicates that document at `http://www.cancer.org/desc.html` can be reached from `http://www.virtualhealth.com/` by three different paths. A user may explicitly initiate the elimination of duplicate tuples (first and the second tuples in this case). The web table created after the removal of identical web tuples is shown in Figure 4(b). ∎

2 Preliminaries

With the enormous amount of data stored in the World Wide Web, it is increasingly important to develop powerful web warehousing and web data mining tools. The key objective of our web warehousing project, called WHOWEDA (*W*are*h*ouse *o*f *We*b *Da*ta), is to design and implement a web warehouse that materializes and manages useful information from the Web [15].

WHOWEDA is a data repository of useful, relevant web information, available for querying and analysis. As relevant information becomes available in the WWW, these information are coupled from various sources, mapped into a common web data model and integrated with existing data in WHOWEDA. In the next section, we briefly describe WICS.

2.1 Web Information Coupling System

The primary components of WICS is a web data model and an algebra for retrieving information from the Web and

[2] Note that in all the figures in this paper, the boxes and directed lines correspond to web documents and hyperlinks respectively. The dashed arrows signifies the existence of *unbound node* and/or *link variables*. Observe that some of the boxes and directed lines have keywords imposed on them. These keywords express the contents of the documents or hyperlinks.

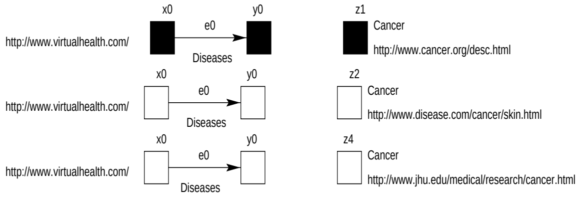
x0
e0
y0
z1
http://www.virtualhealth.com/
Diseases
Cancer
http://www.cancer.org/desc.html
x0
e0
y0
z2
http://www.virtualhealth.com/
Diseases
Cancer
http://www.disease.com/cancer/skin.html
x0
e0
y0
z4
http://www.virtualhealth.com/
Diseases
Cancer
http://www.jhu.edu/medical/research/cancer.html

Figure 5. Web table after duplicate elimination.

manipulating these information to derive additional useful information.

2.1.1 Web Data Model

Web Objects

It consists of a hierarchy of web objects. The fundamental objects are *Nodes* and *Links*. Nodes correspond to HTML or plain text documents and links correspond to hyper-links interconnecting the documents in the World Wide Web. We define a Node type and a Link type to refer to these two sets of distinct objects. These objects consist of a set of attributes such as: `Node = [url, title, format, size, date, text]` and `Link = [source-url, target-url, label, link-type]`. Note that hyperlinks in the WWW may be characterized into three types: *interior*, *local*, and *global* [14].

The next higher level of abstraction is a `web tuple`. A web tuple is a set of connected directed graphs each consisting of a set of `nodes` and `links` which are instances of `Node` and `Link` respectively. A collection of web tuples is called a `web table`. If the table is materialized, we associate a `name` with the table. There is a *schema* (see next section) associated with every web table. A `web warehouse` consists of a set of web tables.

Web Schema

A web schema contains meta-information that binds a set of web tuples in a web table. Formally, a web schema is a 4-tuple $S = \langle X_n, X_\ell, C, P\rangle$ where X_n is a set of *node variables*, X_ℓ is a set of *link variables*, C is a set of *connectivities* (in Disjunctive Normal Form), and P is a set of *predicates* (in Disjunctive Normal Form). We illustrate the concept of web schema with an example. Please refer to [15] for detailed exposition.

Let us revisit the Example 1. The query is specified as a query graph as shown in Figure 1. The query graph is evaluated using global web coupling operator [8] and a set of results in the form of web tuples is materialized in web table Cancer. The four components of the web schema of Cancer are then created from the query graph. For example, the boxes in Figure 1 with identifiers x, y and z are the node variables of the web schema. The arrows between the boxes (denoted by e, f) represent the link variables. These variables denote arbitrary instances of `Node` or `Link`. These variables are either *bound* or *unbound*. Bound node or link variables have keywords imposed on them. For example, x, y and z are bound node variables. These keywords are constraints imposed over the nodes and links variables and are defined in the form of a set of predicates of the web schema. Unbound variables are not defined by any predicates of the web schema. For example, the node variable between y and z is unbound. The connectivities between the node variables are expressed by directed arrows.

Note that the web schema of a web table can be generated by any one of the following two ways. First, if a web table is constructed by retrieving set of inter-linked documents from the Web using the global web coupling operator then the X_n, X_ℓ, C and P components of the query graph are assigned as the corresponding components of the schema of the web table. Second, if a new web table is generated from the existing web table(s) by using different web operators then the schema of the resultant web table is generated automatically by manipulating the web schema(s) of the input web table(s).

2.1.2 Web Algebra

The web algebra provides a formal foundation for data representation and manipulation for the web warehouse. The basic algebraic operators include global and local web coupling, web select, web project, web join, etc [7].

In this section, we briefly describe the web project operator. Then, we introduce the concept of *web bag*, a by-product of web project operation. A complete description of web project and web bag is given in [5].

Web Project

The web project operation on a web table extracts portions of a web tuple satisfying certain *project conditions*. These conditions are expressed as node and link variables and/or connectivities between the node variables. The web project is used to isolate data of interest in a web table, allowing subsequent web queries to execute over smaller web table, perhaps having more complete web schema.

Given a web table W with schema $S = \langle X_n, X_\ell, C, P\rangle$, a web project on W computes a new web table W_p or a web bag W_b with schema $S_p = \langle X_{n_p}, X_{\ell_p}, C_p, P_p\rangle$. The components of S_p depends on the project condition(s). Note that, unlike relational project, the web project operation does not remove duplicate web tuples automatically. The projected collection of web tuples may contain identical web tuples. In this case, it is called a web bag. Formally, we define web project as $W_b = \pi_{\langle project\_condition(s)\rangle}(W)$

where π is the web project operator. The duplicate elimination process is then initiated explicitly by the user and is performed by the following operation: $W_p =$ **Distinct(**W_b**)** where W_b is a web bag and W_p is the projected web table with distinct web tuples. Note that if a web bag is not created after a web project operation then $W_p = \pi_{\langle project\_condition(s)\rangle}(W)$. Note that in web project, we specify the node variables to be eliminated in the project conditions, as opposed to relational project, where we specify the attributes to be projected from a relation.

A user may explicitly specify any one of the conditions or any combination of the three conditions identified below to initiate a web project operation.

- **Set of node variables:** A set of node variables to eliminate from the web table.
- **Start-node variable and end-node variable:** To eliminate all the instances of node variables between two given node variables.
- **Node variable and depth of links:** This condition restricts the set of nodes to be eliminated within a limited number of links starting from the specified node variable.

Web Bag

Informally, a web bag is a web table containing multiple occurrences of *identical web tuples*[5]. Recall that a web tuple is a set of inter-linked documents retrieved from the WWW which satisfies a query graph. A web bag may be created by eliminating some of the nodes from web tuples of a web table using the web project operator. A web bag contains different collections of identical web tuples. We call each collection of such identical web tuples a *multiplet*. A web bag may have one or more multiplets. Note that a multiplet is a special type of bag in which all the web tuples are identical. For example, consider the collection of web tuples in Figure 4. The first, second and fourth web tuples (denoted by t_1, t_2 and t_4) are identical, i.e., $t_1 = t_2 = t_4$. Thus, the collection may be considered as a web bag and $\langle\{t_1, t_2, t_4\}\rangle$ forms a multiplet.

3 Reduction Ratio

In this section, we discuss some issues related to the storage of web table or web bag resulted from a web project operation. The reduction ratio is the ratio of the *size* of the web bag or projected web table (after the removal of identical web tuples) to the size of the input web table. The size of a web table depends on the number of web tuples in a web table and the number of nodes in each web tuple in a web table. We define two flavors of reduction ratios; *tuple* and *node* reduction ratios, to quantify the size of the web bag or projected web table compared to the input web table. In the next section, we will show how the reduction ratios are used to quantify the cost associated with a web bag. Note that due to space limitations we have omitted the proofs of the propositions discussed in this paper. Please refer to [6] for further details.

3.1 Tuple Reduction Ratio

Definition 1 *The* **tuple reduction ratio***, denoted as φ, is the ratio of the total number of web tuples in the web bag or projected web table to the total number of web tuples in the input web table. Formally, let W be the input web table, and W_b and W_p be the web bag and projected web table after elimination of identical web tuples respectively. Then, $\varphi_b = \frac{|W_b|}{|W|}$ and $\varphi_w = \frac{|W_p|}{|W|}$ where φ_b and φ_w denotes the tuple reduction ratio of web bag and projected web table after duplicate elimination respectively.* ■

Proposition 1 *Let M be a set of multiplets created after a web project operation on web table W, then $\varphi_b = 1$ and*

$$\varphi_w = 1 - \frac{\sum_{r=1}^{|M|} \mathsf{count}(\mathsf{M_{b_r}}) - |M|}{\mid W \mid} \tag{1}$$

where M_{b_r} is a multiplet in the web bag W_b. ■

Observation 1 The tuple reduction ratio varies from 0 to 1 and indicates the existence of identical web tuples in a web table. For a given web table and percentage of identical web tuples in the web table, the tuple reduction ratio increases with the number of multiplets in the web bag. The value of tuple reduction ratio increases (closer to 1) as the total number of identical web tuples in a web bag for a given number of multiplets decreases. However, as the size of the web table increases for a given number of multiplets and identical web tuples the tuple reduction ratios becomes almost identical. In the next section, we will see how the storage cost of projected web table increases with the increase in tuple reduction ratio. ■

Example 2 Consider the web table Cancer in Figure 2. After eliminating the node variables between y and z, the web bag and the projected web table are shown in Figures 4 and 5 respectively. Note that $\mid W \mid = 5$, $|M| = 1$ and $\sum_{r=1}^{|M|}(\mathsf{count}(\mathsf{M_{b_r}}) - |M|) = 3 - 1 = 2$. Thus, $\varphi_w = 1 - 2/5 = 0.6$. It indicates that the number of web tuples in the projected web table (Figure 5) is reduced by 40% in comparison to the original web table. ■

3.2 Node Reduction Ratio

Definition 2 *The* **node reduction ratio***, denoted as τ, is the ratio of the total number of nodes (web documents) in* *the web bag or projected web table to the total number of nodes in the input web table. Formally, let $N_w(t_i)$ be total number of nodes in the web tuple t_i where $t_i \in W$. Then,*

$$\tau_b = \frac{\sum_{j=1}^{|W_b|} N_{w_b}(t_j)}{\sum_{i=1}^{|W|} N_w(t_i)} \text{ and } \tau_w = \frac{\sum_{j=1}^{|W_p|} N_{w_p}(t_j)}{\sum_{i=1}^{|W|} N_w(t_i)}$$

where τ_b and τ_w denotes the node reduction ratio of the web bag W_b and projected web table W_p respectively. ∎

Proposition 2 *Let $N_w(t_i)$ be total number of nodes in the web tuple t_i where $t_i \in W$, and $P_w(t_i)$ be total number of nodes eliminated from the web tuple t_i after a web project operation, then*

$$\tau_w = \left(|W| - \sum_{r=1}^{|M|} \mathsf{count}(\mathsf{M}_{\mathsf{b_r}}) + |M|\right) \times \left(\frac{\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i)}{|W| \times \sum_{i=1}^{|W|} N_w(t_i)}\right) \quad (2)$$

$$\tau_b = \frac{\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i)}{\sum_{i=1}^{|W|} N_w(t_i)} \quad (3)$$

Observation 2 The node reduction ratio varies from 0 to 1. From Equation 2, we may infer that the node reduction ratio depends on the total number of eliminated nodes ($\sum_{i=1}^{|W|} P_w(t_i)$) and the tuple reduction ratio of projected web table i.e., φ_w. The total number of nodes removed from a given input web table depends only on the project conditions. Thus, for a given project condition(s) on a given web table, $\sum_{i=1}^{|W|} P_w(t_i)$ is constant. In this case, τ_w is directly proportional to φ_w; if φ_w decreases then τ_w also decreases. φ_w decrease if the value of $\left(\sum_{r=1}^{|M|} \mathsf{count}(\mathsf{M}_{\mathsf{b_r}}) - |M|\right)$ is large. This implies that τ_w decreases if the size of the multiplets in a web bag is large compared to the total number of multiplets in a web bag. ∎

Example 3 Continuing with Example 2, the node reduction ratio for the projected web table (Figure 5) is

$$\begin{aligned}\tau_w &= (5 - 3 + 1) \times \frac{\sum_{i=1}^{5}(N_w(t_i) - P_w(t_i))}{5 \times \sum_{i=1}^{5} N_w(t_i)} \\ &= 0.33\end{aligned}$$

That is, the total number of nodes is reduced by 67% after the web project operation. However, if we materialize duplicate tuples then the node reduction ratio for the web bag (Figure 5) is

$$\begin{aligned}\tau_b &= \frac{\sum_{i=1}^{5}(N_w(t_i) - P_w(t_i))}{\sum_{i=1}^{5} N_w(t_i)} \\ &= 0.56\end{aligned}$$

That is, the there is 44% reduction in the number of nodes in the web bag created due to the project operation. ∎

Symbol	Meaning
C_w	Total cost without web bags
C_b	Total cost with web bags
c_{q_w}	Query processing cost in C_w
c_{q_b}	Query processing cost in C_b
c_{t_w}	Transmission cost in C_w
c_{t_b}	Transmission cost in C_b
c_{s_w}	Storage cost in C_w
c_{s_b}	Storage cost in C_b
c_{d_w}	Duplicate elimination cost in C_w
c_{p_w}	Cost for projecting nodes in C_w
c_{p_b}	Cost for projecting nodes in C_b
X	Benefit

Table 1. Symbols used in cost-benefit analysis.

4 Cost-Benefit Analysis

In this section, we perform a cost-benefit analysis of materializing web bags. We discuss issues and implication of materializing web bags as opposed to web tables containing distinct web tuples. First, we define different types of cost which we consider for our analysis. Then, we describe the cost benefit analysis of a web bag.

4.1 Storage Cost

It is the cost of storing a web table. It is defined in terms of the number of memory blocks needed.

Proposition 3 *Let u be the number of bytes stored per memory block. Let y_n and y_ℓ be the average number of bytes needed to store a node and a link respectively. Furthermore, let c_{blk} be the cost of storing a memory block. Then, the storage cost of a web table, denoted as c_s is*

$$c_s = \frac{c_{blk}}{u}\left(y_n \tau \sum_{i=1}^{|W|} N_w(t_i) + y_\ell \left(\tau \sum_{i=1}^{|W|} N_w(t_i) - \varphi \mid W \mid\right)\right) \quad (4)$$

∎

4.2 Transmission Cost

We now discuss the cost for transmitting a web table from one server to another server over the fixed network. This cost will incur in case the warehouse is distributed over many geographically separated servers.

Proposition 4 *Let L_{max} be the maximum size of a message, c_{dist} be the cost of transmitting each distinct byte and $C_{message}$ be the associated message cost per message. Then, the transmission cost, denoted as c_t is*

$$c_t = \left[c_{dist} + \frac{C_{message}}{L_{max}}\right] \times \left(y_n \tau \sum_{i=1}^{|W|} N_w(t_i) + y_\ell \left(\tau \sum_{i=1}^{|W|} N_w(t_i) - \varphi \mid W \mid\right)\right) \quad (5)$$

∎

4.3 Operational Cost

Since, our intention is to investigate the cost effectiveness of materializing a web bag, the operational cost depends on the operation of projecting nodes from the input web table and eliminating duplicate web tuples from the web bag. It depends on the number of accesses performed on nodes and links.

Let c_{q_w} and c_{q_b} be the operational costs of distinct web table and web bag respectively. Let c_{p_w} and c_{p_b} be the cost for project operation for distinct web table and web bag respectively. Furthermore, let c_{d_w} be the cost for duplicate elimination. Then,

$$\begin{aligned} c_{q_w} &= c_{p_w} + c_{d_w} \\ c_{q_b} &= c_{p_b} \end{aligned}$$

Note that $c_{p_w} = c_{p_b}$, since the cost of project operation is always equal regardless of the existence of duplicate web tuples.

Projection Cost

We now discuss the projection cost. It is the cost incurred in eliminating the nodes from each web tuple based on the project conditions.

Proposition 5 *Let c_{acc} be the access cost for each nodes and links which are to be eliminated. Then, the projection cost, denoted as c_p, is*

$$c_p = c_{acc} \times \left(2 \sum_{i=1}^{|W|} P_w(t_i) - |W| \right) \quad (6)$$

$\blacksquare$

Duplicate Elimination Cost

We now calculate the cost of duplicate elimination in a web project operation. Our cost model is based on the worst case scenario. We compare each web tuple in a web table with the other to determine the existence of duplicate web tuples. This approach is a brute force approach and can be improved significantly using more efficient algorithm. However, our intention is to calculate the cost associated with this naive approach which we believe will be the maximum cost associated with duplicate elimination.

Proposition 6 *Let c_{acc} be the cost associated with each access of node and link information. Let d_i be the total number of duplicate tuples for web tuple t_i in the web table W. For example, in Figure 4, for the first web tuple $d_1 = 2$, since the first web tuple has two duplicates (second and the fourth web tuples). Moreover, for distinct web tuple $d_i = 0$. For example, $d_3 = 0$ for the third web tuple in Figure 4. The duplicate elimination cost, denoted by c_{d_w}, is*

$$\begin{aligned} c_{d_w} &= c_{acc} \left(\frac{2\tau_b \sum_{i=1}^{|W|} N_w(t_i)}{|W|} - 1 \right) \times \\ & \sum_{i=1}^{k} \left(\mid W_b \mid - \sum_{j=0}^{k-1} d_j - k \right) \quad (7) \end{aligned}$$

$\blacksquare$

4.4 Cost-Benefit Analysis

Let C_w and C_b be the cost associated with projected web table with distinct web tuples and web bag containing multiplets respectively. Then,

$$\begin{aligned} C_w &= c_{q_w} + c_{s_w} + c_{t_w} \\ C_b &= c_{q_b} + c_{s_b} + c_{t_b} \end{aligned}$$

Let X be the difference of cost associated with web bag and projected web table with distinct web tuples. Then,

$$\begin{aligned} X &= C_w - C_b \\ &= -(c_{q_b} - c_{q_w}) - (c_{s_b} - c_{s_w}) - (c_{t_b} - c_{t_w}) \\ &= -\Delta c_q - \Delta c_s - \Delta c_t \end{aligned}$$

where $\Delta c_q = c_{q_b} - c_{q_w}$, $\Delta c_s = c_{s_b} - c_{s_w}$ and $\Delta c_t = c_{t_b} - c_{t_w}$. Note that $c_{q_w} = c_{p_w} + c_{d_w}$ and $c_{q_b} = c_{p_b}$. Further, the cost of eliminating nodes from web tuples in the input web table is equal for both the cases, i.e., $c_{p_w} = c_{p_b}$. Thus, $\Delta c_q = -c_{d_w}$ and $X = c_{d_w} - \Delta c_s - \Delta c_t$. This implies that in order to prove that the materialization of web bag is cost effective as opposed to web table with distinct web tuples, the following inequality must hold:

$$\begin{aligned} X &> 0 \\ \Delta c_s + \Delta c_t &< c_{d_w} \quad (8) \end{aligned}$$

We now quantify Δc_s and Δc_t in the above equation.

Calculation of Δc_s

Proposition 7 *The difference between the storage cost of a web bag and projected web table with distinct web tuples, denoted by Δc_s, is expressed by the following formula*

$$\begin{aligned} \Delta c_s &= \frac{c_{blk}}{u}(1 - \varphi_w) \times \\ & \left[\left(\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i) \right) \times (y_n + y_\ell) - y_\ell \mid W \mid \right] \quad (9) \end{aligned}$$

$\blacksquare$

Observation 3 Note that in the above equation, $(y_n + y_\ell) > y_\ell$, $\left(\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i) \right) \geq |W|$ and $0 < \varphi_w \geq 1$. Thus, $\Delta c_s \geq 0$. This implies that the value of Δc_s is always positive. Thus, the storage cost of web bags is always higher than the storage cost of projected web table with distinct web tuples. Note that in order to satisfy Equation 8, the value of Δc_s should be minimized. Note that in Equation 9, $\sum_{i=1}^{|W|} P_w(t_i)$ depends only on the project condition. Thus, for a given project condition Δc_s varies with φ_w. The value of Δc_s decreases as φ_w increases. This implies that the difference between the storage cost of a web bag and projected web table decreases if the difference between the total number of identical web tuples in web bag and the total number of multiplets decreases. ■

Calculation of Δc_t

Proposition 8 *The difference between the transmission cost of a web bag and projected web table with distinct web tuples, denoted by Δc_t, is expressed by the following*

$$\begin{aligned}\Delta c_t \;=\; & \left(c_{dist} + \frac{c_{message}}{L_{max}}\right)(1-\varphi_w) \times \\ & \left[\left(\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i)\right)(y_n + y_\ell) - y_\ell \mid W \mid\right] \quad (10)\end{aligned}$$

■

Observation 4 Analogous to Δc_s, the value of Δc_t is always positive. Furthermore, the difference between the transmission cost of a web bag and projected web table increases if the number of identical web tuples in each multiplet in a web bag is large compared to the total number of multiplets in a web bag. However, for a given web table and project condition, the difference between the transmission cost is always greater than the difference between the storage cost by the factor: $u\left(c_{dist} + \frac{C_{message}}{L_{max}})/(c_{blk})\right)$. ■

4.5 Is $\Delta c_s + \Delta c_t < c_{d_w}$?

We now investigate the validity of the inequality in Equation 8. We first calculate the left hand side of the inequality ($\Delta c_s + \Delta c_t$) and then compare it with c_{d_w}.

Using Equations 9 and 10 we can express ($\Delta c_s + \Delta c_t$) as the following:

$$\begin{aligned}\Delta c_s + \Delta c_t \;=\; & \left(\frac{c_{blk}}{u} + c_{dist} + \frac{c_{message}}{L_{max}}\right)(1-\varphi_w) \times \\ & \left[\left(\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i)\right) \times (y_n + y_\ell) - y_\ell \mid W \mid\right] \quad (11)\end{aligned}$$

Next we express the left hand side of the inequality as

$$\Delta c_s + \Delta c_t \;=\; A \times y_\ell (1-\varphi_w)\left[2B - |W|\right] \quad (12)$$

where

$$A = \left(\frac{c_{blk}}{u} + c_{dist} + \frac{c_{message}}{L_{max}}\right)$$
$$B = \left(\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i)\right)$$

and $y_n \leq y_\ell$. Thus the maximum value of $\left(\frac{y_n}{y_\ell} + 1\right)$ is 2. Now consider the right hand side of the inequality and express c_{d_w} as

$$\begin{aligned}c_{d_w} \;=\; & c_{acc}\left(\frac{2\tau_b \sum_{i=1}^{|W|} N_w(t_i)}{|W|} - 1\right) \times \\ & \sum_{i=1}^{k}\left(\mid W_b \mid - \sum_{j=0}^{k-1} d_j - k\right) \\ =\; & c_{acc}\left(2 \times \frac{\left(\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i)\right)}{|W|} - 1\right) \times \\ & \sum_{i=1}^{k}\left(\mid W_b \mid - \sum_{j=0}^{k-1} d_j - k\right) \\ =\; & c_{acc}\left(\frac{2B}{|W|} - 1\right) \times \sum_{i=1}^{k}\left(\mid W_b \mid - \sum_{j=0}^{k-1} d_j - k\right) \quad (13)\end{aligned}$$

For simplicity, we approximate $\sum_{i=1}^{k}\left(\mid W_b \mid - \sum_{j=0}^{k-1} d_j - k\right)$ by $k \times |W|$. Then, Equation 13 can be simplified to the following equation:

$$\begin{aligned}c_{d_w} \;=\; & c_{acc}\left(\frac{2B}{|W|} - 1\right) k|W| \\ =\; & c_{acc}\left(2B - |W|\right) k \quad (14)\end{aligned}$$

Now consider the inequality in Equation 8. Replacing $\Delta c_s + \Delta c_t$ and c_{d_w} of Equation 8 with Equation 12 and Equation 13 respectively, we get the following inequality

$$\begin{aligned}\Delta c_s + \Delta c_t \;&<\; c_{d_w} \\ A y_\ell (1-\varphi_w)\left[2B - |W|\right] \;&<\; c_{acc}\left(2B - |W|\right) k \\ \left[2B - |W|\right]\left[A y_\ell (1-\varphi_w) - k c_{acc}\right] \;&<\; 0 \quad (15)\end{aligned}$$

Thus, materialization of web bag is cost effective if the above inequality is satisfied. The inequality in Equation 15 holds if the following conditions are true:

1. $[2B - |W|] > 0$ and $[A \times y_\ell (1-\varphi_w) - k c_{acc}] < 0$.
2. $[2B - |W|] < 0$ and $[A \times y_\ell (1-\varphi_w) - k c_{acc}] > 0$.

We now analyze each of the above conditions in detail.

Condition 1 The materialization of web bag is cost effective as opposed to web table with distinct web tuples if $[2B - |W|] > 0$ and $[A \times y_\ell (1-\varphi_w) - k c_{acc}] < 0$. This implies that $B > |W|/2$ or $\left(\sum_{i=1}^{|W|} N_w(t_i) - \sum_{i=1}^{|W|} P_w(t_i)\right) > |W|/2$ and $\varphi_w > 1 - (k \times c_{acc})/(A \times y_\ell)$. That is, if the difference between the total number of nodes in a web table and total number of eliminated nodes is greater than half the total number of web tuples, then for materialization of web bag to be cost effective, φ_w must be greater than $(1 - (k \times c_{acc})/(A \times y_\ell))$ where k is the total number of passes made on the web bag to eliminate duplicate web tuples. Since $0 \leq \varphi_w \leq 1$, $k < Ay_\ell/c_{acc}$, thus, the total number of passes made on the web bag should vary between 1 and Ay_ℓ/c_{acc}. Furthermore, if k decreases then φ_w decreases. This implies that k will reduce if the difference between the total number of identical web tuples and total number of multiplets is also decreased. Note that the number of passes k can also be reduced by optimizing the algorithm for duplicate elimination. ■

Condition 2 If $[2B - |W|] < 0$ then φ_w should be less than $(1 - (k \times c_{acc})/(A \times y_\ell))$ for materialization of web bag to be cost effective. To elaborate further, if $B < |W|/2$, i.e., the difference between the total number of nodes in a web table and total number of eliminated nodes is less than half the total number of web tuples in W, then $0 < \varphi_w < (1 - (k \times c_{acc})/(A \times y_\ell))$. ■

5 Summary & Future Work

In this paper, we have performed a cost-benefit analysis with respect to storage, transmission and operational cost of web bags and discussed issues and implication of materializing web bags as opposed to web tables containing distinct web tuples. A web bag helps to discover knowledge related to query traversed path, visible documents or web sites, luminous documents or web sites, etc. Currently, we have implemented web bag in our web warehouse. In this paper, we have provided an analytical approach for measuring the benefits associated with web bag. As part of future work, we plan to perform experimental analysis to validate the accuracy of our analytical analysis when compared to actual runs on both synthetic and real web data.

References

[1] http://www.cais.ntu.edu.sg:8000/˜whoweda/.

[2] S. Abiteboul, D. Quass, J. McHugh, J. Widom, J. Weiner. The Lorel Query Language for Semistructured Data. *Journal of Digital Libraries*, 1(1):68-88, April 1997.

[3] G. Arocena, A. Mendelzon. WebOQL: Restructuring Documents, Databases and Webs. *Proceedings of ICDE 98*, Orlando, Florida, February 1998.

[4] T. Bray. Measuring the Web. *Proceedings of the 5th International World Wide Web Conference (WWW)*, Paris, France, 1996.

[5] S. Bhowmick, S. K. Madria, W.-K. Ng, E.-P. Lim. Web Bags: Are They Useful in A Web Warehouse? *Proceedings of 5th International Conference of Foundation of Data Organization (FODO'98)*, Kobe, Japan, November 1998.

[6] S. Bhowmick, S. K. Madria, W.-K. Ng, E.-P. Lim. Cost-Benefit Analysis of Web Bag in a Web Warehouse: An Analytical Approach. *Technical Report*, CAIS-TR-98-23, Center for Advanced Information Systems, Nanyang Technological University, Singapore, 1999.

[7] S. Bhowmick, S. K. Madria, W.-K. Ng, E.-P. Lim. Web Warehousing: Design and Issues. *Proceedings of International Workshop on Data Warehousing and Data Mining (DWDM'98) (in conjunction with ER'98)*, Singapore, 1998.

[8] S. Bhowmick, W.-K. Ng, E.-P. Lim. Information Coupling in Web Databases. *Proceedings of the 17th International Conference on Conceptual Modeling (ER'98)*, Singapore, 1998.

[9] S. Bhowmick, S. K. Madria, W.-K. Ng, E.-P. Lim. Bags in A Web Warehouse. *Technical Report*, CAIS-TR-98-13, Center for Advanced Information Systems, Nanyang Technological University, Singapore, 1998. Submitted for publication.

[10] P. Buneman, S. Davidson, G. Hillebrand, D. Suciu. A query language and optimization techniques for unstructured data. *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Canada, June 1996.

[11] D. Florescu, A. Levy, A. Mendelzon. Database Techniques for the World-Wide Web: A Survey. *SIGMOD Records*, Vol 27, No. 3, Sept 1998.

[12] D. Konopnicki, O. Shmueli. W3QS: A Query System for the World Wide Web. *Proceedings of the 21st International Conference on Very Large Data Bases*, Zurich, Switzerland, 1995.

[13] L.V.S. Lakshmanan, F. Sadri., I.N. Subramanian A Declarative Language for Querying and Restructuring the Web. *Proceedings of the Sixth International Workshop on Research Issues in Data Engineering*, February, 1996.

[14] A. O. Mendelzon, G. A. Mihaila, T. Milo. Querying the World Wide Web. *Proceedings of the International Conference on Parallel and Distributed Information Systems (PDIS'96)*, Miami, Florida.

[15] W.-K. Ng, E.-P. Lim, C.-T. Huang, S. Bhowmick, F.-Q. Qin. Web Warehousing: An Algebra for Web Information. *Proceedings of IEEE International Conference on Advances in Digital Libraries (ADL'98)*, Santa Barbara, California, April 22–24, 1998.

[16] A. Woodruff, P. Aoki, E. Brewer, P. Gauthier, L. Rowe An Investigation of Documents from the WWW. *Proceedings of the 5th International World Wide Web Conference (WWW)*, Paris, France, 1996.