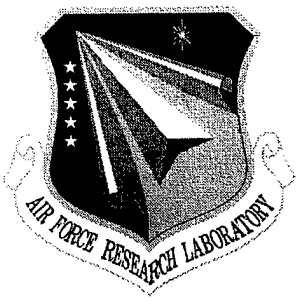
AFRL-IF-RS-TR-2000-168
Final Technical Report
December 2000



MULTI-DATABASES: REMOVAL OF REDUNDANT INFORMATION

The Research Foundation of State University of New York at Binghamton

Nicholas G. Bourbakis and Weiyi Meng

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2000-168 has been reviewed and is approved for publication.

APPROVED: Stanley E. Borek

STANLEY E. BOREK
Project Engineer

FOR THE DIRECTOR: John V. McNamara

JOHN V. MCNAMARA, Technical Advisor
Information and Intelligence Exploitation Division
Information Directorate

If your address has changed or if you wish to be removed from the Air Force Research Laboratory Rome Research Site mailing list, or if the addressee is no longer employed by your organization, please notify AFRL/IFED, 32 Brooks Road, Rome, NY 13441-4114. This will assist us in maintaining a current mailing list.

Do not return copies of this report unless contractual obligations or notices on a specific document require that it be returned.

REPORT DOCUMENTATION PAGE	Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	DECEMBER 2000	Final Apr 98 - Jun 99

4. TITLE AND SUBTITLE
MULTI-DATABASES: REMOVAL OF REDUNDANT INFORMATION

5. FUNDING NUMBERS
C - F30602-98-C-0072
PE - 63260F
PR - 3481
TA - 00
WU - P1

6. AUTHOR(S)
Nicholas G. Bourbakis and Weiyi Meng

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Research Foundation of State University of New York at Binghamton
Office of Research and Sponsored Programs
Binghamton University
P.O. Box 6000
Binghamton NY 13902-6000

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/IFED
32 Brooks Road
Rome NY 13441-4114

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2000-168

11. SUPPLEMENTARY NOTES
Air Force Research Laboratory Project Engineer: Stanley E. Borek/IFED/(315) 330-2095

12a. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This report presents the result of a research effort implemented by Dr. Bourbakis and his research group at the State University of New York (SUNY), Binghamton University. In particular, this effort included the partial development of a search engines for multimedia web documents and the complete implementation of a prototype methodology for removing (partially or totally) redundant information from multiple documents in an effort to synthesize new documents. A typical multimedia document contains free text and images and additionally has associating well-structured data. An SQL-like query language, WebSSQL, has been used to retrieve these types of documents. The main differences between Web SSQL and other proposed SQL extensions for retrieving web documents are that Web SSQL is similarity-based and supports conditions on images. This report also describes a software methodology for the detection and removal of redundant information (text paragraphs and images) from multiple retrieved documents. Documents reporting the same or related events and stories may contain substantial redundant information. The removal of the redundant information and the synthesis of these documents into a single document can not only save a user's time to acquire the information , but also storage space to archive the data. The methodology reported here consists of techniques for analyzing text paragraphs and images as well as a set of similarity criteria used to detect redundant paragraphs and images. The methodology developed in this project has the ability either to work independently with text paragraphs and images, or to combine both in one synthetic document.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Search Engine, Multimedia Documents, Redundant Information Removal, Synthesized Documents, Similar Text Documents, Similar Images, Image Differences, Synthetic DOCS, SQL-Like Query Language	36
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

TABLE OF CONTENTS

List of Figures

List of Tables

1.0 Introduction

The World Wide Web is a vast information resource and is being used by millions of people daily. To help users find information on the Web, many search engines have been created and some of the best known ones are HotBot, Alta Vista, InfoSeek, Lycos, and Excite. In addition, numerous small-scale search engines are operational at the web sites of many organizations.

Most existing search engines are strictly keywords based. A typical user query submitted to such a search engine consists of one or more words with possibly some Boolean operators like "and" or "or". The search result is a list of web pages (or their URLs) in descending similarities (ranking scores) with the query. A common problem with these search engines is that often too many web pages are retrieved for each query and most of them are not relevant to the query. A main cause for this is that user queries are often too general and are not sufficiently precise. One possible solution to this problem is to enrich the queries so that user desired information can be expressed more precisely.

A careful examination of web pages reveals that in addition to words that appear in each web page, there are also other related information that could be used to describe users' search needs more precisely. Such information includes (1) well defined (structured) information about each web page such as its URL and title; (2) metadata associated with each web page such as its size and the time it was last modified; (3) images in a web page; and (4) the links that connect different web pages and images.

In the first part of this report, we describe a new type of search engine that is capable of utilizing the above information for more precise retrieval of multimedia web pages. This search engine has three main components. The first component is an indexer whose main functionality is to prepare data associated with web pages to facilitate efficient and effective retrieval. The indexer has a built-in robot that fetches appropriate web pages from the Web. Next, pre-determined structured data and metadata associated with each web page and image are extracted and stored in tables in a relational database. Text web pages are indexed based on the terms in them as in a traditional information retrieval system. Images are indexed based on their contents (color distributions, textures and content descriptions) as in a typical image database. The second component is a web interface to be used by users to specify queries. A query language WebSSQL that extends SQL to support similarity-based retrieval of multimedia web pages is proposed. The third component is a query processor for evaluating user queries. Since three types of data, namely structured data stored in a relational database, text data and image data, are involved, query processing can be complex, as to be discussed later.

Recently, there has been active research in developing data models and query languages for structured and semi-structured documents (e.g., HTML pages) [1, 6, 7]. These models and query languages explore internal structures of documents and are mainly interested in finding rather specific information within documents. In contrast, we are primarily interested in finding

information at the higher level, namely, our search engine, like other search engines on the Web, is designed to find desired documents and images. However, our search engine supports the specification of more precise conditions than regular search engines. W3QL [3] is an SQL-like query language designed for W3QS. Its emphasis is more on utilizing Unix utilities and supporting dynamic view maintenance. No database tables are utilized by W3QL. The work that is closest to our work is WebSQL [4] which also uses database tables to support the specification of queries. In fact, the information in the tables used in WebSQL is nearly identical to that in two of our three database tables. The main difference between WebSQL and WebSSQL is that the latter supports similarity-based retrieval while the former does not. This is also a major difference between W3QL and WebSSQL. As a result, our search engine can rank retrieved documents (or images) based on how well they match the user query while WebSQL and W3QL cannot. Other differences between WebSQL and WebSSQL are (1) WebSQL uses only virtual tables while our tables are real tables populated by our indexer; (2) WebSQL distinguishes local and remote documents and uses the distinction to improve query evaluation while we have only local documents; (3) WebSSQL treats images as an important type of data while WebSQL does not. WebSQL has a formal query semantics. Providing a formal semantics for WebSSQL queries is research we are currently undertaking.

Document processing also is an important research area, where several techniques have been developed for separating text-paragraphs from images and drawings [16-22]. However, the reconstruction of a new document using a number of different documents on the same subject is still an open and challenging problem that requires further study. Thus, the second part of this report contributes to the effort of reconstructing new documents from a group of old ones by removing the existing redundant information. In particular, we present a methodology that removes redundant information (images, text paragraphs) from retrieved multimedia documents. Each document consists of two main parts stored in different DBs. The first part of a document represents text paragraphs, the second part consists of the images and drawings related with the text paragraphs. The information reduction methodology examines first the text paragraphs of each document related with a specific topic, and removes the redundant information, such as same or similar paragraphs, by keeping pointers useful for future reconstruction of the original documents. The remaining text paragraphs and the set of pointers are used to compose the first version of a new document. The methodology also examines all the images related with the set of original documents and removes the same or similar images while keeping pointers that could assist a future reconstruction of the original documents. At this point, the methodology merges text-paragraphs and images and creates the first synthesis of the new document.

The rest of the report is organized as follows. In Section 2, we describe how data is prepared for our search engine. In Section 3, we present WebSSQL. Both the syntax and semantics of WebSSQL will be informally discussed. In section 4, we present an approach for evaluating WebSSQL queries. Section 5 presents the IRR scheme. Section 6 describes the redundancy removal scheme and the report's conclusion and discussion are presented in Section 7.

2.0 Data Preparation

The indexer component of the search engine first fetches a collection of web pages and images from the Web. For each web page fetched, the following information is obtained (most are provided by the HTTP protocol):

- file_id: the id number of the web page; it is generated by the indexer.
- url: the url of the web page.
- title: the title of the web page.
- size: the size of the web page in bytes.
- type: the type of the web page (html, text, ...).
- fetch_time: the date and time when the web page is fetched.
- last_modified: the date and time when the web page was last modified.

The information obtained for all web pages is stored in a table, **Webpages**, in a relational database. Each information item, such as file_id, becomes an attribute of the table. Each row in the table corresponds to one web page.

For each image fetched, the following information is obtained:

- image_id: the id number of the image; it is generated by the indexer.
- url: the url of the image.
- title: the title or the caption of the image.
- size: the size of the image in bytes.
- type: the type of the image (jpg, gif, ...).
- fetch_time: the date and time when the image is fetched.
- last_modified: the date and time when the image was last modified.
- color: the distribution of the main color(s) of the image. An example of the color distribution of an image is (red, 40; blue, 30; green, 30), that is, there are 40% of red and 30% of each of blue and green. The color distribution can be obtained by standard image processing techniques.
- description: the description of the contents of the image. It is typically a short paragraph. There are a number of ways to obtain the description for an image. Some images already have associated descriptions. We can also apply sophisticated image understanding techniques to achieve this. Since the current emphasis of the project is on the query language and related processing techniques, the description is provided manually in our current implementation.

Again, we want to store the above information in a table, **Images**, in a relational database with one row in the table corresponding to one image. Since each image may have multiple colors, the color information should be stored in a separate table to be fully utilized when only first normal form (1NF) tables are allowed. Our current implementation simplified this by keeping only the main colors as a single character string and omitting the percentages. For example, with the simplification, (red, 40; blue, 30; green, 30) becomes "red blue green". Consequently, only the "like" operator is supported in our current WebSSQL. For instance, condition "color

like '%red%'" is used to find images that have red as one of the main colors, where the percentage sign % is a wild card matching zero or more characters in SQL.
Another table, **ChildURLs**, that keeps track of the link relationships between fetched web pages and images is also created and stored in the database. Each row in this table has two URLs (url, child_url), indicating that there is a link from the web page identified by "url" to the web page or image identified by "child_url". Typically, when child_url is the URL of an image, the image will be displayed in the web page identified by the corresponding url in the same row.

The fetched web pages are indexed based on the terms in them as in traditional information retrieval [9]. Conceptually, each web page will be represented as a vector of weights (w_1, w_2, ..., w_k), where w_i is the weight or significance of the ith term in representing the contents of the web page. Usually, non-content words such as "the", "of", etc. are excluded from consideration (i.e., they are not considered as "terms"). The weight of a term usually depends on the number of occurrences of the term in the web page (relative to the total number of occurrences of all terms in the page) [8, 10]. It may also depend on the number of pages having the term relative to the total number of pages that are fetched. In order to facilitate efficient query processing, an inverted file index is typically created. For a given term, such an index can be used to find the weights of those documents containing the term quickly.

3.0 WebSSQL

In this section, we present an SQL-like query language called **WebSSQL** (Web and Similarity based SQL). WebSSQL has a basic four-clause structure. The *select clause* indicates what are to be retrieved. Based on our objective, either web pages or images can be selected. The *from clause* lists the database tables that are to be involved in the query. As a requirement, if web pages are to be searched, then the table Webpages must be in the from clause. Similarly, if images are to be retrieved, then the table Images must be in the from clause. The main novel feature of WebSSQL is its *where clause* which specifies the conditions to be satisified (in terms of similarity) by returned web pages (or images). More detail about this clause will be provided shortly. The fourth clause indicates the maximum number of results (web pages or images) that are to be returned. For a large search engine, a large number web pages may satisfy a query to some extent (i.e., has a positive similarity). The user may choose to retrieve only the desired number of top ranked web pages (or images). A default number for results can be used.

The where clause of WebSSQL may contain up to three types of conditions as explained below:
1. Conditions on structured data. This includes conditions on all attributes in tables Webpages and ChildURLs, and all attributes in table Images except the description attribute. In addition, for the color attribute of Images, the operator is restricted to only "like" - the SQL substring matching operator.
2. Condition on text. The only text operator that is supported in the current WebSSQL is the **similar_to** operator. This operator computes the similarity between "a text query" and each web page using a similarity function to be described shortly. Boolean conditions can be supported easily. Note that most search engines on the Web support condition on text only.

3. Condition on the description attribute of Images. Again, similar_to is the only operator allowed in our current implementation. In this case, this operator computes the similarity between "a user description of an image" and the description of each image.

For each web page and image, the evaluation of each type of conditions will result in a similarity between 0 and 1. In particular, for a given web page, the evaluation of the conditions on structured data will result in a similarity of 0 if the web page does not satisfy the conditions, or 1 if the web page satisfies the conditions; the evaluation of any condition on text will also result in a similarity between 0 and 1. The same is true for each image, except that the condition on text will be replaced by condition on the description attribute.

The similarity function for both text and the description attribute is the *normalized dot product function* (or **Cosine function** as known in the information retrieval community [8, 10]). Let $p = (w_1, w_2, ..., w_n)$ be the vector representation of a web page (or the description of an image), where w_i is the weight of term t_i and n is the number of distinct terms in all web pages (or image descriptions). Let $q = (q_1, q_2, ..., q_n)$ be the vector representation of a query against the web page text (or an image description), where q_i is the weight (typically the frequency) of term t_i in the query. Then the similarity between p and q is defined to be:

$$sim(p, q) = \frac{\sum_{i=1}^{n} w_i * q_i}{|p| * |q|}$$

where |p| and |q| represent the *norms* of the vectors p and q, respectively. The norm of a vector $(x_1, x_2, ..., x_n)$ is the square root of $(x_1^2 + ... + x_n^2)$.

A query for retrieving images may contain conditions on structured data (i.e., all fields in Images except the description field) and a condition on the description field. For a given image i and a given query q against i, if the similarity due to evaluating the conditions on structured data is s_1 and the similarity due to evaluating the condition on the description field is s_2, then the final similarity between i and q is defined to be:

$$sim(i, q) = \min\{s_1, s_2\}.$$

Note that using the minimum function to combine the similarities is only one of several possibilities. Other possibilities will be explored in the near future. If one type of condition is absent in a query, the corresponding similarity of any image with respect to this type of condition can be considered to be 1. In the above discussion, if there was no condition on structured data, then $s_1 = 1$ would be used. This is consistent with intuition.

A query for retrieving web pages may contain up to three types of conditions as mentioned earlier. The similarity between a web page p and a query q is determined as follows. Suppose s_1 is the similarity due to evaluating the conditions on structured data (the similarity will be either

1 or 0 depending on whether or not the conditions are satisfied) and s_2 is the similarity due to the condition on the text. Notice that a web page may contain zero or more images. Suppose the images contained in p are $i_1, ..., i_k$. Suppose further that the similarities between the image i_j and (the related portion of) the query is:

$$sim(i_j, q) = ss_j, j = 1, ..., k.$$

Let

$$s_3 = \max\{ss_1, ..., ss_k\}.$$

That is, s_3 is the similarity between the query (the portion concerning the image data) and the best matching image in web page p. Finally, the final similarity between the web page p and the query q is defined to be:

$$sim(p, q) = \min\{s_1, s_2, s_3\}.$$

The web pages (their links) in the output will be listed in descending order of similarities with respect to the query and subject to the number of desired results. Only web pages with positive similarities will be displayed. Again, if any one of the three types of conditions is absent in a query, then the corresponding similarity (i.e., s_1, s_2 or s_3) will be considered to be 1.

We use the following example to illustrate the syntax and semantics of WebSSQL queries.

Example 1: *Find web pages that were fetched after August 13 of 1998 from the "usaf.romelab.mil" domain, whose text is similar to "air fighters used in the Gulf War" and it contains an image whose description is similar to "flying helicopter". The WebSSQL query may look like:*

```
select Webpages.url
from Webpages, ChildURLs, Images
where Webpages.fetch_time > 'Aug-13-1998' and Webpages.url like '%usaf.romelab.mil%'
 and Webpages.url = ChildURLs.url and ChildURLs.childurl = Images.url
 and text similar_to 'air fighters used in the Gulf War'
 and Images.description similar_to 'flying helicopter'
```

Condition "Webpages.url = ChildURLs.url and ChildURLs.childurl = Images.url" is used to ensure that only images that are directly referenced by the web page are considered.

Suppose five web pages ($p_1, ..., p_5$) satisfy the conditions on fetch_time and url. Therefore, for the five pages, their $s_1 = 1$. Suppose for p_1, the similarity between its text and "air fighters used in the Gulf War" is 0. This means that p_1 will not be retrieved as its final similarity will be 0 according to our formula. Suppose for the remaining four web pages, the similarities between their texts and "air fighters used in the Gulf War" are 0.4, 0.6, 0.3 and 0.7, respectively. Suppose p_2 does not contain any image and p_3 contains one image but the similarity between its description and "flying helicopter" is 0. Consequently, p_2 and p_3 will eventually have zero similarity with the query and will not be output. Suppose p_4 has one image and the similarity

between its description and "flying helicopter" is 0.5. Suppose p_5 has two images i_1 and i_2; the similarity between the description of i_1 and "flying helicopter" is 0.5, and the similarity between the description of i_2 and "flying helicopter" is 0.6. Based on our discussion, only i_2 in p_5 will be used. The final similarity between p_4 and the query is min{1, 0.3, 0.5} = 0.3. The final similarity between p_5 and the query is min{1, 0.7, 0.6} = 0.6. Therefore, only p_4 and p_5 will be retrieved with p_5 being displayed ahead of p_4.

4.0 Query Processing

Each user query will be processed in three steps. First, the query will be decomposed into a number of subqueries. Second, the subqueries will be processed in a certain efficient way. Finally, the results of the subqueries will be assembled to produce the final result. These three steps are described below.

4.1 Query Decomposition

When a user query is received, it will first be decomposed by a query decomposition algorithm. Consider a query q for retrieving web pages (for queries for retrieving images, the discussion will be similar and simpler). In general, q could be decomposed into up to three subqueries, q_1, q_2, and q_3 depending on the number of different types of conditions involved. Subquery q_1 will be a standard SQL query against the tables in the relational database (i.e., Webpages, ChildURLs and Images, except the description field of Images). This subquery will be generated only when some data stored in the three database tables are referenced in q. Subquery q_2 will be a query against the text documents and will be generated only when the "text similar_to" condition appears in the where clause of q. Finally, subquery q_3 will be a query against the description field of the Images table and will be generated only when the "Images.description similar_to" condition is present in the where clause of q. In addition, if "text similar_to" is in q, then Webpages.file_id will be added to the select clause of q_1. This is to enable the matching of web page records in Webpages and their corresponding text indexes. If "Images.description similar_to" is present in q, then Images.image_id and Images.description will be added to the select clause of q_1. The former is used to determine which images appear in which web pages and the latter is to retrieve the descriptions of appropriate images to feed into the subquery q_3.

4.2 Subquery Processing

We consider only the case when a user query q has been decomposed into three subqueries q_1, q_2 and q_3 (see above) as other cases are simpler. Subqueries q_1 and q_2 can be processed in any order or in parallel. However, based on the above decomposition algorithm, q_3 should be evaluated after q_1 is processed because q_3 needs the data (i.e., Images.description) from the result of q_1.

The processing of q1 is done by the employed relational database system. After q_1 is processed, a set of quadruplets (Webpages.url, Webpages.file_id, Images.image_id, Images.description) will be obtained. For each quadruplet, a triplet (Webpages.url, Webpages.file_id, Images.image_id) will be extracted and the set of triplets will be used by the next step for

assembling the final result. The fourth component of each quadruplet (i.e., Images.description) will be extracted and be used by subquery q_3.

Subquery q2 is a query against a collection of web pages. This is exactly the kind of query seen in most search engines and in traditional document retrieval systems [8]. The standard approach for evaluating such a query is to use the inverted file index for the document collection [8, 10]. The result of evaluating q_2 is a set of pairs (Webpages.file_id, similarity), where similarity > 0 is the similarity of the web page identified by the file id with the query (the description following "text similar_to"). This set of pairs will be used by the next step for assembling the final result.

As mentioned earlier, subquery q_3 will be processed after subquery q_1 has been processed. A benefit of this order is that only image descriptions whose corresponding image ids are returned by q_1 need to be used to process q_3. Note that there is no inverted file index for the descriptions of images. As a result, the similarities between the image descriptions returned by q_1 and the image description in the query have to be computed one by one. The evaluation of q_3 produces a set of pairs (Images.image_id, similarity), where similarity > 0 is the similarity of the description of the image identified by the image id with the image description in the user query. Again, this set of pairs will be used by the next step for assembling the final result.

4.3 Result Assembling

As described above, the result of evaluating q_1 is a set of triplets (Webpages.url, Webpages.file_id, Images.image_id), the result of evaluating q_2 is a set of pairs (Webpages.file_id, similarity), and the result of evaluating q_3 is a set of pairs (Images. image_id, similarity). We now discuss how to generate the final result to the user query from these triplets and pairs. The result assembling is accomplished by the following algorithm:

1. Sort the triplet file based on the url field. Sort the two pair files based on the file_id and the image_id fields, respectively.
2. For each triplet, say (url_1, w_id_1, i_id_1), do

- Use w_id_1 to find a pair in (Webpages.file_id, similarity) such that w_id_1 = Webpages.file_id. Note that at most one such pair can be found as file id is unique for web pages. It is possible that no such pair can be found. This corresponds to the case where the web page with the file id has a similarity of zero with the text query in q_2. In this case, discard the triplet and continue with the next triplet. With loss of generality, suppose there is one matching pair and it is (w_id_1, sim_{11}).
- Use i_id_1 to find a pair in (Images.image_id, similarity) such that i_id_1 = Images.image_id. Again, at most one such pair can be found. If no such pair is found, discard the triplet and continue with the next triplet. Again, suppose there is one matching pair and it is (i_id_1, sim_{12}).
- Let $sim_1 = \min\{sim_{11}, sim_{12}\}$. Note that sim_1 is the combined similarity of the web page based on the image identified by i_id_1. Return pair (url_1, sim_1).

3. If a url appears in several returned pairs (url, sim), keep the pair with the largest sim and discard other pairs for the url. Note that several pairs with the same url may be returned from

step 2 if the corresponding web page contains multiple images that satisfy the relevant conditions in the query.

4. Display the remaining urls in descending sim values.

We illustrate the result assembling algorithm using the following example.

Example 2: *Suppose Table 1 contains the triplets from evaluating q_1. Note that the web page with url_3 contains 2 images with ids 33 and 44. Suppose Tables 2 and 3 contain the pairs from evaluating q_2 and q_3, respectively.*

Table 1: Result from q_1

Webpages.url	Webpages.file_id	Images.image_id
Url1	123	11
Url2	234	22
Url3	345	33
Url3	345	44

Table 2: Result from q_2

Webpages.file_id	Similarity
123	0.6
234	0.2
345	0.5
456	0.7
567	0.2

Table 3: Result from q_3

Images.image_id	Similarity
11	0.3
22	0.6
33	0.7
44	0.4

By following the four steps of the algorithm, we have:

1. *All tables have been sorted.*
2. *For (url_1, 123, 11), we obtain (123, 0.6) and (11, 0.3). The result is (url_1, 0.3).*
   For (url_2, 234, 22), we obtain (234, 0.2) and (22, 0.6). The result is (url_2, 0.2).
   For (url_3, 345, 33), we obtain (345, 0.5) and (33, 0.7). The result is (url_3, 0.5).
   For (url_3, 345, 44), we obtain (345, 0.5) and (44, 0.4). The result is (url_3, 0.4).
3. *url_3 has two similarities, 0.5 and 0.4. Keep (url_3, 0.5) and discard (url_3, 0.4).*
4. *Final output:*

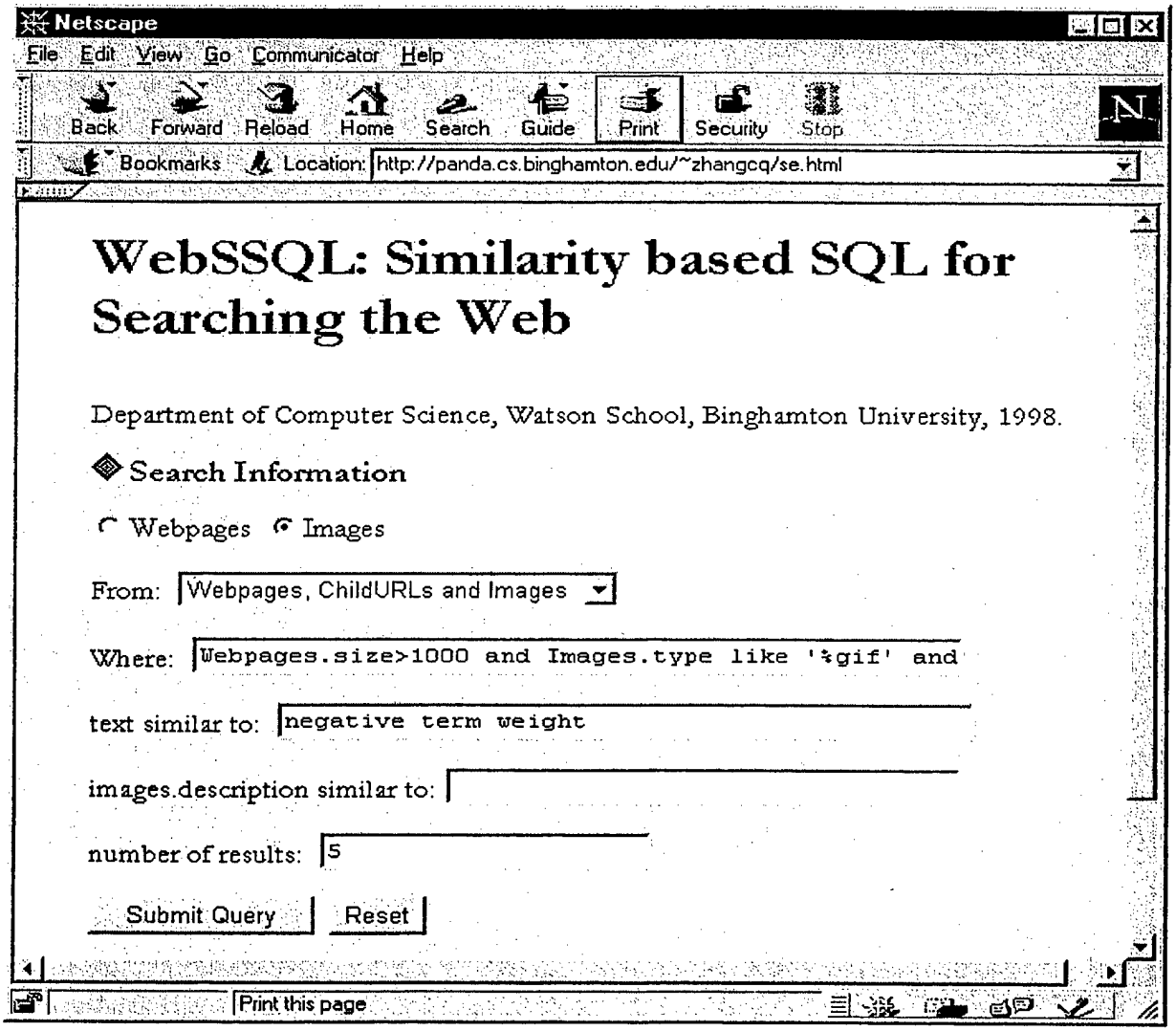
url_3	0.5
url_1	0.3
url_2	0.2

Based on the design principles described in the previous sections, an operational search engine supporting WebSSQL has been implemented:

`http://panda.cs.binghamton.edu/~zhangcq/se.html`

A Sybase relational database system is used to store the three tables Webpages, Images, and ChildURLs. The Web interface of this search engine is shown in Figure 1.

Figure 1: Web Interface of WebSSQL



5.0 The Information Redundancy Removal Scheme

The architectural configuration of the methodology presented here is shown in Figure 2. The original documents retrieved by the search engine are stored into the user's workstation [23], where the Information Redundancy Removal (IRR) software scheme processes the input pieces of text and image information to create the new document.

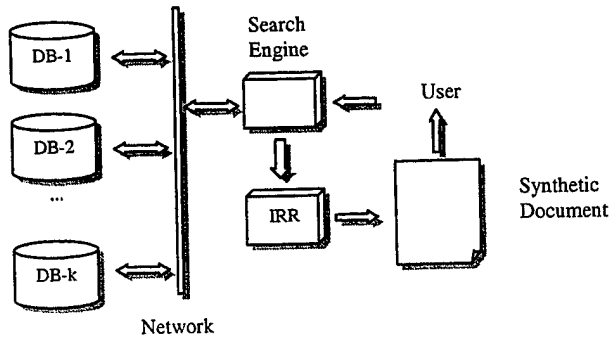


Figure 2.

6.0 Removal of Redundant Information

In this section we present the removal of information redundancies and the synthesis of pieces of information used in this report. In particular, the information retrieved from different resources will be stored temporarily in the user's workstation. This information is composed by text, images and data. Each piece (text, image, data) of this information is stored into a different memory space in order to be efficiently and independently processed. The process used here includes two major parts:

(1) removal of the existing redundancies in text and images
(2) first stage document synthesis

6.1 Text Redundancy Removal

With redundancy in text we mean the duplication of certain large parts of a text paragraph, or the duplication of an entire paragraph In this case, all the text pieces are organized into paragraphs (P), sentences (S) without loosing their referenced pointers to other items (images, data). Then, each sentence, or paragraph is analyzed and compared with the other sentences and paragraphs from different documents in order for a possible redundancy to be discovered.

Discovery of text redundancy

Each text paragraph will be analyzed by the IRR methodology and important statistical features (f) will be extracted. These features are:

- The size of the paragraph (Ps) in text characters;
- Character histogram, i.e. the number of A's, B's, C'c, ..., etc.
- The number of sentences (Sn)
- The number of words in a sentence (Sw)
- Histogram of words
- The starting word (Ws) of each sentence in a paragraph;
- The ending (or stop) word (We) of the paragraph

If two paragraphs P1 and P2 have the same features described above, then P1 and P2 are considered as similar with a probability p(f) of removal. This means that one of these two paragraphs has to be removed as redundancy under the condition that both have the same reference pointers (or ids) to other items, such as images, data, or tables. If the pointers are different then a more detailed analysis takes place on the examined paragraphs and the removal operation is postponed until an analytical examination will take place at the corresponding images and data parts. In addition, if the paragraphs have been placed in a different order in a text-paragraph, two additional features (1 - the starting word of a new sentence (W2) and 2 - the length of each sentence (SL)) will be used for a more accurate matching of two paragraphs.

The algorithmic steps used of matching two text-paragraphs are described in Figure 3.

At this point we present an example illustrating the operation of the IRR scheme by using five synthetic documents (a.doc, b.doc, c.doc, d.doc, e.doc) with same, similar and different text-paragraphs and images. We also consider that these document have been preprocessed in a way that the text-paragraphs and images have been stored in different memory locations. In particular, the text-paragraphs form five text-based documents (a.txt, b.txt, c.txt, d.txt, e.txt) and the images are saved in three different files (a.jpg, b.jpg, c.jpg). Thus, the IRR process starts with the text based documents below:

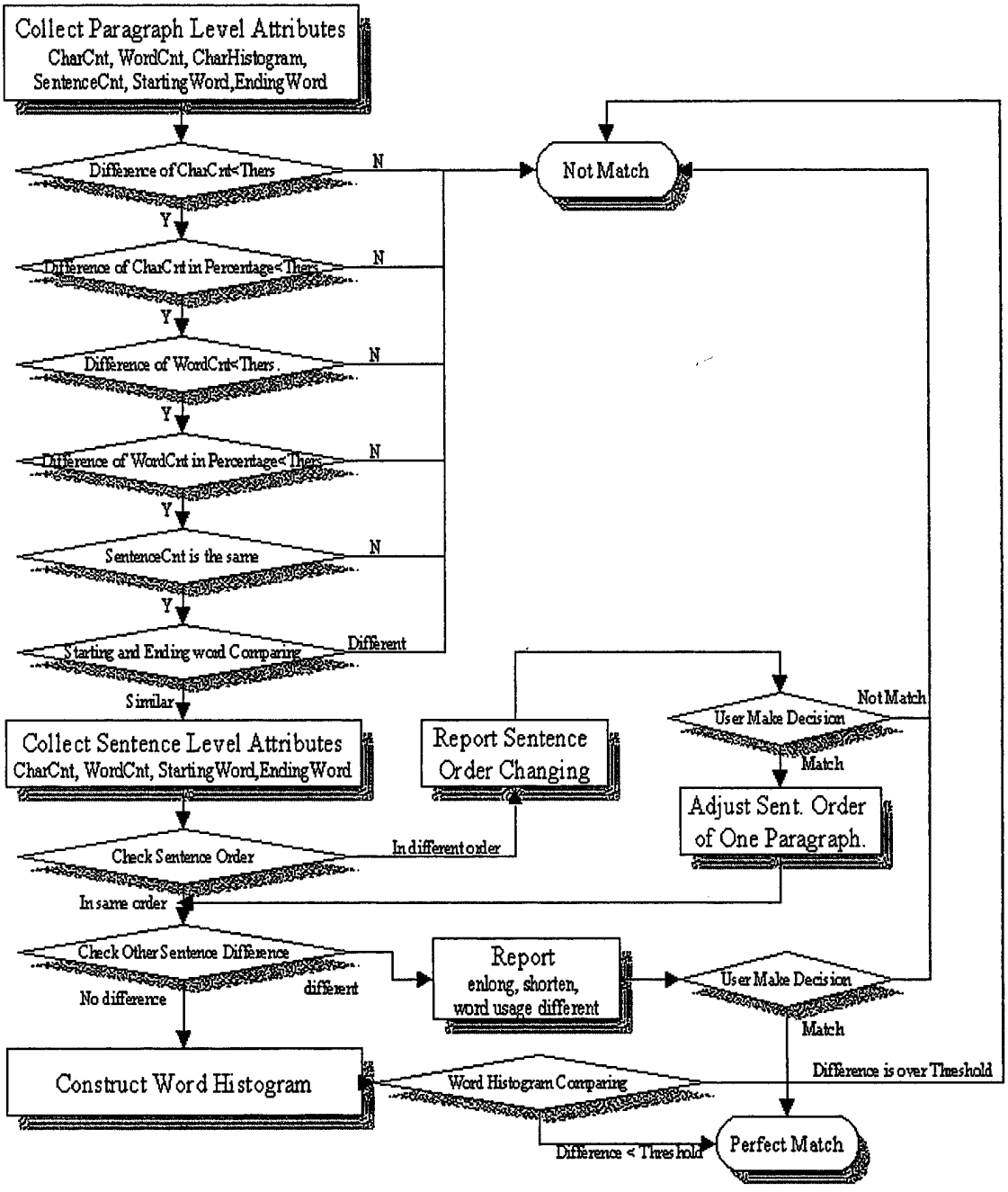

Collect Paragraph Level Attributes
CharCnt, WordCnt, CharHistogram,
SentenceCnt, StartingWord,EndingWord
Difference of CharCnt<Thers
N
Not Match
Y
Difference of CharCnt in Percentage<Thers
N
Y
Difference of WordCnt<Thers
N
Y
Difference of WordCnt in Percentage<Thers
N
Y
SentenceCnt is the same
N
Y
Starting and Ending word Comparing
Different
Similar
Collect Sentence Level Attributes
CharCnt, WordCnt, StartingWord,EndingWord
Report Sentence
Order Changing
User Make Decision
Not Match
Match
Adjust Sent. Order
of One Paragraph.
Check Sentence Order
In different order
In same order
Check Other Sentence Difference
different
No difference
Report
enlong, shorten,
word usage different
User Make Decision
Match
Construct Word Histogram
Word Histogram Comparing
Difference is over Threshold
Difference < Threshold
Perfect Match

Figure 3.

Illustrative Example:

a.txt

Binghamton University, fig.1 embarks on its inaugural Division II year at home as it renews old rivalries with SUNY Albany and begins new ones with members of the New England Collegiate Conference (NECC) on Saturday, September 12.

Five teams will be in action at home for the NCAA Division II kick-off celebration. Activities begin at 10 a.m. with women's volleyball vs. UMass Lowell inside the West Gym, and youth clinics in baseball, tennis and soccer on the fields and courts near the West Gym.

President Lois B. DeFleur will offer a welcome and preside over a cake cutting at 12:30 p.m., adjacent to the soccer field as Binghamton makes its official entrance at home into the NECC--considered one of the nation's top Division II conferences.

The "America's Best Colleges" issue and guidebook will be available on newsstands Monday, August 24. U.S. News will release its "Best College Values" rankings in an upcoming issue of the magazine, available on newsstands Monday, August 31.

The Class of 2002 continues to reflect the diversity of the state. Students of Asian background comprise 18.2 percent of the class, followed by Hispanic/Latino and African-American at 6.5 percent each and Native American at .1 percent. There are 36 international students enrolled as entering freshmen.

b.txt

Approximately 2,037 freshmen and 816 transfer students arrive on campus Thursday, August 27, toting boxes and suitcases full of belongings. The freshman class is about 200 students larger than last year's and represents a return to the University's plan for growth that was put on hold during years of budget cuts.

Binghamton University, fig.1 embarks on its inaugural Division II year at home as it renews old rivalries with SUNY Albany and begins new ones with members of the New England Collegiate Conference (NECC) on Saturday, September 12.

With an average combined SAT score of 1206, about 200 points above the national average, the new arrivals maintain Binghamton's reputation as a selective school. Of those from high schools that rank their graduates, 21 percent come from the top five percent, 49 percent from the top 10 percent and 86 percent from the top fifth of their graduating class. The mean high school average for the freshman class is 92.1.

The "USA's Best Colleges" issue and guidebook will be available on newsstands Monday, Aug 24. U.S. News will release its "Best College Values" rankings in an upcoming issue of the magazine, available on newsstands Monday, August 31.

A geographic profile of the incoming class shows the vast majority of freshman--94.8 percent--are New York state residents. Nearly six percent are from Broome and Tioga counties, with 62.5 percent from the New York City area. In all, the first-year class includes students from 55 of New York's 62 counties.

The Class of 2002 continues to reflect the diversity of the state. Students of Asian background comprise 18.2 percent of the class, followed by Hispanic/Latino and African-American at 6.5 percent each and Native American at .1 percent. There are 36 international students enrolled as entering freshmen.

Approximately 2,037 freshmen and 816 transfer students arrive on campus Thursday, August 27, toting boxes and suitcases full of belongings. The freshman class is about 200 students larger than last year's and represents a return to the University's plan for growth that was put on hold during years of budget cuts.

U.S. News calculated scores based on academic reputation, retention rates, faculty resources, student selectivity, financial resources, graduation rates and alumni giving rates to compile its rankings. According to the magazine the rankings are reliable, objective and fair with "each school's rank based on the same set of quality measures".

Binghamton University embarks on its inaugural Division II year at home as it renews old rivalries with SUNY Albany and begins new ones with members of the New England Collegiate Conference (NECC) on Saturday, September 12.

The mean high school average for the freshman class is 92.1. With an average combined SAT score of 1206, about 200 points above the national average, the new arrivals maintain Binghamton's reputation as a selective school. Of those from high schools that rank their graduates, 21 percent come from the top five percent, 49 percent from the top 10 percent and 86 percent from the top fifth of their graduating class.

A geographic profile of the incoming class shows the vast majority of freshman--94.8 percent--are New York state residents. Nearly six percent are from Broome and Tioga counties, with 62.5 percent from the New York City area. In all, the first-year class includes students from 55 of New York's 62 counties.

The Class of 2002 continues to reflect the diversity of the state. Students of Asian background comprise 18.2 percent of the class, followed by Hispanic/Latino and African-American at 6.5 percent each and Native American at .1 percent. There are 36 international students enrolled as entering freshmen.

Binghamton University, fig.1 is again among the elite Top 25 public universities in the nation according to U.S. News and World Report. The magazine's 12th annual "America's Best Colleges" issue and guidebook lists Binghamton 21st in its list of top public universities.

U.S. News calculated scores based on academic reputation, retention rates, faculty resources, student selectivity, financial resources, graduation rates and alumni giving rates to compile its rankings. According to the magazine the rankings are reliable, objective and fair with "each school's rank based on the same set of quality measures".

U.S. News notes the rankings are also helpful to those choosing a college for several reasons: they are based on accepted measures of academic quality chosen based on U.S. News' experience in reporting on education and on research about measuring educational outcomes, after consultation with experts; they have been developed independently of any particular institution; they are comparable and complete; they are the single best source of information because they allow readers to compare the strengths and weaknesses of different schools; and they condense a great deal of information, making it easier to compare institutions.

"Binghamton University has been recognized consistently by national publications for its quality and value," said President Lois B. DeFleur. "Appearing in the U.S. News ranking of the Top 25 public universities for two years in a row is another highly visible validation of the efforts our

faculty and staff make in educating our students and of our success in doing so".

The "America's Best Colleges" issue and guidebook will be available on newsstands Monday, August 24. U.S. News will release its "Best College Values" rankings in an upcoming issue of the magazine, available on newsstands Monday, August 31.

d.txt

The "America's Best Colleges" issue and guidebook will be available on newsstands Monday, August 24. U.S. News will release its "Best College Values" rankings in an upcoming issue of the magazine, available on newsstands Monday, August 31.

The grant will be used to help equip and furnish the Decker School's new home in the Academic I complex, which is scheduled to be completed later this year. The estimate for fully equipping the Decker School's new home is approximately 1.2 million dollors.

"This grant will help to provide laboratory, computer and medical equipment to support our outstanding academic programs in nursing," said President Lois B. DeFleur. "An investment of this size by the Decker Foundation will enable us to continue the Decker School's innovative efforts in rural and community health, family nursing and gerontology education.

"We are thrilled that we have received this wonderful gift from the Decker Foundation," said Mary Collins, dean of the Decker School. "The gift puts us well on the way to providing us with the resources that reflect the quality educational programs we provide and will allow us to meet the learning needs of students well into the next century".

The Class of 2002 continues to reflect the diversity of the state. Students of Asian background comprise 18.2 percent of the class, followed by Hispanic/Latino and African-American at 6.5 percent each and Native American at .1 percent. There are 36 international students enrolled as entering freshmen.

Binghamton University is again among the elite Top 25 public universities in the nation according to U.S. News and World Report. The magazine's 12th annual "America's Best Colleges" issue and guidebook lists Binghamton 21st in its list of top public universities.

Five teams will be in action at home for the NCAA Division II kick-off celebration. Activities begin at 10 a.m. with women's volleyball vs. UMass Lowell inside the West Gym, and youth clinics in baseball, tennis and soccer on the fields and courts near the West Gym.

e.txt

P[0, a] is the same as P[1, b], P[2, c] and P[3, d].

P[3, a] is similar to P[3, b] with "August" is changed to "Aug", "America" is changed to "USA".

P[4, b] is similar to p[4, c] with S1S2 is reversed to S2S1.

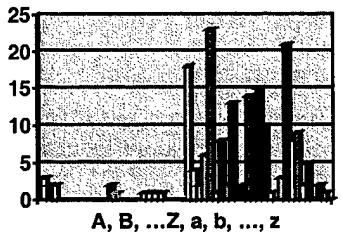
P[6, c] is the same as P[5, d].

P[1, c] is similar to P[0, d] with "are reliable" changed to "can be trusted".

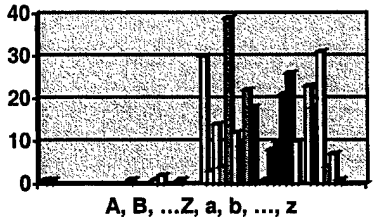
P[1, a] is the same as P[6, d].

P[4, a] is the same as P[5, c] and P[4, d].

Below we present some of the results obtained by using the IRR scheme on these five text-documents:



Letter histogram from paragraph 0, a.txt



Letter histogram from paragraph 2, b.txt

. a.txt file size 1269	
(paragraph No. 0)	
Number of characters	228
Effective character number	223
Number of words	36
Number of sentences	1
Starting word for S1	Binghamton
Ending word for S1	12
...	
Statistical Features	

. b.txt file size 1784	
...	
(paragraph No.2)	
Number of characters	419
Effective character number	413
Number of words	69
Number of sentences	2
Starting word for S1	With
Ending word for S1	class
Starting word for S2	The
Ending word for S2	1
...	
Statistical Features	

. a.txt word histogram
(paragraph No. 0)

on	2
Binghamton	1
old	1
University	1
(Any other word appears only once)	

...

. b.txt word histogram

...

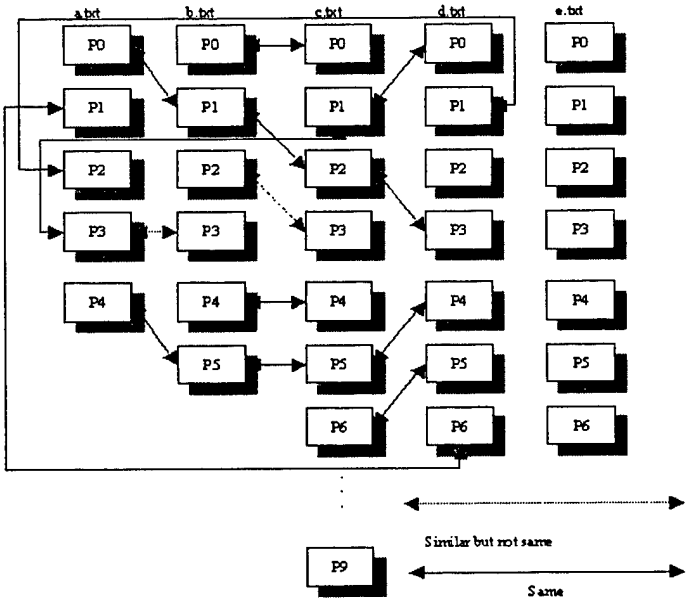
(paragraph No.2)

the	6
percent	4
top	3
from	3
average	2
of	2
school	2
class	2
their	2
with	1
an	1
freshman	1
(Any other word appears only once)	

...

Word Histograms

Graph of same and similar paragraphs



6.2 Image Redundancy Removal

With redundancy in images we mean the occurrence of the same image more than twice, with the same or different resolution, size and/or color.

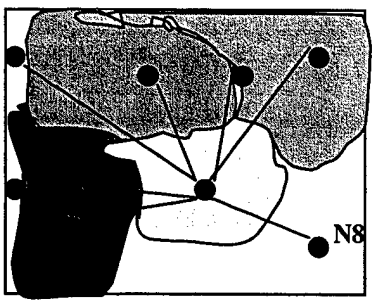
Discovery of image redundancy

Each image will be analyzed and a number of statistical characteristics (c) will be extracted from it. These characteristics are:

- Number of image regions (nr)
- Histogram of colors
- Relative size of the regions (sr)
- Shapes of regions (shr)
- Texture of regions (tr)
- Weighted regions graph (G)

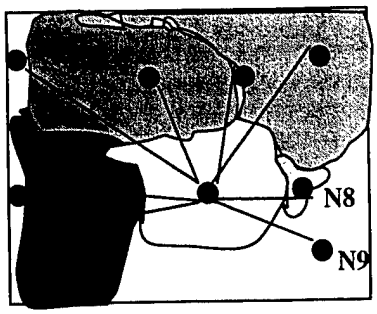
If two images I1 and I2 have the same features described above, then I1 and I2 are considered similar or same with a probability p'(f) of removal. In this case one of these two images will be removed under the condition that both have the same pointers (or ids) to other forms, such as text, and/or data. If the pointers are different then a more detailed analysis takes place on the examined images and the removal operation is postponed until an analytical examination will take place at the corresponding text and data parts. For ambiguous cases, an image understanding process will be used to make clear the final decision of removing or not one of the examined images. Figure 4 illustrates graphically the generation of the weighted graph of an image [24]. This means that the comparison of two images is mainly based on the comparison of their features and especially their regions weighted graphs, which carry all the information needed for each region.

IMAGE -A
Image regions and the graph of gravity



$$G(A_{(N1)}) = (N_1 R_{12} N_2)\ \Phi_{23}\ (N_1 R_{13} N_3)\ \Phi_{34}\ (N_1 R_{14} N_4) \ldots \Phi_{67}\ (N_1 R_{17} N_7)\ \Phi_{78}\ (N_1 R_{18} N_8)\ \Phi_{81}$$

IMAGE -B
Image regions and the graph of gravity

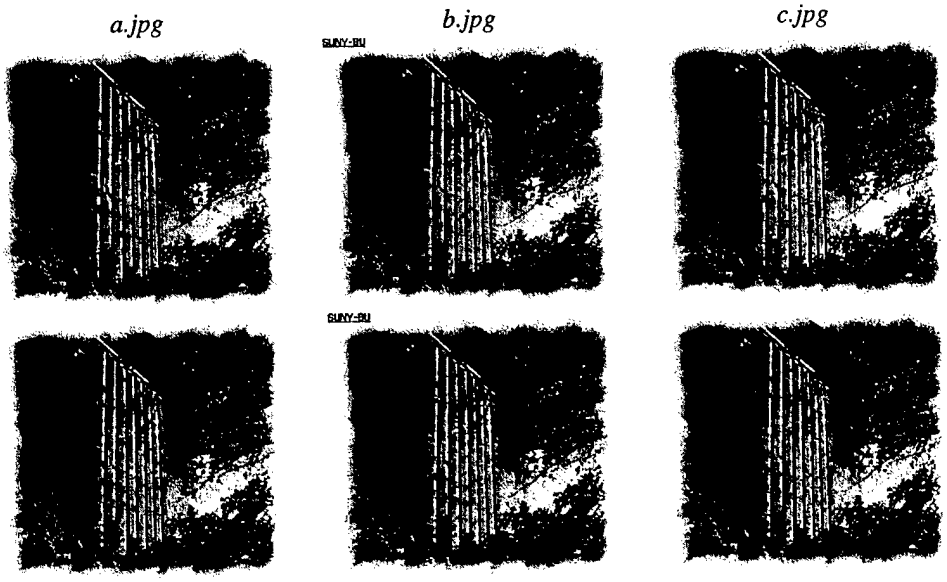


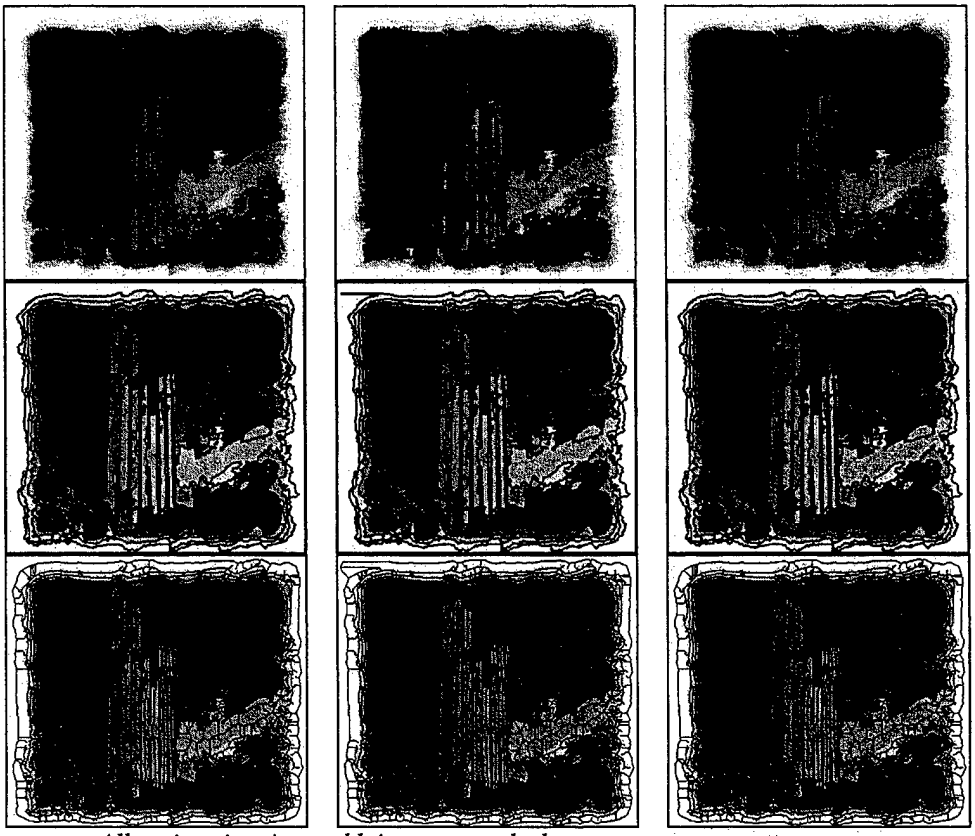
Comparison A and B
7/8 region relationships same
5/7 angles same

$$G(A_{(N1)}) = (N_1R_{12}N_2)\ \Phi_{23}\ (N_1R_{13}N_3)\ \Phi_{34}\ (N_1R_{14}N_4)\ \ldots$$
$$\Phi_{67}\ (N_1R_{17}N_7)\ \Phi_{78}\ (N_1R_{18}N_8)\ \Phi_{89}\ (N_1R_{19}N_9)\ \Phi_{91}$$

Figure 4: Note that Ni represents the vector or record of an image region, Rij represents the relative distance between the regions Ni and Nj, and Φ represents the relative direction or angle between two regions.

The IRR scheme performs image processing-analysis on the three images, extracts the regions features, generates their weighted graphs and compares them for defining their similarities:





All regions in a.jpg and b.jpg are matched

The different regions in
a.jpg and c.jpg
Similarity=0.95
(in black color)

The results from the IRR scheme show that these three images are actually the same with minor differences. Thus, the selection of one of them is sufficient to represent the others in the new synthetic document.

6.3 Information Synthesis

The synthesis of text and image information takes place after the removal of redundancies from both text and image parts. The synthesis process combines text paragraphs and their associated images and will generate a new kind of document by reassigned numbers in

paragraphs and images. This information will be compared with the "caption" of a particular image. In case of a match the image will be placed after the examined paragraph and an appropriate number will be given to it. In addition, all the numbers related with captions will be reassigned. Note that the synthetic document produced by the IRR scheme carries all the information needed to reconstruct any of the original documents , if necessary.

Example: The synthetic document produced by the IRR scheme applied on the five documents mentioned above:

Synthesis of a New Document

<Start of File: example/a.txt>
<"example/a.txt", Paragraph 0>
<Same as/similar to "example/b.txt", Paragraph 1 >
<Same as/similar to "example/c.txt", Paragraph 2 >

Binghamton University, fig.1 embarks on its inaugural Division II year at home as it renews old rivalries with SUNY Albany and begins new ones with members of the New England Collegiate Conference (NECC) on Saturday, September 12.

<"example/a.txt", Paragraph 1>
<Same as/similar to "example/d.txt", Paragraph 6 >

Five teams will be in action at home for the NCAA Division II kick-off celebration. Activities begin at 10 a.m. with women's volleyball vs. UMass Lowell inside the West Gym, and youth

Fig.1

clinics in baseball, tennis and soccer on the fields and courts near the West Gym.

<"example/a.txt", Paragraph 2>

President Lois B. DeFleur will offer a welcome and preside over a cake cutting at 12:30 p.m., adjacent to the soccer field as Binghamton makes its official entrance at home into the NECC--considered one of the nation's top Division II conferences.

<"example/a.txt", Paragraph 3>
<Same as/similar to "example/b.txt", Paragraph 3 >
<Same as/similar to "example/c.txt", Paragraph 1 >
<Same as/similar to "example/d.txt", Paragraph 0 >

The "America's Best Colleges" issue and guidebook will be available on newsstands Monday, August 24. U.S. News will release its "Best College Values" rankings in an upcoming issue of the magazine, available on newsstands Monday, August 31.

<"example/a.txt", Paragraph 4>
<Same as/similar to "example/b.txt", Paragraph 5 >
<Same as/similar to "example/c.txt", Paragraph 5 >
<Same as/similar to "example/d.txt", Paragraph 4 >

The Class of 2002 continues to reflect the diversity of the state. Students of Asian background comprise 18.2 percent of the class, followed by Hispanic/Latino and African-American at 6.5 percent each and Native American at .1 percent. There are 36 international students enrolled as entering freshmen.`<End of File: example/a.txt>
<Start of File: example/b.txt>
<"example/b.txt", Paragraph 0>
<Same as/similar to "example/c.txt", Paragraph 0 >

Approximately 2,037 freshmen and 816 transfer students arrive on campus Thursday, August 27, toting boxes and suitcases full of belongings. The freshman class is

about 200 students larger than last year's and represents a return to the University's plan for growth that was put on hold during years of budget cuts.

<"example/b.txt", Paragraph 1>
<Same As "example/a.txt", Paragraph 0 >
<"example/b.txt", Paragraph 2>
<Same as/similar to "example/c.txt", Paragraph 3 >
With an average combined SAT score of 1206, about 200 points above the national average, the new arrivals maintain Binghamton's reputation as a selective school. Of those from high schools that rank their graduates, 21 percent come from the top five percent, 49 percent from the top 10 percent and 86 percent from the top fifth of their graduating class. The mean high school average for the freshman class is 92.1.

<"example/b.txt", Paragraph 3>
<Same As "example/a.txt", Paragraph 3 >
<"example/b.txt", Paragraph 4>
<Same as/similar to "example/c.txt", Paragraph 4 >
A geographic profile of the incoming class shows the vast majority of freshman--94.8 percent--are New York state residents. Nearly six percent are from Broome and Tioga counties, with 62.5 percent from the New York City area. In all, the first-year class includes students from 55 of New York's 62 counties.

<"example/b.txt", Paragraph 5>
<Same As "example/a.txt", Paragraph 4 >
<End of File: example/b.txt>
<Start of File: example/c.txt>
<"example/c.txt", Paragraph 0>
<Same As "example/b.txt", Paragraph 0 >
<"example/c.txt", Paragraph 1>
<Same As "example/a.txt", Paragraph 3 >
<"example/c.txt", Paragraph 2>
<Same As "example/a.txt", Paragraph 0 >
<"example/c.txt", Paragraph 3>
<Same As "example/b.txt", Paragraph 2 >
<"example/c.txt", Paragraph 4>
<Same As "example/b.txt", Paragraph 4 >
<"example/c.txt", Paragraph 5>
<Same As "example/a.txt", Paragraph 4 >
<"example/c.txt", Paragraph 6>
<Same as/similar to "example/d.txt", Paragraph 5 >
Binghamton University is again among the elite Top 25 public universities in the nation according to U.S. News and World Report. The magazine's 12th annual "America's Best Colleges" issue and guidebook lists Binghamton 21st in its list of top public universities.

<"example/c.txt", Paragraph 7>
U.S. News calculated scores based on academic reputation, retention rates, faculty resources, student selectivity, financial resources, graduation rates and alumni giving rates to compile its rankings. According to the magazine the rankings are reliable, objective and fair with "each school's rank based on the same set of quality measures".

<"example/c.txt", Paragraph 8>
U.S. News notes the rankings are also helpful to those choosing a college for several reasons: they are based on accepted measures of academic quality chosen based on U.S. News' experience in reporting on education and on research about measuring educational outcomes, after consultation with experts; they have been developed independently of any particular institution; they are comparable and complete; they are the single best source of information because they allow readers to compare the strengths and weaknesses of different schools; and they condense a great deal of information, making it easier to compare institutions.

<"example/c.txt", Paragraph 9>
"Binghamton University has been recognized consistently by national publications for its quality and value," said President Lois B. DeFleur. "Appearing in the U.S. News ranking of the Top 25 public universities for two years

```
in a row is another highly visible validation of the efforts our faculty and staff make in educating our students and of our success in doing so".

<End of File: example/c.txt>
<Start of File: example/d.txt>
<"example/d.txt", Paragraph 0>
<Same As "example/a.txt", Paragraph 3 >
<"example/d.txt", Paragraph 1>
       The grant will be used to help equip and furnish the Decker School's new home in the Academic I complex, which is scheduled to be completed later this year. The estimate for fully equipping the Decker School's new home is approximately 1.2 million dollors.

<"example/d.txt", Paragraph 2>
       "This grant will help to provide laboratory, computer and medical equipment to support our outstanding academic programs in nursing," said President Lois B. DeFleur. "An investment of this size by the Decker Foundation will enable us to continue the Decker School's innovative efforts in rural and community health, family nursing and gerontology education.

<"example/d.txt", Paragraph 3>
       "We are thrilled that we have received this wonderful gift from the Decker Foundation," said Mary Collins, dean of the Decker School. "The gift puts us well on the way to providing us with the resources that reflect the quality educational programs we provide and will allow us to meet the learning needs of students well into the next century".

<"example/d.txt", Paragraph 4>
<Same As "example/a.txt", Paragraph 4 >
<"example/d.txt", Paragraph 5>
<Same As "example/c.txt", Paragraph 6 >
<"example/d.txt", Paragraph 6>
<Same As "example/a.txt", Paragraph 1 >
<End of File: example/d.txt>
<Start of File: example/e.txt>
<"example/e.txt", Paragraph 0>
       P[0, a] is the same as P[1, b], P[2, c] and P[3, d].

<"example/e.txt", Paragraph 1>
       P[3, a] is similar to P[3, b] with "Auguest" is changed to "Aug", "America" is changed to "USA".

<"example/e.txt", Paragraph 2>
       P[4, b] is similar to p[4, c] with S1S2 is reversed to S2S1.

<"example/e.txt", Paragraph 3>
       P[6, c] is the same as P[5, d].

<"example/e.txt", Paragraph 4>
       P[1, c] is samilar to P[0, d] with "are reliable" changed to "can be trusted".

<"example/e.txt", Paragraph 5>
       P[1, a] is the same as P[6, d].

<"example/e.txt", Paragraph 6>
       P[4, a] is the same as P[5, c] and P[4, d].

<End of File: example/e.txt>
```

7.0 Conclusions and Discussions

In this report, a new type of search engine for multimedia web pages has been presented . The most novel aspect of this search engine is its query language WebSSQL. Due to the similarity-based feature, WebSSQL retains the ability to rank retrieval results of regular search engines while at the same time supporting the specification of more precise queries using additional information collected in advance. Allowing users to

express their search needs more precisely can significantly improve the retrieval effectiveness of search engines.

As an on-going project, the proposed search engine can be improved in many directions. For example, a more general yet more rigorous syntax for WebSSQL is needed so that the formal query semantics of WebSSQL can be provided. As another example, manual indexing of images in the current implementation is not practical. We are working on incorporating into the search engine a program that can obtain image color histograms automatically. In addition, we are exploring automatic methods to generate text representations for images. This is possible because the images we consider are not isolated images but images that appear in web pages in HTML syntax. Associating text descriptions of images through referencing anchors and other means is widely used. We would like to find ways to extract these text descriptions for image representation.

A methodology that removes redundant information in received documents has also been presented. More specifically, the development of text and image redundancy detection and removal methods for removing duplicate information due to the retrieval process, and a document synthesis mechanism to integrate the remaining information and generate a new "integrated document" were described. The new document carries information able to reconstruct any of the original documents used for its creation. Results from the first stage text-paragraphs reduction and illustration of the image reduction approach were also provided. Note that from the illustrative example presented in this report, the information reduction ratio was 1.67 (original = 8998 characters, synthetic = 5386 characters). In addition, the three original pictures contained in the original documents (a.doc, b.doc, c.doc) were reduced into one in the new synthetic document. This implies that an additional information reduction took place by making the overall removal of redundant information more effective. Extensions to this methodology are the use of natural language and image understanding to filter and edit the synthetic document by creating a human like new document. This is a real advantage in the document processing field, where new documents will be automatically generated from a sizable set of original ones.

In summary, the advantages of this first version of the software prototype are:

1. The removal of redundant information from similar text-documents.
2. The comparison of similar images and record their differences.
3. The creation of synthetic documents (text and images) as a combination of a set of documents that contain similar or same pieces of information by removing the redundant information.
4. Regeneration of all the original documents used for the creation of a synthetic document.

The main disadvantages of this first version of the software prototype are:

1. It is unable to detect redundant information in text paragraphs that contain sentences with different words but with the same or similar semantic meaning.
2. It is unable to provide a Natural Language (NL) understanding of the text paragraphs.
3. It is unable to rank images based on their degree of similarity.

The second version of this software prototype will address and include several of these missing capabilities.

8.0 Acknowledgement

This project is partially supported by an Air Force Research Laboratory (AFRL) Grant through the Information Directorate's (IF) Global Information Base Branch (IFED) at the Rome Research Site, Rome, NY.

9.0 References

[1] S. Abiteboul, D. Quass, J. McHugh, J. Widom, and J. L. Wiener. *The Lorel Query Language for Semistructured Data.* International Journal on Digital Libraries, 1:1, pp.68-88, April 1997.

[2] C. Buckley, G. Salton, and J. Allan. *Automatic Retrieval with Locality Information Using Smart.* First TREC Conference, Gaithersburg, MD, pp.59-72, 1993.

[3] D. Konopnicki, and O. Shmueli. *W3QS: A Query System for the World Wide Web.* Very Large Data Bases Conference, 1995.

[4] A. Mendelzon, G. Mihaila, and T. Milo. *Querying the World Wide Web.* International Journal on Digital Libraries, 1:1, pp.54-67, April 1997.

[5] W. Meng, C. Yu, W. Wang, and N. Rishe. *Performance Analysis of Three Text-Join Algorithms.* IEEE Trans on Knowledge and Data Engineering, 10:3, pp.477-492, 1998.

[6] G. Navarro, and R. Baeza-Yates. *A Model to Query Documents by Contents and Structure.* ACM SIGIR Conference, pp.93-101, 1995.

[7] D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. Querying Semistructured Heterogeneous Information. DOOD International Conference, pp.319-344, 1995.

[8] G. Salton and M. McGill. *Introduction to Modern Information Retrieval.* New York: McCraw-Hill, 1983.

[9] G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.* Addison Wesley, 1989.

[10] C. Yu, and W. Meng. *Principles of Database Query Processing for Advanced Applications.* Morgan Kaufmann, San Francisco, 1998.

[11] D. Beech, P. Chellone and C. Ellis. *An ADT Approach to Full Text.* ISO/IEC JTC1/SC21/WG3 DBL CBR-57, 1992.
[12] W. Croft, L. Smith and H. Turtle. *A Loosely-Coupled Integration of a Text Retrieval System and an Object-Oriented Database System.* SIGIR, 1992.
[13] M. Flickner, et al. *Query by Image and Video Content: The QBIC System.* IEEE Computer, September 1995.
[14] V. Ogle and M. Stonebraker. *Chabot: Retrieval from a Relational Database of Images.* IEEE Computer, September 1995.
[15] L. Saxton and V. Raghavan. *Design of an Integrated Information Retrieval/Database Management System.* IEEE TKDE, pp. 210-219, June 1990.
[16] R. Srihari. *Automatic Indexing and Content-based Retrieval of Captioned Images.* IEEE Computer, September 1995.
[17] L.O'Gorman,The document spectrum for page layout analysis"IEEE-T-PAMI,15,1993
[18] N.Bourbakis ,"A method of separating text from images",IEEE Symp.ISNL,Nov.1996
[19] L.A.Fletcer and R.Kasturi,"A robust algorithm for text string separation from text graphics images", IEEE T-PAMI, 10, 910-918, 1988
[20] F.Wahl K.Wong and R.Casey,"Block segmentation and text extraction in mixed text image documents", CVGIP, 20,1989
[21]N. Bourbakis, "Information synthesis using SPNGs", TR-1996, IEEE T-SMC sub.
[22] R. Samir and N. Bourbakis,"Distributed Multimedia Information Systems", Int. Conf. On Design and Process Evaluation, TX, Dec. 1996
[23] N.Bourbakis, W.Meng, Z.Wu, J.Salerno and S.Borek,"Redundancy removal in documents retrieved from different resources, IEEE ICTAI, 1998, Taiwan, pp. 112-119
[24] N.Bourbakis, Retrieving images by using regions attributed graphs" , GMU-ECE-TR-1987.

MISSION
OF
AFRL/INFORMATION DIRECTORATE (IF)

The advancement and application of Information Systems Science and Technology to meet Air Force unique requirements for Information Dominance and its transition to aerospace systems to meet Air Force needs.