The WBR-99 Collection

Description of the WBR-99 collection data-structures and file formats

Pável Pereira Calado
pavel@dcc.ufmg.br

LATIN - Laboratório para o Tratamento de Informação
Departamento de Computação
Universidade Federal de Minas Gerais
http://www.dcc.ufmg.br/latin

Contents

1 Introduction

The WBR-99 collection was built from a set of documents collected from the Brazilian Web in November 1999. It was taken from the database of TodoBR, a search engine for the Brazilian Web[1], and offered to the LATIN laboratory, for research in Information Retrieval problems. Experiments with the WBR-99 collection have already been used in several docotrate and master thesis, and published in works such as [1], and [6].

The collection contains about 6 million HTML documents in an already indexed format. It also contains the complete set of queries submitted to TodoBR during November 1999. For fifty of those queries a set of relevant documents is available. All the information in the collection was obtained by crawling the Brazilian Web, with the crawler described in [5].

This manual is intentded as a description of the data structures and file formats used in the WBR-99 collection, so that other researchers can use it to test their solutions for Web Information Retrieval. Please note that we assume the reader is familiar with some common Information Retrieval data structures and models, such as *inverted lists* [7], and the *vector-space model* [4].

2 Statistics

Table 1 presents some useful statistics on the contents of the WBR-99 collection. *Total size* refers to the size of the collection, all files included; *Text size* refers to total size of the text from the documents in the collection (without HTML tags); *Number of terms* refers to the total number of words in all documents; *Number of links* refers to the total number of links between all documents; *Number of available queries* refers to the number of queries in the November 1999 TodoBR log; *Number of evaluated queries* refers to the number of queries for which human users have selected a set of relevant/non-relevant documents; *Number of evaluated documents* refers to the number of documents examined by human users for the evaluation of the queries mentioned in the previous item.

[1]The TodoBR search engine is available at `http://www.todobr.com.br`.

Total size	20G
Text size	16G
Number of documents	5 939 061
Number of terms	2 669 965
Number of links	40 871 504
Number of available queries	33 154
Number of evaluated queries	50
Number of evaluated documents	4 117

Table 1: Statistics on the WBR-99 collection

3 File Formats

The WBR-99 collection comprises of a set of files, each containing information on a specific part of the collected documents. Table 2 presents a summarized description of all the files. The following sections describe in detail the internal structure of each file. These descriptions should allow the programming of simple functions to access the contents of the WBR-99 collection.

3.1 General Structure

In the WBR-99 collection, to each document is assigned a number, henceforth called the *document ID*. This number corresponds to the position of the document in the collection files. Equally, to each term in the collection is also assigned a number, henceforth called the *term ID*.

In almost all files, both terms and documents are referred to by their IDs. Thus, for instance, the list of terms for a given document, which can be found in the 'wbrinvlist' files, is a sequence of IDs. Values are always stored as longs (4 bytes, always unsigned), floats (four bytes) or chars (1 byte).

All the '.idx' files store only he position offsets in the files where the real data is stored. Thus, the 'wbrinvlist.idx' file stores, for each term in the collection, the position on the 'wbrinvlist' file where the inverted list for that term is stored. This will be made clear in the following sections, where each file is explained in detail.

Please note that, in the figures illustrating the following explanations, all values shown are hypothetical, used only as examples, and do not correspond to the real values stored in the files.

Main directory	
wbr99.pdf	this file
readme.txt	a file with important information
copyright.txt	copyright information
wbrtext[0–7]	text of all documents (without HTML tags)
wbrtext.idx	index to the text files
wbrinvlist[0–1]	inverted list files
wbrinvlist.idx	index to the inverted list files
wbrnorm	norms of the vectors of all documents
wbrvoc	list of all terms in the collection
wbridf	IDF value of all terms in the collection
wbrinlink	in-links of all documents
wbrinlink.idx	index to the in-link file
wbroutlink	out-links of all documents
wbroutlink.idx	index to the out-link file
wbrgeral0	general information on all documents
wbrgeral.idx	index to the general information file
'consultas' directory	
log_consultas_nov_1999.log	queries submitted to TodoBR in Nov/99
consultas.txt	Textual description of the evaluated queries
consultas.dat	Terms of the evaluated queries
'consultas/relevantes' directory	
cons_[1–50].rel	Evaluated documents for each query

Table 2: The WBR-99 collection files.

3.2 The Text Files

The files 'wbrtext0' through 'wbrtext7' contain the text of each HTML document collected. All HTML tags and structure were removed. The files consist simply of a sequence of characters, with the text for each document separated by the '\0' (zero code) character.

Figure 1 illustrates the relation between the index file and the text files. . The 'wbrtext.idx' file is the index file for the documents text. Each record stores the offset in one of the 'wbrtext' files for the position where the corresponding document text starts.

Each record in 'wbrtext.idx' has two fields, a short (2 bytes) and a long

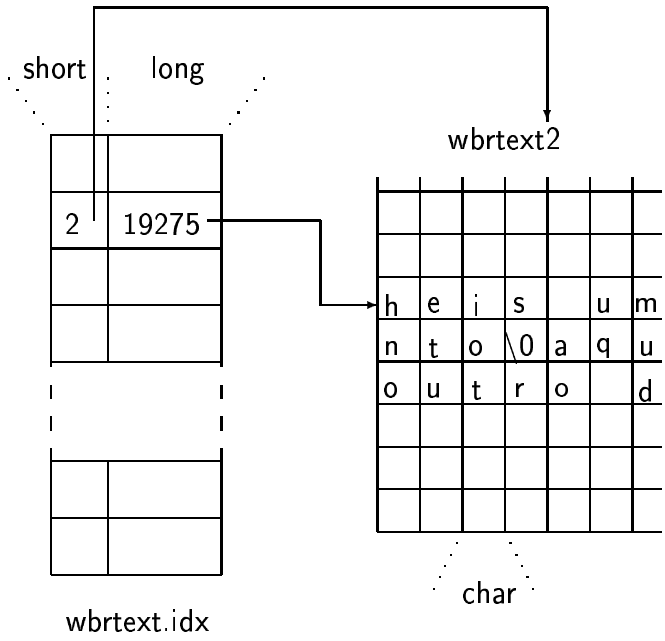

Figure 1: The text files.

(4 bytes). The record corresponding to document d, where d is a document ID, is, therefore, stored at position $d \times 6$. The first field indicates which of the 'wbrtext' file contains the text of document d. The second field contains the offset position in the indicated 'wbrtext' file where the text of document d starts. The text is ended by the '\0' character.

In the example of Figure 1, the record indicates that the text for document 2 is stored in the file 'wbrtext2', and that it starts at position 19275.

3.3 The Inverted List Files

Documents in the WBR-99 collection were already indexed into inverted list structures. File 'wbrinvlist.idx' contains the index to the inverted list and files 'wbrinvlist0' and 'wbrinvlist1' contain the actual list of documents. Figure 2 illustrates their structure.

Each record in 'wbrinvlist.idx' has two fields, a short (2 bytes) and a long (4 bytes). The record corresponding to term t, where t is a term ID, is,

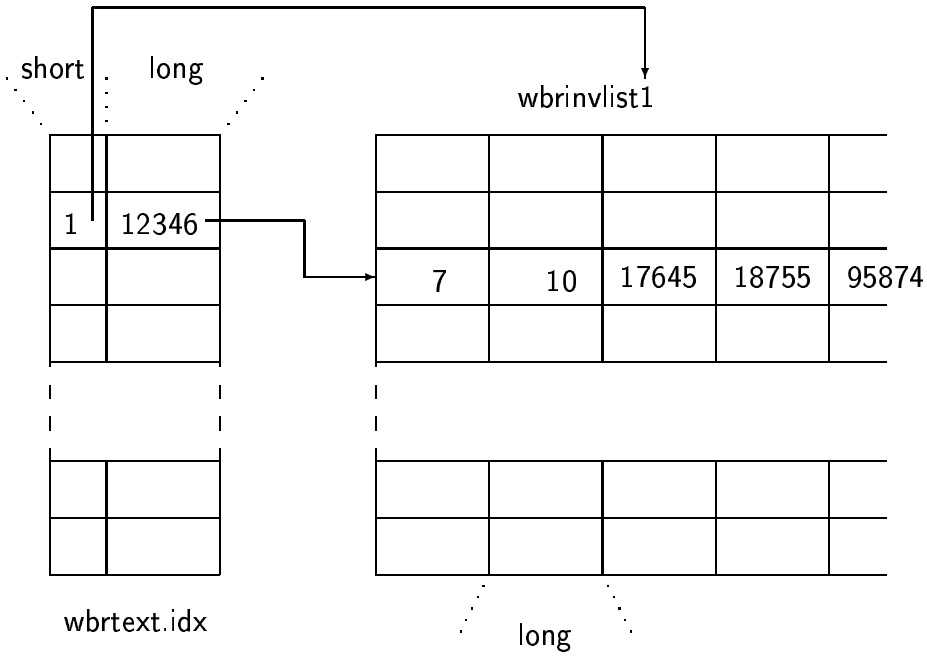

Figure 2: The inverted list files.

therefore, stored at position $t \times 6$. The first field indicates which 'wbrinvlist' file contains the inverted list of term t. The second field contains the offset position in the indicated 'wbrinvlist' file where the inverted list of term t starts.

Each record in a 'wbrinvlist' starts with a long value (4 bytes), indicating the frequency of term t in the following list of documents. The next value is a long (4 bytes) indicating the size of the list of documents where term t occurs with the given frequency. Following is a list of long values (4 bytes each) containing the IDs of the documents in which term t occurs with the given frequency. The inverted list for term t consists of a sequence of these frequency/length/list-of-documents triples, which are repeated until a frequency value of 0 (zero) is found.

In the example of Figure 2, the record indicates that the text for term 2 is stored in the file 'wbrinvlist1', and that it starts at position 12346. The inverted list indicates that term 2 appears with frequency 7 in 10 documents: 17645, 18755, 95874, etc. After the 10 documents, a new frequency should

appear, followed by a new length, and so on, until we find a frequency value of 0.

3.4 The Norms and IDF Files

The 'wbrnorm' and 'wbridf' files are simply a sequence of float values (4 bytes each). The first, 'wbrnorm', contains the norms of the vectors for each document in the collection. The second, 'wbridf', contains the IDF values for each term in the collection. Figure 3 illustrates the structure of the norms file. The structure of the IDFs file is equivalent.

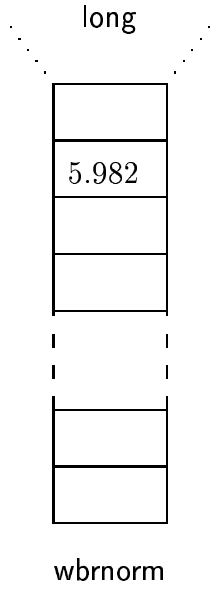

Figure 3: The norms file.

The following formula was used to compute the norm of each document d:

$$norm_d = \sqrt{\sum_{t \in d} (TF_t \times IDF_t)^2} \quad (1)$$

where TF_t is the frequency of term t in document d, which is stored in the 'wbrinvlist' files, and IDF_t is the IDF value of term t, which is stored in

'wbridf' file. The IDF values are computed as:

$$IDF_t = \ln(N/F_t) \tag{2}$$

where N is the total number of documents in the collection and F_t is the number of documents where term t occurs.

3.5 The Vocabulary File

All the terms extracted from the documents in the WBR-99 collection are stored in the 'wbrvoc' file. Its structure is illustrated in Figure 4.

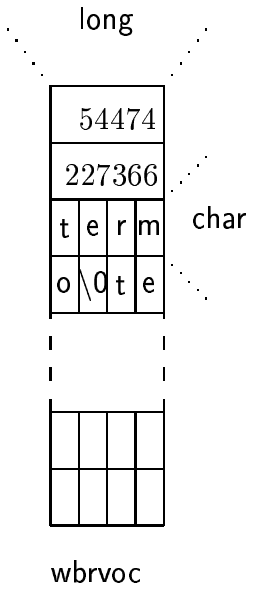

Figure 4: The vocabulary file.

The file starts with two long values (4 bytes each). The first indicates the number of terms in the file and the second indicates the size, in bytes, of the file. Following the two longs, the file consists of a sequence of chars, containing the actual terms, all separated by the '\0' character.

3.6 The Link Structure Files

The links between the HTML documents in the collection are stored in the 'wbrinlink' and 'wbroutlink' files. Similarly to the inverted list files, described in Section 3.3, the 'wbrinlink.idx' and 'wbroutlink.idx' files contain the offsets of the positions where the lists of links start.

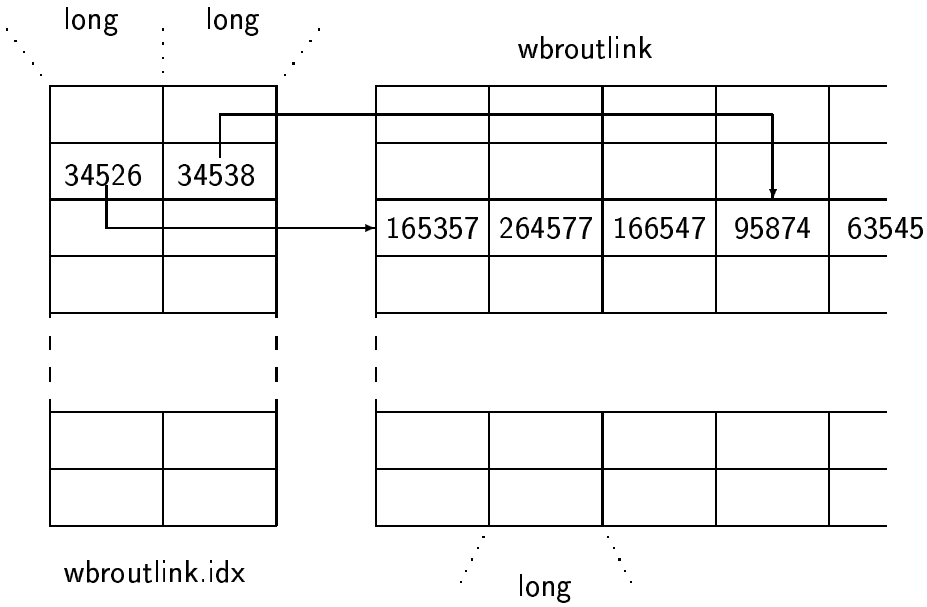

Figure 5: The link structure files.

Figure 5 shows the structure of the 'wbroutlink.idx' and 'wbroutlink' files. As shown, each record in the 'wbroutlink.idx' file is composed of two longs (4 bytes each). The record corresponding to document d is, therefore, stored at position $d \times 8$. The first long contains the offset position in the 'wbroutlink' file for the start of the list of external links. The second contains the offset position for the start of the list of internal links. The 'wbroutlink' file contains the actual lists of links. Each list is a sequence of long values (4 bytes each).

An external link is a link to a page on different site. We consider sites to be different if they are under a different domain, where by domain we mean

the first part of the URL, from 'http://' through the first '/' symbol. Also, some heuristics that compare the text of the HTML documents where used to determine if the same site appears under different domains. Conversely, an internal link is a link to a page on the same site.

The 'wbrinlink.idx' and 'wbrinlink' files are equal in structure to the 'wbroutlink.idx' and 'wbroutlink' files but containing in-link information.

In the example of Figure 5, the record indicates that the list of external links from document 2 starts at position 34526. The list of internal links starts at position 34538. Thus, document 2 to links to documents 165357, 264577, and 166547, on a different site, and to documents 95874, 63545, etc., on the same site. If Figure 5 represented the 'wbrinlink.idx' and 'wbrinlink' files, we would say that document 2 is linked by documents 165357, 264577, and 166547, on a different site, and by documents 95874, 63545, etc., on the same site.

We should note that the '.idx' files contain an extra record, with the ending position of the list of internal links for the last document. This is to avoid having to check for the occurrence of end-of-file, when reading the last list of links.

3.7 The General Information Files

File 'wbrgeral0' contains the date, size, type, URL, and title of each document in the WBR-99 collection. As before, the file 'wbrgeral.idx' contains the starting positions of each record in 'wbrgeral0'.

As illustrated in Figure 6, Each record in 'wbrgeral.idx' has two fields, a short (2 bytes) and a long (4 bytes). The record corresponding to document d is, therefore, stored at position $d \times 6$. The first value indicates which 'wbrgeral' file contains the information of document d. The second value contains the offset position in the indicated 'wbrinvlist' file where the information for document d starts.

Each record in a 'wbrgeral' starts with a long value (4 bytes), indicating the date document d was collected. The date is simply the number of seconds since 00:00:00 UTC, January 1, 1970. The next value is a long (4 bytes) indicating the size, in bytes, of document d. Following is a char (1 byte), indicating the type of document. A document may be of type ???????. Next is a sequence of chars, containing the URL of the document, terminated by the '\0' char. Finally, we have another sequence of chars, containing the title of the document, also terminated by the '\0' character.

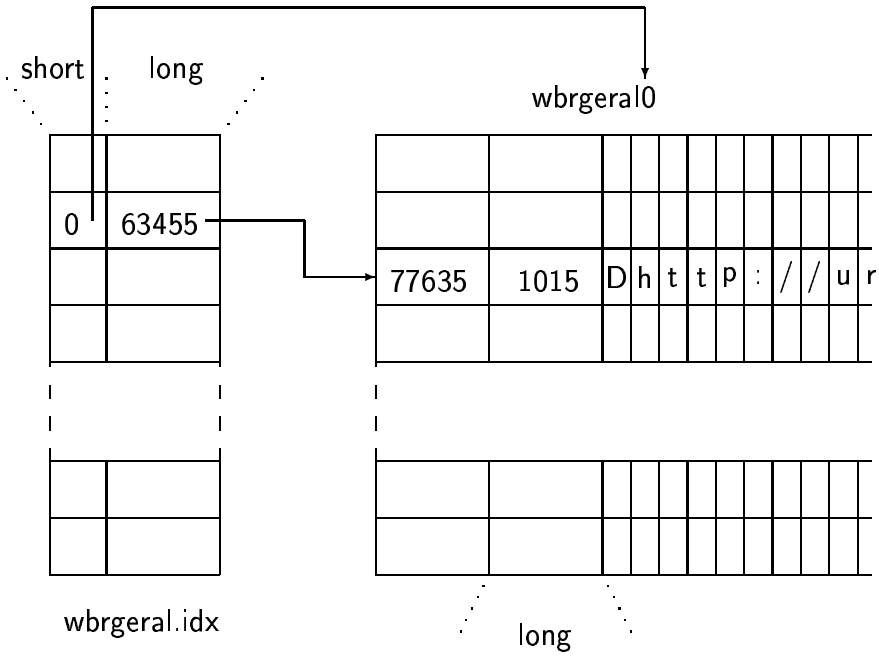

Figure 6: The general information files.

3.8 The Queries

Together with the documents extracted from the Web, the WBR-99 collection has a set of queries, extracted from a query log of the TodoBR search engine. File 'log_consultas_nov_1999.log' is a text file containing all the queries submitted to TodoBR in November 1999. Each line contains, on the first column, the submission frequency of the query, i.e., how many times it was submitted to TodoBR during November, and on the second column, the actual query terms.

File 'consultas.txt' contains 50 queries selected from the most frequent queries in 'log_consultas_nov_1999.log'. Each line contains the query and a textual description of its meaning. File 'consultas.dat' contains the IDs of the query terms for the queries in 'consultas.txt'.

Finally, files 'cons_1.rel' through 'cons_50.rel' contain all the evaluated documents for the queries in 'consultas.txt'. To evaluate the documents, for each of the 50 queries, a pool formed by the top 20 documents generated

by the set of algorithms described in [1] was formed. All documents in each query pool were submitted to a manual evaluation by a group of 29 users, all of them familiar with Web searching. Users were allowed to follow links and evaluated the pages according not only to their textual content, but also to their linked pages and graphical content. The average number of relevant pages per query pool is 36. We adopted the same pooling method used for the Web-based collection of TREC [2, 3].

The evaluated documents were given one of three different values: -1, meaning that the document was invalid (for instance, an error page); 0, meaning that the document was not relevant to the topic of the given query; 1, meaning that the document, although not completely relevant, contained related information or links; 2, meaning that the document contained related information or links to relevant documents; and 3, meaning that the document contained relevant information.

4 Copyright and Terms of Use

The WBR-99 collection was built and indexed for the TodoBR search engine (`http://www.todobr.com.br`) developed by Akwan Information Technologies (`http://www.akwan.com.br/`), who holds the copyright.

The WBR-99 collection is available to scholars without fee for educational and research purposes, on request. Potential users may not use the WBR-99 collection unless they agree to the following conditions:

1. The user acknowledges that the WBR-99 is subject to copyright restrictions and agrees to abide by them. The user acknowledges that violations of copyright restrictions may result in legal liability.

2. The user agrees to notify all associates who access the downloaded copy of the WBR-99 of the copyright restrictions and of their obligation to respect the provisions of this agreement as additional users.

3. The user will make no commercial use of the WBR-99.

4. The user will not redistribute the WBR-99 to others except in limited passages under the ordinary standards of scholarly citation.

5. The user will acknowledge the WBR-99 in any written work or oral presentations based on research using this material.

6. The user acknowledges that the creators and distributors of the WBR-99 make no warranties, express or implied, concerning the WBR-99, including but not limited to their ownership, merchantability, or fitness for a particular purpose. The creators and distributors shall not be liable for any direct, consequential, punitive or other damages suffered by the user or any other person resulting from the use of the distributed material.

5 Acknowledgments

We would like to thank Akwan Information Technologies, who kindly made available the WBR-99 for academic research.

References

[1] Pável Calado, Berthier Ribeiro-Neto, Nivio Ziviani, Edleno Moura, and Ilmério Silva. Local versus global link information in the Web. *ACM Transactions On Information Systems*, 21(1):42–63, January 2003.

[2] David Hawking, Nick Craswell, and Paul B. Thistlewaite. Overview of TREC-7 very large collection track. In *The Seventh Text REtrieval Conference (TREC-7)*, pages 91–104, Gaithersburg, Maryland, USA, November 1998.

[3] David Hawking, Nick Craswell, Paul B. Thistlewaite, and Donna Harman. Results and challenges in web search evaluation. *Computer Networks*, 31(11–16):1321–1330, May 1999. Also in Proceedings of the 8th International World Wide Web Conference.

[4] Gerard Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[5] Altigran Silva, Eveline Veloso, Paulo Golgher, Berthier Ribeiro-Neto, Alberto Laender, and Nivio Ziviani. CobWeb - a crawler for the brazilian web. In *Proceedings of the String Processing and Information Retrieval Symposium (SPIRE'99)*, pages 184–191, Cancun, Mexico, September 1999.

[6] Ilmério Silva, Berthier Ribeiro-Neto, Pável Calado, Edleno Moura, and Nívio Ziviani. Link-based and content-based evidential information in a belief network model. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 96–103, Athens, Greece, July 2000.

[7] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images.* Morgan Kaufmann Publishers, 2nd edition, 1999.