

# CIRB030 (Chinese Information Retrieval Bench, version 3.0)

Kuang-hua Chen  
Department of Library and Information Science  
National Taiwan University  
Taipei 10617, Taiwan  
khchen@ntu.edu.tw

Hsin-Hsi Chen  
Department of Computer Science and Information Engineering  
National Taiwan University  
Taipei 10617, Taiwan  
hh\_chen@csie.ntu.edu.tw

An information retrieval (IR) test collection is used to evaluate the performance of IR systems. It is a helpful and powerful tool for investigation of the developing systems and the developed systems. CIRB030 (Chinese Information Retrieval Benchmark, version 3.0) test collection is such kind of test collection, which is designed to be used for evaluation of Chinese document retrieval. There are 4 folders and 10 files in CIRB030 CD-ROM. Please take a look at the Figure 1.















 AnswerSet	CIRB030 Answer Set
 DocSet	CIRB030 Document Set
 TopicSet	CIRB030 Topic Set
 TREC_EVAL	Trec_Eval program
 CIRB030DocStatistics.pdf	Statistics of CIRB030 Document Set
 CIRB030OverviewCH.doc	Overview of CIRB030 (Chinese version, WORD format)
 CIRB030OverviewCH.pdf	Overview of CIRB030 (Chinese version, PDF format)
 CIRB030OverviewEN.doc	Overview of CIRB030 (English version, WORD format)
 CIRB030OverviewEN.pdf	Overview of CIRB030 (English version, PDF format)
 CIRB030ReadmeCH.doc	Readme of CIRB030 CD-ROM (Chinese version, WORD format)
 CIRB030ReadmeCH.pdf	Readme of CIRB030 CD-ROM (Chinese version, PDF format)
 CIRB030ReadmeEN.doc	Readme of CIRB030 CD-ROM (English version, WORD format)
 CIRB030ReadmeEN.pdf	Readme of CIRB030 CD-ROM (English version, PDF format)
 CLIRNTCIR3ReportFinal.pdf	Report of CLIR Task in NTCIR3 Workshop

Figure 1: Content of CIRB030 CD-ROM

This document (CIRB030ReadmeEN) introduces CIRB030 CD-ROM in brief.

## 1. DocSet (CIRB030 Document Set)

CIRB030 is a little different from the CIRB020 and CIRB011 used in NTCIR workshop [1]. Basically, The CIRB030 contains the documents of CIRB020 and CIRB011, but we correct some errors in the documents and delete some “title-only” documents. As a result, the number of documents in CIRB030 is less than the sum of CIRB011’s documents and CIRB020’s documents. The CIRB011 contains the same documents in CIRB010 but with different tag format. CIRB020 is denoted with the same tag set as the CIRB011 been tagged. In order to make it much clearer, Figure 2 shows the version evolution. Another point should be noted is the news articles in CIRB030 are combined into 7 files, which are cdn1998-1999 (Central

Daily News), chd1998-1999 (Chinese Daily News), ctc1998-1999 (Commercial Times), cte1998-1999 (China Times Express), cts1998-1999 (China Times), udn1998 (UDN.COM 1998), and udn1999 (UDN.COM 1999) in the **DocSet** folder of CD-ROM. In contrary, each news article in CIRB011 and CIRB020 is a single file.

## 2. TopicSet (CIRB030 Topic Set)

At first the topic set consists of 50 topic. 8 topics are screened out after the evaluation carried out in NTCIR 3 Workshop. These topics might be too difficult or too simple. The topic IDs of these qualified topics are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 32, 33, 34, 35, 36, 37, 38, 39, 40, 42, 43, 45, 46, 47, 48, 49, and 50. As a result, you will find some topics are missed in **TopicSet** of CIRB030 CD-ROM, since the unqualified topics are not listed in CIRB030 Topic Set.

## 3. AnswerSet (CIRB030 Answer Set)

Four categories of relevance are identified: “Highly Relevant”, “Relevant”, “Partially relevant”, and “Irrelevant.” Each kind of relevance is assigned a relevance score. “Highly relevant” is 3, “Relevant” is 2, “Partially relevant” is 1, and “Irrelevant” is 0. Therefore, we have two sets of relevance judgments. One is “Rigid Relevance”; the other is “Relaxed Relevance”. The “Highly Relevant” and “Relevant” are regarded as relevance in “Rigid Relevance”. The “Highly Relevant”, “Relevant”, and “Partially Relevant” are regarded as relevance in “Relaxed Relevance”.

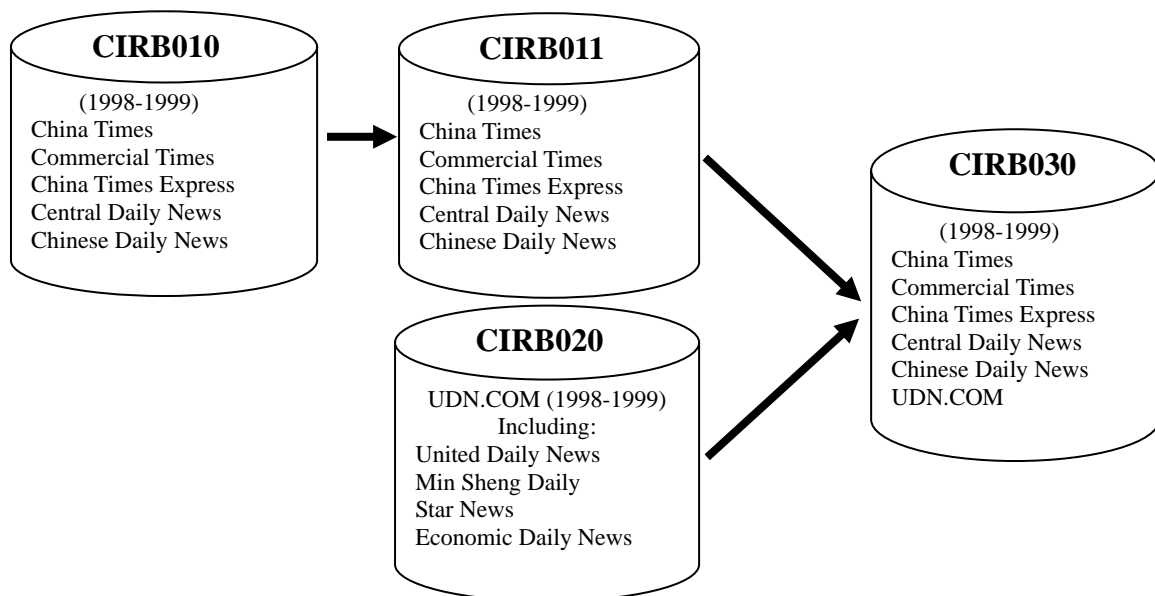


Figure 2: CIRB Version Evolution

## 4. TREC\_EVAL

TREC\_EVAL folder contains an evaluation program, TREC\_EVAL. This program is created by Buckley, which could be used to evaluate performance for information retrieval system based on binary relevance judgment. There are a few documents describing the usage and the output of this program.