Data Structures and Algorithms (11)

Instructor: Ming Zhang
Textbook Authors: Ming Zhang, Tengjiao Wang and Haiyan Zhao
Higher Education Press, 2008.6 (the "Eleventh Five-Year" national planning textbook)

https://courses.edx.org/courses/PekingX/04830050x/2T2014/

Contents

- Basic Concepts
- 11.1 Linear Indexing
- 11.2 Static Indexing
- 11.3 Inverted Indexing
- 11.4 Dynamic Indexing
- 11.5 Bit Indexing
- 11.6 Red-Black Tree

11.3 Inverted Indexing

- 11.3.1 Attribute based inverted indexing
  - Require that certain attribute meet certain conditions (do not search by keys)
  - Construct index according to attribute

- 11.3.2 Text based inverted indexing
  - Construct index according to words in the text

Table of Teacher database

EMP#	NAME	Department	Profession	Specialty	Address
0155	Alice	Math	Prof	Algebra	C105
0421	Bob	Language	TA	English	E310
0208	Cathy	Physics	TA	Mechanics	C211
0211	Danny	Physics	Lecturer	Atomic Physics	D508
0132	Eddie	Math	TA	Geometry	E220
0119	Frank	Math	Lecturer	Algebra	B102
0330	Gary	IT	Prof	Software	A108
0455	Henry	Language	Lecturer	French	A225
0310	Irene	IT	Lecturer	English	B423
0341	Jack	IT	TA	Computer Science	F406
……					

11.3.1 Attribute based inverted indexing

- Construct index according to attribute, i.e., inverted indexing.
- (attr, ptrList)
  - (attribute,pointers pointing to records of such attribute)
  - pointers could be key, or locations of records
- Reverse the order of the records, thus called inverted indexing
- Attribute are usually discrete
  - use B tree instead for continuous indexing
- Inverted file: file with inverted indexing

Inverted Table of Teacher Database

Department list	EMP#
Math	0155, 0132, 0119
Physics	0208, 0211
IT	0330, 0310, 0341
Language	0421, 0455
Profession list	**EMP#**
Prof	0155, 0330
Lecturer	0211, 0119, 0455, 0310
TA	0421, 0208, 0132, 0341
Specialty list	**EMP#**
Algebra	0155, 0119
Geometry	0132
Mechanics	0208
Atomic Physics	0211
Software	0330, 0341
English	0421, 0310
French	0455

Pros and Cons

- Pros:
  - Search can be performed efficiently by attribute based searching
- Cons:
  - more storage cost
  - slow down the updating operation

11.3.2 Text based inverted indexing

- Text Indexing means "constructing an index to search text efficiently"
- Approaches
  - word index
  - full-text index

Word Index

- Basic Idea
  - Treat the text as a set of words and symbols, extract keywords from it, and construct an index for fast indexing by using these keywords.
- Suitable for many types of text, especially those that could easily be parsed to a set of words
  - English
  - Oriental languages like Chinese require segmentation.

Full-Text Index

- Basic idea:
  - Treat the text as a long string
  - Store the starting locations of substrings in the index
  - Searching can be performed for every substring
- Construct index for each character and indexing are not limited to key words
- More space required

Word index is the most common used among inverted index

- Word index is the most common used
- A sorted list of keywords
  - Each keyword points to an inverted index posting list)
    - points to the document set in which the keyword appears
    - locations in the documents

Examples of Inverted Indexing

Text: consists of 6 documents, each of which is a long string

ID	Content
1	Pease porridge hot, please porridge cold,
2	Pease porridge in the pot,
3	Nine days old.
4	Some like it hot, some like it cold,
5	Some like it in the pot,
6	Nine days old.

文档编号	文本内容
1	Pease porridge hot, please porridge cold,
2	Pease porridge in the pot,
3	Nine days old.
4	Some like it hot, some like it cold,
5	Some like it in the pot,
6	Nine days old.

Inverted Index

ID	Word	(doc id , loc)
1	cold	(1,6)
2	days	
3	hot	(1,3)
4	in	
5	it	
6	like	
7	nine	
8	old	
9	pease	(1,1) (1,4) (2,1)
10	porridge	(1,2) (1,5)
11	pot	
12	some	
13	the	

Deal with the words in each document in turn

Constructing Inverted Index File

1. Divide the text into several documents.

Division depends on requirements of programs

Fixed length of blocks, paragraghs, or even a set of documents

Constructing Inverted Index File

2. Assign a group of keywords to each record

Extract keywords from records manually or automatically

Stopwords

Stemming

Segmentation

Chinese Segmentation

我知道你不知道我知道你不知道我知道你不知道

- 我知道，你不知道。我知道，你不知道我知道，你不知道
- 我知道你，不知道我。知道你不知道我，知道你不知道
- 我，知道你不知道我知道。你，不知道我知道你不知道

中科院计算所汉语词法分析系统

教育部部长周济说，今后若干年内，毕业生总量将会持续增加，每年都有数十万的增量，对毕业生就业工作无疑是巨大的挑战。2005年高校毕业生总量大、增幅高，各地和高校工作进展情况差异较大，并且还存在许多深层次的矛盾和问题。在全社会就业形势十分严峻的形势下，2005年高校毕业生就业工作压力十分突出。

操作选项

- (•) 词语切分
- ( ) 一级标注
- ( ) 二级标注

输出格式

- (•) 北大标准
- ( ) 973标准
- ( ) XML

运行　处理文件...　退出　关于...

1　结果　平滑参数：0　当前结果评分：-101.911148　处理用时：10 ms

教育部 部长 周济 说 ， 今后 若干 年 内 ， 毕业生 总量 将 会 持续 增加 ， 每年 都 有 数十万 的 增量 ， 对 毕业生 就业 工作 无疑 是 巨大 的 挑战 。 2005年 高校 毕业生 总量 大 、 增幅 高 ， 各地 和 高校 工作 进展 情况 差异 较 大 ， 并且 还 存在 许多 深 层次 的 矛盾 和 问题 。 在 全 社会 就业 形势 十分 严峻 的 形势 下 ， 2005年 高校 毕业生 就业 工作 压力 十分 突出 。

Constructing Inverted Index File

3. Constructing inverted index and inverted index file
   - Construct the set of keywords
   - Construct inverted index for each keyword
   - Write all the inverted indexes into files

Searching keywords

- First, search the word in the inverted index file
- Second, if the keyword is found, then get the corresponding inverted table and finally get the record.
- Usually other indexing structures(such as dictionary) are used to index on keywords efficiently.
  - Trie
  - Hash

Pros and Cons

- Efficient indexing, suitable for text database system
- Indexing type that it supports is limited
  - Key value (only can be keywords of documents)
  - Inverted index might not be efficient
  - High space cost

Discussion

- How to organize inverted index effectively?

- If a keyword appears more than once in the same document, is it ok to merge its entries in the inverted index?

Data Structures and Algorithms

Thanks

the National Elaborate Course (Only available for IPs in China)
http://www.jpk.pku.edu.cn/pkujpk/course/sjjg/
Ming Zhang, Tengjiao Wang and Haiyan Zhao
Higher Education Press, 2008.6 (awarded as the "Eleventh Five-Year" national planning textbook)