

Note:

期中 / 期末 考试

平时 10% / 大作业 30% / 期中 20% / 期末 40%

Introduction

[Information retrieval](#) (Wikipedia)

Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing.

广义：满足人们对信息需求的学问

狭义：信息搜索技术

Main Task

1. 搜索 (Ad-hoc Search)
2. 过滤 (Filtering)
3. 分类
4. 问题回答

Basic Problems

- 相关性问题
- 性能问题
- 测试问题 (如何衡量准确性)
- 人机交互问题

Boolean Retrieval

用户的信息需求通过词和布尔代数表示

Example

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

Figure 1.1 A term-document incidence matrix. Matrix element (t, d) is 1 if the play in column d contains the word in row t , and is 0 otherwise.

To answer the query Brutus AND Caesar AND NOT Calpurnia, we take the vectors for Brutus, Caesar and Calpurnia, complement the last, and then do a bitwise AND:

$$110100 \text{ AND } 110111 \text{ AND } 101111 = 100100$$

The answers for this query are thus *Antony and Cleopatra* and *Hamlet* (Figure 1.2).

缺陷

关系矩阵稀疏造成空间浪费

改进

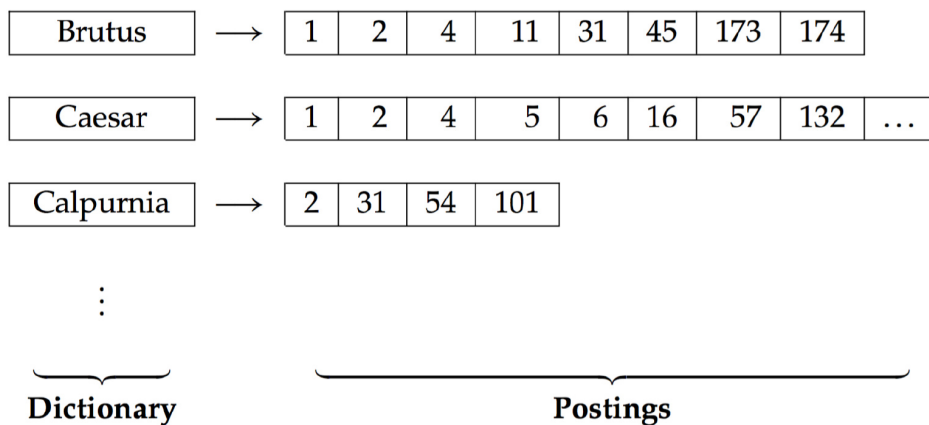


Figure 1.3 The two parts of an inverted index. The dictionary is commonly kept in memory, with pointers to each postings list, which is stored on disk.

在此基础上进行集合运算以进行文档检索

PS: 取交集过程——边遍历边合并 [多个Postings的时候, 先取长度短的相互扫描]

倒排索引的创建

Doc 1

I did enact Julius Caesar: I was
killed i' the Capitol; Brutus killed
me.

Doc 2

So let it be with Caesar. The noble
Brutus hath told you Caesar was
ambitious:

term	docID	term	docID
I	1	ambitious	2
did	1	be	2
enact	1	brutus	1
julius	1	brutus	2
caesar	1	capitol	1
I	1	caesar	1
was	1	caesar	2
killed	1	caesar	2
i'	1	did	1
the	1	enact	1
capitol	1	hath	1
brutus	1	I	1
killed	1	I	1
me	1	i'	1
so	2	it	2
let	2	julius	1
it	2	killed	1
be	2	killed	1
with	2	let	2
caesar	2	me	1
the	2	noble	2
noble	2	so	2
brutus	2	the	1
hath	2	the	2
told	2	told	2
you	2	you	2
caesar	2	was	1
was	2	was	2
ambitious	2	with	2

term	doc. freq.	→	postings lists
ambitious	1	→	2
be	1	→	2
brutus	2	→	1 → 2
capitol	1	→	1
caesar	2	→	1 → 2
did	1	→	1
enact	1	→	1
hath	1	→	2
I	1	→	1
i'	1	→	1
it	1	→	2
julius	1	→	1
killed	1	→	1
let	1	→	2
me	1	→	1
noble	1	→	2
so	1	→	2
the	2	→	1 → 2
told	1	→	2
you	1	→	2
was	2	→	1 → 2
with	1	→	2