Information Retrieval (5LN712)

Boolean Retrieval

Ali Basirat

Department of Linguistics and Philology
Uppsala University

March 30, 2020

Table of contents

Boolean Retrieval

- Index documents with their terms
- Take a Boolean query and process it
- Return a set of relevant documents

Indexing

- Construct a term document incidence matrix
- Each column is a binary vector representing a document
- Each row is a binary vector representing a term

Example

	Anthony 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
...						

Boolean Query

- A query is what an end user conveys to the computer to communicate an information need
- A Boolean query consists of terms and Boolean (logical) operators (AND, OR, NOT, etc.) between them

Processing Boolean Queries

Algorithm 1 Finding relevant documents for a Boolean query

Input: A Boolean query q and a term-doc. incidence matrix M

Output: A set of document ids

1: Select row vectors of M corresponding to the terms of q
2: Apply Boolean operators defined by q on the selected rows
3: Return document ids whose values in the final Boolean vector is 1

Processing Boolean Queries

Example

- q: `Brutus` AND `Caesar` AND NOT `Calpurnia`
- M:

	Anthony 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
...						

Processing Boolean Queries

Example

- q: `Brutus` AND `Caesar` AND NOT `Calpurnia`
- M:

	Anthony 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
...						

Processing Boolean Queries

Example

- q: `Brutus` AND `Caesar` AND NOT `Calpurnia`
- M:

	Anthony 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
...						

Processing Boolean Queries

Example

- q: `Brutus` AND `Caesar` AND NOT `Calpurnia`
- M:

	Anthony 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
...						

Processing Boolean Queries

Example

- q: `Brutus` AND `Caesar` AND NOT `Calpurnia`
- M:

	Anthony 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
...						

- $[110100] \wedge [110111] \wedge \neg[010000] = [100100]$

Processing Boolean Queries

Example

- q: `Brutus` AND `Caesar` AND NOT `Calpurnia`
- M:

	Anthony 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
...						

- $[110100] \wedge [110111] \wedge \neg[010000] = [100100]$
- **return** {Anthony and Cleopatra, Hamlet}

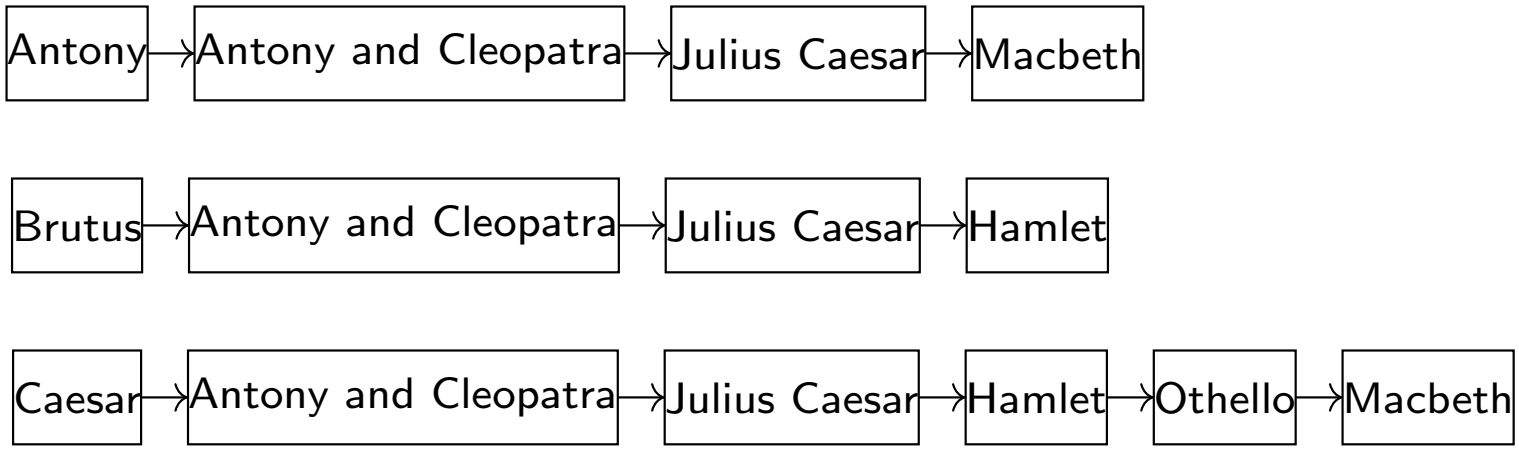
Term-document Incidence Matrix

- A very large and sparse matrix
- Only ones are stored
- Terms are associated with the documents they occur in (inverted index)

Table of contents

Inverted Index

Each term is associated with a list, called a posting list, whose elements are the documents in which the term is occurred



...

Inverted Index

The major steps to construct an inverted index:

1. collect the documents to be indexed
2. tokenize the text
3. do linguistic pre-processing (token normalisation)
4. index the documents

Inverted Index

- **Doc1**: I did enact Julius Caesar: I was killed i' the Capitol; Brutus killed me.
- **Doc2**: So let it be with Caesar. The noble Brutus hath told you Caesar was ambitious:

Inverted Index

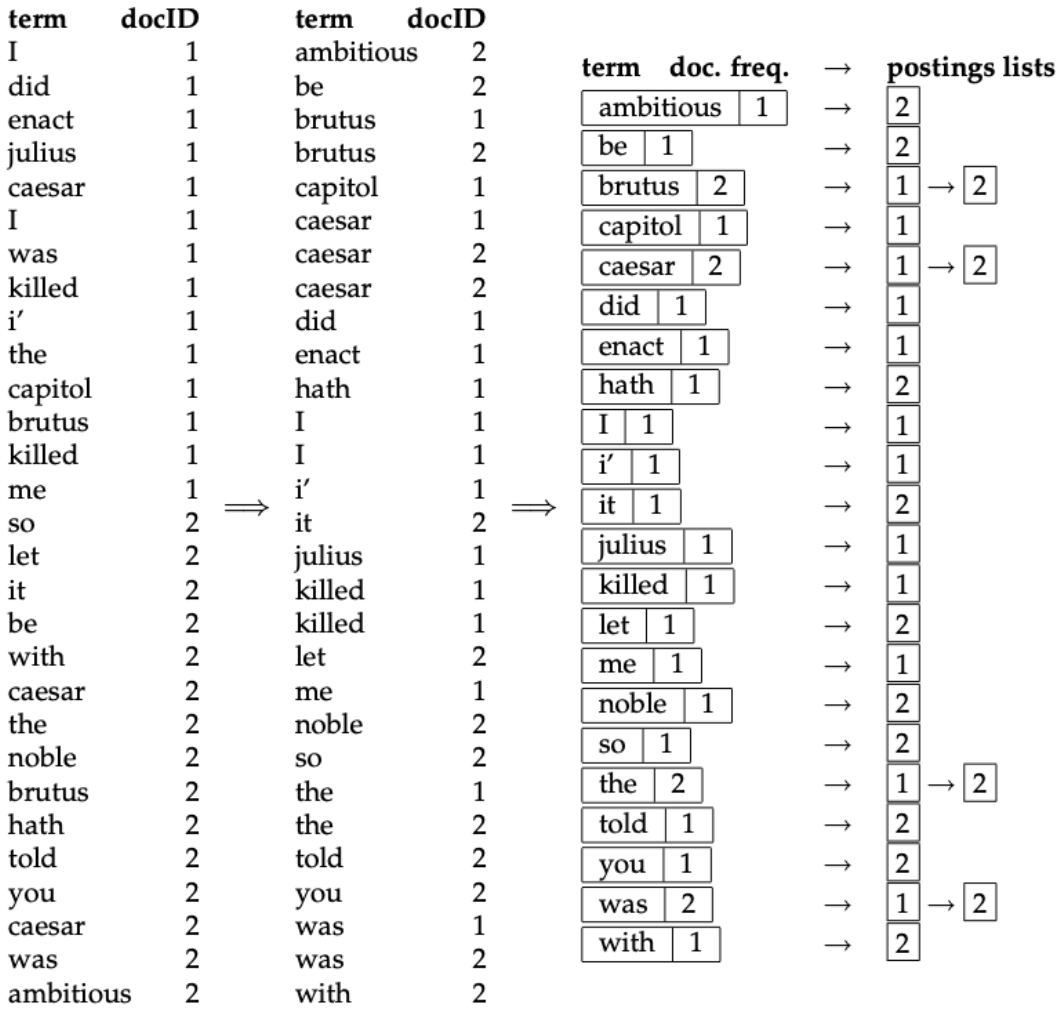


Figure: An example of the inverted index [1]

Table of contents

Processing Boolean Queries

The use of an inverted index instead of a term-document matrix entails a move from Boolean algebra to the set theory, which are related to each other.

Example

- $\wedge \implies \bigcap$ (intersect)
- $\vee \implies \bigcup$ (union)
- $\neg \implies -$ (subtract)

Processing Boolean Queries

Definition

Conjunctive query: a query consisting of terms and logical AND ($\wedge$) operator in between

Problem

How to process a conjunctive query using inverted indices?

Solution

- *locate terms in the dictionary*
- *retrieve their postings*
- *intersect the postings*

Processing Boolean Queries

How to intersect two postings?

```
INTERSECT(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then p1 ← next(p1)
 9                else p2 ← next(p2)
10  return answer
```

Figure: The intersect algorithm. It is assumed that the postings are sorted in advance.

Processing Boolean Queries

How to process a conjunctive query consisting of an arbitrary number of terms?

```
INTERSECT(⟨t_1, ..., t_n⟩)
1  terms ← SORTBYINCREASINGFREQUENCY(⟨t_1, ..., t_n⟩)
2  result ← postings(first(terms))
3  terms ← rest(terms)
4  while terms ≠ NIL and result ≠ NIL
5  do result ← INTERSECT(result, postings(first(terms)))
6     terms ← rest(terms)
7  return result
```

Figure: The extended intersect algorithm.

Processing Boolean Queries

Skip-pointers in postings lists

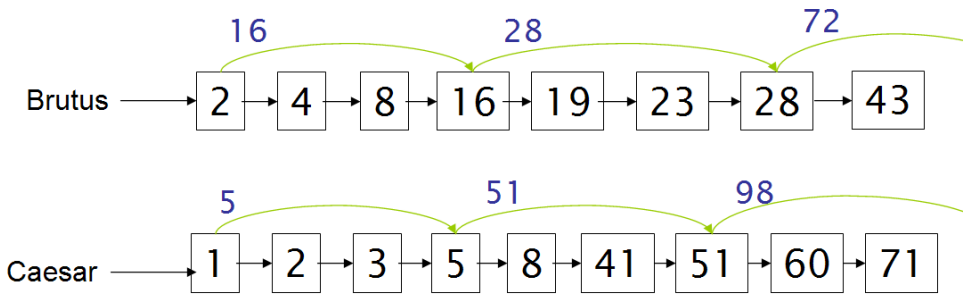


Figure: Postings lists with skip pointers

Processing Boolean Queries

The intersection of two postings lists augmented with skip-pointers

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

Figure: Postings lists intersection with skip pointers.

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

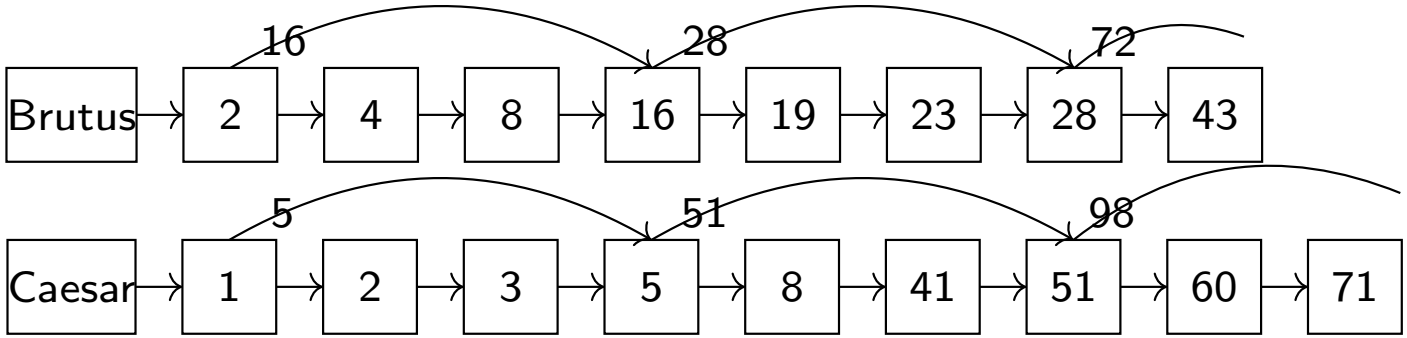


Figure: Postings lists with skip pointers

answer={}

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

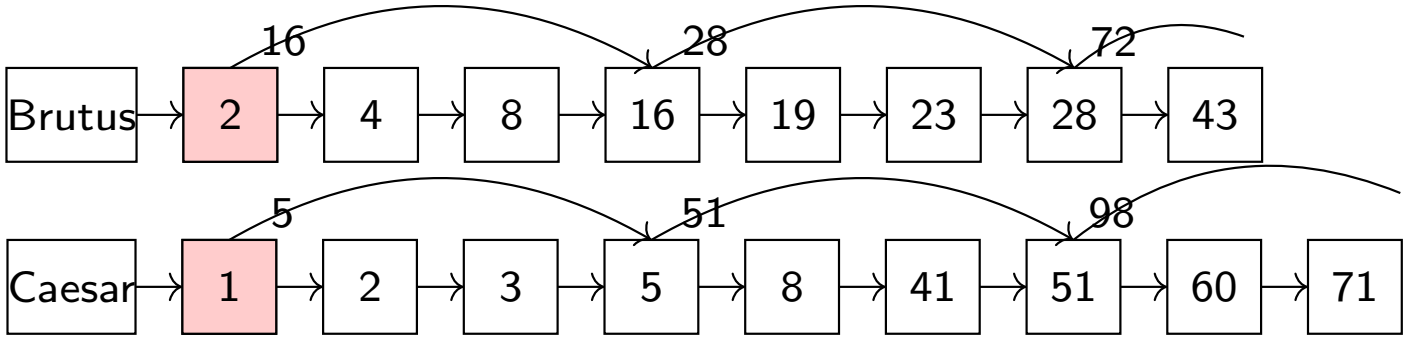


Figure: Postings lists with skip pointers

answer={}

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

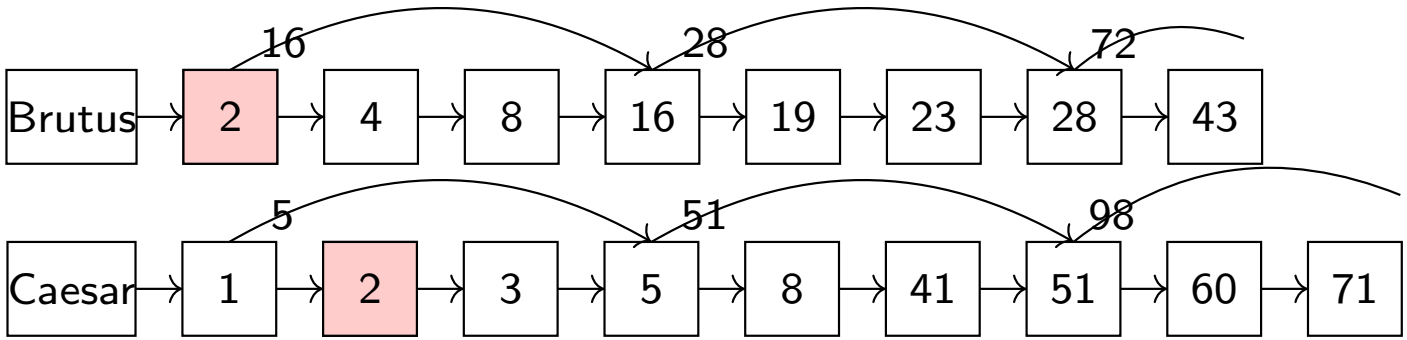


Figure: Postings lists with skip pointers

answer={}

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

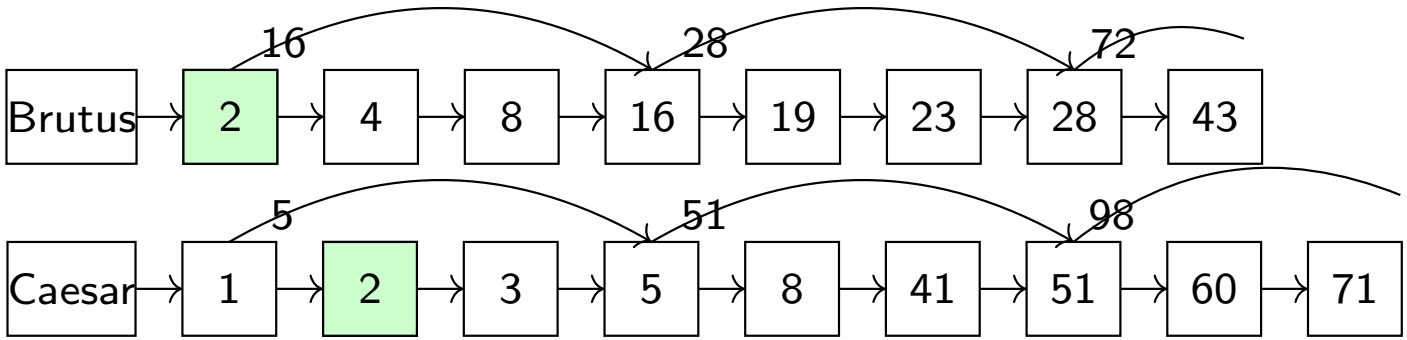


Figure: Postings lists with skip pointers

answer=$\{2\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

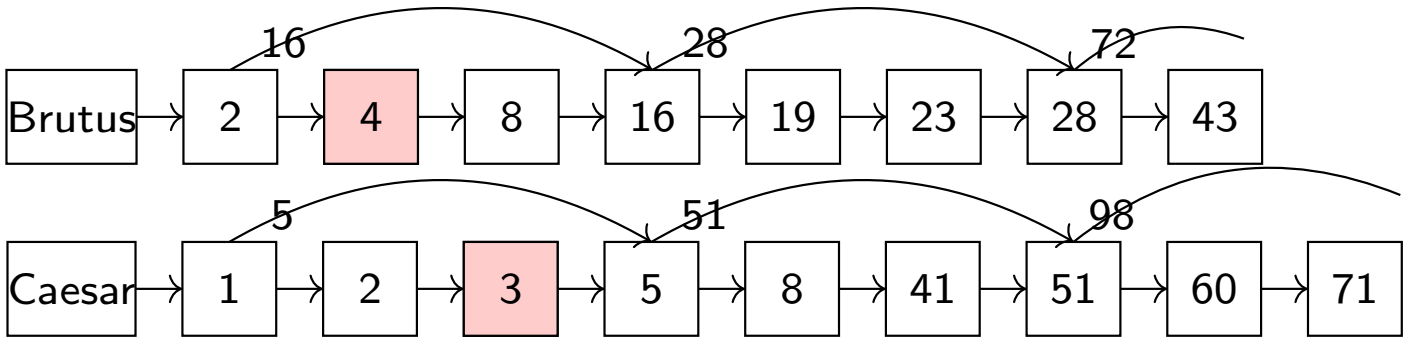


Figure: Postings lists with skip pointers

answer=$\{2\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

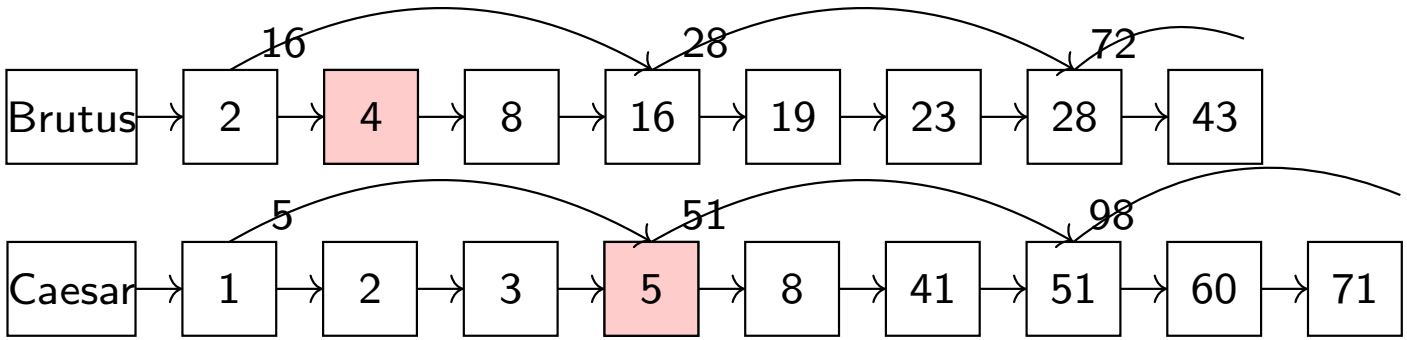


Figure: Postings lists with skip pointers

answer=$\{2\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

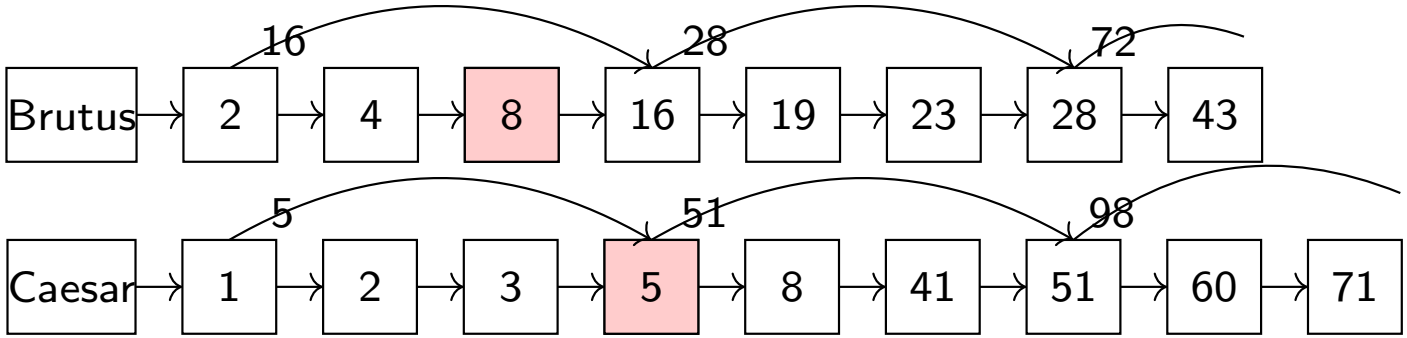


Figure: Postings lists with skip pointers

answer=$\{2\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

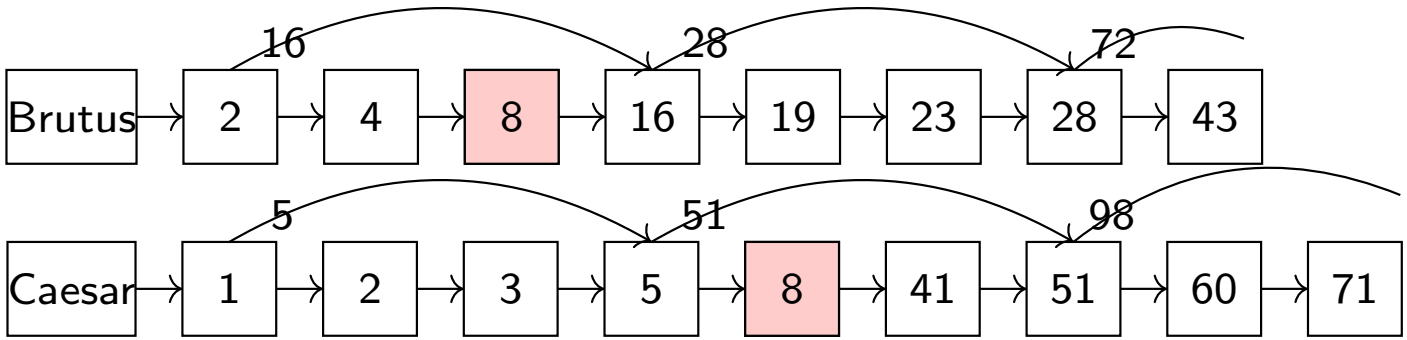


Figure: Postings lists with skip pointers

answer=$\{2\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

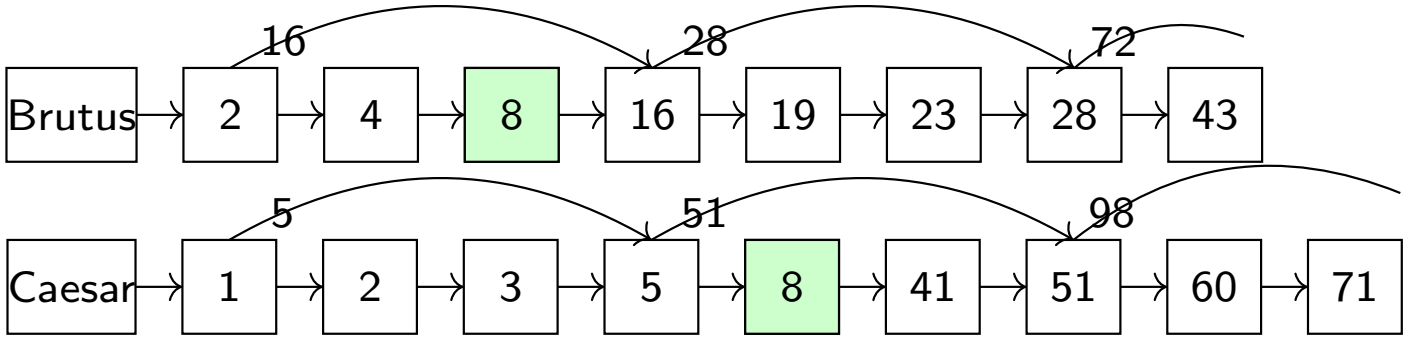


Figure: Postings lists with skip pointers

answer=$\{2, 8\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

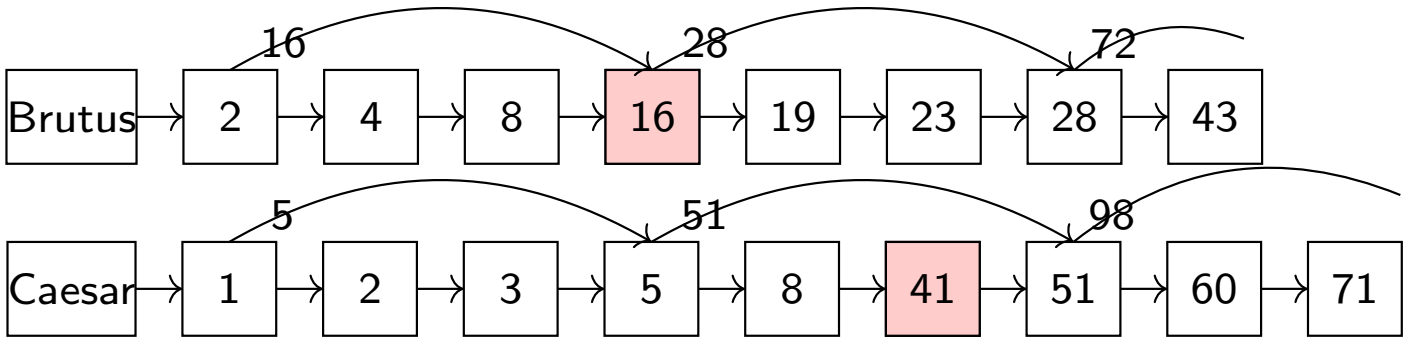


Figure: Postings lists with skip pointers

answer=$\{2, 8\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

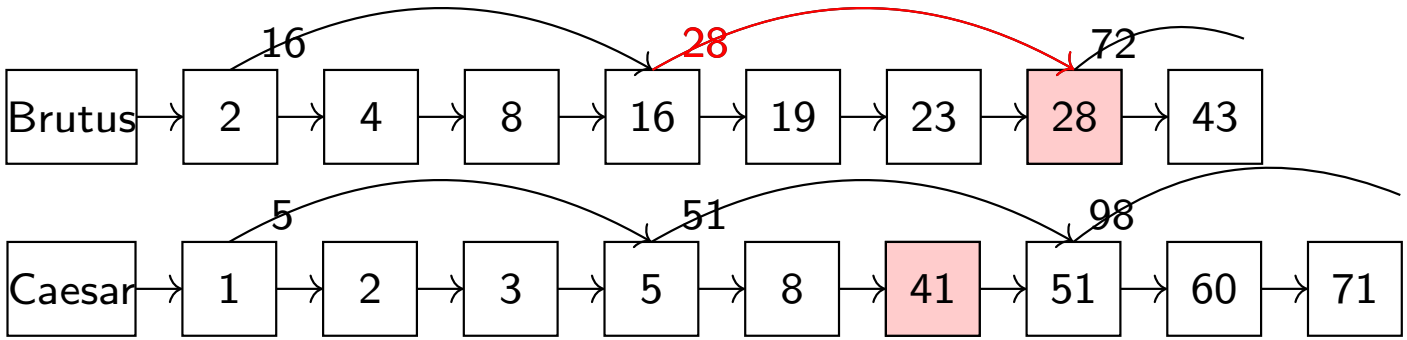


Figure: Postings lists with skip pointers

answer=$\{2, 8\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

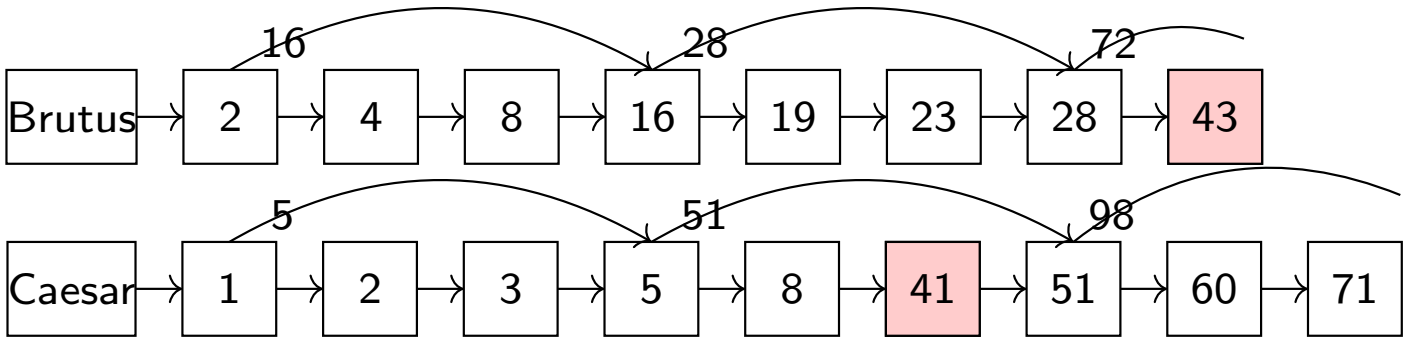


Figure: Postings lists with skip pointers

answer=$\{2, 8\}$

Processing Boolean Queries

```
INTERSECTWITHSKIPS(p1, p2)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then ADD(answer, docID(p1))
 5             p1 ← next(p1)
 6             p2 ← next(p2)
 7        else if docID(p1) < docID(p2)
 8                then if hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
 9                     then while hasSkip(p1) and (docID(skip(p1)) ≤ docID(p2))
10                          do p1 ← skip(p1)
11                     else p1 ← next(p1)
12                else if hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
13                     then while hasSkip(p2) and (docID(skip(p2)) ≤ docID(p1))
14                          do p2 ← skip(p2)
15                     else p2 ← next(p2)
16  return answer
```

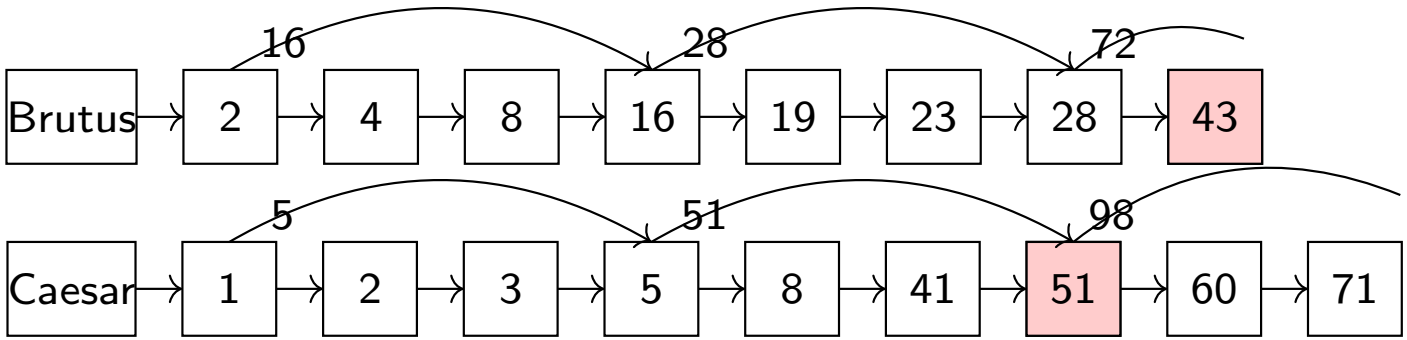


Figure: Postings lists with skip pointers

answer=$\{2, 8\}$

Processing Boolean Queries

Where to put the skip pointers?

- More skip pointers means shorter skip span, and we are more likely to skip.
- It also means lots of comparisons and lots of spaces to store pointers.
- Fewer pointers means long skip span, and that we are less likely to skip.
- Fewer comparisons and smaller space storing the pointers.
- Use $\sqrt{P}$ evenly-spaced pointers for a postings list of length P

Processing Boolean Queries

Definition

Disjunctive query: a query consisting of terms and logical OR ($\vee$) operator in between

Problem

How to process a disjunctive query using inverted indices?

Solution

- *locate terms in the dictionary*
- *retrieve their postings*
- *union the postings*

Processing Boolean Queries

How to union two postings?
How to process a disjunctive query consisting of an arbitrary number of terms?

Processing Boolean Queries

Definition

Conjunction normal form (CNF): a Boolean expression is in conjunction normal form if it is a conjunction of some disjunction literals (i.e., it is **AND** of **OR**s)

Corollary

Every Boolean expression can be converted to CNF.

Processing Boolean Queries

How to convert a Boolean expression consisting of ANDs, ORs, and NOTs into the CNF format?

1. Push negations inward by repeatedly applying De Morgan's Law:
   - $\neg(P \vee Q) = \neg P \wedge \neg Q$
   - $\neg(P \wedge Q) = \neg P \vee \neg Q$
   - $\neg\neg P = P$
2. Push ORs inward over ANDs by repeatedly apply the distributive law
   - $P \vee (Q \wedge R) = (P \vee Q) \wedge (P \vee R)$

Processing Boolean Queries

If we can process CNF expressions then we can handle arbitrary Boolean queries.
Can we process CNF expressions?

- CNF expressions are conjunctions of disjunctions
- we know how to process a conjunctive query
- we know how to process a disjunctive query

Summary

- Boolean information retrieval
- Indexing mechanism
- What is a Boolean query
- How to process Boolean queries
- Term-document Incidence Matrix
- Inverted Index
- Skip-pointers
- Toward processing arbitrary Boolean queries

Exercises

- Write out a postings merge algorithm for x OR y.
- How should the Boolean query x AND NOT y be handled?
- *Query optimization* is the process of selecting how to organize the work of answering a query so that the least total amount of work needs to be done by the system. Recommend a query processing order for: `(tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes)` given the following postings

list sizes:

term	**postings size**
eyes	213312
kaleidoscope	87009
marmalade	107913
skies	271658
tangerine	46653
trees	316812

References I

Manning, Christopher D. and Raghavan, Prabhakar and Schütze, Hinrich.
Introduction to Information Retrieval.
Cambridge University Press, 2008.

Wikipedia Article.
Conjunctive normal form
https://en.wikipedia.org/wiki/Conjunctive_normal_form