

# Similarity Search

## The q-Gram Distance

Nikolaus Augsten

nikolaus.augsten@sbg.ac.at  
Department of Computer Sciences  
University of Salzburg



WS 2021/22

Version October 26, 2021

## Outline

- 1 Filters for the Edit Distance
  - Motivation
  - Lower Bound Filters
  - Length Filter
  - q-Grams: Count Filter
  - q-Grams: Position Filtering
  - Experiments
- 2 The q-Gram Distance
- 3 Conclusion

## Outline

- 1 Filters for the Edit Distance
  - Motivation
  - Lower Bound Filters
  - Length Filter
  - q-Grams: Count Filter
  - q-Grams: Position Filtering
  - Experiments
- 2 The q-Gram Distance
- 3 Conclusion

## Application Scenario

- Scenario:
  - A company offers a number of services on the Web.
  - You can subscribe for each service independently.
  - Each service has its own database (no unique key across databases).

- Example: customer tables of two different services:

A			B		
ID	name	...	ID	name	...
1023	Frodo Baggins	...	948483	John R. R. Tolkien	...
21	J. R. R. Tolkien	...	153494	C. S. Lewis	...
239	C.S. Lewis	...	494392	Fordo Baggins	...
863	Bilbo Baggins	...	799294	Biblo Baggins	...
...	...	...	...	...	...

- Task: Created unified customer view!

## The Join Approach

- **Solution:** Join customer tables on name attribute (Q1):

```
SELECT * FROM A,B
WHERE A.name = B.name
```

- **Exact Join:** Does not work!
- **Similarity Join:** Allow  $k$  errors...
  - (1) Register UDF (User Defined Function) for the edit distance:

$$\text{ed}(x, y)$$

returns the union cost edit distance between the strings  $x$  and  $y$ .

- (2) Rewrite query Q1 as similarity join (Q2):

```
SELECT * FROM A,B
WHERE ed(A.name, B.name) <= k
```

## Effectiveness and Efficiency of the Approximate Join

- **Effectiveness:** Join result for  $k = 3$ :

ID	name	ID	name
1023	Frodo Baggins	494392	Fordo Baggins
21	J. R. R. Tolkien	948483	John R. R. Tolkien
239	C.S. Lewis	153494	C. S. Lewis
863	Bilbo Baggins	799294	Biblo Baggins

⇒ very good (100% correct)

- **Efficiency:** How does the DB evaluate the query?
  - (1) compute  $A \times B$
  - (2) evaluate UDF on each tuple  $t \in A \times B$
- **Prohibitive runtime!**

## Using a Filter for Search Space Reduction

- **Search space:**  $A \times B$  ( $\Rightarrow |A| \cdot |B|$  edit distance computations)
- **Filtering (Pruning):** Remove tuples that can not match, without actually computing the distance.

## Outline

- 1 Filters for the Edit Distance
  - Motivation
  - Lower Bound Filters
  - Length Filter
  - $q$ -Grams: Count Filter
  - $q$ -Grams: Position Filtering
  - Experiments
- 2 The  $q$ -Gram Distance
- 3 Conclusion

## Filter Properties

- Error Types:

		Correct Result	
		positive	negative
Filter Test	positive	true positive	false positive
	negative	false negative	true negative

- Example: "Are  $x$  and  $y$  within edit distance  $k$ ?"
  - Correct result: compute edit distance and test  $\text{ed}(x, y) \leq k$
  - Filter test: give answer without computing edit distance
  - False negatives:  $x$  and  $y$  are pruned although  $\text{ed}(x, y) \leq k$ .
  - False positives:  $x$  and  $y$  are not pruned although  $\text{ed}(x, y) \not\leq k$ .
- Good filters have
  - no false negatives (i.e., miss no correct results)
  - few false positive (i.e., avoid unnecessary distance computations)

## Lower Bound Filters

- Lower bound (lb) for distance  $\text{dist}(x, y)$ :

$$\text{dist}(x, y) \geq \text{lb}_{\text{dist}}(x, y)$$

- Query Q3 with Lower Bound Filter :
 

```
SELECT * FROM A,B
WHERE lb(A.name, B.name) <= k AND
      ed(A.name, B.name) <= k
```

  - $\text{lb}(A.name, B.name)$  is a cheap function
  - database will optimize query: compute  $\text{ed}(A.name, B.name)$  only if  $\text{lb}(A.name, B.name) \leq k$
- No false negatives!

## Outline

- Filters for the Edit Distance
  - Motivation
  - Lower Bound Filters
  - Length Filter
    - $q$ -Grams: Count Filter
    - $q$ -Grams: Position Filtering
    - Experiments
- The  $q$ -Gram Distance
- Conclusion

## Length Filtering

Theorem (Length Filtering [GIJ<sup>+</sup>01])

If two strings  $x$  and  $y$  are within edit distance  $k$ , their lengths cannot differ by more than  $k$ :

$$\text{ed}(x, y) \geq \text{abs}(|x| - |y|)$$

- Proof: At least  $\text{abs}(|x| - |y|)$  inserts are needed to bring  $x$  and  $y$  to the same length.  $\square$
- Query Q4 with Length Filtering:
 

```
SELECT * FROM A,B
WHERE ABS(LENGTH(A.name)-LENGTH(B.name)) <= k AND
      ed(A.name, B.name) <= k
```

## Example: Length Filtering

- Execute query without/with length filter ( $k = 3$ ):

A		B	
ID	name	ID	name
1023	Frodo Baggins <sub>13</sub>	948483	John R. R. Tolkien <sub>18</sub>
21	J. R. R. Tolkien <sub>16</sub>	153494	C. S. Lewis <sub>11</sub>
239	C.S. Lewis <sub>10</sub>	494392	Fordo Baggins <sub>13</sub>
863	Bilbo Baggins <sub>13</sub>	799294	Biblo Baggins <sub>13</sub>

- Without length filter: 16 edit distance computations
- With length filter ( $k = 3$ ): 12 edit distance computations
  - J. R. R. Tolkien  $\leftrightarrow$  C. S. Lewis is pruned
  - all pairs ( $\dots$ , John R. R. Tolkien) except (J. R. R. Tolkien, John R. R. Tolkien) are pruned

## Outline

- Filters for the Edit Distance
  - Motivation
  - Lower Bound Filters
  - Length Filter
  - $q$ -Grams: Count Filter
  - $q$ -Grams: Position Filtering
  - Experiments
- The  $q$ -Gram Distance
- Conclusion

What is a  $q$ -Gram?

- Intuition:
  - slide window of length  $q$  over string  $x \in \Sigma^*$
  - characters covered by window form a  $q$ -gram
  - where window extends string: fill with dummy character  $\# \notin \Sigma$

- Example:  $x = \text{Frodo}$ ,  $q = 3$

extended: `##Frodo##`

$q$ -grams: `##F`

`#Fr`

`Fro`

`rod`

`odo`

`do#`

`o##`

- $q$ -Gram Profile  $G_x$ : bag of all  $q$ -grams of  $x$
- Profile size:  $|G_x| = |x| + q - 1$

Single Edit Operations and Changing  $q$ -Grams

- Intuition: Strings within small edit distance share many  $q$ -grams.
- How many  $q$ -grams ( $q = 3$ ) change/remain?

$x$	$ G_x $	$y$	$ G_y $	$ G_x \bowtie G_y $
peter	7	meter	7	4
peter	7	peters	8	5
peter	7	peer	6	4

- $\text{ed}(x, y) = 1 \Rightarrow |G_x \bowtie G_y| = \max(|G_x|, |G_y|) - q$

Multiple Edit Operations and Changing  $q$ -Grams

- $\text{ed}(x, y) = 1 \Rightarrow |G_x \bowtie G_y| = \max(|G_x|, |G_y|) - q$
- What if  $\text{ed}(x, y) = k > 1$ ?

$x$	$ G_x $	$y$	$ G_y $	$ G_x \bowtie G_y $
peter	7	meters	8	2
peter	7	petal	7	3

- Multiple edit operations may affect the same  $q$ -gram:

peter  $\rightarrow G_x = \{\#\#p, \#pe, \text{pet}, \text{ete}, \text{ter}, \text{er}\#, \text{r}\#\}$   
 petal  $\rightarrow G_y = \{\#\#p, \#pe, \text{pet}, \text{eta}, \text{tal}, \text{al}\#, \text{l}\#\}$

- Each edit operation affects at most  $q$   $q$ -grams.

## Count Filtering

Theorem (Count Filtering [GIJ<sup>+</sup>01])

Consider two strings  $x$  and  $y$  with the  $q$ -gram profiles  $G_x$  and  $G_y$ , respectively. If  $x$  and  $y$  are within edit distance  $k$ , then the cardinality of the  $q$ -gram profile intersection is at least

$$|G_x \bowtie G_y| \geq \max(|G_x|, |G_y|) - kq$$

- Proof (by induction):
  - true for  $k = 1$ :  $|G_x \bowtie G_y| \geq \max(|G_x|, |G_y|) - q$
  - $k \rightarrow k + 1$ : each additional edit operation changes at most  $q$   $q$ -grams.  $\square$

Implementation of  $q$ -Grams

- Given: tables  $A$  and  $B$  with schema  $(id, name)$ 
  - $id$  is the key attribute
  - $name$  is string-valued
- Compute auxiliary tables  $QA$  and  $QB$  with schema  $(id, qgram)$ :
  - each tuple stores one  $q$ -gram
  - string  $x$  of attribute  $name$  is represented by its  $|x| + q - 1$   $q$ -grams
  - $QA.id$  is the key value ( $A.id$ ) of a tuple with  $A.name = x$
  - $QA.qgram$  is one of the  $q$ -grams of  $x$
- Example:

$A$		$QA$	
$id$	$name$	$id$	$qgram$
1023	Frodo Baggins	1023	\#\#F
21	J. R. R. Tolkien	1023	\#Fr
239	C.S. Lewis	...	...
863	Bilbo Baggins	21	\#\#J
		21	\#J.
		...	...

## Count Filtering Query

- Query Q5 with Count Filtering:
 

```
SELECT  A.id, B.id, A.name, B.name
FROM    A, QA, B, QB
WHERE   A.id = QA.id AND
        B.id = QB.id AND
        QA.qgram = QB.qgram AND
        ABS(LENGTH(A.name)-LENGTH(B.name)) <= k
GROUP BY A.id, B.id, A.name, B.name
HAVING  COUNT(*) >= LENGTH(A.name)-1-(k-1)*q AND
        COUNT(*) >= LENGTH(B.name)-1-(k-1)*q AND
        ed(A.name, B.name) <= k
```

## Problem with Count Filtering Query

- Previous query Q5 works fine for  $kq < \max(|G_x|, |G_y|)$ .
- However: If  $kq \geq \max(|G_x|, |G_y|)$ , no  $q$ -grams may match even if  $\text{ed}(x, y) \leq k$ .
- Example ( $q = 3, k = 2$ ):  
WHERE-clause prunes  $x$  and  $y$ , although  $\text{ed}(x, y) \leq k$   
 $x = \text{IBM} \quad G_x = \{\#\#I, \#IB, \text{IBM}, \text{BM}\#, \text{M}\#\# \} \quad |G_x| = 5$   
 $y = \text{BMW} \quad G_y = \{\#\#B, \#BM, \text{BMW}, \text{MW}\#, \text{W}\#\# \} \quad |G_y| = 5$
- False negatives:
  - short strings with respect to edit distance (e.g.,  $|x| = 3, k = 3$ )
  - even if within given edit distance, matches tend to be meaningless (e.g.,  $\text{abc}$  and  $\text{xyz}$  are within edit distance  $k = 3$ )

## Fixing Count Filtering Query

- Fix query to avoid false negatives [GIJ<sup>+</sup>03]:
  - Join pairs  $(x, y)$  with  $kq \geq \max(|G_x|, |G_y|)$  using only length filter.
  - Union results with results of previous query Q5.
- Query Q6 **without false negatives** (extends previous query Q5):  

```

...
UNION
SELECT A.id, B.id, A.name, B.name
FROM   A, B
WHERE  LENGTH(A.name)+q-1 <= k*q AND
        LENGTH(B.name)+q-1 <= k*q AND
        ABS(LENGTH(A.name) - LENGTH(B.name)) <= k AND
        ed(A.name,B.name) <= k

```
- **Note:** We omit this part in subsequent versions of the query since it remains unchanged.

## Outline

- 1 Filters for the Edit Distance
  - Motivation
  - Lower Bound Filters
  - Length Filter
  - q-Grams: Count Filter
  - q-Grams: Position Filtering
  - Experiments
- 2 The q-Gram Distance
- 3 Conclusion

## Positional q-Grams

- Enrich  $q$ -grams with position information:
  - extended string: prefix and suffix string  $x$  with  $q - 1$  characters  $\#$
  - slide window of length  $q$  over extended string  $x'$
  - characters covered by window after shifting it  $i$  times form the  $q$ -gram at position  $i + 1$
- **Example:**  $x = \text{Frodo}$   
 extended string: `##Frodo##`  
 positional  $q$ -grams:
 

```

(1,##F)
(2,#Fr)
(3,Fro)
(4,r od)
(5,o do)
(6,d o#)
(7,o ##)

```

## Computing Positional q-Grams in SQL

- Given: table  $N$ 
  - $N$  has a single attribute  $i$
  - $N$  is filled with numbers from 1 to  $max$  ( $max$  is the maximum string length plus  $q - 1$ )
- Positional  $q$ -grams for table  $A$  in SQL (Q7):

```
CREATE TABLE QA AS
  SELECT A.id, N.i AS pos,
         SUBSTRING(CONCAT(
           SUBSTRING('#..#', 1, q - 1),
           LOWER(A.name),
           SUBSTRING('#..#', 1, q - 1)),
         N.i, q) AS qgram
  FROM A, N
  WHERE N.i <= LENGTH(A.name) + q - 1
```

## Corresponding q-Grams

- Corresponding  $q$ -gram:
  - Given: positional  $q$ -grams  $(i, g)$  of  $x$
  - transform  $x$  to  $y$  applying edit operations
  - $(i, g)$  "becomes"  $(j, g)$  in  $y$
  - We define:  $(i, g)$  *corresponds* to  $(j, g)$
- Example:
  - $x' = \# \# \text{abaZabaabaaba} \# \#$ ,  $y' = \# \# \text{abaabaabaaba} \# \#$
  - edit distance is 1 (delete Z from  $x$ )
  - $(7, \text{aba})$  in  $x$  corresponds to  $(6, \text{aba})$  in  $y$
  - ... but not to  $(9, \text{aba})$

## Position Filtering

### Theorem (Position Filtering [GIJ<sup>+</sup>01])

*If two strings  $x$  and  $y$  are within edit distance  $k$ , then a positional  $q$ -gram in one cannot correspond to a positional  $q$ -gram in the other that differs from it by more than  $k$  positions.*

- Proof:
  - each increment (decrement) of a position requires an insert (delete);
  - a shift by  $k$  positions requires  $k$  inserts/deletes.  $\square$

## Position Filtering

- Query Q8 with Count and Position Filtering:

```
SELECT  A.id, B.id, A.name, B.name
FROM    A, QA, B, QB
WHERE   A.id = QA.id AND
        B.id = QB.id AND
        QA.qgram = QB.qgram AND
        ABS(LENGTH(A.name)-LENGTH(B.name)) <= k AND
        ABS(QA.pos-QB.pos) <= k
GROUP BY A.id, B.id, A.name, B.name
HAVING  COUNT(*) >= LENGTH(A.name)-1-(k-1)*q AND
        COUNT(*) >= LENGTH(B.name)-1-(k-1)*q AND
        ed(A.name, B.name) <= k
```

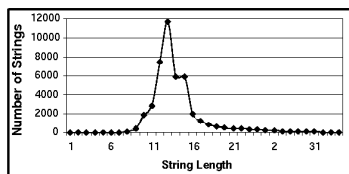
## Outline

- 1 Filters for the Edit Distance
  - Motivation
  - Lower Bound Filters
  - Length Filter
  - $q$ -Grams: Count Filter
  - $q$ -Grams: Position Filtering
  - Experiments
- 2 The  $q$ -Gram Distance
- 3 Conclusion

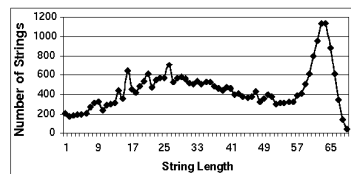
## Experimental Data

- All experimental results taken from [GIJ<sup>+</sup>01]
- Three string data sets:
  - set1: 40K tuples, average length: 14 chars
  - set2: 30K tuples, average length: 38 chars
  - set3: 30K tuples, average length: 33 chars

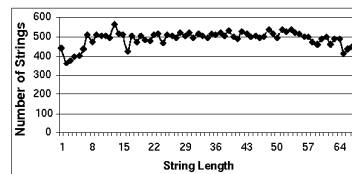
## String Length Distributions



Set 1



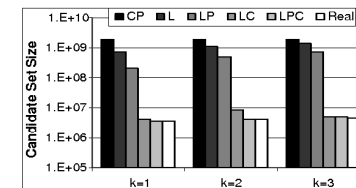
Set 2



Set 3

## Candidate Set Size

- Question: How many edit distances do we have to compute?
- Show candidate set size for different filters (small is good).
- $q = 2$
- Caption:
  - CP: cross product
  - L: length filtering, P: position filtering, C: count filtering
  - Real: number of real matches

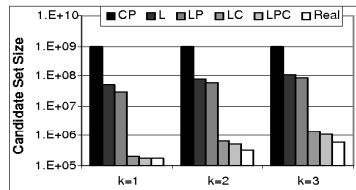


Set 1



## Candidate Set Size

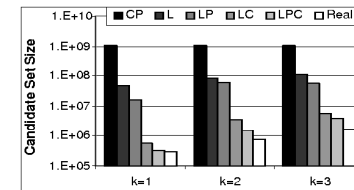
- Question: How many edit distances do we have to compute?
- Show candidate set size for different filters (small is good).
- $q = 2$
- Caption:
  - CP: cross product
  - L: length filtering, P: position filtering, C: count filtering
  - Real: number of real matches



Set 2

## Candidate Set Size

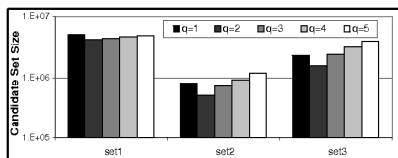
- Question: How many edit distances do we have to compute?
- Show candidate set size for different filters (small is good).
- $q = 2$
- Caption:
  - CP: cross product
  - L: length filtering, P: position filtering, C: count filtering
  - Real: number of real matches



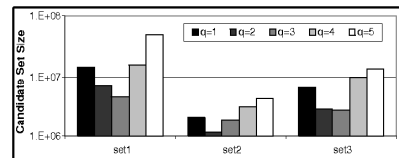
Set 3

## Various $q$ -Gram Lengths

- Question: How does the choice of  $q$  influence the filter effectiveness?
- Show candidate set size for different  $q$  values (small is good).



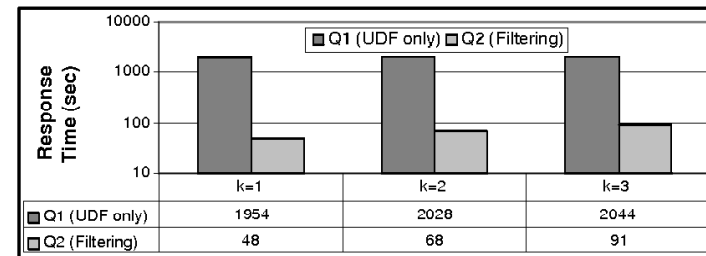
Edit Distance Threshold  $k = 2$



Edit Distance Threshold  $k = 3$

## Response Time

- Approximate self-join on small sample of 1000 tuples (set 1) (full dataset > 3 days without filters!)
- Measure response time (small is good).
- Caption:
  - $k$ : edit distance threshold
  - Q1: edit distance without filters
  - Q2: edit distance with filters



## The q-Gram Distance

### Definition (q-Gram Distance [Ukk92])

Let  $G_x$  and  $G_y$  be the  $q$ -gram profiles of the strings  $x$  and  $y$ , respectively. The *q-gram distance* between two strings is the number of  $q$ -grams in  $G_x$  and  $G_y$  that have no match in the other profile,

$$\text{dist}_q(x, y) = |G_x \uplus G_y| - 2|G_x \cap G_y|.$$

- Example:  $q = 2, x = \text{abab}, y = \text{abcb}$

$$G_x = \{\#a, ab, ba, ab, b\# \}$$

$$G_y = \{\#a, ab, bc, ca, ab, b\# \}$$

$$G_x \uplus G_y = \{\#a, ab, ba, ab, b\#, \#a, ab, bc, ca, ab, b\# \}$$

$$G_x \cap G_y = \{\#a, ab, ab, b\# \}$$

$$\text{dist}_q(x, y) = |G_x \uplus G_y| - 2|G_x \cap G_y| = 11 - 2 \cdot 4 = 3$$

## Pseudo Metric q-Gram Distance

- The  $q$ -gram distance is a **pseudo metric**:

For all  $x, y, z \in \Sigma^*$

- $\text{dist}_q(x, y) + \text{dist}_q(y, z) \geq \text{dist}_q(x, z)$  (triangle inequality)
- $\text{dist}_q(x, y) = \text{dist}_q(y, x)$  (symmetric)
- $\text{dist}_q(x, y) = 0 \Leftrightarrow x = y$

- **Note:** Identity condition relaxed:  $\text{dist}_q(x, y) = 0 \not\Rightarrow x = y$   
i.e., the  $q$ -gram distance between two *different* strings can be 0

- **Example:**

$$\text{dist}_q(\text{axybxcxyd}, \text{axycbxybxyd}) = 0$$

$$G_x = G_y = \{\#\#a, \#ax, axy, xyb, ybx, bxy, xyc, ycx, cxy, xyd, yd\#, d\#\# \}$$

## Distance Normalization (1/3)

- What is a good threshold?

$$\text{ed}(\text{International Business Machines Corporation}, \text{International Business Machine Corporation}) = 2$$

$$\text{ed}(\text{IBM}, \text{BMW}) = 2$$

$$\text{ed}(\text{Int. Business Machines Corp.}, \text{International Business Machines Corporation}) = 17$$

- Problem: Absolute numbers not always meaningful...
- Solution: Compute error relative to string length!

## Distance Normalization (2/3)

- Normalize distance such that  $\delta(x, y) \in [0..1]$
- **Edit Distance:**  $0 \leq \text{ed}(x, y) \leq \max(|x|, |y|)$
- **Normalized Edit Distance:**  $0 \leq \text{norm-ed}(x, y) \leq 1$

$$\text{norm-ed}(x, y) = \frac{\text{ed}(x, y)}{\max(|x|, |y|)}$$

- **q-Gram Distance:**  $0 \leq \text{dist}_q(x, y) \leq |G_x \uplus G_y| - |G_x \cap G_y|$
- **Normalized q-Gram Distance:**  $0 \leq \text{norm-dist}_q(x, y) \leq 1$

$$\text{norm-dist}_q(x, y) = \frac{\text{dist}_q(x, y)}{|G_x \uplus G_y| - |G_x \cap G_y|} = 1 - \frac{|G_x \cap G_y|}{|G_x \uplus G_y| - |G_x \cap G_y|}$$

- Dividing by  $|G_x \uplus G_y|$  also normalizes to  $[0..1]$ , but the metric properties (triangle inequality) get lost [ABG10].

## Distance Normalization (3/3)

- Normalized edit distance:

$\text{norm-ed}(\text{International Business Machines Corporation}, \text{International Bussiness Machine\_Corporation}) = 0.047$

$\text{norm-ed}(\text{IBM}, \text{BMW}) = 0.66$

$\text{norm-ed}(\text{Int. Business Machines Corp.}, \text{International Business Machines Corporation}) = 0.4$

- Normalized  $q$ -gram distance ( $q = 3$ ):

$\text{norm-dist}_q(\text{International Business Machines Corporation}, \text{International Bussiness Machine\_Corporation}) = 0.089$

$\text{norm-dist}_q(\text{IBM}, \text{BMW}) = 1.0$

$\text{norm-dist}_q(\text{Int. Business Machines Corp.}, \text{International Business Machines Corporation}) = 0.36$

Edit Distance vs.  $q$ -Gram Distance

- Edit distance can not handle block-moves well:

$x = \text{Nikolaus Augsten}$   $y = \text{Augsten Nikolaus}$

$\text{norm-ed}(x, y) = 1.0$

$\text{norm-dist}_q(x, y) = 0.39$  ( $q = 3$ )

- $q$ -Gram distance may be too strict:

$x = +39-06-46-74-22$   $y = (39 06 467422)$

$\text{norm-ed}(x, y) = 0.4$

$\text{norm-dist}_q(x, y) = 1.0$  ( $q = 3$ )

## Summary

- Approximate join with edit distance inefficient.
- Edit distance filters speed up join:
  - Length filter: based on the string length
  - Count filter: based on  $q$ -Grams
  - Position filter: based on positional  $q$ -Grams

-  Nikolaus Augsten, Michael Böhlen, and Johann Gamper. The  $pq$ -gram distance between ordered labeled trees. *ACM Transactions on Database Systems (TODS)*, 35(1):1–36, 2010.
-  Luis Gravano, Panagiotis G. Ipeirotis, H. V. Jagadish, Nick Koudas, S. Muthukrishnan, and Divesh Srivastava. Approximate string joins in a database (almost) for free. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 491–500, Roma, Italy, September 2001. Morgan Kaufmann Publishers Inc.
-  Luis Gravano, Panagiotis G. Ipeirotis, H. V. Jagadish, Nick Koudas, S. Muthukrishnan, and Divesh Srivastava. Approximate string joins in a database (almost) for free — Erratum. Technical Report CUCS-011-03, Department of Computer Science, Columbia University, 2003.
-  Esko Ukkonen. Approximate string-matching with  $q$ -grams and maximal matches.

*Theoretical Computer Science*, 92(1):191–211, January 1992.