Clustering of Short Strings in Large Databases

TIR 2009

M. Kazimianec (FUB)

A. Mazeika (MPII)

Outline

- Background (String Similarity, Proximity Graph (PG), GPC Method)
- Problem of Clustering Short Strings
- CLOSS. Milestones
  - Border Identification
  - Center Optimization
  - PG Smoothing

String Similarity

Each string is represented as a set (bag) of unordered q-grams;

One string is chosen as a counting point (center c);

Overlap O of the string s with the center c is computed;

Overlap O is accepted as a string similarity measure.

$$c = string,\ s_1 = strip,\ s_2 = triad;\ q = 2.$$

$$O(c, s_1) = \{\#s, st, tr, ri, in, ng, g\$\} \cap \{\#s, st, tr, ri, ip, p\$\} = 4.$$

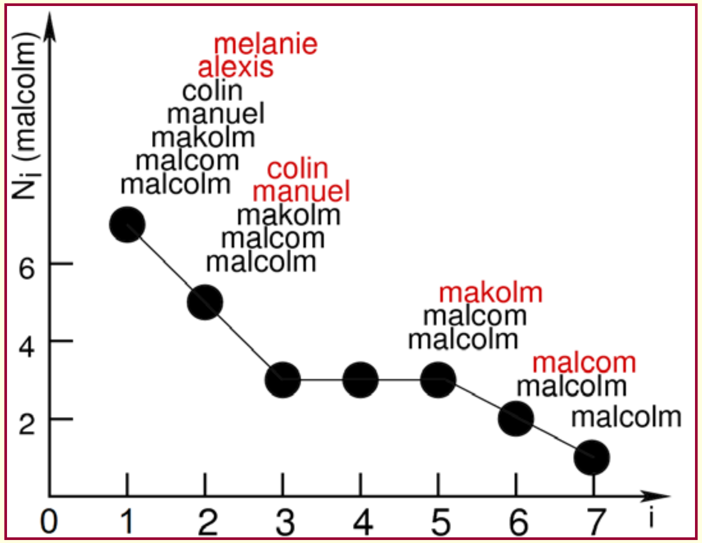
$$O(c, s_2) = \{\#s, st, tr, ri, in, ng, g\$\} \cap \{\#t, tr, ri, ia, ad, d\$\} = 2.$$

“strip” is more similar to “string” than “triad”.

Proximity Graph

Proximity graph (PG) is a discrete numerical decreasing function depending on overlap threshold expressed by the integer value i.

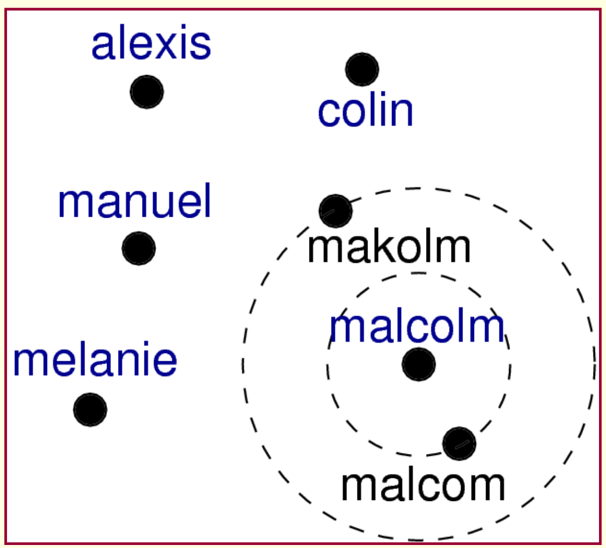
In the point i PG value is a number of strings that have overlap O with the center c not exceeding the given threshold i.

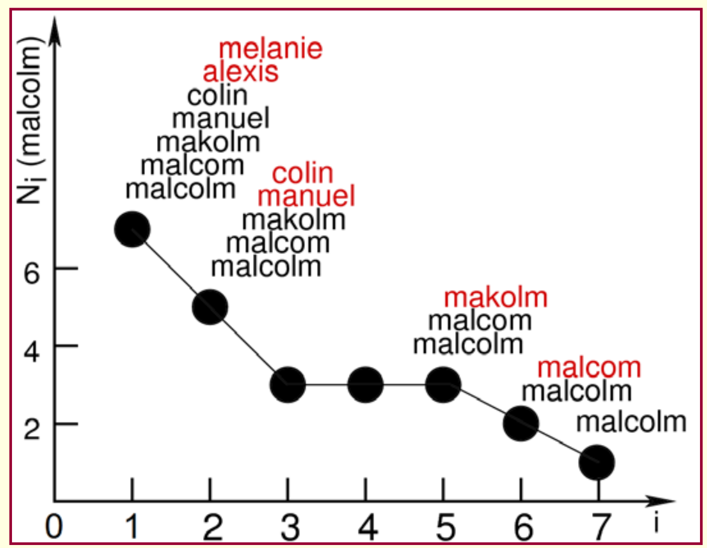


T I R 2009

GPC Method for String Clustering

GPC takes a center string and examines the shape of the proximity graph. If there is a horizontal line (overlaps 3,4,5) then GPC declares the cluster border in the extreme right point of the line (border = 5, cluster = {malcolm, malcom, makolm}).



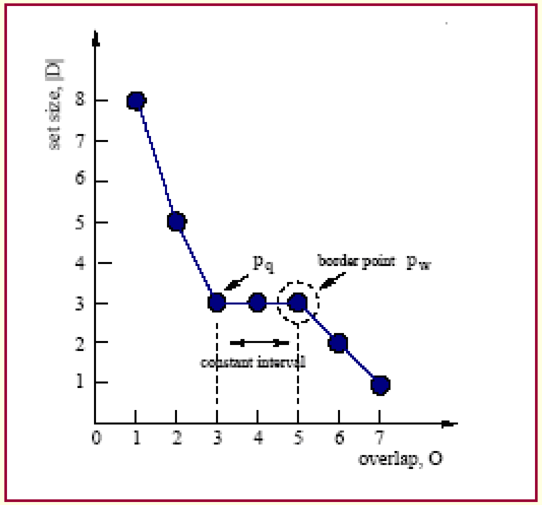


What Are the GPC Disadvantages?

GPC is weak if:

- horizontal line is not present in the PG (short strings),
- there are multiple horizontal lines in the PG (long and middle strings),
- dataset is not ordered by string length.

GPC application is cut down by following PG model:

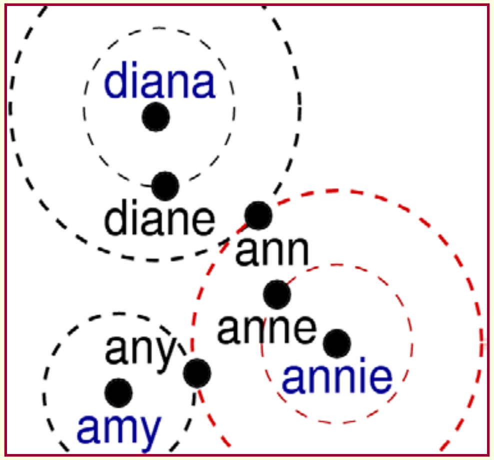


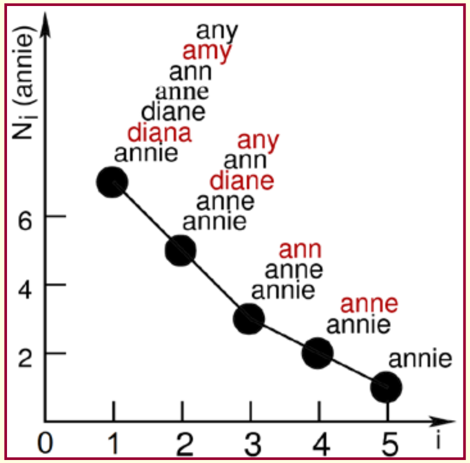
T
I
R

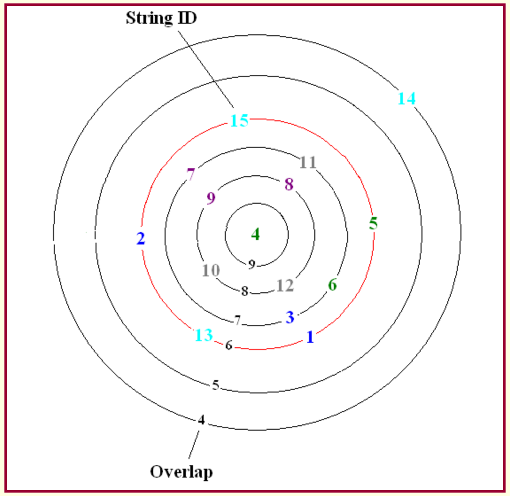
2
0
0
9

Problems of Clustering Short Strings

- Touching Clusters – PG has no horizontal lines,
- Overlapping Clusters – PG has multiple horizontal lines.







Border Identification

Oxford Dataset Sample

Overlap value

	String	q-length	s-border	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	countenance	12	8	12	8	10	6	5	6	6	6	6	5	5	6	5	3	6
2	contanance	11	8	8	11	9	6	6	6	6	6	6	6	6	6	6	4	4
3	contenance	11	9	10	9	11	6	7	7	7	7	7	7	7	7	7	4	7
4	conscience	11	6	6	6	7	11	6	7	7	8	8	8	7	8	7	4	6
5	conceince	10	6	5	6	6	6	10	7	6	6	6	6	8	5	6	3	6
6	conciance	10	7	6	7	7	7	7	10	7	6	6	6	7	6	6	4	6
7	convenience	12	9	6	6	7	7	8	6	12	9	11	7	7	7	8	5	9
8	convience	10	9	6	6	7	8	6	6	9	10	10	6	7	6	6	4	6
9	convienience	13	10	6	6	7	8	6	6	11	10	13	7	7	7	8	5	9
10	consequence	12	10	6	6	7	8	6	6	7	7	7	12	10	10	7	4	7
11	concequence	12	9	6	6	7	7	8	7	7	6	7	10	12	9	7	4	7
12	consiquence	12	9	6	6	7	8	6	6	7	7	7	10	9	12	7	4	7
13	convalescence	14	8	6	6	7	8	7	6	8	8	7	7	7	7	14	8	11
14	convalaces	11	7	3	4	4	4	4	4	4	4	4	4	4	4	8	11	7
15	convalensence	14	7	6	6	7	7	6	6	7	7	8	8	7	7	11	7	14

Blue color marks out subjective (true) clusters. Red color shows alien strings for s-border. The last is minimal overlap preserving all misspellings.

How we solve…

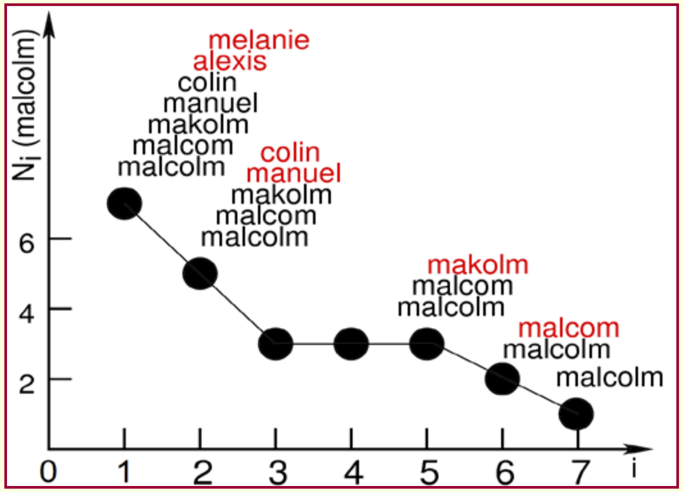
The task is to minimize the number of alien strings in the cluster maximally preserving misspellings.

The solution is related to the CLOSS method (Clustering of Short Strings)

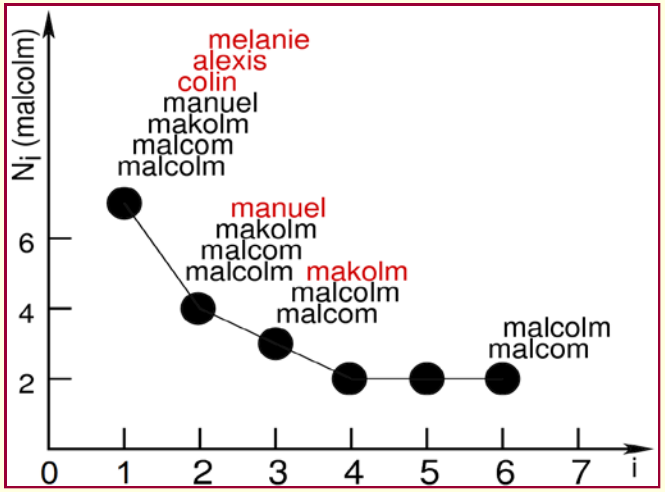
- Center optimization (by string ordering)
- Border identification
- Resolving of multiple PG lines

CLOSS. Dataset Ordering

The choice of the shorter center may lead to a PG shape without horizontal line even for long strings:



Center is malcolm



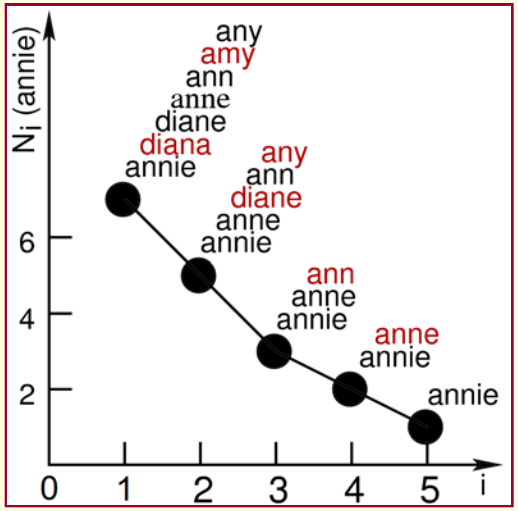
Center is malcom

Ordering by string length and clustering starting from the longest strings resolve this problem.

CLOSS. Border Interval

- Border interval $[i', i'']$ is found by means of PG interpolation by the polynomial *f(x)*.
- Starting point i' is set to the overlap value, where the curvature of *f(x)* is maximal”:

$$i' = \arg\max_i \left\{ \frac{\left| f^{(2)}(i) \right|}{\left[1 + (f^{(1)}(i))^2 \right]^{\frac{3}{2}}} \right\}.$$



- Ending point i'' is set to be $k \cdot q$ numbers of *q*-grams away from the maximal overlap

$$i'' = length(s) + q(1-k) - 1.$$

CLOSS. Border Point

Let $[i', i'']$ be the border interval. Then $i_b \in [i', i'']$ is the cluster border iff

(i) $i_b = \arg\min_i \{\Delta_i(s), i = i', i'+1, \ldots, i''\}$, where $\Delta_i(s) = N_i(s) - N_{i+1}(s) \geq 0$ is the PG neighborhood decrease at the overlap threshold i;

(ii) $\nexists$ j $< i_b$, $j \in [i', i'']: \Delta_j(s) = \Delta_{i_b}(s)$.

Defined border exists independently of the PG shape.

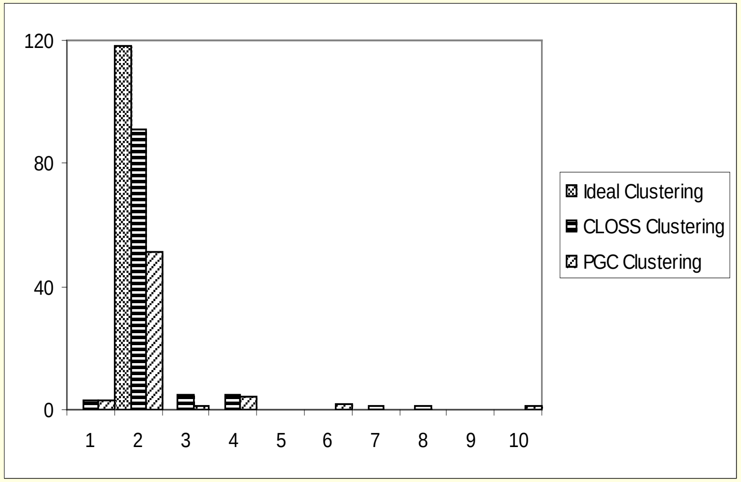
Algorithm

Input:
$D = \{s_1, s_2, ..., s_n\}$: database of ordered strings,
q: size of q-grams, r: range of smoothing.
Output:
$Clusters = \{C_1, ..., C_d\}$: clusters of strings.
Body:
1. Initialize the clustered strings
$Clustered\_Strings = \emptyset$; $Clusters = \{\emptyset\}$.
2. Scan database strings. For each $s \in D$ Do
2.1. If $s \notin Clustered\_Strings$ Then
2.1.1. Compute the proximity graph
$PG(s) = \{(1, N_1(s)), ..., (l, N_l(s))\}$ (see [1]).
2.1.2. Compute the smoothed proximity graph
$PG^{sm}(s) = Smooth(PG(s), 1, l, r)$ (see Figure 5).
2.1.3. Find the interval $[i', i'']$
2.1.4. Find the border $i_b \in PG^{sm}(s)$, $i_b \in [i', i'']$,
by computing the PG jump $\Delta_{i_b}^{sm}(s)$:
$\Delta_{i_b}^{sm}(s) = \min_i\{\Delta_i^{sm}(s)\}$, where
$\Delta_i^{sm}(s) = N_i^{sm}(s) - N_{i+1}^{sm}(s)$.
2.2 Update the clustered strings:
$Clustered\_Strings = Clustered\_Strings \cup C_{i_b}$,
C_{i_b} is such that $\forall s' \in C_{i_b} : o(s, s') \geq i_b$.
2.4 Insert a new cluster to the set of clusters:
$Clusters = Clusters \cup \{C_{i_b}\}$.
3. Return $Clusters$.

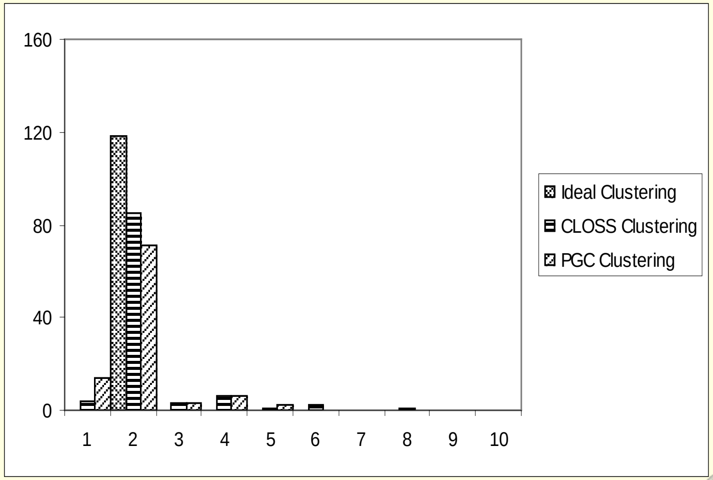
Evaluation. Clustering of the Cyclone Name Dataset

CLOSS and GPC (improved by string ordering) were compared by applying them to the cyclone name dataset (www.nhc.noaa.gov/aboutnames.html) artificially corrupted by introducing

one mistake

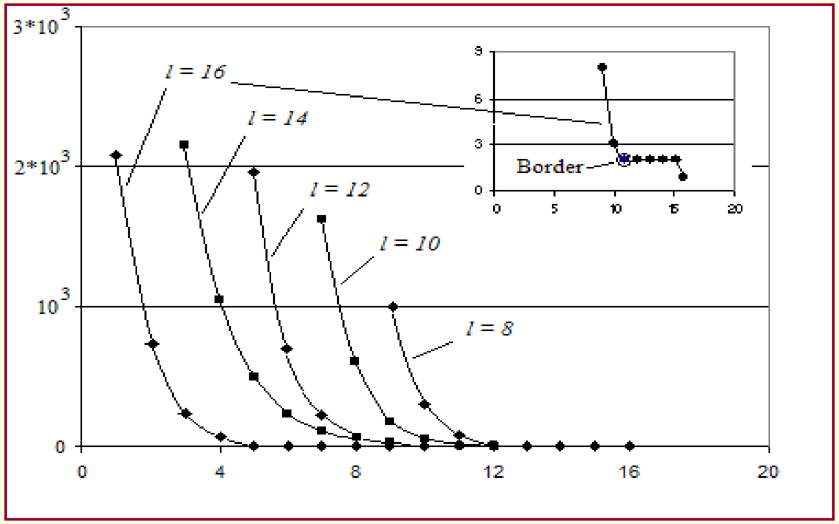


many (up to 3) mistakes

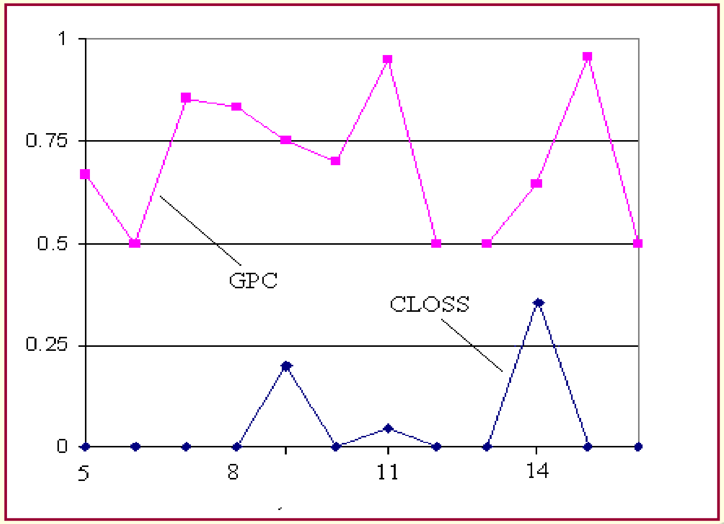


Evaluation. Text Retrieval Using Oxford Misspellings

CLOSS was used to enhance text retrieval by means of misspellings. File “birkbeck” (http://ota.ahds.ac.uk/), containing 36133 misspellings of 6136 words, was considered as a misspelling source.



PG Shapes



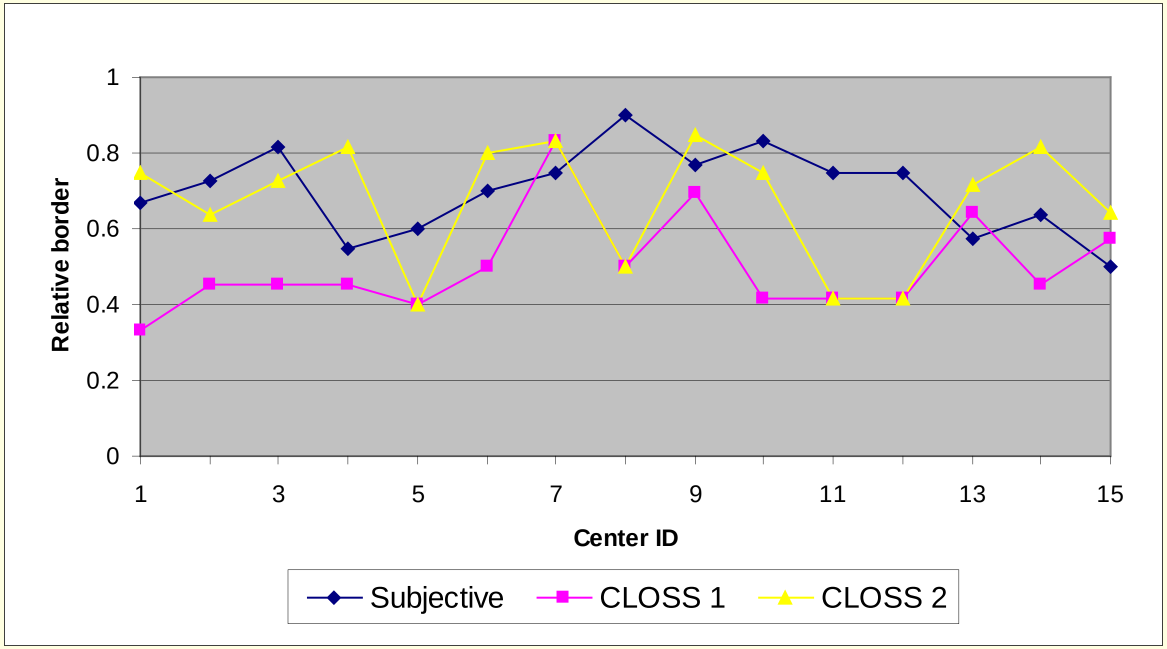
$$\Delta_{CLOSS/GPC} = 1 - \frac{C^{corr}}{C}$$

CLOSS and Subjective Clustering

			Overlap number														CLOSS Border	
String	q-length	s-border	1	2	3	4	5	6	7	8	9	10	11	12	13	14	CLOSS1	CLOSS2
countenance	12	8	15	15	15	14	14	10	10	3	2	2	1	1	0	0	4	9
contanance	11	8	15	15	15	15	13	13	3	3	2	1	1	0	0	0	5	7
contenance	11	9	15	15	15	15	14	14	13	3	3	2	1	0	0	0	5	8
conscience	11	6	15	15	15	15	14	14	9	5	1	1	1	0	0	0	5	9
conceince	10	6	15	15	15	14	14	12	3	2	1	1	0	0	0	0	4	4
conciance	10	7	15	15	15	15	14	14	7	1	1	1	0	0	0	0	5	8
convenience	12	9	15	15	15	15	15	14	11	6	4	2	2	1	0	0	10	10
convience	10	9	15	15	15	15	14	14	6	4	3	2	0	0	0	0	5	5
convienience	13	10	15	15	15	15	15	14	10	6	4	3	2	1	1	0	9	11
consequence	12	10	15	15	15	15	14	14	10	4	3	3	1	1	0	0	5	9
concequence	12	9	15	15	15	15	14	14	11	4	3	2	1	1	0	0	5	5
consiquence	12	9	15	15	15	15	14	14	10	4	3	2	1	1	0	0	5	5
convalescence	14	8	15	15	15	15	15	15	12	6	2	2	2	1	1	1	9	10
convalaces	11	7	15	15	15	14	3	3	3	2	1	1	1	0	0	0	5	9
convalensence	14	7	15	15	15	15	15	15	11	4	2	2	2	1	1	1	8	9

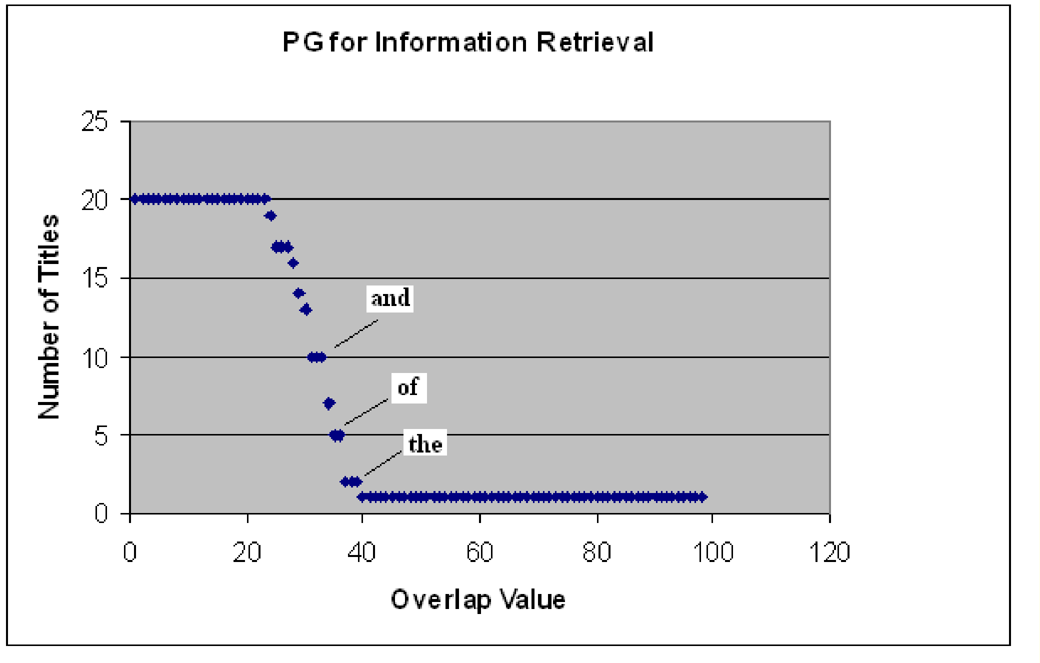
CLOSS and Subjective Clustering

Preserving misspellings CLOSS reduces the number of alien strings.



Multiple Horizontal Lines Problem

Typical example is the DBLP dataset of paper titles.

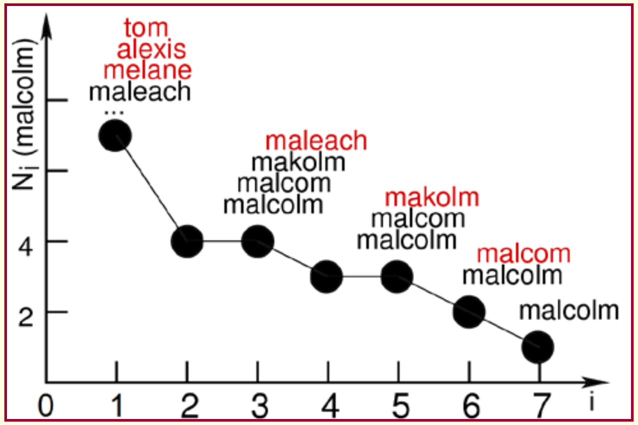


Multiple horizontal lines arise because of the common words (and their parts) in the titles.

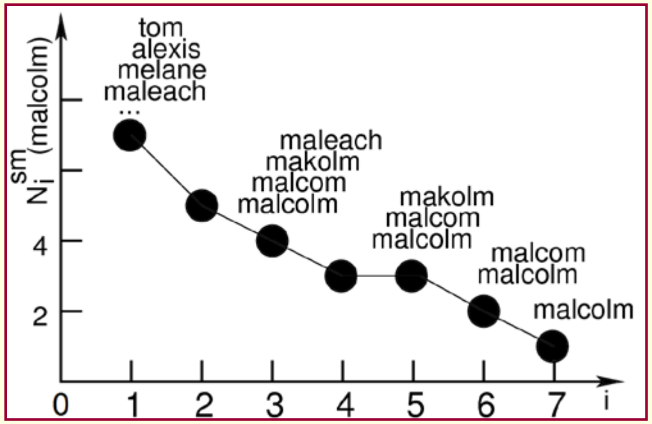
CLOSS. Smoothing

Smoothing modifies the PG shape by using moving averages. This allows to identify cluster border for the case of multiple lines that take place in datasets containing long and short/long strings.

$$N_i^{sm}(s) = \frac{1}{2r+1} \sum_{j=i-r}^{i+r} N_j(s)$$



PG without smoothing



Smoothed PG

Resume

- Proposed method is intended to cluster strings in textual databases of different origin. It uses dataset ordering, string representation by q-grams, novel border identification technique as well as proximity graph smoothing (for the case of multiple horizontal lines).
- Evaluation shows CLOSS efficiency for datasets with strings of different length, even if cluster border is not prominent (short strings).

Future Investigations

- It is observed that if PG has multiple horizontal lines then clustering quality varies depending on string length and smoothing interval. In the nearest future we suppose to stabilize the quality applying adaptive smoothing that takes into account string length dispersion in each point of the proximity graph.

Questions

?