On Cosine and Tanimoto Near Duplicates Search among Vectors with Domains Consisting of Zero, a Positive Number and a Negative Number

Marzena Kryszkiewicz
Institute of Computer Science
Warsaw University of Technology

1

ZPN-vectors

- *A ZPN-vector* is a vector each domain of which may contain at most three values: zero, a positive value and a negative value.

2

Example application of *ZPN*-vectors…

- Search of documents that cite similar papers in a similar way – a paper cited as **valuable** could be graded with **a positive value**; the paper cited as **invaluable** could be graded with **a negative value**; **not cited** paper could be graded with **0**.

Paper	Cited {-2,0,1}
#1	1
#2	1
#3	-2
#4	0
#5	1

3

Example application of *ZPN*-vectors

- Some teachers grade answers to test questions in three ways - a **positive answer** is graded with a **positive value**, a **negative answer** is graded with a **negative value** and **lack of an answer** is graded with **0**. *Such grading might discourage students from guessing answers.*

Student	Question 1 {-1,0,1}	Question 2 {-4,-,4}	Question 3 {-2,0,5}
#1	1	4	5
#2	1	4	-2
#3	-1	-4	5
#4	0	0	-2
#5	1	-4	0

4

Goal

To derive bounds on lengths of *ZPN*-vectors such that:

- $cosSim(u, v) \geq \varepsilon$,
- $T(u, v) \geq \varepsilon$,

for $\varepsilon > 0$.

5

Bounds on Length of Neighbor Vectors…

Let the domain of an i-th dimension be equal to $\{0, a, b\}$, where $a > 0$ and $b < 0$. Then:

u_i	v_i	$u_i\, v_i$	u_i^2
0	0	0	0
0	a	0	0
0	b	0	0

u_i	v_i	$u_i\, v_i$	u_i^2
a	0	0	a^2
a	a	a^2	a^2
a	b	ab	a^2

u_i	v_i	$u_i\, v_i$	u_i^2
b	0	0	b^2
b	a	ab	b^2
b	b	b^2	b^2

Proposition. For any *ZPN*-vectors u and v:

- $u_i\, v_i \leq u_i^2$ for any dimension i;
- $u \cdot v \leq |u|^2$.

Deriving Bounds on Lengths of Cosine Similar *ZPN*-Vectors

Theorem. Let u and v be non-zero *ZPN*-vectors, $cosSim(u, v) \geq \varepsilon$ and $\varepsilon \in (0,1]$. Then:

$$|v| \in \left[\varepsilon |u|, \frac{|u|}{\varepsilon}\right] \quad \text{and} \quad |v|^2 \in \left[\varepsilon^2 |u|^2, \frac{|u|^2}{\varepsilon^2}\right].$$

Proof. Since $cosSim(u, v) \geq \varepsilon$ and $u \cdot v \leq |u|^2$, then

$\varepsilon \leq cosSim(u, v) = \frac{u \cdot v}{|u||v|} \leq \frac{|u|^2}{|u||v|} = \frac{|u|}{|v|}$. So, $\varepsilon \leq \frac{|u|}{|v|}$.

Analogously, one may derive that $\varepsilon \leq \frac{|v|}{|u|}$.

Hence, $|v| \in \left[\varepsilon |u|, \frac{|u|}{\varepsilon}\right]$. □

Dense and sparse representations of an example data set

Id	1	2	3	4	5	6	7	8	9
v1	-3,0	4,0			3,0	5,0	3,0	6,0	
v2	3,0	-2,0				5,0		6,0	
v3									
v4		-2,0							
v5									
v6									
v7									
v8		4,0							
v9									
v10		-2,0							

$NZD(v2) = \{1,2,6,8\}$.

Id	(non-zero dimension, value) pairs
v1	{(1,-3.0), (2, 4.0), (5, 3.0), (6, 5.0), (7, 3.0), (8, 6.0)}
v2	{(1, 3.0), (2,-2.0), (6, 5.0), (8, 6.0)}
v3	{(3, 6.0), (4, 4.0)}
v4	{(2,-2.0), (4, 4.0), (6, 5.0), (8,-5.0)}
v5	{(4, 4.0), (5,-3.0), (7, 3.0)}
v6	{(3,-9.0), (4, 4.0), (9, 5.0)}
v7	{(3, 6.0), (4, 4.0)}
v8	{(2, 4.0), (4, 4.0), (9, 5.0)}
v9	{(4,-2.0), (5, 3.0), (7, 3.0), (9, 5.0)}
v10	{(2,-2.0), (3,-9.0)}

Example: Using bounds on lengths for searching cosine similar *ZPN*-vectors

Id	(non-zero dimension, value) pairs	length
v5	{(4, 4.0), (5,-3.0), (7, 3.0)}	5.83
v9	{(4,-2.0), (5, 3.0), (7, 3.0), (9, 5.0)}	6.86
v3	{(3, 6.0), (4, 4.0)}	7.21
v7	{(3, 6.0), (4, 4.0)}	7.21
v8	{(2, 4.0), (4, 4.0), (9, 5.0)}	7.55
v2	{(1, 3.0), (2,-2.0), (6, 5.0), (8, 6.0)}	8.60
v4	{(2,-2.0), (4, 4.0), (6, 5.0), (8,-5.0)}	8.37
v10	**{(2,-2.0), (3,-9.0)}**	**9.22**
v1	**{(1,-3.0), (2, 4.0), (5, 3.0), (6, 5.0), (7, 3.0), (8, 6.0)}**	**10.20**
v6	**{(3,-9.0), (4, 4.0), (9, 5.0)}**	**11.05**

Vector length ∈ [8.67, 12.00]

Let us consider vector $u = v1$ and let $\varepsilon = 0.85$. Then, only vectors the lengths of which belong to the interval $\left[\varepsilon |u|, \frac{|u|}{\varepsilon}\right] = \left[0.85 \times 10.20, \frac{10.20}{0.85}\right] \approx [8.67, 12.00]$ have a chance to be sought near cosine duplicates of u; that is vectors $v1$, $v6$ and $v10$. □

The Tanimoto Similarity of Vectors

The Tanimoto similarity between vectors u and v is denoted by *T(u, v)* and is defined as; that is,

$$T(u,v) = \frac{u \cdot v}{u \cdot u + v \cdot v - u \cdot v} = \frac{u \cdot v}{|u|^2 + |v|^2 - u \cdot v}$$

where:

- $u \cdot v$ is a standard vector dot product of u and v and equals $\sum_{i=1..n} u_i v_i$;
- $|u|$ is the length of vector u and equals $\sqrt{u \cdot u}$.

Property [Willett, Barnard, Downs]. $T(u, v) \in [-1/3, 1]$.

10

The Tanimoto Similarity of Binary Vectors

Property. In the case of binary vectors with domains restricted to {0, 1}, the Tanimoto similarity between two vectors determines the ratio of the number of non-zero dimensions shared by both vectors to the number of non-zero dimensions occurring in either vector.

Then, the **Tanimoto similarity (*T*)** coincides with the **Jaccard coefficient (*J*)**.

Example. Let u = **[01101]** and v = **[10101]**. Then:

$$T(u,v) = \frac{u \cdot v}{u \cdot u + v \cdot v - u \cdot v} = \frac{u \cdot v}{|u|^2 + |v|^2 - u \cdot v} = \frac{2}{3 + 3 - 2} = \frac{2}{4}.$$

Eq., let U={**bce**} and V={**ace**}. Then, $J(U, V) = \frac{|U \cap V|}{|U \cup V|} = \frac{|U \cap V|}{|U| + |V| - |U \cap V|} = \frac{2}{4}$.

11

Deriving Bounds on Lengths of Tanimoto Similar *ZPN*-Vectors

Theorem. Let u and v be non-zero *ZPN*-vectors, $T(u, v) \geq \varepsilon$ and $\varepsilon \in (0,1]$. Then:

$$|v|^2 \in \left[\varepsilon |u|^2, \frac{|u|^2}{\varepsilon}\right] \text{ and } |v| \in \left[\sqrt{\varepsilon}\, |u|, \frac{|u|}{\sqrt{\varepsilon}}\right].$$

Proof. Follows from the fact that $u \cdot v \leq |u|^2$. □

Namely, $\varepsilon \leq T(u,v) = \frac{u \cdot v}{|u|^2 + |v|^2 - u \cdot v} \leq \frac{|u|^2}{|u|^2 + |v|^2 - |u|^2} = \frac{|u|^2}{|v|^2}$. So, $\varepsilon \leq \frac{|u|^2}{|v|^2}$.

Analogously, $\varepsilon \leq \frac{|v|^2}{|u|^2}$.

Therefore, $|v|^2 \in [\varepsilon |u|^2, \frac{|v|^2}{\epsilon}]$.

Example: Using bounds on lengths for searching Tanimoto similar *ZPN*-vectors

Id	(non-zero dimension, value) pairs	length
v5	{(4, 4.0), (5,-3.0), (7, 3.0)}	5.83
v9	{(4,-2.0), (5, 3.0), (7, 3.0), (9, 5.0)}	6.86
v3	{(3, 6.0), (4, 4.0)}	7.21
v7	{(3, 6.0), (4, 4.0)}	7.21
v8	{(2, 4.0), (4, 4.0), (9, 5.0)}	7.55
v2	{(1, 3.0), (2,-2.0), (6, 5.0), (8, 6.0)}	8.60
v4	{(2,-2.0), (4, 4.0), (6, 5.0), (8,-5.0)}	8.37
v10	{(2,-2.0), (3,-9.0)}	9.22
v1	**{(1,-3.0), (2, 4.0), (5, 3.0), (6, 5.0), (7, 3.0), (8, 6.0)}**	**10.20**
v6	**{(3,-9.0), (4, 4.0), (9, 5.0)}**	**11.05**

Vector length ∈ [9.40, 11.07]

Let us consider vector $u = v1$ and let $\varepsilon = 0.85$. Then, only vectors the lengths of which belong to the interval $\left[\sqrt{\varepsilon}|u|, \frac{|u|}{\sqrt{\varepsilon}}\right] \subseteq [9.40, 11.07]$ have a chance to belong to ε-neighbourhood of vector u; that is, vectors: $v1$ and $v6$. □

Bounds on lenghts for *ZPN*-vectors with {-1, 0, +1} dimensions' domains

One may easily note that if u is a *ZPN*-vector whose each dimension has domain {0, 1, -1}, then $|u|^2 = |NZD(u)|$. This observation allows us to obtain:

Corollary. Let u and v be non-zero *ZPN*-vectors whose each dimension has domain {0, 1, -1} and $\varepsilon \in (0,1]$. Then:

If $cosSim(u, v) \geq \varepsilon$, then $|NZD(v)| \in \left[\varepsilon^2 |NZD(u)|, \frac{|NZD(u)|}{\varepsilon^2}\right]$.

If $T(u, v) \geq \varepsilon$, then $|NZD(v)| \in \left[\varepsilon |NZD(u)|, \frac{|NZD(u)|}{\varepsilon}\right]$.

References

- Kryszkiewicz, M.: On Cosine and Tanimoto Near Duplicates Search among Vectors with Domains Consisting of Zero, a Positive Number and a Negative Number. FQAS 2013: 531-542
- Willett, P., Barnard, J.M., Downs, G.M.: Chemical similarity searching. J. Chem. Inf. Comput. Sci., 38 (6), pp. 983–996 (1998)

15