

CLAIMS

1. A method for analyzing addresses, the method comprising:
 - obtaining a first address and a second address, the first address comprising n morphemes and the second address comprising m morphemes, wherein a morpheme comprises a smallest semantic unit in the address;
 - determining multiple first correlation values between the n morphemes and the m morphemes;
 - obtaining a second correlation value between the first address and the second address based on the multiple first correlation values and a preset algorithm; and
 - analyzing a correlation between the first address and the second address based on the second correlation value.
2. The method of claim 1, wherein the step of determining multiple first correlation values between the n morphemes and the m morphemes comprises determining first correlation values between each of the n morphemes and the m morphemes to obtain $n \times m$ first correlation values.
3. The method of claim 2, wherein the step of obtaining a second correlation value between the first address and the second address based on the multiple first correlation values and a preset algorithm comprises:
 - selecting a second morpheme best matching a first morpheme in the n morphemes from the m morphemes based on the $n \times m$ first correlation values and the preset algorithm, and recording a target correlation value between the first morpheme and the second morpheme until n target correlation values are recorded; and
 - obtaining a second correlation value between the first address and the second address based on the n target correlation values.
4. The method of claim 3, wherein obtaining a second correlation value between the first address and the second address based on the n target correlation values comprises summing the n target correlation values.

5. The method of claim 3 or 4, wherein the step of selecting a second morpheme best matching with a first morpheme in the n morphemes from the m morphemes based on the $n \times m$ first correlation values and the preset algorithm comprises:

constructing an $n \times m$ matrix based on the $n \times m$ first correlation values;

pre-processing the $n \times m$ matrix based on the preset algorithm; and

selecting the second morpheme best matching with the first morpheme in the n morphemes from the m morphemes based on the pre-processed $n \times m$ matrix.

6. The method of any one of claims 1 to 5, wherein the first correlation value comprises an edit distance value, a hamming distance value, a Jaccard distance value, an N-Gram distance value, a JW distance value, or a cosine distance value.

7. The method of claim 6, wherein the first correlation value comprises an edit distance value, and wherein:

selecting a second morpheme best matching with a first morpheme in the n morphemes comprises solving an optimal matching problem of the edit distance value; and

pre-processing the $n \times m$ matrix based on the preset algorithm comprises subtracting the smallest element of each row of the matrix from the respective row of elements and/or subtracting the smallest element of each column from the respective column of elements.

8. The method of any one of claims 1 to 7, wherein the first address or second address comprise an English address, and wherein the morphemes comprise words included in the first address.

9. The method of any one of claims 1 to 8, wherein the first address or second address comprise a Chinese address, and wherein the morphemes comprise terms included in a standardization of the Chinese address, and optionally wherein:

the method further comprises transforming the first or second address into n or m groups of phonetic alphabets or n or m groups of strokes, and

determining first correlation values comprises:

determining first correlation values between the n groups of phonetic alphabets

included in the first address and m groups of phonetic alphabets included in the second address, or determining first correlation values between the n groups of strokes included in the first address and m groups of strokes included in the second address.

10. The method of any one of claims 1 to 19, wherein the preset algorithm comprises a Hungarian algorithm or an exhaustion algorithm.

11. The method of any one of claims 1 to 10, wherein the second correlation value comprises a difference value, and wherein a greater second correlation value between the first address and the second address indicates a smaller similarity between the first address and the second address.

12. The method of any one of claims 1 to 11, wherein the second correlation value comprises a similarity value, and wherein the first address and the second address are more similar to each other if the similarity value between these two addresses is closer to 1, and the first address and the second address are less similar to each other if the similarity value between these two addresses is closer to 0.

13. The method of claim 1, wherein n and m comprise natural numbers.

14. The method of claim 1, further comprising providing the analyzed correlation between the first address and the second address for cluster analysis, term frequency analysis or address standardization.

15. An apparatus for address analysis, the apparatus comprising an input unit, a determination unit, an acquisition unit, and an analysis unit, the apparatus configured to perform operations according to any one of claims 1 to 14.