CLAIMS

What is claimed is:

1. A computerized method of compressing symbolic information organized into a plurality of documents, each document having a plurality of symbols, the method comprising:

with a first document of the plurality of documents as an input document, automatically with a computer:

(a) identifying a plurality of symbol pairs, each symbol pair consisting of two sequential or non-sequential symbols in the input document, one or more symbol pairs consisting of two non-sequential symbols in the input document;

(b) for each unique symbol pair of the plurality of symbol pairs, updating a count identifying the number of appearances of the unique symbol pair; and

(c) producing a compressed document by causing the compressed document to include, at each position associated with one of the plurality of symbol pairs from the input document, including one or more symbol pairs consisting of two non-sequential symbols, (i) a replacement symbol associated by a compression dictionary with the unique symbol pair matching the one of the plurality of symbol pairs, if the count for the unique symbol pair exceeds a threshold, and (ii) for at least those symbol pairs consisting of two non-sequential symbols, indicia indicating a distance between locations of the non-sequential symbols of the pair in the input document.

2. The method of claim 1 wherein the plurality of symbol pairs includes only those pairs of non-sequential symbols for which the distance between locations of the non-sequential symbols of the pair in the input document is less than a numeric distance cap.

3. The method of claim 2 wherein the numeric distance cap is ten symbols.

4. The method of claim 1 wherein the threshold is two.

5. The method of claim 1 further comprising sorting the compression dictionary in order of the count for the respective unique symbol pairs.

6. The method of claim 5 further comprising excluding or removing from the compression dictionary all unique symbol pairs having a count that does not exceed the threshold and making an association in the compression dictionary

between each unique symbol pair having a count that exceeds the threshold and a replacement symbol.

7. The method of claim 1 further comprising performing an additional pass by repeating parts (a), (b), and (c) using as the input document the compressed document.

8. The method of claim 7 further comprising performing additional passes, until the compressed document created by one of the passes is unaltered from the compressed document created by the immediately previous pass.

9. The method of claim 1 wherein each symbol from the first document indicates a word.

10. The method of claim 9 further comprising, before part (a), creating the first document by replacing each word of a text document with a corresponding one of the plurality of symbols.

11. The method of claim 1 wherein each symbol in the first document indicates an alphanumeric character.

12. The method of claim 1 wherein each replacement symbol comprises an address of an entry in the compression dictionary, which entry stores the unique symbol pair associated with replacement symbol.

13. The method of claim 1 further comprising performing parts (a), (b), and (c) on each of the other documents of the plurality of documents, wherein the respective counts for the symbol pairs identifies the number of previous appearances of that symbol pair in any of the plurality of documents.

14. The method of claim 13 wherein parts (a) and (b) are performed on each of the plurality of documents before part (c) is performed on any of the plurality of documents.

15. The method of claim 14 further comprising, before part (c) is performed on any of the plurality of documents, making an association in the compression dictionary between each unique symbol pair and a replacement symbol if and only if the unique symbol pair has a count that exceeds the threshold.

16. The method of claim 13 further comprising sorting the compression dictionary in order of the count for the respective unique symbol pairs.

17. The method of claim 16 further comprising removing from the compression dictionary all unique symbol pairs having a count that does not exceed the threshold and making an association in the compression dictionary

between each unique symbol pair having a count that exceeds the threshold and a replacement symbol.

18. The method of claim 13 further comprising repeating parts (a), (b), and (c) on each of the other documents to make at least one additional pass by repeating parts (a), (b), and (c), using as the input document the compressed document previously produced by the process, until the compressed document created by one of the passes is unaltered from the compressed document created by the immediately previous pass.

19. The method of claim 18 further comprising assigning the first document at least one of a plurality of attributes based on the frequency with which a replacement symbol associated with a unique symbol pair appears in an output document associated with the first document.

20. The method of claim 19 further comprising suggesting a related document based the at least one attribute of the first document and the at least one attribute of the related document.

21. A computer system having at least one processor capable of compressing symbolic information organized into a plurality of documents, each document having a plurality of symbols, the computer system programmed to perform automatically, with the processor, the method of any one of claims 1 through 20 with a first document of the plurality of documents as an input document.

22. A computer-readable storage medium that is not a transitory propagating signal storing a set of computer instructions for compressing symbolic information organized into a plurality of documents, each document having a plurality of symbols, wherein the set of computer instructions, when executed on the computer, causes the computer to perform automatically the method of any one of claims 1 through 20 with a first document of the plurality of documents as an input document.

CLAIMS

What is claimed is:

1. A computer system comprising (a) at least one processor configured to compress symbolic information organized into a plurality of documents, and (b) a computer-readable storage medium configured to store a set of computer instructions for compressing symbolic information organized into a plurality of documents, wherein each of the plurality of documents has a plurality of symbols, and wherein the computer system is configured to perform the following process:

utilizing a first document of the plurality of documents as an input document, automatically utilizing the computer system to:

(a) identify a plurality of symbol pairs, wherein each symbol pair consists of two sequential or two non-sequential symbols in the input document, and at least one of the symbol pairs consists of the two non-sequential symbols in the input document;

(b) for each unique symbol pair of the plurality of symbol pairs, updating a count of the number of appearances of the unique symbol pair; and

(c) producing a compressed document to include, at each position associated with one of the plurality of symbol pairs from the input document: (i) a replacement symbol associated by a compression dictionary with the unique symbol pair matching the one of the plurality of symbol pairs, if the count for the unique symbol pair exceeds a threshold, and (ii) for at least the symbol pairs consisting of two non-sequential symbols, indicia indicating a distance between locations of the non-sequential symbols of the pair in the input document.

2. The computer system of claim 1 wherein the plurality of symbol pairs includes only the pairs of non-sequential symbols for which the distance between locations of the non-sequential symbols of the pair in the input document is less than a numeric distance cap.

3. The computer system of claim 2, wherein the numeric distance cap is ten symbols.

4. The computer system of any one of claims 1-3, wherein the threshold is two.

5. The computer system of any one of claims 1-4 that is further configured to sort the compression dictionary in order of the count for the respective unique symbol pairs.

6. The computer system of any one of claims 1-5 that is further configured to exclude or remove from the compression dictionary all unique symbol pairs having a count that does not exceed the threshold.

7. The computer system of any one of claims 1-6 that is further configured to make an association in the compression dictionary between each unique symbol pair having a count that exceeds the threshold and a replacement symbol.

8. The computer system of any one of claims 1-7 that is further configured to perform an additional pass by repeating parts (a), (b), and (c) using as the input document the compressed document.

9. The computer system of claim 8 that is further configured to perform additional passes until the compressed document created by one of the passes is unaltered from the compressed document created by an immediately previous pass.

10. The computer system of any one of claims 1-9, wherein each symbol from the first document indicates a word.

11. The computer system of any one of claims 1-9, wherein at least one symbol from the first document indicates a word.

12. The computer system of any one of claims 10-11 that is further configured to, before part (a), creating the first document by replacing each word of a text document with a corresponding symbol.

13. The computer system of any one of claims 1-9, wherein each symbol in the first document indicates an alphanumeric character.

14. The computer system of any one of claims 1-9 or 11, wherein at least one symbol in the first document indicates an alphanumeric character.

15. The computer system of any one of claims 13-14 that is further configured to, before part (a), creating the first document by replacing each alphanumeric character with a corresponding symbol.

16. The computer system of any one of claims 1-15, wherein each replacement symbol comprises an address of an entry in the compression dictionary, wherein the entry stores the unique symbol pair associated with replacement symbol.

17. The computer system of any one of claims 1-16 that is further configured to perform parts (a), (b), and (c) on each of the plurality of documents other than the first document, wherein the respective counts for the symbol pairs identifies the number of previous appearances of that symbol pair in any of the plurality of documents other than the first document.

18. The computer system of claim 17 that is configured to perform parts (a) and (b) on each of the plurality of documents before part (c) is performed on any of the plurality of documents.

19. The computer system of any one of claims 1-18 that is further configured to, before part (c) is performed on any of the plurality of documents, making an association in the compression dictionary between each unique symbol pair and a replacement symbol if the unique symbol pair has a count that exceeds the threshold.

20. The computer system of any one of claims 1-19 that is further configured to sort the compression dictionary in order of the count for the respective unique symbol pairs.

21. The computer system of any one of claims 1-20 that is further configured to remove from the compression dictionary all unique symbol pairs having a count that do not exceed the threshold and making an association in the compression dictionary between each unique symbol pair having a count that exceeds the threshold and a replacement symbol.

22. The computer system of any one of claims 14-18 that is further configured to repeat parts (a), (b), and on each of the plurality of documents other than the first document in order to make an additional pass by repeating parts (a), (b), and (c), using as the input document a compressed document previously produced by performing parts (a), (b), and (c), until a second compressed document created by one of the passes is unaltered from the compressed document created by the immediately previous pass.

23. The computer system of any one of claims 1-19 that is further configured to assign the first document at least one of a plurality of attributes based on the frequency with which a replacement symbol associated with a unique symbol pair appears in an output document associated with the first document.

24. The computer system of claim 23 that is further configured to suggest a related document based the at least one of the plurality of attributes of the first document and at least one attribute of the related document.