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:

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

(i) identifying a plurality of symbol pairs, each symbol pair consisting of two sequential symbols in the input document; and

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

(b) performing part (a) on each of the other documents of the plurality of documents, wherein the respective counts for the symbols pairs identifies the number of previous appearances of that symbol pair in any of the plurality of documents; and

(c) after part (b), for at least one of the plurality of documents, 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, 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.

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

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

4. The method of claim 3 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.

5. The method of claim 1 further comprising performing an additional pass on the at least one of the plurality of documents by repeating parts (a), and (c) using as the input document the compressed document.

6. The method of claim 5 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.

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

8. The method of claim 7 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.

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

10. The method of claim 1 further comprising performing part (c) on each of the plurality of documents.

11. The method of claim 1 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.

12. The method of claim 1 further comprising, for each of the plurality of documents, performing at least one additional pass by repeating parts (a) and (c), using as the input document the compressed document, until the compressed document created by one of the passes is unaltered from the compressed document created by the immediately previous pass.

13. The method of claim 1 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.

14. The method of claim 13 further comprising suggesting a related document of the other documents based on the at least one attribute of the first document and a parallel at least one attribute assigned to the related document based on the frequency with which a replacement symbol associated with a unique symbol pair appears in an output document associated with the related document.

15. 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 the method of any of claims 1-14.

16. 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 a computer, causes the computer, automatically to perform the method of any of claims 1-14.