

**METHODS AND APPARATUS FOR
PERFORMING SPELLING CORRECTIONS
USING ONE OR MORE VARIANT HASH
TABLES**

FIELD OF THE INVENTION

[0001] The present invention relates generally to techniques for real-time spelling correction of a term against a dictionary of valid words and more particularly, to techniques for real-time spelling correction of a term using one or more hash tables.

BACKGROUND OF THE INVENTION

[0002] A number of techniques exist for automatically detecting and correcting spelling errors. Suppose that a spell checking algorithm is given a word, G, and attempts to find one or more other words from a list of candidate words (such as validly spelled words) that are within a given edit distance from G. The edit distance between two words is the smallest number of operations that transform the candidate word into the given word (with each operation consisting of removing one letter (deletion), adding one letter (insertion), replacing one letter with another letter (replacement), or transposing two letters (transposition)).

[0003] Two words are said to have a distance (or “edit distance”) of zero between them if they are identical. The two words are said have a distance one separation if one can get from one word to the other word, by: (1) transposing one pair of adjacent characters; (2) replacing a single character with any other character; (3) deleting any one character; or (4) inserting an arbitrary character at any position in the original word. Likewise, words are a distance two apart if two moves of the type described above are required to get from the first word to the second word. More generally, two words are a distance N apart if N moves are required to get from the first word to the second.

[0004] U.S. Pat. No. 6,616,704 B1, assigned to the assignee of the present invention and entitled “Two Step Method for Correcting Spelling of a Word or Phrase in a Document,” discloses a method for correcting the spelling of a word or phrase in a document. The disclosed method proceeds in two steps: first an initial approximate method eliminates most candidate words from consideration (without computing the exact edit distance between the given word whose spelling is to be corrected and any candidate word), and then a “slow method” computes the exact edit distance between the word whose spelling is to be corrected and each of the few remaining candidate words. For a dictionary of size D and a maximum word length W, the disclosed two step method is said to run in time on the order of (D), if the number of exact edit distance calculations is small, and on the order of (D*W²) otherwise.

[0005] While such existing techniques for real-time spelling correction of a term against a dictionary of valid words provide an effective mechanism for detecting and correcting spelling errors, they suffer from a number of limitations, which if overcome, could further improve the efficiency, utility and reliability of spell checking functions. More particularly, a number of existing techniques generate an excessive amount of false positives. In addition, for the detection of certain errors, existing techniques are said to run in time on the order of the dictionary size, D, or on the order of log(D), the log of the size of the dictionary.

[0006] A need therefore exists for improved techniques for real-time spelling correction of a term against a dictionary of valid words.

SUMMARY OF THE INVENTION

[0007] Generally, methods and apparatus are provided for performing spelling corrections using one or more variant hash tables. According to one aspect of the invention, the spelling of at least one candidate word is corrected by obtaining at least one variant dictionary hash table based on variants of a set of known correctly spelled words, wherein the variants are obtained by applying one or more of a deletion, insertion, replacement, and transposition operation on the correctly spelled words; obtaining from the candidate word one or more lookup variants using one or more of the deletion, insertion, replacement, and transposition operations; evaluating one or more of the candidate word and the lookup variants against the at least one variant dictionary hash table; and indicating a candidate correction if there is at least one match in the at least one variant dictionary hash table.

[0008] In an exemplary “distance one” implementation, a dictionary hash table is also employed, where the dictionary hash table and the at least one variant dictionary hash table are based on a dictionary of correctly spelled words and are comprised of at least one distance one variation for each dictionary entry, wherein the distance one variation comprises one or more of a deletion, insertion, replacement, and transposition operation performed on the entries. The step of evaluating one or more of the candidate word and the lookup variants against the at least one variant dictionary hash table further comprises the step of evaluating one or more distance one variants against the at least one variant dictionary hash table.

[0009] A more complete understanding of the present invention, as well as further features and advantages of the present invention, will be obtained by reference to the following detailed description and drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0010] FIG. 1 is a flow chart illustrating the overall flow of an exemplary distance one spelling correction algorithm; [0011] FIG. 2 is a flow chart illustrating an exemplary process of testing variants of the candidate word against hash tables derived from the dictionary for distance one misspellings in accordance with the present invention; [0012] FIG. 3 is a flow chart illustrating the overall flow of the distance two spelling correction algorithm; [0013] FIG. 4 is a flow chart illustrating the process of testing variants of the candidate word against hash tables derived from the dictionary for distance two misspellings; [0014] FIG. 5 is a flow chart illustrating the overall flow of the “soft” distance two spelling correction algorithm; and [0015] FIG. 6 describes the process of testing variants of the candidate word against hash tables derived from the dictionary for “soft” distance two misspellings.

DETAILED DESCRIPTION OF PREFERRED
EMBODIMENTS

[0016] The present invention provides improved techniques for real-time spelling correction of a term against a dictionary of valid words (including all word forms). The dictionary can be multi-lingual, i.e., it can be composed of