CLAIMS

We claim:

1. A method for automatic ingestion and filing of documents in a database, the database having a plurality of folder levels within a hierarchy and a plurality of file locations of previously stored documents, wherein each file location of the plurality of file locations is associated with any one of the plurality of folder levels, the method comprising:

providing a plurality of stored keywords, based on the previously stored documents, wherein each of the plurality of stored keywords is associated with at least one of the plurality of file locations, wherein a selected keyword in the plurality of stored keywords is associated with at least two of the plurality of file locations, wherein a first location of the at least two of the plurality of file locations has a first location-specific weighting indicating a relevance of the first location to the selected keyword, and wherein a second location of the at least two of the plurality of file locations has a second location-specific weighting indicating a relevance of the second location to the selected keyword;

receiving an electronic file comprising a document;

identifying text data in the document;

indexing the text data to identify one or more document keywords;

comparing the one or more document keywords to the plurality of stored keywords based on the comparing, generating a ranked plurality of suggested file locations for the document, wherein the generating the ranked plurality comprises, for each element of the one or more document keywords;

generating respective keyword score for the element, wherein the respective keyword score is based on a similarity measure between any selected one of the plurality of stored keywords and the element;

adding the at least one file location associated with the selected one to the ranked plurality in association with the element; and

ranking the at least one file location associated with the element in the ranked plurality according to the respective keyword score of the

element and, when the selected one is the selected keyword, modifying the ranking of the first location based on the first location-specific weighting, and modifying the ranking of the second location based on the second location-specific weighting.

2. The method of claim 1, further comprising

identifying a plurality of pages in the document;

determining a plurality of page markers for each page;

determining that the document comprises a plurality of distinct documents; and

assigning each page to one of the distinct documents by grouping the plurality of pages into the distinct documents by comparing the page markers for the plurality of pages.

3. The method of claim 2, wherein the page markers comprise image-based page markers derived from a visual appearance of the page.

4. The method of any one of claims 2 to 3, wherein the page markers comprise text-based page markers derived from the text data in the document.

5. The method of any one of claims 1 to 4 further comprising

determining a keyword coefficient for each element of the one or more document keywords, the keyword coefficient indicating a measure of importance of the element to the document, wherein the respective keyword score is further based on the keyword coefficient.

6. The method of claim 5, wherein the measure of importance of the element to the document is determined by

identifying keyword text attributes for the element, the keyword text attributes including at least one of a text size, a text location and a text format, wherein

determining the keyword coefficient for the element is based on the keyword text attributes.

7. The method of any one of claims 1 to 6, wherein identifying the text data comprises performing optical character recognition on the document to identify the text data.

8. The method of claim 5, further comprising determining a recommended file name for the documents based on the keyword coefficient of the one or more document keywords.

9. A computer program product for automatic ingestion and filing of documents in a database, the database having a plurality of folder levels within a hierarchy and a plurality of file locations of previously stored documents, wherein each file location of the plurality of file locations is associated with any one of the plurality of folder levels, the computer program product comprising a non-transitory computer readable storage medium and computer-executable instructions stored on the computer readable storage medium, the instructions for configuring a processor to:

provide a plurality of stored keywords, based on the previously stored documents, wherein each of the plurality of stored keywords is associated with at least one of the plurality of file locations, wherein a selected keyword in the plurality of stored keywords is associated with at least two of the plurality of file locations, wherein a first location of the at least two of the plurality of file locations has a first location-specific weighting indicating a relevance of the first location to the selected keyword, and wherein a second location of the at least two of the plurality of file locations has a second location-specific weighting indicating a relevance of the second location to the selected keyword;

receive an electronic file comprising a document;

identify text data in the document;

index the text data to identify one or more document keywords;

compare the one or more document keywords to the plurality of stored keywords;

based on the comparing, generate a ranked plurality of suggested file locations for the document, wherein the generating the ranked

plurality comprises, for each element of the one or more document keywords:

generating a respective keyword score for the element, wherein the respective keyword score is based on a similarity measure between any selected one of the plurality of stored keywords and the element;

adding the at least one file location associated with the selected one to the ranked plurality in association with the element; and

ranking the at least one file location associated with the element in the ranked plurality according to the respective keyword score of the element and, when the selected one is the selected keyword, modifying the ranking of the first location based on the first location-specific weighting, and modifying the ranking of the second location based on the second location-specific weighting.

10. The computer program product of claim 9, further comprising instructions for configuring the processor to:

identify a plurality of pages in the document;

determine a plurality of page markers for each page;

determine that the document comprises a plurality of distinct documents; and

assign each page to one of the distinct documents by grouping the plurality of pages into the distinct documents by comparing the page markers for the plurality of pages.

11. The computer program product of claim 10, wherein the page markers comprise image-based page markers derived from a visual appearance of the page.

12. The computer program product of any one of claims 10 to 11, wherein the page markers comprise text-based page markers derived from the text data in the document.

13. The computer program product of any one of claims 9 to 12 further comprising instructions for configuring the processor to determine a keyword coefficient for each

element of the one or more document keywords, the keyword coefficient indicating a measure of importance of the element to the document, wherein the respective keyword score is further based on the keyword coefficient.

14. The computer program product of claim 13, further comprising instructions for configuring the processor to determine the measure of importance of the element to the document by identifying keyword text attributes for the element, the keyword text attributes including at least one of a text size, a text location and a text format, wherein determining the keyword coefficient for the document keyword is based on the keyword text attributes.

15. The computer program product of any one of claims 9 to 14, further comprising instructions for configuring the processor to identify the text data by performing optical character recognition on the document.

16. The computer program product of claim 13, further comprising instructions for configuring the processor to determine a recommended file name for the documents based on the keyword coefficient for the one or more document keywords.

17. The method of claim 1, wherein indexing the text data to identify the one or more document keywords comprises determining whether an occurrence level for one or more words in the document is greater than a keyword threshold.

18. The method of claim 1, wherein the plurality of stored keywords are generated based on at least one of: file names of previously stored documents, document keywords identified in previously stored documents, file location names in the file directory, folder location names in the file directory, contents of files in the file directory and user-defined keywords associated with specific files and folders in the file directory.

19. The computer program product of claim 9, wherein indexing the text data to identify the one or more document keywords comprises determining whether an occurrence level for one or more words in the document is greater than a keyword threshold.

20. The computer program product of claim 9, wherein the plurality of stored keywords are generated based on at least one of: file names of previously stored documents, document keywords identified in previously stored documents, file location names in the file directory, folder location names in the file directory, contents of files in the file directory and user-defined keywords associated with specific files and folders in the file directory.