Related U.S. Application Data

(50) Provisional application No. 60/956,880, filed on Aug. 20, 2007.

Publication Classification

(51) Int. Cl.
G06F 1/730 (2006.01)

(52) U.S. Cl. .................................... 707/5; 707/E17.071

(57) ABSTRACT

Passages in a digital corpus are scored and ranked based at least in part on characteristics of instances of the passages occurring in the corpus. Such characteristics include the popularity of the author, the characteristics of the words introducing and following the similar passage, frequency of appearance of the passage in the digital corpus, the length of the similar passage, the words of the similar passage, the usage of punctuation with the similar passage, and the diffusion of the similar passage within the digital corpus. The characteristics are scored and weighted to produce ranking scores for the associated passages. The ranking scores are used for purposes including selecting passages to display in association with a document and ranking passages displayed in response to a search.
Receive Similar Passage Instances

Calculate individual scores

Determine ranking score

Store score in association with passage

FIG. 4
FIG. 5

Client

Send Request 502

Receive Request 504

Server

Determine results 506

Rank results 508

Receive Results 510

Display Results 512
“That’s one small step for a man. One giant leap for mankind.”

See also: Apollo 11, Buzz Aldrin, Edwin Aldrin, Eastern Daylight, Michael Collins, July 20, Eagle has landed, Surveyor program, Apollo program, Project Apollo.
Appears in 313 books from 1966 to 2007 or Timothy Racine pg. 156

“HERE MEN FROM THE PLANET EARTH FIRST SET FOOT UPON THE MOON JULY 1969, AD WE CAME IN PEACE FOR ALL MANKIND
Pacific Ocean near Hawaii on December 27.”

See also: Apollo 11, Buzz Aldrin, Richard Nixon, space wind, Space Flight Network, Michael Collins, Frank Borman, Edwin E., space race, Eagle has landed.
Appears in 136 books from 1966 to 2007 World Book, Inc. pg. 179

“We choose to go to the moon! We choose to go to the moon in this decade, and do the things, not because they are easy, but because they are hard.”

See also: nationalism, Fort Management, President Kennedy, Houston, NASA, John F.,
Appears in 31 books from 1872 to 2007 United States National Aeronautics and Space Administration pg 416

FIG. 6
RANKING SIMILAR PASSAGES

CROSS REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Patent Provisional Application No. 60/936,880, filed Aug. 20, 2007, the contents of which are hereby incorporated by reference.


BACKGROUND

[0003] 1. Field of Art

[0004] This invention pertains, in general, to scoring similar passages in digital text documents and, in particular, to ranking similar passages based on characteristics of the similar passages occurring in the digital text documents.

[0005] 2. Description of the Related Art

[0006] Advancement in digital technology has changed the way people acquire information. For example, people can now view electronic documents that are stored in a predominantly text corpus such as a digital library that is accessible via the Internet. Such a digital text corpus is established, for example, by scanning paper copies of documents including books and newspapers, and then applying an optical character recognition (OCR) process to produce computer-readable text from the scans. The corpus can also be established by receiving documents and other texts already in machine-readable form.

[0007] Many of these electronic documents contain similar passages or quotations that appear multiple times within the corpus. Users may search for documents in the digital corpus based on various search queries. Additionally, users may search for the documents based on known or popular quotations or phrases contained in the documents. However, these types of searches may yield thousands of matching results and the most relevant results may not initially be displayed making it difficult for users to locate the documents or passages most relevant to their queries.

SUMMARY

[0008] The problems described above are addressed by a computer-implemented method, computer program product, and computer system for calculating a score for a passage having a plurality of instances occurring in a digital corpus. Embodiments of the method comprise calculating at least one score based at least in part on characteristics of instances of the passage occurring in the digital corpus and generating a ranking score associated with the passage based at least in part on the calculated at least one score. The method further comprises storing the ranking score in association with the passage in a computer-readable medium. Embodiments of the computer program product and computer system comprise computer code for performing similar functions.

[0009] The features and advantages described in the specification are not all inclusive and, in particular, many additional features and advantages will be apparent to one of ordinary skill in the art in view of the drawings, specification, and claims. Moreover, it should be noted that the language used in the specification has been principally selected for readability and instructional purposes, and may not have been selected to delineate or circumscribe the disclosed subject matter.

BRIEF DESCRIPTION OF DRAWINGS

[0010] The disclosed embodiments have other advantages and features which will be more readily apparent from the detailed description, the appended claims, and the accompanying figures (or drawings). A brief introduction of the figures is below.

[0011] FIG. 1 shows an environment adapted to support ranking similar passages according to one embodiment.

[0012] FIG. 2 is a high-level block diagram illustrating a functional view of a typical computer for use as one of the entities illustrated in the environment of FIG. 1 according to one embodiment.

[0013] FIG. 3 is a high-level block diagram illustrating modules within the scoring engine according to one embodiment.

[0014] FIG. 4 is a flow chart illustrating steps performed by the scoring engine according to one embodiment.

[0015] FIG. 5 is a flow chart illustrating the interaction between the client device and the web server, the scoring engine, and the ranking engine according to one embodiment.

[0016] FIG. 6 is an exemplary web page showing ranked search results according to one embodiment.

[0017] The figures depict various embodiments of the present invention for purposes of illustration only. One skilled in the art will readily recognize from the following discussion that alternative embodiments of the structures and methods illustrated herein may be employed without departing from the principles of the invention described herein.

DETAILED DESCRIPTION

[0018] The Figures (FIGS.) and the following description describe embodiments by way of illustration only. It should be noted that from the following discussion, alternative embodiments of the structures and methods disclosed herein will be readily recognized as viable alternatives that may be employed without departing from the principles of what is claimed.

[0019] FIG. 1 shows an environment adapted to support ranking similar passages according to one embodiment. The environment 100 includes a data store 110 for storing a corpus 112 and a similar passage database 114, a passage mining engine 116 for identifying similar passages in the corpus, a scoring engine 128 for assigning scores to similar passages, and a ranking engine 130 for ranking similar passages. The environment also includes a client 118 for requesting and/or viewing information from the data store 110, and a web server 120 for interacting with the client and providing interfaces allowing the client to access the information in the data store. A network 122 enables communications between and among the data store 110, passage mining engine 116, scoring engine 128, ranking engine 130, client 118, and web server 120.

[0020] Not all the entities shown in FIG. 1 are required to be connected to the network 122 at the same time for the functionalities described herein to be realized. In one embodiment, passage mining engine 116 and/or scoring engine 128 are connected to the network 122 periodically. When it is online, the engines 116 and 128 only need to communicate with the data store 110 in order to score similar passages in the corpus 112 and store the passage data in the passage database.
114. The engines 116 and 128 do not need to interact with the client 118 or the web server 120 according to one embodiment. Once identifying similar passages is finished, the passage mining engine 116 may be off-line, and the web server 120 supports passage navigating by interacting with the client 118 and the data store 110 to retrieve information from the data store that is requested by the client. Similarly, once the scoring of the similar passages is done, the scoring engine 128 may be off-line, and the web server 120 supports retrieval of ranking information by interacting with the client 118 and data store 110 to retrieve information from the data store that is requested by the client. In another embodiment, the scoring engine 128 is connected to the network 122 periodically. When it is online, the scoring engine 128 communicates with the passage mining engine 116 or data store 110 in order to identify which similar passage instances to rank. The scoring engine 128 does not need to interact with the client 118 or the web server 120 according to one embodiment. Moreover, different embodiments of the environment 100 include different and/or additional entities than the ones shown in FIG. 1, and the entities are organized in a different manner.

[0021] The data store 110 stores the corpus 112 of information and the similar passage database 114. It also stores data utilized to support the functionalities or generated by the functionalities described herein. The data store 110 can also store other corpora and data. The data store 110 receives requests for information stored in it and provides the information in return. In a typical embodiment, the data store 110 is comprised of multiple computers and/or storage devices configured to collectively store a large amount of information.

[0022] The corpus 112 stores a set of information in one embodiment, the corpus 112 stores the contents of a large number of digital documents. As used herein, the term “document” refers to a written work or composition. This definition includes, for example, conventional books such as published novels, and collections of text such as newspapers, new stories, magazines, journals, pamphlets, letters, articles, web pages, and other electronic documents. The document contents stored by the corpus 112 include, for example, the document text represented in a computer-readable format, images from the documents, scanned images of pages from the documents, etc. As used herein, the term “word” refers to a token containing a block of structured text. The word does not necessarily have meaning in any language, although it will have meaning in most cases.

[0023] In addition, the corpus 112 stores metadata about the documents within it. The metadata are structured data that describe the documents. Examples of metadata include metadata about a book such as the author, publisher, year published, number of pages, edition, and libraries that carry the book. The metadata stored in the corpus is associated with the similar passages stored in the similar passage database 114.

[0024] The similar passage database 114 stores data describing similar passages in the corpus 112. The similar passage database 114 also stores the ranking score of the similar passage once a ranking score is assigned by the scoring engine 128. More details describing the function of the scoring engine 128 are provided below.

[0025] As used herein, the phrase “similar passage” refers to a passage in a source document that is found in a similar form in one or more different target documents. Occurrences of the same similar passage are referred to as “instances” of that passage. Oftentimes, the similar passage instances are identical. Nevertheless, the passages are referred to as “similar” because there might be slight differences among the passage instances in the different documents. When a source document is said to have multiple “similar passages,” it means that multiple passages in the source document are also found in other documents. This phrase does not necessarily mean that the “similar passages” within the source document are similar to each other. Similar passages are also referred to as “quotations,” “shared passages,” “popular passages,” and “related passages.”

[0026] In one embodiment, the passage database 114 is generated by the passage mining engine 116 to store information obtained from passage mining. In some embodiments, the passage mining engine 116 constructs the passage database 114 by copying existing quotation collections such as Bartlett’s, and searching and indexing the instances of quotations and their variations that appear in the corpus 112. In some embodiments, the passage mining engine 116 constructs the passage database 114 by copying existing text appearing in a quoted form, such as delimited by quotation marks, from the corpus, and searching and indexing the instances of the text in the corpus 112. Further, in some embodiments the passage mining engine 116 constructs the passage database 114 by copying each group of words, such as sentences, from the corpus, and searching and indexing the instances of the group of words in the corpus 112. In one embodiment, the database 114 stores similar passages, document identifiers (Doc IDs) identifying the documents in which the passages exist, position identifiers (Pos IDs) identifying the location in the documents at which the passages appear, passage ranking results, etc. Further, in some embodiments, the database 114 also stores the documents or portions of the documents that have the similar passages.

[0027] The passage mining engine 116 includes one or more computers adapted to analyze the texts of documents in the corpus 112 in order to identify similar passages. For example, the passage mining engine 116 may find that the passage “I read somewhere that everybody on this planet is separated by only six other people” from the book “Six Degrees of Separation” by John Guare, also appears in 13 other books published between 2000 and 2006. The passage mining engine 116 may store, in the similar passage database 114, the passage, its location in the “Six Degrees of Separation” book, Doc IDs of the 13 other books, Pos IDs indicating the locations of the passage instances in the 13 other books, and its ranking relative to other similar passages in the “Six Degrees of Separation” book or relative to other similar passages in the corpus 112. More detail regarding the passage mining engine 116 is described in the related application, U.S. patent application Ser. No. 11/781,213, filed Jul. 20, 2007, and titled “Identifying and Linking Similar Passages in a Digital Text Corpus.” Passage mining may be performed off-line, asynchronously of any queries made by the client 118 against the data store 110. In one embodiment, the passage mining engine 116 runs periodically to process all the text information in the corpus 112 from scratch and generate similar passage data for storing in the similar passage database 114, disregarding any information obtained from prior passage mining. In another embodiment, the passage mining engine 116 is used periodically to incrementally update the data stored in the similar passage database 114, for example, as new documents are added to the corpus 112.

[0028] The scoring engine 128 includes one or more computers adapted to assign scores to the similar passages iden-
tified by the passage mining engine 116 and stored in the similar passages database 114. In one embodiment, the scoring engine 128 analyzes the characteristics of the similar passages and the documents containing the similar passages stored in the similar passage database 114 and assigns ranking scores to the similar passages. Scoring may be performed on-line when the scoring engine is connected to network 122 and may also be performed off-line, asynchronously of any queries made by client 118 against the data store 110. In one embodiment, the scoring engine 128 runs periodically to process all of the content from the data store 110 from scratch and assigns a score associated with a similar passage for storing in the similar passage database 114. In another embodiment, scoring engine 128 is used periodically to incrementally update the ranking information stored in the similar passage database 114, for example, as new similar passages are found and added to the similar passage database.

[0029] The ranking engine 130 ranks a set of similar passages to be displayed on the client 118. The ranking engine 130 ranks the set of similar passages based on the associated ranking scores of the similar passages. The set of similar passages can be displayed on the client 118 in the ranked order.

[0030] For purposes of illustration, FIG. 1 shows the passage mining engine 116, the scoring engine 128, and the ranking engine 130 as discrete servers. However, in various embodiments, any or all of these engines can be combined. This allows a single server to perform the functions of one or more of the above-described engines.

[0031] In one embodiment, the client 118 is an electronic device having a web browser for interacting with the web server 120 via the network 122, and it is used by a human user to access and obtain information from the data store 110. It can be, for example, a notebook, desktop, or handheld computer, a mobile telephone, personal digital assistant (PDA), mobile email device, portable game player, portable music player, computer integrated into a vehicle, etc.

[0032] The web server 120 interacts with the client 118 and the ranking engine 130 to provide information from the data store 110. In one embodiment, the web server 120 includes a User Interface (UI) module 124 that communicates with the client’s 118 web browser to receive and present information. The web server 120 also includes a searching module 126 that searches for information in the data store 110. For example, the UI module 124 may receive a query from the web browser issued by a user of the client 118, and the searching module 126 may execute the query against the corpus 112 and the similar passage database 114, and retrieve information including similar passages information that satisfies the query. The similar passages are displayed and listed in accordance with a ranking order provided by the ranking engine 130.

[0033] The network 122 represents communication pathways between the data store 110, passage mining engine 116, client 118, web server 120, the scoring engine 128, and the ranking engine 130. In one embodiment, the network 122 is the Internet. The network 122 can also utilize dedicated or private communications links that are not necessarily part of the Internet. In one embodiment, the network 122 uses standard communications technologies, protocols, and/or interprocess communications techniques. Thus, the network 122 can include links using technologies such as Ethernet, 802.11, integrated services digital network (ISDN), digital subscriber line (DSL), asynchronous transfer mode (ATM), etc. Similarly, the networking protocols used on the network 122 can include the transmission control protocol/Internet protocol (TCP/IP), the hypertext transport protocol (HTTP), the simple mail transfer protocol (SMTP), the file transfer protocol (FTP), the short message service (SMS) protocol, etc. The data exchanged over the network 122 can be represented using technologies and/or formats including the hypertext markup language (HTML), the extensible markup language (XML), etc. In addition, all or some of links can be encrypted using conventional encryption technologies such as the secure sockets layer (SSL), HTTP over SSL (HTTPS), and/or virtual private networks (VPNs). In another embodiment, the nodes can use custom and/or dedicated data communications technologies instead of, or in addition to, the ones described above.

[0034] FIG. 2 is a high-level block diagram illustrating a functional view of a typical computer 200 for use as one or more of the entities illustrated in the environment 100 of FIG. 1 according to one embodiment. Illustrated are at least one processor 202 coupled to a bus 204. Also coupled to the bus 204 are a memory 206, a storage device 208, a keyboard 210, a graphics adapter 212, a pointing device 214, and a network adapter 216. A display 218 is coupled to the graphics adapter 212.

[0035] The processor 202 may be any general-purpose processor such as an INTEL x86 compatible-CPU. The storage device 208 is any device capable of holding data, like a hard drive, compact disk read-only memory (CD-ROM), DVD, or a solid-state memory device. The memory 206 holds instructions and data used by the processor 202 and may be, for example, firmware, read-only memory (ROM), non-volatile random access memory (NVRAM), and/or RAM. The pointing device 214 may be a mouse, track ball, or other type of pointing device, and is used in combination with the keyboard 210 to input data into the computer system 200. The graphics adapter 212 displays images and other information on the display 218. The network adapter 216 couples the computer system 200 to the network 122.

[0036] As is known in the art, the computer 200 is adapted to execute computer program modules. As used herein, the term “module” refers to computer program logic and/or data for providing the specified functionality. A module can be implemented in hardware, firmware, and/or software. In one embodiment, the modules are stored on the storage device 206, loaded into the memory 206, and executed by the processor 202 as one or more processes.

[0037] The types of computers used by the entities of FIG. 1 can vary depending upon the embodiment and the processing power utilized by the entity. For example, the client 118 typically requires less processing power than the passage mining engine 116, scoring engine 128, ranking engine 130, and web server 120. Thus, the client 118 system can be a standard personal computer or a mobile telephone. The passage mining engine 116, scoring engine 128, ranking engine 130, and web server 120, in contrast, may comprise processes executing on more powerful computers, logical processing units, and/or multiple computers working together to provide the functionality described herein. Further, the passage mining engine 116, scoring engine 128, ranking engine 130, and web server 120 might lack devices that are not required to operate them, such as displays 218, keyboards 210, and pointing devices 214.

[0038] Embodiments of the entities described herein can include other and/or different modules than the ones
described here. In addition, the functionality attributed to the modules can be performed by other or different modules in other embodiments. Moreover, this description occasionally omits the term “module” for purposes of clarity and convenience.

In one embodiment, frequency of appearance of the similar passage, or the number of similar passage instances in the digital corpus 112, is a characteristic that is examined. The characteristics analysis module 302 examines and identifies the frequency of appearance of the similar passage in the digital corpus 112. If the similar passage appears in fewer documents, the characteristics analysis module 302 assigns a lower score to that similar passage. If the similar passage appears in many documents, the characteristics analysis module 302 assigns a higher score to that similar passage.

In some embodiments, there are certain similar passages that tend to appear very frequently and the characteristics analysis module 302 adjusts the score downward as a result. For example, a cliché or overused slogan may be identified as a similar passage and may be very prevalent throughout the digital corpus 112. In those instances, the cliché or slogan may be assigned a lower score because the high frequency of occurrence does not necessarily indicate that the passage has great significance.

In some embodiments, the length of the similar passage may be a factor in determining a score based on the frequency of appearance of the similar passage. For example, a very short similar passage (for example, one that including less than five or six words) may appear frequently. However, since this passage is shorter than the average length of a passage, it is assigned a lower score. Conversely, if the similar passage is long (for example, more than ten words in length), it would still be assigned a high score if the frequency of appearance of the similar passage within the digital corpus 112 is high. In one embodiment, the score associated with the frequency of appearance of the similar passage in the digital corpus 112 is represented by F(Q).

In one embodiment, the length of the similar passage is a characteristic that is separately examined and scored by the characteristics analysis module 302. The characteristics analysis module 302 assigns a lower score to a very short passage (for example, one that including less than five or six words) and assigns a higher score to a long passage (for example, more than ten words in length). In one embodiment, the score associated with the length of the similar passage in the digital corpus 112 is represented by L(Q).

In one embodiment, the variation of words and grammar of the similar passage are characteristics that are examined. The characteristics analysis module 302 examines the words of the similar passage and assigns a score to the similar passage in response. The characteristics analysis module 302 assigns a lower score to a similar passage that contains repeating words or numbers and assigns a higher score to a passage that contains few repeating words or numbers. In some embodiments, if the similar passage is a chart, or another table-like presentation of words (i.e. words with no verbs), then the characteristics analysis module 302 assigns a lower score to that similar passage.

In some embodiments, the characteristics analysis module 302 applies one or more language models to analyze the words of the similar passage. For example, language models may be used to determine whether the words of the similar passage demonstrate usage of proper grammar or whether the words contain too many numbers. In such embodiments, a high score is assigned to a passage that demonstrates use of proper grammar and a low score is assigned to a passage that demonstrates use of improper grammar. Additionally, the score of a passage that contains too many numbers is lowered. In one embodiment, the score associated
with the word analysis of the similar passage in the digital corpus is represented by \( W(Q) \).

[0049] In one embodiment, the usage of punctuation associated with the similar passage is identified and examined by the characteristics analysis module 302. For example, the use of quotation marks surrounding a similar passage is an indication that the similar passage is a quotation and therefore the passage is assigned a higher score. In one embodiment, the score associated with the use of punctuation marks is represented by \( P(Q) \).

[0050] In one embodiment, the document that contains a similar passage instance is a characteristic that is identified and examined by the characteristics analysis module 302. Similar to the analysis of the author of the document, the characteristics analysis module 302 compares the identified document to a list or database of previously-identified famous or known documents. In one embodiment, each document in the list or database has an associated score. In such embodiments, when the characteristics analysis module 302 compares the identified document to the list or database of documents, and the identified document is found therein, the module 302 assigns the score associated with that document to the similar passage instance. If the identified document is not found in the database, the module 302 assigns a low score or a score of zero. In some embodiments, the documents in the list or database do not have associated scores. In such embodiments, the module 302 assigns a score to the similar passage instance based on whether the identified document was found therein. In one embodiment, the assigned score is represented by \( B(Q) \).

[0051] In one embodiment, the set of words introducing a similar passage and the set of words following a similar passage is a characteristic that is examined. In some embodiments, these words are known as speech acts. For example, words such as “Person X says” or “Person X wrote” are indications that a similar passage is to follow. As another example, speech acts such as “said Person X” are indications that a similar passage appeared before the exemplary speech act phrase. A higher score is assigned to a similar passage that is introduced by or followed by a speech act. In one embodiment, the assigned score is represented by \( S(Q) \).

[0052] In one embodiment, a diffusion of the similar passage in the digital corpus 112 is examined by the characteristics analysis module 302. In one embodiment, the assigned score is represented by \( D(Q) \) and is calculated by first calculating entropy scores as explained below.

[0053] In one embodiment, the variation of the authors, or number of different authors, of the documents containing a particular similar passage is a component of the diffusion score. The characteristics analysis module 302 examines the authors of the documents containing the instances of a particular similar passage in order to determine the number of different authors. The characteristics analysis module 302 assigns a higher score to a similar passage that is associated with many different authors, and assigns a lower score to a similar passage that is associated with fewer different authors. In one embodiment, the score is calculated using the following entropy equation:

\[
E(A) = - \sum_{x \in A} p(x) \log_2(p(x))
\]

[0054] As shown in the exemplary equation above, the entropy of the authors (\( E(A) \)), is calculated by taking the negative summation of the product of \( p(x) \) and the log of \( p(x) \), where \( p(x) \) is the probability that author \( x \) will occur in a given set of examined documents and is expressed as a fraction. For example, when calculating \( E(A) \), the individual probabilities correspond to the probability that a particular author will appear as an author of a document among the set of examined documents containing a particular similar passage. Using the equation above, if ten documents containing instances of a particular similar passage were examined and all ten documents were associated with the same author, \( p(x) \) would be one, and the entropy of the author (\( E(A) \)) would be zero. However, if some of the documents were associated with different authors, the entropy of the author (\( E(A) \)) would be greater than zero. If a large number of documents were examined and all the documents were associated with different authors, the value of the entropy of the authors would be high. For example, if ten documents were examined and ten authors were identified (each document corresponding to a different author), \( p(x) \log_2(p(x)) \) for each author is \(-0.332 \) and the negative summation is \( 3.32 \).

[0055] In one embodiment, the variation of the publishers of the documents associated with the particular similar passage is a component of the diffusion score. The publishers of the documents containing instances of the particular similar passage are examined and identified. Similar to the calculation for authors, the characteristics analysis module 302 calculates an entropy of the publishers (\( E(P) \)) by using a formula similar to the one above, but in this case \( p(x) \) corresponds to the probability of the occurrence of a particular publisher. Therefore, similar to the analysis of the authors, the characteristics analysis module 302 assigns a higher score to a similar passage that is associated with many different publishers, and assigns a lower score to a similar passage that is associated with fewer different publishers.

[0056] In one embodiment, the variation of the libraries that carry copies of the documents containing instances of the particular passage is a component of the diffusion score that is identified by the characteristics analysis module 302. Similar to the calculation for authors and publishers, the characteristics analysis module 302 calculates an entropy of the libraries (\( E(L) \)). In this case, \( p(x) \) corresponds to the probability of the appearance of a particular library that carries a copy of a document containing a particular similar passage. Therefore, similar to the analysis of the authors and publishers, the characteristics analysis module 302 assigns a higher score to a similar passage that is appears in a document that is held in a collection of many different libraries, and assigns a lower score to a similar passage that appears in a document that is held in a collection of fewer different libraries.

[0057] In one embodiment, the variation of the parts of documents in which the similar passage instances appear is a component of the diffusion score. The characteristics analysis module 302 examines and identifies parts of the documents in which the similar passage appears. In some embodiments, a document is divided into a number of parts. For example, a document may be divided into three parts: a first third (the beginning part of the document), a second third (the middle part of the document), and a last third (the end part of the document). Among the documents containing the similar passage instances, the characteristics analysis module 302 makes a determination as to which parts of the documents the similar passage instances appear. Similar to the calculations above,
the characteristics analysis module 302 calculates an entropy of the parts of the documents (E(Q)) using a similar formula. In this case, the p(x) corresponds to the probability of the appearance of a passage instance in a particular part of a document. Therefore, the characteristics analysis module 302 assigns a higher score to a similar passage that appears in different parts of documents, and assigns a lower score to a similar passage that appears in the same part, or mostly the same part, of the documents.

[0050] The characteristics analysis module 302 combines the entropies calculated above (E(A), E(P), E(L), and E(Q)) in order to calculate a total diffusion (D(Q)) of the similar passage throughout the corpus. Depending upon the embodiment, the characteristics analysis module 302 calculates D(Q) as a sum of its components, as a weighted linear combination, as a weighted geometric mean or using another technique. The characteristics analysis module 302 assigns the total diffusion score D(Q) to the similar passage. In some embodiments, the total diffusion score is stored in association with the similar passage in the similar passage database 114.

[0059] An embodiment of the score calculation module 306 combines the individual scores described above (A(Q), F(Q), L(Q), W(Q), P(Q), B(Q), S(Q), and D(Q)) to determine the total score assigned to a similar passage. In one embodiment, the total score is calculated by summing the individual scores. In some embodiments, certain individual characteristics are more important or more relevant than others. Therefore, the characteristics analysis module 302 weights scores for certain characteristics more than scores for other characteristics. In some embodiments, the total score is determined by a weighted linear combination of the individual scores. In other words, each individual score is assigned a weight and is multiplied by its assigned weight to yield a weighted score. The weighted scores are summed in order to yield the total score. In other embodiments, the total is determined by a weighted geometric mean. In other words, each score is assigned a weight. Each score is then raised to the power of the weight to yield a weighted score. The weighted scores are then multiplied together to yield the total score. In some embodiments, the sum of the weights equals one. Therefore, if one weight is increased by a certain amount the total of the other weights is decreased by the same amount such that the sum of the weights remains one.

[0060] The total score serves as the ranking score for the passage. In some embodiments, the score calculation module 306 aggregates a subset of the scores described above to produce the ranking score for a similar passage. Information about the similar passage and its associated ranking score are stored in the similar passage database 114.

[0061] FIG. 4 is a flow chart illustrating steps performed by the scoring engine 128 according to one embodiment. Other embodiments may perform different or additional steps than the ones shown in FIG. 4.

[0062] The scoring engine 128 receives 402 a set of similar passage instances for a passage in the digital corpus 112 to be analyzed. The scoring engine 128 calculates 404 the individual scores (A(Q), F(Q), L(Q), W(Q), P(Q), B(Q), S(Q), and D(Q)) for the examined characteristics. The scoring engine 128 then determines 406 a ranking score for the identified passage. In one embodiment, the individual scores are summed in order to produce a total score that serves as the ranking score for the identified passage. The scores can also be combined using one or more of the weighting techniques described above. The ranking score is associated with the passage and stored 408 in the similar passage database 114. This process can be performed for each similar passage in the similar passage database 114.

[0063] FIG. 5 is a flow chart illustrating the interaction between the client device 118 and web server 120, scoring engine 128 and ranking engine 130 according to one embodiment. Other embodiments may perform different or additional steps than the ones shown in FIG. 5.

[0064] A client device 118 sends 502 a request to the web server 120. The request from the client device 118 may be a search query entered by a user. In some embodiments, the request from the client device 118 may be created when the user selects a hypertext link presented on the client device. The web server 120 receives 504 the request and determines 506 a set of results from the similar passage database 114. The set of results is a set of similar passages. The ranking engine 130 ranks 508 the similar passages based on the ranking scores associated with the similar passages, thereby determining the order in which to display the similar passages. The search results are received 510 by the client device 118 and displayed 512 in the ranked order.

[0065] FIG. 6 is an exemplary web page 600 showing ranked search results according to one embodiment. In the example shown in FIG. 6, the page 600 displays search results 604 that are displayed when a user enters the search query “space race” in the search field 602 of the web page 600. The search results 604 identify three books that relate to the query “space race.” For each book, the web page 600 displays an image 606, a passage 608, and related terms and other information associated with the book/passage 610.

[0066] In FIG. 6, the books in the search results 604 are ranked based at least in part on the ranking score of the passage. The ranking score can be used to influence both the order of the books displayed in the search results and the selection of a particular passage from a book. For example, the first search result 604A displays the passage 608A “That’s one small step for a man. One giant leap for mankind.” This passage is highly quoted and thus would have received a very high ranking score relative to other passages. As a result, a book that contains this passage is presented first in the ranked order of books, and the passage itself is displayed in association with the book (as opposed to other passages appearing in the book that have lower ranking scores).

[0067] As used herein any reference to “one embodiment” or “an embodiment” means that a particular element, feature, structure, or characteristic described in connection with the embodiment is included in at least one embodiment. The appearances of the phrase “in one embodiment” in various places in the specification are not necessarily all referring to the same embodiment.

[0068] As used herein, the terms “comprises,” “comprising,” “includes,” “including,” “has,” “having” or any other variation thereof, are intended to cover a non-exclusive inclusion. For example, a process, method, article, or apparatus that comprises a list of elements is not necessarily limited to only those elements but may include other elements not expressly listed or inherent to such process, method, article, or apparatus. Further, unless expressly stated to the contrary, “or” refers to an inclusive or and not to an exclusive or. For example, a condition A or B is satisfied by any one of the following: A is true (or present) and B is false (or not present), A is false (or not present) and B is true (or present), and both A and B are true (or present).
In addition, use of the “a” or “an” are employed to describe elements and components of the embodiments herein. This is done merely for convenience and to give a general sense of the invention. This description should be read to include one or at least one and the singular also includes the plural unless it is obvious that it is meant otherwise.

Upon reading this disclosure, those of skill in the art will appreciate still additional alternative structural and functional designs for a system and a process for ranking similar passages through the disclosed principles herein. Thus, while particular embodiments and applications have been illustrated and described, it is to be understood that the disclosed embodiments are not limited to the precise construction and components disclosed herein. Various modifications, changes and variations, which will be apparent to those skilled in the art, may be made in the arrangement, operation and details of the method and apparatus disclosed herein without departing from the spirit and scope defined in the appended claims.

What is claimed is:

1. A computer-implemented method for calculating a score for a passage having a plurality of instances occurring in a digital corpus, comprising:
   calculating at least one score based at least in part on characteristics of instances of the passage occurring in the digital corpus;
   generating a ranking score associated with the passage based at least in part on the calculated at least one score; and
   storing the ranking score in association with the passage in a computer-readable medium.

2. The method of claim 1, wherein a plurality of scores are calculated based on a plurality of characteristics of the instances of the passage occurring in the digital corpus, and wherein generating the ranking score comprises combining the plurality of scores to form the ranking score.

3. The method of claim 1, wherein calculating the at least one score comprises:
   accessing a database identifying authors and having associated author scores;
   determining whether an author of a document in the digital corpus in which a passage instance occurs is found in the database and
   responsive to the author being found in the database, calculating the score based at least in part on the author score associated with the author in the database.

4. The method of claim 1, wherein calculating the at least one score comprises:
   accessing a database identifying documents and having associated document scores;
   determining whether a document in the digital corpus in which a passage instance occurs is found in the database and
   responsive to the document being found in the database, calculating the score based at least in part on the document score associated with the document in the database.

5. The method of claim 1, wherein calculating the at least one score comprises:
   identifying a frequency that the passage instances appear in the digital corpus; and
   calculating the score based at least in part on the frequency.

6. The method of claim 1, wherein calculating the at least one score comprises:
   determining a length of the passage; and
   calculating the score based at least in part on the length.

7. The method of claim 1, wherein calculating the at least one score comprises:
   determining an amount of variation of words of the passage; and
   calculating the score based at least in part on the amount of variation of words of the passage.

8. The method of claim 1, wherein calculating the at least one score comprises:
   applying one or more language models to analyze words within the passage; and
   calculating the score based at least in part on the application of the one or more language models.

9. The method of claim 1, wherein calculating the at least one score comprises:
   determining a usage of punctuation associated with the passage; and
   calculating the score based at least in part on the usage of punctuation associated with the passage.

10. The method of claim 1, wherein calculating the at least one score comprises:
    identifying words introducing the passage and/or following the passage in a document in the digital corpus containing an instance of the passage;
    ascertaining whether the words introducing and/or following the passage denote a speech act; and
    calculating the score based at least in part on whether the words introducing and/or following the similar passage denote a speech act.

11. The method of claim 1, wherein calculating the at least one score comprises:
    identifying a characteristic of the plurality of passage instances occurring in the digital corpus;
    examining the plurality of passage instances to determine an amount of variation in the identified characteristic over the plurality of passage instances; and
    calculating the at least one score based at least in part on the amount of variation in the characteristic.

12. The method of claim 11, wherein an identified characteristic is an author of a document in which a passage instance appears.

13. The method of claim 11, wherein an identified characteristic is a publisher of a document in which a passage instance appears.

14. The method of claim 11, wherein an identified characteristic is a library containing a document in which a passage instance appears.

15. The method of claim 11, wherein an identified characteristic is a part of a document in which a passage instance appears.

16. The method of claim 1, wherein a plurality of ranking scores are calculated for a plurality of different passages occurring in the digital corpus and further comprising:
   ranking the plurality of different passages in an order responsive to the ranking scores calculated for the passages.

17. A computer-readable storage medium containing executable program code for calculating a score for a passage having multiple occurrences in a digital corpus, the program code comprising code for:
   calculating at least one score based at least in part on characteristics of instances of the passage occurring in the digital corpus;
generating a ranking score associated with the passage based at least in part on the calculated at least one score; and
storing the ranking score in association with the passage in a computer-readable medium.

18. The computer-readable storage medium of claim 17, wherein a plurality of scores are calculated based on a plurality of characteristics of the instances of the passage occurring in the digital corpus, and wherein generating the ranking score comprises combining the plurality of scores to form the ranking score.

19. The computer-readable storage medium of claim 17, wherein calculating the at least one score comprises:
identifying a characteristic of the plurality of passage instances occurring in the digital corpus;
examining the plurality of passage instances to determine an amount of variation in the identified characteristic over the plurality of passage instances; and
calculating the at least one score based at least in part on the amount of variation in the characteristic.

20. The computer-readable storage medium of claim 17, wherein a plurality of ranking scores are calculated for a plurality of different passages occurring in the digital corpus and further comprising:
ranking the plurality of different passages in an order responsive to the ranking scores calculated for the passages.

21. A computer system for calculating a score for a passage having multiple occurrences in a digital corpus, the system comprising:
a computer-readable storage medium containing executable program code for calculating a score for a passage having multiple occurrences in a digital corpus, the program code comprising code for:
calculating at least one score based at least in part on characteristics of instances of the passage occurring in the digital corpus;
generating a ranking score associated with the passage based at least in part on the calculated at least one score; and
storing the ranking score in association with the passage in a computer-readable medium.

22. The computer system of claim 21, wherein a plurality of scores are calculated based on a plurality of characteristics of the instances of the passage occurring in the digital corpus, and wherein generating the ranking score comprises combining the plurality of scores to form the ranking score.

23. The computer system of claim 21, wherein calculating the at least one score comprises:
identifying a characteristic of the plurality of passage instances occurring in the digital corpus;
examining the plurality of passage instances to determine an amount of variation in the identified characteristic over the plurality of passage instances; and
calculating the at least one score based at least in part on the amount of variation in the characteristic.

24. The computer system of claim 21, wherein a plurality of ranking scores are calculated for a plurality of different passages occurring in the digital corpus and further comprising:
ranking the plurality of different passages in an order responsive to the ranking scores calculated for the passages.

* * * * *