

METHOD AND SYSTEM FOR STORING INFORMATION ABOUT DOCUMENTS TO PROVIDE DISPLAYS

Field of the Invention

5 This invention relates to processes for storing information about documents generated over a period of time to provide comparative information about the documents. More particularly, this invention relates to processes for storing and providing information about documents generated over a period time to show the change of the data in the documents. Still more particularly, this invention relates to processes for storing information
10 from documents retrieved over the Internet to show the change in occurrences of specific terms and phrases in the documents over a period of time.

Prior Art

 The Internet has become a very popular medium for dispersing information.
15 Recently, the rise in popularity of blogs and social networks has allowed more users to share views and opinions on all types of topics. As the amount of information on the Internet has proliferated, users have become interested in knowing what topics are popular with other users. To address this desire, Internet search engines have begun to provide information about the most popular searches requested by users and the most commonly accessed
20 sites. This type of information is often referred to as trending.

 However, the type of information provided by trending often does not provide details about the information about the content of the web pages and does not provide a historical perspective of how the trends have changed over time. This type of information may be
25 useful so that a user can determine the buzz and sentiments of others on certain topics and/or determine the issues that are commonly being discussed about the topic. One problem with providing this type of information is the amount data that must be analyzed to provide such information. A second problem is that a great amount of data would typically need to be stored in order to provide such information. Especially, if the user wants to see
30 the changes in the terms being used over time to see how the issues about the topic has changed. As such, those skilled in the art are constantly striving to provide systems that can provide this type of information about a topic over time that limits both the amount of analyzing needed and the amount of data that must be stored.

Summary of the Invention

35 The above and other problems are solved and an advance in the art is made by the method and system provided in accordance with this invention. A first advantage of a

method and system in accordance with this invention is that it provides data to provide displays that show the change in content of documents generated over a period time. A second advantage of a method and system in accordance with this invention is that the data to generate the displays may be produced without the need of retrieving and searching each document. A third advantage of a method and system in accordance with this invention is that the amount of data that must be stored to provide the information for the displays is reduced. This allows the method and/or system in accordance with this invention to be implemented using a device with limited memory and more efficiently.

In accordance with embodiments of this invention, information for providing information about data in documents generated over a time period to generate a display is provided in the following manner. A search matrix for information about the data in the documents is maintained in a memory of one or more devices. The search matrix includes a term index and an entity index. The term index stores a count for each term for each time unit in the time period. The count of each term represents a number of documents generated during a particular time unit that include the term. The entity index stores a count for each entity for each time unit of the time period. The count for each entity represents a number of documents generated during a particular time unit that include the entity. For purposes of this discussion, a term is a word or a phrase and; an entity is a name of a person, place, business, or thing. To provide the desired information, the counts of one or more terms and one or more entities over any number of time units of time is retrieved from term index and entity index of the search matrix to generate count information. The count information is then provided to generate a display.

In accordance with some embodiments of this invention, information about a document is added to the search matrix to maintain the search matrix. In accordance with some of these embodiments, the addition of information about a document to the search matrix begins by receiving the document from a source. A record for the document is then generated in a document index. The data from the document is then pre-processed. Each term and entity in the text of the document is then determined from the pre-processed data. The search matrix is then updated with the determined information for the document.

In accordance with some of these embodiments, the record in the document index is generated in the following manner. A record is created in the document index. The document is then assigned an identifier. The identifier is stored in the record. A pointer to a location storing the document is then stored in the record. The document may be stored

within the system or the pointer may refer to URL or other address identifier of the source of the document.

5 In accordance with some of these embodiments of this invention, the pre-processing of the data from the document is performed in the following manner. The process converts all text in the data to lower case letters. The text is then parsed. Each stop word is then removed from the parsed text of the document. In accordance with some of these embodiments, a term extractor is applied to the string in the pre-processed text data to identify each term in the text. Each term identified by the term extractor is then compared to
10 entries in a semantic database to identify each entity in the identified terms.

In accordance with some of the embodiment of this invention, the search matrix further includes an entity-document map, a term-document map, a document term map, and a time unit-document map. The entity-document map stores a record for each entity
15 indicating each of the documents that includes the entity. The term-document map stores a record for each term indicating the documents that include the term. The document-term map stores a record for each document indicating each term included in the document. The time unit-document map stores a record for each time units indicating the documents generated during the time unit. In accordance with some of these embodiments, the process
20 of updating search matrix includes updating the term index, the entity index, the time unit-document map, the entity-document map, the term-document map, and the document-term map when a document is added.

In accordance with some embodiments of this invention, the time unit-document map
25 is updated in the following manner. A field for a last document included in a time unit is updated with the identifier of the document being added and a counter indicating the number of documents generated during the time unit is incremented by one. In accordance with some of these embodiments, the process begins by determining if the time unit currently has a record in the time unit-document map and creates an entry for the time unit if no record
30 exists. The identifier of the document is then stored in the fields for the first and last document and the counter is set to one in the created entry.

In accordance with some embodiments of this invention, the term index is updated in the following manner. The number of occurrences or frequency of each term identified in the
35 document is determined. The frequency for each identified term is then compared to a threshold. If the frequency of a term is greater than or equal to the threshold, the frequency count of an entry for the term in the term index is updated. In accordance with some of

these embodiments, the term index is searched in response to a particular term having frequency greater than or equal to the threshold and an entry in the term index is created for the term and updated appropriately if no entry exists. In accordance with some of these embodiments of this invention, the term index is searched for the time unit in response to
5 determining the frequency of a term is greater than or equal to the threshold and an entry is created for the time unit if no entry exists.

In accordance with some embodiments of this invention, the entity index is updated in the following manner. The number of occurrences or frequency of each entity identified in the
10 document is determined. The frequency for each identified entity is then compared to a threshold. If the frequency of an entity is greater than or equal to the threshold, the frequency count of an entry for the entity in the entity index is updated. In accordance with some of these embodiments, the entity index is searched in response to a particular entity having a frequency greater than or equal to the threshold; and an entry in the entity index is
15 created for the entity and updated appropriately if no entry exists. In accordance with some of these embodiments of this invention, the entity index is searched for the time unit in response to determining a frequency of an entity is greater than or equal to the threshold and an entry is created for the time unit if no entry exists.

In accordance with some embodiments, the entity-document map is updated in the following manner. The document identifier assigned to the added document is inserted into the entry for each entity identified in the added document. In accordance with some of these
20 embodiments, the entity document map may first be searched for an entry for each entity identified in the added document. For each particular entity that an entry is not found, an
25 entry is created.

In accordance with some embodiments, the term-document map is updated in the following manner. The document identifier assigned to the added document is inserted into the entry for each term identified in the added document. In accordance with some of these
30 embodiments, the term-document map may first be searched for an entry for each term identified in the added document. For each particular term that an entry is not found, an entry is created.

In accordance with some embodiments of this invention, the document-term map is
35 updated in the following manner. An entry in the document-term map is created from the added document. The entry includes the identifier assigned to the added document and a list of each term identified in the added document.

In accordance with some embodiments of the invention, the document-entity map is updated in the following manner. An entry is generated in the document-entity map for the added document that includes the identifier assigned to the added document and a list of each entity identified in the added document.

In accordance with some embodiments of this invention, the search matrix further includes a term index and a search index for a second time period. The second time period is made up of time units that are the length of the time period. In accordance with some of these embodiments, the term index of the second time period is updated in the following manner. An entry for each term and the time unit of the time period is updated with sums of the counts for each time unit of the time period for each term. In accordance with some of these embodiments, the updating of the term index for the second time period is performed in the following manner. When the time period expires, an entry for the time period is created in the term index of the second time period. The entries for each term in the term index of the time period are then read. The counts read from the entries for each term are then summed and the sum is inserted into the entry for the term and the time period. The process may first determine if a particular term has an entry in the term index of the second time period and create an entry for the particular term if the entry does not exist before adding the sum into the entry for the particular term and time period. In accordance with some of these embodiments, the entity index of the second time period is updated in the following manner. An entry for each entity and the time unit of the time period is updated with sums of the counts for each time unit of the time period for each term. In accordance with some of these embodiments, the updating of the entity index for the second time period is performed in the following manner. When the time period expires, an entry for the time period is created in the entity index of the second time period. The entries for each entity in the entity index of the time period are then read. The counts read from the entries for each entity are then summed and the sum is inserted into the entry for the term and the time period. The process may first determine if a particular entity has an entry in the entity index of the second time period and create an entry for the particular entity if the entry does not exist before adding the sum into the entry for the particular entity and time period.

In accordance with some embodiments of the invention, the information provided to generate the display is generated in the following manner. A query is received. A determination is then made about the information requested by the query. In accordance with some embodiments, a query may request information about a specific entity. In accordance with these embodiments, information about the specific entity is retrieved from

the search matrix and provided. In accordance with some of these embodiments, the information is retrieved from the search matrix in the following manner. The count from each time unit from the specific entity is read from entity index of the time period. A list of the counts for the time units is generated. The entries for each time unit including a count for the entity are read from time unit-document map. The entry in the entity document map for the specific entity is then read and the list of documents including the entity is retrieved. The retrieved list of documents is then compared to the entries for each time unit to determine each document generated within each time unit of the time period. Each document determined to be in the range of time units is then added to a list of documents. The list of counts and the list of documents for the specific entity are then returned to be provided in the information.

In accordance with some embodiments, a query may request information about a specific term. In accordance with these embodiments, information about the specific term is retrieved from the search matrix and provided. In accordance with some of these embodiments, the information is retrieved from the search matrix in the following manner. The count from each time unit for the specific term is read from term index of the time period. A list of the counts for the time units is generated. The entries for each time unit including a count for the term are read from time unit-document map. The entry in the term document map for the specific term is then read and the list of documents including the term is retrieved. The retrieved list of documents is then compared to the entries for each time unit to determine each document generated within each time unit of the time period. Each document determined to be in the range of time units is then added to a list of documents. The list of counts and the list of documents for the specific term are then returned to be provided in the information.

In accordance with some embodiments, the query may request information for a range of time units of a particular time period. In accordance with some of these embodiments, the information for the range of time units of the particular time period is retrieved in the following manner. First, the term and entity indexes for the proper time period are determined. Term count information for the range is then retrieved by retrieving the count for each of the terms for each time unit in the range from term index for the particular time period. Entity count information for the range is then retrieved by retrieving the count for each of entities for each time units in the range from the entity index of the particular time period. In accordance with some embodiments, the term count information may be limited to the counts of a pre-determined number of terms and/or the entity count may be limited to a pre-determined number of entities.

In accordance with some embodiments, the query may include a request for information about terms associated with a particular entity. In accordance with some of these embodiments, the information provided to generate a display includes a list of terms
5 that is generated in the following manner. The process begins by an entry for the particular entity from the entity-document map. An entry in the document term map listed in the entry read for each of the documents listed in the entry is read from the entity-document map and the terms from the list of terms for each document is added to a term list that is included in the information provided to generate the display. In accordance with some embodiments,
10 the term list may be limited to a pre-determined number of terms.

In accordance with some embodiments, the query may request information about entities associated with a particular term. In accordance with some of these embodiments a list of associated entities is returned. In accordance with some of these embodiments, the
15 list of associated entities is generated in the following manner. The process begins by reading an entry for the particular term from the term-document map. The entity-document map is then searched for each entry that includes a document in the list of document in the entry for the term in the term-document map. Each entity associated with an entry that includes one of the documents is then added to the entity list. The entity list may then be
20 limited to a pre-determined number of entities.

Brief Description of the Drawings

The above and other features and advantages of a method and/or system in accordance with this invention are described in the following Detailed Description and are
25 shown in the following drawings:

Figure 1 illustrates devices communicatively connected to a network to provide or perform a system and/or method in accordance with embodiments of this invention;

Figure 2 illustrates a processing system of a device performing processes to provide a system in accordance with this invention;

30 Figure 3 illustrates a flow diagram of a process for storing information about data from a source document in accordance with an embodiment of this invention;

Figure 4 illustrates a flow diagram for generating a record for storing information about the source document in accordance with an embodiment of this invention;

35 Figure 5 illustrates a flow diagram of a process for pre-processing data from the source document in accordance with an embodiment of this invention;

Figure 6 illustrates a flow diagram of a process for updating a search matrix with information about a source document in accordance with an embodiment of this invention;

Figure 7 illustrates a flow diagram of a process for updating information in a source-document map of the search matrix with information about the source document in accordance with an embodiment of this invention;

5 Figure 8 illustrates a flow diagram of a process for updating information in a term index for a first time period of the search matrix with information about a source document in accordance with an embodiment of this invention;

Figure 9 illustrates a flow diagram of a process for updating information in an entity index for a first time period of the search matrix with information about a source document in accordance with an embodiment of this invention;

10 Figure 10 illustrates a flow diagram of a process for updating information in an entity-document map of the search matrix in accordance with an embodiment of this invention;

Figure 11 illustrates a flow diagram of a process for updating information in a term-document map of the search matrix in accordance with an embodiment of this invention;

15 Figure 12 illustrates a flow diagram of a process for updating information in a Document Term map of the search matrix in accordance with an embodiment of this invention

Figure 13 illustrates a flow diagram of a process for updating information in a document-entity map of the search matrix in accordance with an embodiment of this invention;

20 Figure 14 illustrates a flow diagram of for updating information in a document-term map of the search matrix in accordance with an embodiment of this invention;

Figure 15 illustrates a flow diagram of a process for updating information in a term index for a second time period made of time units of the first time period of the search matrix in accordance with an embodiment of this invention;

25 Figures 16 and 17 illustrate a flow diagram of a process for responding to a query in accordance with an embodiment of this invention;

Figure 18 illustrates a flow diagram of a process for retrieving information from the search matrix about an entity in accordance with an embodiment of this invention;

30 Figure 19 illustrates a flow diagram of a process for retrieving information from the search matrix about a term in accordance with embodiments of this invention;

Figure 20 illustrates a flow diagram of a process for retrieving information from the search matrix about terms associated with a specific entity in accordance with embodiments of this invention;

35 Figure 21 illustrates a flow diagram of a process for retrieving information from the search matrix about entities associated with a term in accordance with an embodiment of this invention;

Figure 22 illustrates a flow diagram of a process for providing information from the search matrix about the occurrence of terms in the documents generated over a period time;

Figure 23 illustrates a flow diagram of a process for providing information from the search matrix about the occurrence of entities in the documents generated over a period
5 time;

Figure 24 illustrates a representation of a display with the tags that may be generated with the information from a search matrix in accordance with embodiments of this invention;

Figure 25 illustrates a representation of the data structure of the term index in accordance with embodiments of this invention;

10 Figure 26 illustrates a representation of the data structure of the entity index in accordance with embodiments of this invention;

Figure 27 illustrates a representation of the data structure of an entity-document map in accordance with an embodiment of this invention;

15 Figure 28 illustrates a representation of the data structure of a term-document map in accordance with an embodiment of this invention;

Figure 29 illustrates a representation of the data structure of a document-term map in accordance with an embodiment of this invention; and

Figure 30 illustrates a representation of the data structure of a document-entity map in accordance with an embodiment of this invention.

Detailed Description

This invention relates to processes for storing information about document generated over a period of time to provide comparative information about the documents. More particularly, this invention relates to processes for storing and providing information about
25 documents generated over a period time to show the change of the data in the documents. Still more particularly, this invention relates to processes for storing information from documents retrieved over the Internet to show the change in occurrences of specific terms and phrases in the documents over a period of time.

30 It is envisioned that a system and/or method in accordance with embodiments of this invention may be used to provide information about documents read and/or received over the internet over a period of time. The documents may include "tweets", blogs, social network pages, and/or any other document having a Universal Resource Locator (URL). Figure 1 illustrates network 100 including provider server 105 and client devices 110,115,
35 and 120. Network 100 is a network such as the Internet that allows processing systems to communicate with one another. Provider server 105 is server or group of servers that store data and execute processes that provide content to user devices over network 100. Device

110 is a traditional desktop computer system that is connected to network 100 via either a wired or wireless connection and accesses the content of a provider via a web browser application performed by device 110. Device 115 is mobile device such a smartphone or Person Digital Assistant (PDA) that is connected to network 100 via a wireless and performs
5 a web browser or other specific application to access the content provided by the provider. Device 120 is a conventional laptop computer that is connected to network 100 via a wired or wireless connection and that accesses the provider content via a web browser or other application. One skilled in the art will recognize that devices 110, 115, and 120 are only provided as examples and other types of user devices may be used to access the content of
10 the provider without departing from this invention.

Processes stored as instructions in a media that are executed by a processing system in a user computer system; one or more connected servers; or other connected devices provide the method and/or system in accordance with this invention. The
15 instructions may be stored as firmware, hardware, or software. Figure 2 illustrates processing system 200 such as the processing system in a user's computer system, server, or other device that execute the instructions to perform the processes for providing a method and/or system in accordance with this invention. One skilled in the art will recognize that the exact configuration of each processing system may be different and the exact configuration
20 of the processing system in each device may vary. Thus, processing system 200 shown in Figure 2 is given by way of example only.

Processing system 200 includes Central Processing Unit (CPU) 205. CPU 205 is a processor, microprocessor, or any combination of processors and microprocessors that
25 execute instructions to perform the processes in accordance with the present invention. CPU 205 connects to memory bus 210 and Input/ Output (I/O) bus 215. Memory bus 210 connects CPU 205 to memories 220 and 225 to transmit data and instructions between the memories and CPU 205. I/O bus 215 connects CPU 205 to peripheral devices to transmit data between CPU 205 and the peripheral devices. One skilled in the art will recognize that
30 I/O bus 215 and memory bus 210 may be combined into one bus or subdivided into many other busses and the exact configuration is left to those skilled in the art.

A non-volatile memory 220, such as a Read Only Memory (ROM), is connected to memory bus 210. Non-volatile memory 220 stores instructions and data needed to operate
35 various sub-systems of processing system 200 and to boot the system at start-up. One skilled in the art will recognize that any number of types of memory may be used to perform this function.

A volatile memory 225, such as Random Access Memory (RAM), is also connected to memory bus 210. Volatile memory 225 stores the instructions and data needed by CPU 205 to perform software instructions for processes such as the processes required for providing a system in accordance with this invention. One skilled in the art will recognize that any number of types of memory may be used as volatile memory and the exact type used is left as a design choice to those skilled in the art.

I/O device 230, keyboard 235, display 240, memory 245, network device 250 and any number of other peripheral devices connect to I/O bus 215 to exchange data with CPU 205 for use in applications being executed by CPU 205. I/O device 230 is any device that transmits and/or receives data from CPU 205. Keyboard 235 is a specific type of I/O that receives user input and transmits the input to CPU 205. Display 240 receives display data from CPU 205 and display images on a screen for a user to see. Memory 245 is a device that transmits and receives data to and from CPU 205 for storing data to a media. Network device 250 connects CPU 205 to a network for transmission of data to and from other processing systems.

In accordance with this invention, a system and method are provided for maintaining a search matrix that stores information about documents that were generated during a time period. The information in the maintained search matrix may be provided to a system for generating displays that allow a user to analyze the content of the documents over the time period. In accordance with this invention, the displays may display "streaming" tags representing information in the documents over the time period. The tags may include terms or entities. For purposes of this discussion, a term is a phrase or word; and an entity is a name of a person, corporate entity, business, place, thing, or other object. The tags may be displayed in various sizes and colours indicate the importance of the term or entity in the tag with respect to the documents. As user moves along a time line, the tags change showing the change in the importance of terms and entities in the tags over a period of time. An example of a display with the tags that may be generated with the information from a search matrix in accordance with embodiments of this invention is shown in display 2400 of Figure 24. The information in the search matrix includes information about the documents generated during a specific time unit of a time period. For purposes of this discussion, a time period is a specified amount of time that is divided into multiple time units. Time units are a period of time that is smaller than the time period and is a divisible portion of the time period. For example, in some embodiments of this invention, time period may be a month with days and/or weeks as time units. Furthermore, a time period may be a time unit for a

second time period that is greater than the first time period. For example, a month may be a time unit for a year.

To provide the information required for such displays, the search matrix must include
5 information about the occurrence of terms and entities over the time period. The object of the search matrix is minimizing the amount of data that must be stored to provide the system and method in accordance with this invention. By minimizing the amount of data that must be stored, the system and method in accordance with this invention may be provided by systems that have limited memory and/or processing resources. To do so, the search matrix
10 includes a term index and an entity index. The term index maintains a count of the occurrence of each term for each time unit of the time period and the entity index maintains a count of the occurrence of each entity for each time unit of the time period. For purpose of this invention, the count of occurrences may be a count of the number of documents generated during the time unit that includes a particular term or entity or may be the total
15 number of times the particular term or entity appears in the documents. Figure 25 illustrates a representation of the data structure of the term index. Term index 2500 includes array of time units 2505 along a column and array of terms 2510 along a row. The intersection of the two arrays stores the occurrence count of a particular term for a particular time unit. One skilled in the art will recognize that any number of types of data structures including double
20 linked lists, arrays, objects and the likes can be used to generate term index 2500 without departing from this invention. Figure 26 illustrates a representation of the data structure of the entity index. Similar to term index 2500, entity index 2600 includes array of time units 2605 along a column and array of terms 2610 along a row. The intersection of the two arrays stores the occurrence count of a particular entity for a particular time unit. One skilled
25 in the art will recognize that any number of types of data structures including double linked lists, arrays, objects and the likes can be used to generate entity index 2600 without departing from this invention.

In addition to the term and entity indexes, the search matrix may include the following
30 items to quickly provide other information associated with the search: a document index, an entity-document map, a term-document map, a time-unit document map, a document-entity map, and a document-term map. The document index stores information about a document which may include an identifier assigned by the system, a time of generation, and a pointer to a location storing the document. One skilled in the art will recognize that the document
35 index may be a database or equivalent data structures that include a field for each item listed for each entry.

The entity-document map stores an entry for each entity found in the documents. The entry includes a field for the entity and a list of documents that includes an occurrence of the entity. Figure 27 illustrates a representation of the data structure of an entity document map in accordance with an embodiment of this invention. As shown, entity-document map 2700 includes a field for the entity and a list of documents that include the entity in the entity field. One skilled in the art will recognize that data structure may be realized as a database, objects with map or/and list structures, linked list objects, or the likes without departing from this invention.

The term-document map stores an entry for each term found in the documents. The entry includes a field for the term and a list of documents that includes an occurrence of the term. Figure 28 illustrates a representation of the data structure of a term document map in accordance with an embodiment of this invention. As shown, term-document map 2800 includes a field for the entity and a list of documents that include the term in the term field. One skilled in the art will recognize that data structure may be realized as a database, objects with map or/and list structures, linked list objects, or the likes without departing from this invention.

The document-term map stores an entry for each entity found in the documents. The entry includes a field for the identifier assigned to a document and a list of terms that occur in the document. Figure 29 illustrates a representation of the data structure of a document-term map in accordance with an embodiment of this invention. As shown, document-term map 2900 includes a field for the identifier of the document and a list of terms that occur in the document. One skilled in the art will recognize that data structure may be realized as a database, objects with map or/and list structures, linked list objects, or the likes without departing from this invention.

The document-entity map stores an entry for each entity found in the documents. The entry includes a field for the identifier assigned to a document and a list of entity that occur in the document. Figure 30 illustrates a representation of the data structure of a document-entity map in accordance with an embodiment of this invention. As shown, document-entity map 3000 includes a field for the identifier of the document and a list of entity that occur in the document. One skilled in the art will recognize that data structure may be realized as a database, objects with map or/and list structures, linked list objects, or the likes without departing from this invention.

In order to maintain the search matrix, the system and method must populate and update the search matrix with information from documents received and analyzed. Figure 3 illustrate process 300 that is a process for adding information about documents to the search matrix in accordance with an embodiment of this invention. Process 300 begins in step 305
5 when a new document is received. In step 310, process 300 generates a record for the document in a document index. An example of a process for generating the record for the document is discussed below with respect to Figure 4.

In step 315, the data from the received document is pre-processed. An example of a
10 process for pre-processing the data from a document is described below with respect to figure 5. In step 320, the terms that occur in the document are identified. This may be performed by applying a term extractor to the pre-processed data. The number of times a term occurs in the document is also counted and an occurrence count is stored for later use. After the terms are extracted, the entities that occur in the document are determined in step
15 325. The entities may be identified by comparing the terms identified in the document to entries in a semantic database that stores words and/or strings that are identified as entities.

In step 330, the time of generation of the document is identified and converted to a standard format. The time may be resolved to any time unit preferable. In accordance with
20 some embodiments, the smallest time unit is a day and the time is converted to a format yyyyymmdd, where yyyy is the year, mm is the month, and dd is the date the document is generated. In step 335, the search matrix is updated. Examples of the processes for updating the search matrix are described below with respect to Figures 6-15.

Figure 4 illustrates process 400 that is an embodiment for generating a document
25 record in a document index in accordance with an embodiment of this invention. Process 400 begins in step 405 by creating a document record in the document index. A document identifier is then assigned to the document in step 410. The document identifier is assigned by the system to provide a uniform means for identifying the documents that have been analyzed. In accordance with some embodiments, the format of the identifier is Lnnnnnnn,
30 where L is a letter typically starting with the letter 'A' and nnnnnn is a number. The number is incremented for each document received until reaching 99999 then the letter is incremented in alphabetical order. This, ensure a standard form for identifying documents that is easily searched. The assigned identifier is then stored in an identifier field of the
35 record for the document in step 415.

A pointer to the data content of a copy of the source document is then stored in the record for the document in step 420. The pointer may be an address in memory of the device providing the system in accordance with embodiments where the document is stored by the device. In other embodiments, the pointer may be the URL from which the document was obtained. Storing the URL, reduces the amount of memory needed for the system as a copy of each document need not be stored in the memory of the device providing the system. Process 400 then ends. One skilled in the art will recognize that other types of information about the document may be stored in the records of the document index by adding fields to the records without departing from this invention.

Figure 5 illustrates process 500 that is a process for pre-processing a received document in accordance with this invention. Process 500 begins in step 505 by converting the text in the documents to all lower case letters. The all lower cased letter is then parsed in step 510. The stop words are then removed from the parsed data in step 515 and process 500 ends.

Figure 6 illustrates process 600 that is an example for updating a search matrix in accordance with embodiments of this invention. Process 600 begins in step 605 by updating a time unit-document map in step 605. An example of a process for updating a time unit-document map is described below with reference to Figure 7. In step 610, a term index for the time period is updated. An example of a process for updating a term index in accordance with this invention is described below with reference to Figure 8. In step 615, an entity index for the time period is updated. An example of a process for updating an entity index in accordance with this invention is described below with reference to Figure 9.

An entity-document map is updated in step 620. An example of a process for updating an entity-document map is described below with reference Figure 10. A term-document map is updated in step 625. An example of a process for updating a term-document map is described below with reference Figure 11. A document-term map is updated in step 630. An example of a process for updating a document-term map is described below with reference Figure 12. A document-entity map is updated in step 635. An example of a process for updating a document-entity map is described below with reference Figure 13. Process 600 then ends. One skilled in the art will recognize that the steps in process 600 may be performed in any order without departing from this invention.

Figure 7 illustrates process 700 that is a process for updating a time-unit document map in accordance with some embodiments of this invention. Each entry in the time unit-

document map has a field for the time unit, a field for the first document in the time unit, a field for the last document generated in the time unit, and count of documents generated on that date. As discussed above, this scheme is possible due to the incremental assignment of identifiers to the documents and the fact that new documents are generally received within an order generated. This is especially relevant to embodiments that are receiving documents that were generated over the Internet where searches return the new documents to add when found. One skilled in the art will recognize that other schemes may need to be used if the documents are not regularly received in a fashioned ordered by the time of creation.

Process 700 begins in step 705 by determining whether the time unit of a received document has an entry in the time-unit document map. In step 710, process 700 determines whether an entry from the time unit was found. If an entry for the time unit is not found, an entry for the time unit is created in the time-unit document map in step 715. The document identifier assigned to the received document is then inserted into the first document field of the entry in step 720.

After step 720 or if there is an entry for the time unit in step 710, the last document field of the entry for the time unit is updated by inserting the identifier assigned to the received document in step 725. Process 700 then ends in step 730 by incrementing the document count of the entry.

Figure 8 illustrates process 800 that is a process for updated a term index for a received document in accordance with an embodiment of this invention having a term index structured in accordance with Figure 25 described above. Those skilled in the art will recognize that some modification may be needed to this process if the term index is maintained in a different manner.

Process 800 begins in step 805 by searching for an entry in the term index for the time unit. In accordance with the term index described above with reference to Figure 25 this may be performed by searching for the time unit in a time unit array. Process 800 determines whether an entry for the time unit is found in step 810. If no entry for the time unit is found in step 810, process 800 generates an entry for the time unit in step 815.

After the entry is generated or if it is determined that an entry for the time unit exists, process 800 selects one of the terms identified in the document in step 820. In step 825, process 800 determines the number of occurrences or frequency of the identified term in the

received document. The frequency is then compared to a threshold in step 835. The threshold should be chosen to limit the terms added to information in the search matrix and may be determined based upon study of the types of documents being reviewed. In a first alternative embodiment, the threshold is set to maintain a predetermined number of terms in the term index. In a second alternative embodiment, the threshold may be the frequency of the term having the smallest frequency in the matrix. This term may then be removed from the matrix if the frequency of the selected term is greater. The lowest frequency of the terms in the updated matrix would then be used as the threshold. This limits the size of the term index to a predetermined number of terms.

If the frequency is not greater than the threshold, process 800 proceeds to step 855. If the frequency of the term is greater than the threshold, process 800 determines whether the term has an entry in the term index in step 840. If there is no entry for the term in the term index, an entry is created in step 845. After the entry is created or found, the count in the entry for the time unit and the term is updated in step 850. If the count is counting the number of occurrences in the documents, the determined number of occurrences determined in step 825 is added to the count. If the count only counts the number of document in which the term occurs, the count is incremented by one for the term. After the count is updated or if the frequency is below the threshold, process determines if there is another identified term from the received document. If so, process 800 repeats from step 820 selecting a new term. If not, process 800 ends.

Figure 9 illustrates process 900 that is a process for updated an entity index for a received document in accordance with an embodiment of this invention having an entity index structured in accordance with Figure 26 described above. Those skilled in the art will recognize that some modification may be needed to this process if the entity index is maintained in a different manner.

Process 900 begins in step 905 by searching for an entry in the entity index for the time unit. In accordance with the entity index described above with reference to Figure 26 this may be performed by searching for the time unit in a time unit array. Process 900 determines whether an entry for the time unit is found in step 910. If no entry for the time unit is found in step 910, process 900 generates an entry for the time unit in step 915.

After the entry is generated or if it is determined that an entry for the time unit exists, process 900 selects one of the entities identified in the document in step 920. In step 925, process 900 determines the number of occurrences or frequency of the selected entity in the

received document. The frequency is then compared to a threshold in step 935. The threshold should be chosen to limit the entities added to information in the search matrix and may be determined based upon study of the types of documents being reviewed. In a first alternative, the threshold is set to maintain a predetermined number of entities in the entity index. Alternatively, the threshold may be the frequency of the entity having the smallest weighted frequency in the matrix. This entity may then be removed from the matrix if the weighted frequency of the selected entity is greater. The lowest weighted frequency of the entity in the updated index would then be used as the threshold. This limits the size of the entity index to a predetermined number of terms.

If the frequency is not greater than the threshold, process 900 proceeds to step 955. If the frequency of the entity is greater than the threshold, process 900 determines whether the entity has an entry in the entity index in step 940. If there is no entry for the entity in the entity index, an entry is created in step 945. After the entry is created or found, the count in the entry for the time unit and the entity is updated in step 950. If the count is counting the number of occurrences in the documents, the determined number of occurrences determined in step 925 is added to the count. If the count only counts the number of document in which the entity occurs, the count is incremented by one for the entity. After the count is updated or if the frequency is below the threshold, process determines if there is another identified entity from the received document. If so, process 900 repeats from step 920 selecting a new entity. If not, process 900 ends.

Figure 10 illustrates process 1000 that is a process for updating an entity-document map for an entity in accordance with an embodiment of this invention that maintains an entity-document map as described above with respect to Figure 27. Process 1000 is performed for each entity that is identified in the received document. Process 1000 begins in step 1005 by searching for an entry for the entity in the entity-document map. Step 1010 determines whether an entry for the entity was found by the search in step 1005. If an entry is not found, process 1000 creates an entry for the entity in step 1015. If an entry is found or after an entry has been created, the identifier assigned to the received document is added to the entry for the entity in step 1020 and process 1000 ends.

Figure 11 illustrates process 1100 that is a process for updating a term-document map for a term in accordance with an embodiment of this invention that maintains a term-document map as described above with respect to Figure 28. Process 1100 is performed for each term that is identified in the received document. Process 1100 begins in step 1105 by searching for an entry for the term in the term-document map. Step 1110 determines

whether an entry for the term was found by the search in step 1105. If an entry is not found, process 1100 creates an entry for the term in step 1115. If an entry is found or after an entry has been created, the identifier assigned to the received document is added to the entry for the term in step 1120 and process 1100 ends.

5

Figure 12 illustrates process 1200 that updates a document-term map in accordance with an embodiment of this invention that maintains a document-term map as described above with reference to Figure 29. Process 1200 begins in step 1205 by creating an entry for the received document and inserting the identifier assigned to the received document into the document field. The list of terms identified in the document is then inserted into the term list file in step 1210 and process 1200 ends.

10

Figure 13 illustrates process 1300 that updates a document-entity map in accordance with an embodiment of this invention that maintains a document-entity map as described above with reference to Figure 30. Process 1300 begins in step 1305 by creating an entry for the received document and inserting the identifier assigned to the received document into the document field. The list of entities identified in the document is then inserted into the entity list filed in step 1310 and process 1300 ends.

15

In accordance with some embodiment of this invention, the search matrix may include entity and term indexes for more than one time period. In particular, these embodiments the search matrix have term and entity indexes for larger time periods that summarize the indexes of smaller time periods that form the time units of the larger time period. For example, an embodiment may maintain term and entity indexes for a month which has daily time units and term and entity indexes for a year which has monthly time units. In the embodiments, the indexes of the smaller time units are summed and added to an entry the smaller time period in the index for the larger time period. In the given example, this is performed by summing the daily entries of a monthly index at the end of a month and adding an entry for the month to the yearly term and entity indexes. One skilled in the art will recognize that such "scoping" of time periods may be done any number of times depending on the length of time periods used without departing from this invention.

20

25

30

Figure 14 illustrates process 1400 for updating a term index of a second, larger time period from the index of a first, smaller time period in accordance with an embodiment of this invention. Process 1400 begins in step 1405 when the first, smaller time period expires. At this time, the index for the time period may be saved for use in retrieving data for the particular time units stored in the term index of the expired time period. In step 1410, an

35

entry is made in the time units for the first, smaller time period that has expired. In accordance with term indexes described above with reference to Figure 26, the first smaller, time period is added to the array of time units of the term index of the second time period.

5 In step 1415, a term is read from the term index of the first smaller time period. Process 1400 then determines whether the term index of the second time period has an entry for the term in step 1420. This is to account for the fact that the usage of terms will change over time and new terms may have been added in just expired smaller first time period. If an entry for the term is not found in the term index of the second time period, an
10 entry for the term is created in step 1425. The time units of the term index of the first time period are then read for the term and the counts from each time unit is summed to generate a total count for the time period in step 1430. The count is then inserted into the entry for term and the first smaller time period in the term index of the second time period in step 1435. Process 1400 then determines whether the term index includes another term in step
15 1440. If so, process 1400 is repeated from step 1415 until all of terms in the term index of the first time period have been handled. Otherwise, process 1400 repeats from step 1405.

Figure 15 illustrates process 1500 for updating an entity index of a second, larger time period from the index of a first, smaller time period in accordance with an embodiment
20 of this invention. Process 1500 begins in step 1505 when the first, smaller time period expires. In step 1510, an entry is made in the time units for the first, smaller time period that has expired. At this time, the index for the time period may be saved for use in retrieving data for the particular time units stored in the entity index of the expired time period. In accordance with entity indexes described above with reference to Figure 27, the first smaller,
25 time period is added to the array of time units of the term index of the second time period.

In step 1510, an entity is read from the entity index of the first smaller time period. Process 1500 then determines whether the entity index of the second time period has an entry for the entity in step 1520. This is to account for the fact that the usage of entities will
30 change over time and new entities may have been added in just expired smaller first time period. If an entry for the entity is not found in the term index of the second time period, an entry for the entity is created in step 1525. The time units of the entity index of the first time period are then read for the term entity and the counts from each time unit is summed to generate a total count for the time period in step 1530. The count is then inserted into the
35 entry for entity and the first smaller time period in the entity index of the second time period in step 1535. Process 1500 then determines whether the entity index includes another entity in step 1540. If so, process 1500 is repeated from step 1515 until all of entities in the entity

index of the first time period have been handled. Otherwise, process 1500 repeats from step 1505.

The indexes and maps maintained as part of the search matrix are used to provide information about documents to provide displays. As stated above, it is envisioned that the information provided to generate the display will include count information about the entities and terms over a time period or a range of time units within a time periods. The following description of Figures 16-23 describes the processes and types of information that may be provided from the search matrix to generate the displays in accordance with embodiments of this invention.

Figures 16 and 17 illustrate process 1600 for receiving a query and providing the desired information. Process 1600 begins in step 1605 by receiving an input query. In step 1610, process 1600 determines whether the query includes a request for information about a particular entity. If so, process 1600 provides count information for the particular entity for the time units from the entity index for the current time period in step 1615. A list of documents that include the entity may also be returned. A process for retrieving this information is described below with respect to Figure 18. In step 1620, process 1600 determines whether the query includes a request for information about a particular term. If so, process 1600 provides count information for the time units from the term index of the current time period. A list of documents that include the term may also be returned. A process for retrieving this information is described below with respect to Figure 19.

In step 1630, process 1600 determines whether the query is for information about a particular entity over a particular range of time unit. If so, process 1600 retrieves frequency information for the particular entity over the range of units, terms associated with the entity over the input range of time units and a list of documents that include the entity over the given range of time units in step 1635. Processes for providing this information are described below with respect to Figures 20 and 23. In step 1640, process 1600 determines whether the query is for information about a particular term over a particular range of time unit. If so, process 1600 retrieves frequency information for the particular terms over the range of units, entities associated with the term over the input range of time units and a list of documents that include the term over the given range of time units in step 1645. Processes for providing this information are described below with respect to Figures 21 and 22. In step 1650, process 1600 determines whether the input query includes a request for information for a particular range of time units for a particular time period. If so, process 1600 retrieves count information for entities and terms from documents generated in the

requested range of time units in step 1655. Examples for process for retrieving the count information the entities and terms are described below with respect to Figures 22 and 23.

5 In step 1660, process 1600 determines whether the query includes a request about a particular entity of a range of first time periods. The input query is requesting information relating to time units in a second time period that are equal to a first time period. If so, process 1600 retrieves frequency information for the particular entity over the range of time periods in the second large time period, terms associated with the entity over the input range of first time periods and a list of documents that include the entity over the given range of time units in step 1665. Processes for providing this information are described below with respect to Figures 20 and 23. In step 1670, process 1600 determines whether the query is for information about a particular term over a particular range of first time periods within the second larger time period. If so, process 1600 retrieves frequency information for the particular terms over the range of time periods, entities associated with the term over the input range of first time periods and a list of documents that include the term over the given range of first time periods in step 1675. Processes for providing this information are described below with respect to Figures 21 and 22. In step 1680, process 1600 determines whether the input query includes a request for information for a particular range of first time periods within a second larger time period. If so, process 1600 retrieves count information for entities and terms from documents generated in the requested range of first time periods in the second larger time period in step 1685. Examples for process for retrieving the count information the entities and terms are described below with respect to Figures 22 and 23.

25 Figure 18 illustrates process 1800 for retrieving information about a particular entity in accordance with embodiments of this invention. Process 1800 may be performed in response to any request for information about a particular entity. One skilled in the art will recognize that the exact steps of the process being performed and information returned may depend on the specific query received. Process 1800 begins in step 1805 by reading the count information for the entity and each time unit in the range of time units from the entity index of the proper time period that includes the time units within the range. In some embodiments, the reading may be performed by finding an entry for the entity in the array and reading the count from each time unit in the time unit array that is within the requested range.

35 In step 1810, a list of counts is generated that indicates the count for the entity for each time unit in the range. The entry of each time unit in the time-unit document map is then read in step 1815. The documents from each entry read are then added to a list of

documents within the range of time unit in step 1820. The entry for the entity is then read from the entity-document map in step 1825. The list of documents for time unit is then compared to the list of documents for the entity and a list of entity-documents containing only those documents in both lists is created in step 1830. The list of counts and the list of entity-documents are then returned in step 1835 and process 1800 ends.

Figure 19 illustrates process 1900 for retrieving information about a particular term in accordance with embodiments of this invention. Process 1900 may be performed in response to any request for information about a particular term. One skilled in the art will recognize that the exact steps of the process being performed and information returned may depend on the specific query received. Process 1900 begins in step 1905 by reading the count information for the term and each time unit in the range of time units from the term index of the proper time period that includes the time units within the range. In some embodiments, the reading may be performed by finding an entry for the term in the array of terms and reading the count from each time unit in the time unit array that is within the requested range.

In step 1910, a list of counts is generated that indicates the count for the term for each time unit in the range. The entry of each time unit in the time unit-document map is then read in step 1915. The documents from each entry read are then added to a list of documents within the range is time unit in step 1920. The entry for the term is then read from the term-document map in step 1925. The list of documents for time unit is then compared to the list of documents for the term and a list of term-documents containing only those documents in both lists is created in step 1930. The list of counts and the list of term-documents are then returned in step 1935 and process 1900 ends.

Figure 20 illustrates process 2000 that is a process for returning terms associated with an entity within a range of time units in accordance with this invention. This process may be performed any time information about the terms associated with a specific entity is request by a query. Process 2000 begins in step 2005 by retrieving the entry for the particular entity from the entity-document map. The entries for at least the first and last time units of the range of time units is read in step 2010 to determine the documents within the range of time units. The entry in the document-term map for each document from entry read from the entity-document map determined to be within the range of time units is then retrieved in step 2015 and the document is added to a document list in step 2020.

The terms from the entries retrieved from the document term map are then determined and added to a list of associated terms in steps 2025 and 2030. Optional step 2035 then retrieves the counts of each term in the list of associated terms for use in truncating the list. This may be performed by a process similar to the process described above with respect to figure 19. The list of terms then may be limited to those terms having a count above a certain threshold or a pre-determined number of terms having the highest counts for the range of time units in optional step 2045. The list of associated terms and possibly the list of documents are then returned in step 2040 after step 2030 or optional step 2045. Process 2000 then ends.

Figure 21 illustrates process 2100 that is a process for returning entities associated with a term within a range of time units in accordance with this invention. This process may be performed any time information about the entities associated with a specific term is request by a query. Process 2100 begins in step 2105 by retrieving the entry for the particular term from the term-document map. The entries for at least the first and last time units of the range of time units is read in step 2110 to determine the documents within the range of time units. The entry in the document-entity map for each document from entry read from the term-document map determined to be within the range of time units is then retrieved in step 2115 and the document is added to a document list in step 2120.

The entities from the entries retrieved from the document-entity map are then determined and added to a list of associated terms in steps 2125 and 2130. Optional step 2135 then retrieves the counts of each entity in the list of associated entities for use in truncating the list. This may be performed by a process similar to the process described above with respect to figure 18. The list of entities then may be limited to those terms having a count above a certain threshold or a pre-determined number of terms having the highest counts for the range of time units in optional step 2145. The list of associated entities and possibly the list of documents are then returned in step 2140 after step 2130 or optional step 2145. Process 2100 then ends.

Figure 22 illustrates process 2200 that is a process for providing count information for terms in a range of time units in accordance an embodiment of this invention. Process 2200 is performed when information is needed about the counts of terms for any particular range of time units. Process 2200 begins in step 2205 by determining the term index of the proper time period or time periods to read. In step 2210, process 2200 selects a time unit. This may be performed by finding the entry of time unit in the array of time units to determine the position of counts for the time units. A term is then selected in step 2215. Process 2200

then determines whether there is a count for the term for the selected time unit in step 2220. This may be performed by determining the position of the term in the array of terms for the term index and determining the position of the count cross referenced with the position of the time unit in the time unit array of the term index. One skilled in the art will recognize that exact processes for performing steps 2210-2220 will determine the exact data structure used to provide the term index.

If the term has a count associated with the selected time unit, process 2200 determines whether a result matrix includes an entry for the term in step 2225. If there is no entry for the term, an entry for the term is created in step 2230. After the entry is created or if the term already has an entry, the count read from the term index for the entry and time unit is inserted into the proper position in the result index in step 2235. One skilled in the art will recognize that other processes may be used depending on the data structure used to provide the result matrix.

In step 2240, process 2200 determines whether the term index includes another term and the process is repeated from step 2215 until the counts for all of the terms are read for a particular time unit. Process 2200 then determines whether there is another time unit in the range in step 2245. If there is another time unit in the range, process 2200 is repeated from step 2210 until the counts of the terms for all of the time units have been read. If all of the time units have been read, process may perform optional step 2250 of limiting the result index to a pre-determined number of terms. This may be performed by summing the counts of each term for all of the time units and comparing the sums to a threshold or only including a predetermined number of terms with the greatest counts. The result index is then returned in step 2255 and process 2200 ends.

Figure 23 illustrates process 2300 that is a process for providing count information for entities in a range of time units in accordance an embodiment of this invention. Process 2300 is performed when information is needed about the counts of entities for any particular range of time units. Process 2300 begins in step 2305 by determining the entity index of the proper time period or time periods to read. In step 2310, process 2300 selects a time unit. This may be performed by finding the entry of time unit in the array of time units to determine the position of counts for the time units. An entity is then selected in step 2315. Process 2300 then determines whether there is a count for the entity for the selected time unit in step 2320. This may be performed by determining the position of the entity in the array of entities for the entity index and determining the position of the count cross referenced with the position of the time unit in the time unit array of the entity index. One skilled in the art will

recognize that exact processes for performing steps 2310-2320 will determine the exact data structure used to provide the entity index.

5 If the entity has a count associated with the selected time unit, process 2300 determines whether a result matrix includes an entry for the entity in step 2325. If there is no entry for the entity, an entry for the entity is created in step 2330. After the entry is created or if the entity already has an entry, the count read from the entity index for the entry and time unit is inserted into the proper position in the result index in step 2335. One skilled in the art will recognize that other processes may be used depending on the data structure used to
10 provide the result matrix.

In step 2340, process 2300 determines whether the entity index includes another entity and the process is repeated from step 2315 until the counts for all of the entities are read from the entity index for a particular time unit. Process 2300 then determines whether
15 there is another time unit in the range in step 2345. If there is another time unit in the range, process 2300 is repeated from step 2310 until the counts of the entities for all of the time units have been read. If all of the time units have been read, process may perform optional step 2350 of limiting the result index to a pre-determined number of entities. This may be performed by summing the counts of each entity for all of the time units and comparing the
20 sums to a threshold or only including a predetermined number of entities with the greatest counts. The result index is then returned in step 2355 and process 2300 ends.

The above is a description of embodiments of a system and method in accordance with this invention. It is envisioned that others can and will design alternative methods and
25 systems as set forth in the following claims based upon this disclosure.

What is claimed is:

1. A product for providing information about data in a plurality of documents generated over a time period to generate a display comprising:

5 instructions for directing said processing unit to:

maintain a search matrix for information about said data in said plurality of documents that includes a term index and an entity index where said term index stores a count for each of a plurality of terms and each of a plurality of time units in said time period wherein each count represents a number of documents of said plurality of documents generated during a particular time unit that include said term and said entity index stores a count for each of a plurality of entities and each of said plurality of time units of said time period wherein each count represents a number of documents of said plurality of documents generated during a particular time unit that include said entity further wherein a term is a word or a phrase and an entity is a name of at least one selected from a group consisting of a person, a place, a business, and a thing,

retrieve said count for each of a portion of said plurality of terms and each of a portion of said plurality of time units from said term index and said count for each of a portion of said plurality of entities and each of said portion of said plurality of time units to generate count information;

provide said count information to provide a display to show a comparison of occurrences of said portion of plurality of terms and said portion of said plurality of entities over said portion of said plurality of time units; and a media readable by said processing unit that stores said instructions.

25

2. The product of claim 1 wherein said instructions to maintain said search matrix comprise:

instructions for directing said processing unit to add information about a one of said plurality of documents to said search matrix.

30

3. The product of claim 1 wherein said search matrix further includes a document index that stores information about each of said plurality of documents and said instructions to add about said one of said plurality of documents comprise:

instructions for directing said processing to:

receive said one of said plurality of documents from a source,
generate a record for said one of said plurality of documents in a document index,

35

pre-process data in said one of said plurality of documents,
identify each term in said data of said one of said plurality of documents,
identify each entity in said data of said one of said plurality of documents, and
update said search matrix.

5

4. The product of claim 3 wherein said instructions to generate said record in said document index comprise:

instructions for directing said processing unit to:

10 generate said document record in said document index for said one of said plurality of documents,

assign an identifier to said one of said plurality of documents,

store said document in said record for said one of said plurality of documents in said identifier, and

15 store a pointer to a location storing said one of said plurality of documents in said record.

5. The product of claim 3 wherein said instructions to pre-process said data of said one of said plurality of documents comprises:

instructions for directing said processing unit to:

20 convert all text in said data to lower case letters,

parse said text, and

remove each stop word from said parsed text.

25 6. The product of claim 5 wherein said instructions to identify each term in said data comprises:

instructions for directing said processing unit to:

apply term extractor to strings to identify each term in said data.

30 7. The product of claim 6 wherein said instructions to identify each entity in said data comprises:

instructions for directing said processing unit to:

compare each said term identified by said term extractor to entries in a semantic database to identify each entity in said terms.

35 8. The product of claim 3 wherein said search matrix further includes an entity-document map which stores a record for each of said plurality of entities indicating each of said plurality of documents that include said entity, a term-document map which stores a

record for each of said plurality of terms indicating each of said plurality of documents that include said term, a document-term map which includes a record for each of said plurality of documents indicating each of said plurality of terms included in each of said plurality of documents, and a time unit-document map which includes a record for each of said plurality
5 of time units indicating each of said plurality of documents generated during each of said plurality of time units and said instructions for updating said search matrix comprise:

instructions for directing said processing unit to:

update said time unit-document map,
update said term index,
10 update said entity index,
update said entity-document map,
update said term-document map, and
update said document-term map.

15 9. The product of claim 8 wherein said instructions to update said time unit-document map comprise:

instructions for directing said processing unit to:

update a field for a last document included in a time unit with said identifier of
said one of said plurality of documents, and

20 increment a counter indicating a number of said plurality of documents
generate during said time unit.

10. The product of claim 9 wherein said instructions to update said time unit-document map further comprise:

25 instructions for directing said processing unit to:

search for said record of said time unit in said time-unit document map,
create an entry for said time unit in said time unit-document map responsive
to a determination that said record is not included in said time unit-document map,
and

30 insert said identifier of said one of said plurality of documents into a field for a
beginning document.

11. The product of 8 wherein said instructions to update said term index comprise:
instructions for directing said processing unit to:

35 determine a number of occurrences of each identified term in said one of said
plurality of documents,

determine whether said number of occurrences of each identified term is greater than or equal to a threshold; and

update a frequency count of an entry of each identified term in said term index having said number of occurrences greater than or equal to said threshold.

5

12. The product of claim 11 wherein said instructions to update said term index comprise:

instructions for directing said processing unit to:

10 determine whether said term index includes an entry for a particular term in response to said particular term having a number of occurrences greater than said threshold; and

add an entry for said particular term to said term index in response to a determination that said term index does not include an entry for said particular term.

15

13. The product of claim 11 wherein said instructions to update said term index comprise:

instructions for directing said processing unit to:

20 determine whether said term index includes an entry for a particular time unit of said one of said plurality of documents; and

add an entry for said particular time unit to said term index in response to a determination that said term index does not include an entry for said particular time unit.

25

14. The product of 8 wherein said instructions to update said entity index comprise:

instructions for directing said processing unit to:

determine a number of occurrences of each identified entity in said one of said plurality of documents,

30 determine whether said number of occurrences of each identified entity is greater than or equal to a threshold; and

update a frequency count of an entry of each identified entity in said entity index having said number of occurrences greater than or equal to said threshold.

35

15. The product of claim 14 wherein said instructions to update said entity index comprise:

instructions for directing said processing unit to:

determine whether said entity index includes an entry for a particular entity in response to said particular entity having a number of occurrences greater than said threshold; and

add an entry for said particular entity to said entity index in response to a determination that said entity index does not include an entry for said particular entity.

16. The product of claim 14 wherein said instructions to update said entity index comprise:

instructions for directing said processing unit to:

determine whether said entity index includes an entry for a particular time unit of said one of said plurality of documents; and

add an entry for said particular time unit to said entity index in response to a determination that said entity index does not include an entry for said particular time unit.

17. The product of claim 8 wherein said instructions to update said entity-document comprise:

instructions for directing said processing unit to:

insert said document identifier of said one of said plurality of documents into an entry of each entity identified in said one of said plurality of documents.

18. The product of claim 17 wherein said instructions to update said entity-document further comprise:

instructions for directing said processing unit to:

determine whether said entity-document map has an entry for each of entity identified in said one of said plurality of documents, and

create an entry for each said entity identified in said one of said plurality of documents that is determined to not have an entry in said entity-document map.

19. The product of claim 8 wherein said instructions to update said term-document map comprise:

instructions for directing said processing unit to:

insert said document identifier of said one of said plurality of documents into an entry of each term identified in said one of said plurality of documents.

20. The product of claim 19 wherein said instructions to update said term-document map further comprise:

instructions for directing said processing unit to:

determine whether said term-document map has an entry for each of term identified in said one of said plurality of documents, and

5 create an entry for each said term identified in said one of said plurality of documents that is determined to not have an entry in said term-document map.

21. The product of claim 8 wherein said instructions for updating said document-term map comprise:

10 instructions for directing said processing unit to generate an entry for said one of said plurality of documents that includes said document identifier and a list of each term identified in said one of said plurality of documents.

22. The product of claim 8 wherein said instructions for updating said document-entity map comprise:

15 instructions for directing said processing unit to generate an entry for said one of said plurality of documents that includes said document identifier and a list of each entity identified in said one of said plurality of documents.

20 23. The product of claim 8 wherein said search matrix further includes a term index and a search index for a second period of time wherein said second period of time is made up of time units of said period time and said instructions further comprise:

instructions for directing said processing unit to:

update said term index for said second time period with information about said plurality of documents generated during said period of time, and

25 update said entity index for said second time period with information about said plurality of documents generated during said period of time.

24. The product of claim 23 wherein said instructions to update said term index of said second period of time comprise:

30 instructions for directing said processing unit to:

determine said period of time has expired,

generate an entry for said period of time in said term index for said second period of time,

35 read entries for each of said plurality of terms in said term index for said period of time,

sum said count from each of said plurality of time units for each of said plurality of terms in said term index of said time period,

determine whether each of said plurality of terms has an entry in said term index for said second period of time,

create an entry in said term index of said second time period for each of said plurality of terms from said term index of said time period determined not to have an entry in said term index of said second time period, and

insert said sum of counts for each particular term of said plurality of terms into said entry for said particular term and said time period in said term index of said second period of time.

25. The product of claim 23 wherein said instructions to update said entity index of said second period of time comprise:

instructions for directing said processing unit to:

determine said period of time has expired,

generate an entry for said period of time in said entity index for said second period of time,

read entries for each of said plurality of entities in said term index for said period of time,

sum said counts from each of said plurality of time units for each of said plurality of entities in said entity index of said time period,

determine whether each of said plurality of entities has an entry in said entity index for said second period of time,

create an entry in said entity index of said second time period for each of said plurality of entities from said entity index of said time period determined not to have an entry in said entity index of said second time period, and

insert said sum of said counts for each particular entity of said plurality of entities into said entry for said particular entity and said time period in said entity index of said second period of time.

26. The product of claim 8 wherein said instructions further comprise:

instructions for directing said processing unit to:

receive an input query,

determine whether said query is for information about a specific entity, retrieve information about said specific entity from said search matrix, and

provide said information to generate a display.

27. The product of claim 26 wherein said instructions to retrieve said information comprise:

instructions for directing said processing unit to:

read frequency information for said specific entity for each said time unit from said entity index,

generate a list of frequencies for time units from said frequency information

5 read entries for each time unit in said list from said time-unit document map

read entry for said specific entity from said entity-document map,

compare documents listed in said entry from said entity-document map to said entries for each time unit to determine each document within said time units and add each document determined to be in range to a list of documents, and

10 return said list of frequencies and said list of documents for said information.

28. The product of claim 26 wherein said instructions further comprise:

instructions for directing said processing unit to:

determine whether said query is for information about a specific term,

15 retrieve information about said specific term from said search matrix, and

provide said information to generate a display.

29. The product of claim 28 wherein said instructions to retrieve said information comprise about said specific term comprise:

20 instructions for directing said processing unit to:

read frequency information for said specific term for each said time unit from said term index,

generate a list of frequencies for each said time unit from said frequency information

25 read entries for each time unit in said list from said time-unit document map

read entry for said specific term from said term-document map,

compare documents listed in said entry from said term-document map to said entries for each time unit to determine each document within said time units and add each document determined to be in range to a list of documents, and

30 return said list of frequencies and said list of documents for said information.

30. The product of claim 26 wherein said instructions further comprise

instructions for directing said processing unit to:

35 determine whether said input query is for information about a range of time units for a particular time period,

retrieve term count information including said count for each of said plurality terms for each time unit in said time units of said particular time period in said range from said term index for said particular time period,

retrieve entity count information including said count for each of said plurality entities for each time unit in said time units of said particular time period in said range from said entity index of said entity index of said particular time period, and

provide said term count information and said entity count information in said information provided to generate said display.

31. The product of claim 30 wherein said instructions to retrieve said term count information comprise:

instructions for directing said processing unit to determine said term index for said particular time period.

32. The product of claim 30 wherein said instructions to retrieve said term count information comprise:

instructions for directing said processing unit to:

limit said term count information to said count for each of a pre-determined number of said plurality of terms.

33. The product of claim 30 wherein said instructions to retrieve said entity count information comprise:

instructions for directing said processing unit to determine said entity index for said particular time period.

34. The product of claim 30 wherein said instructions to retrieve said term count information comprise:

instructions for directing said processing unit to:

limit said entity count information to said count for each of a pre-determined number of said plurality of entities.

35. The product of claim 26 wherein said instructions further comprise: instructions for directing said processing unit to:

determine whether said input query includes a request for terms associated with a particular one of said plurality of entities,

retrieve a term list that includes each of said plurality of terms associated with one of said plurality of entities from said search matrix, and

provide said term list in said information for generating said display.

36. The product of claim 35 wherein said instructions to retrieve said term list comprise:

5 instructions for directing said processing unit to:

read entry of said one of said plurality of entities from said entity document map,

retrieve an entry from each of said plurality of documents in said entry of said one of said plurality of entities from said document term map, and

10 add terms from said entry of each of said plurality of documents in said entry of said one of said plurality of entities from said entity document map to said term list.

37. The product of claim 35 wherein said instructions to retrieve said term list comprise:

15 instructions for directing said processing unit to:

limit said term list to a predetermined number of said plurality of terms.

38. The product of claim 26 wherein said instructions further comprise:
instructions for directing said processing unit to:

20 determine whether said input query includes a request for entities associated with a particular one of said plurality of terms,

retrieve an entity list that includes each of said plurality of entities associated with one of said plurality of terms from said search matrix, and

provide said entity list in said information for generating said display.

25

39. The product of claim 38 wherein said instructions to retrieve said entity list comprise:

instructions for directing said processing unit to:

30 read an entry of said one of said plurality of terms from said term document map,

determine each of entry of said entity-document map that includes a document from said entry read from said term-document map, and

35 add each entity associated with each entry of said entity-document map determined to include a document from said entry read from said term document map.

40. The product of claim 38 wherein said instructions to retrieve said entity list comprise:

instructions for directing said processing unit to:

limit said entity list to a predetermined number of said plurality of entities.

5

41. A system for providing information about data in a plurality of documents generated over a time period to generate a display comprising:

circuitry configured to maintain a search matrix for information about said data in said plurality of documents that includes term index and an entity index where said term index stores a count for each of a plurality of terms and each of a plurality of time units in said time period wherein each count represents a number of documents of said plurality of documents generated during a particular time unit that include said term and said entity index stores a count for each of a plurality of entities and each of said plurality of time units of said time period wherein each count represents a number of documents of said plurality of documents generated during a particular time unit that include said entity further wherein a term is a word or a phrase and an entity is a name of at least one selected from a group consisting of a person, a place, a business, and a thing,

10

15

circuitry configured to retrieve said count for each of a portion of said plurality of terms and each of a portion of said plurality of time units from said term index and said count for each of a portion of said plurality of entities and each of said portion of said plurality of time units to generate count information; and

20

circuitry configured to provide said count information to provide a display to show a comparison of occurrences of said portion of plurality of terms and said portion of said plurality of entities over said portion of said plurality of time units.

25

42. The system of claim 41 wherein said circuitry configured to maintain said search matrix comprises:

circuitry configured to add information about a one of said plurality of documents to said search matrix.

30

43. The system of claim 41 wherein said search matrix further includes a document index that stores information about each of said plurality of documents and said circuitry configured to add about said one of said plurality of documents comprises:

circuitry configured to receive said one of said plurality of documents from a source;

circuitry configured to generate a record for said one of said plurality of documents in a document index;

35

circuitry configured to pre-process data in said one of said plurality of documents;
circuitry configured to identify each term in said data of said one of said plurality of documents;

5 circuitry configured to identify each entity in said data of said one of said plurality of documents; and
circuitry configured to update said search matrix.

44. The system of claim 43 wherein said circuitry configured to generate said record in said document index comprises:

10 circuitry configured to generate said document record in said document index for said one of said plurality of documents;

circuitry configured to assign an identifier to said one of said plurality of documents;

circuitry configured to store said document in said record for said one of said plurality of documents in said identifier; and

15 circuitry configured to store a pointer to a location storing said one of said plurality of documents in said record.

45. The system of claim 43 wherein circuitry configured to pre-process said data of said one of said plurality of documents comprises:

20 circuitry configured to convert all text in said data to lower case letters;

circuitry configured to parse said text; and

circuitry configured to remove each stop word from said parsed text.

46. The system of claim 45 wherein said circuitry configured to identify each term in said data comprises:

circuitry configured to apply term extractor to strings to identify each term in said data.

47. The system of claim 46 wherein said circuitry configured to identify each entity in said data comprises:

30 circuitry configured to compare each said term identified by said term extractor to entries in a semantic database to identify each entity in said terms.

48. The system of claim 43 wherein said search matrix further includes an entity-document map which stores a record for each of said plurality of entities indicating each of said plurality of documents that include said entity, a term-document map which stores a record for each of said plurality of terms indicating each of said plurality of documents that include said term, a document-term map which includes a record for each of said plurality of

documents indicating each of said plurality of terms included in each of said plurality of documents, and a time unit-document map which includes a record for each of said plurality of time units indicating each of said plurality of documents generated during each of said plurality of time units and said circuitry configured to update said search matrix comprises:

- 5 circuitry configured to update said time unit-document map;
- circuitry configured to update said term index;
- circuitry configured to update said entity index;
- circuitry configured to update said entity-document map;
- circuitry configured to update said term-document map;
- 10 circuitry configured to update said document-entity map; and
- circuitry configured to update said document-term map.

49. The system of claim 48 wherein said circuitry configured to update said time unit-document map comprises:

- 15 circuitry configured to update a field for a last document included in a time unit with said identifier of said one of said plurality of documents; and
- circuitry configured to increment a counter indicating a number of said plurality of documents generate during said time unit.

50. The system of claim 49 wherein said circuitry configured to update said time unit-document map further comprises:

- 20 circuitry configured to search for said record of said time unit in said time-unit document map,
- circuitry configured to create an entry for said time unit in said time unit-document
- 25 map responsive to a determination that said record is not included in said time unit-document map; and
- circuitry configured to insert said identifier of said one of said plurality of documents into a field for a beginning document.

51. The system of 48 wherein said circuitry configured to update said term index comprises:

- 30 circuitry configured to determine a number of occurrences of each identified term in said one of said plurality of documents;
- circuitry configured to determine whether said number of occurrences of each
- 35 identified term is greater than or equal to a threshold; and
- circuitry configured to update a frequency count of an entry of each identified term in said term index having said number of occurrences greater than or equal to said threshold.

52. The system of claim 51 wherein said circuitry configured to update said term index comprises:

circuitry configured to determine whether said term index includes an entry for a particular term in response to said particular term having a number of occurrences greater than said threshold; and

circuitry configured to add an entry for said particular term to said term index in response to a determination that said term index does not include an entry for said particular term.

53. The system of claim 51 wherein said circuitry configured to update said term index comprises:

circuitry configured to determine whether said term index includes an entry for a particular time unit of said one of said plurality of documents; and

circuitry configured to add an entry for said particular time unit to said term index in response to a determination that said term index does not include an entry for said particular time unit.

54. The system of 48 wherein said circuitry configured to update said entity index comprise:

circuitry configured to determine a number of occurrences of each identified entity in said one of said plurality of documents;

circuitry configured to determine whether said number of occurrences of each identified entity is greater than or equal to a threshold; and

circuitry configured to update a frequency count of an entry of each identified entity in said entity index having said number of occurrences greater than or equal to said threshold.

55. The system of claim 54 wherein said circuitry configured to update said entity index comprises:

circuitry configured to determine whether said entity index includes an entry for a particular entity in response to said particular entity having a number of occurrences greater than said threshold; and

circuitry configured to add an entry for said particular entity to said entity index in response to a determination that said entity index does not include an entry for said particular entity.

56. The system of claim 54 wherein said circuitry configured to update said entity index comprises:

circuitry configured to determine whether said entity index includes an entry for a particular time unit of said one of said plurality of documents; and

5 circuitry configured to add an entry for said particular time unit to said entity index in response to a determination that said entity index does not include an entry for said particular time unit.

57. The system of claim 48 wherein said circuitry configured to update said entity-
10 document comprise:

circuitry configured to insert said document identifier of said one of said plurality of documents into an entry of each entity identified in said one of said plurality of documents.

58. The system of claim 57 wherein said circuitry configured to update said entity-
15 document further comprise:

circuitry configured to determine whether said entity-document map has an entry for each of entity identified in said one of said plurality of documents; and

circuitry configured to create an entry for each said entity identified in said one of said plurality of documents that is determined to not have an entry in said entity-document map.

20

59. The system of claim 48 wherein said circuitry configured to update said term-document map comprises:

circuitry configured to insert said document identifier of said one of said plurality of documents into an entry of each term identified in said one of said plurality of documents.

25

60. The system of claim 59 wherein said circuitry configured to update said term-document map further comprises:

circuitry configured to determine whether said term-document map has an entry for each of term identified in said one of said plurality of documents; and

30 circuitry configured to create an entry for each said term identified in said one of said plurality of documents that is determined to not have an entry in said term-document map.

61. The system of claim 48 wherein said circuitry configured to update said document-term map comprises:

35 circuitry configured to generate an entry for said one of said plurality of documents that includes said document identifier and a list of each term identified in said one of said plurality of documents.

62. The system of claim 48 wherein said circuitry configured to update said document-entity map comprises:

5 circuitry configured to generate an entry for said one of said plurality of documents that includes said document identifier and a list of each entity identified in said one of said plurality of documents.

63. The system of claim 48 wherein said search matrix further includes a term index and a search index for a second period of time wherein said second period of time is
10 made up of time units of said period time and said system further comprises:

circuitry configured to update said term index for said second time period with information about said plurality of documents generated during said period of time; and

circuitry configured to update said entity index for said second time period with information about said plurality of documents generated during said period of time.

15

64. The system of claim 63 wherein said circuitry configured to update said term index of said second period of time comprises:

circuitry configured to determine said period of time has expired;

20 circuitry configured to generate an entry for said period of time in said term index for said second period of time;

circuitry configured to read entries for each of said plurality of terms in said term index for said period of time;

circuitry configured to sum said count from each of said plurality of time units for each of said plurality of terms in said term index of said time period;

25 circuitry configured to determine whether each of said plurality of terms has an entry in said term index for said second period of time;

circuitry configured to create an entry in said term index of said second time period for each of said plurality of terms from said term index of said time period determined not to have an entry in said term index of said second time period; and

30 circuitry configured to insert said sum of counts for each particular term of said plurality of terms into said entry for said particular term and said time period in said term index of said second period of time.

65. The system of claim 63 wherein said circuitry configured to update said entity
35 index of said second period of time comprises:

circuitry configured to determine said period of time has expired;

circuitry configured to generate an entry for said period of time in said entity index for said second period of time;

circuitry configured to read entries for each of said plurality of entities in said term index for said period of time;

5 circuitry configured to sum said counts from each of said plurality of time units for each of said plurality of entities in said entity index of said time period;

circuitry configured to determine whether each of said plurality of entities has an entry in said entity index for said second period of time;

10 circuitry configured to create an entry in said entity index of said second time period for each of said plurality of entities from said entity index of said time period determined not to have an entry in said entity index of said second time period; and

circuitry configured to insert said sum of said counts for each particular entity of said plurality of entities into said entry for said particular entity and said time period in said entity index of said second period of time.

15

66. The system of claim 48 further comprising:

circuitry configured to receive an input query;

circuitry configured to determine whether said query is for information about a specific entity, retrieve information about said specific entity from said search matrix; and

20 circuitry configured to provide said information to generate a display.

67. The system of claim 66 wherein said circuitry configured to retrieve said information comprises:

25 circuitry configured to read frequency information for said specific entity for each said time unit from said entity index,

circuitry configured to generate a list of frequencies for time units from said frequency information;

circuitry configured to read entries for each time unit in said list from said time-unit document map;

30 circuitry configured to read entry for said specific entity from said entity-document map;

circuitry configured to compare documents listed in said entry from said entity-document map to said entries for each time unit to determine each document within said time units and add each document determined to be in range to a list of documents; and

35 circuitry configured to return said list of frequencies and said list of documents for said information.

68. The system of claim 66 further comprising:

circuitry configured to determine whether said query is for information about a specific term;

circuitry configured to retrieve information about said specific term from said search matrix; and

circuitry configured to provide said information to generate a display.

69. The system of claim 68 wherein said circuitry configured to retrieve said information about said specific term comprises:

circuitry configured to read frequency information for said specific term for each said time unit from said term index;

circuitry configured to generate a list of frequencies for each said time unit from said frequency information;

circuitry configured to read entries for each time unit in said list from said time-unit document map;

circuitry configured to read entry for said specific term from said term-document map;

circuitry configured to compare documents listed in said entry from said term-document map to said entries for each time unit to determine each document within said time units and add each document determined to be in range to a list of documents; and

circuitry configured to return said list of frequencies and said list of documents for said information.

70. The system of claim 66 further comprising:

circuitry configured to determine whether said input query is for information about a range of time units for a particular time period;

circuitry configured to retrieve term count information including said count for each of said plurality terms for each time unit in said time units of said particular time period in said range from said term index for said particular time period;

circuitry configured to retrieve entity count information including said count for each of said plurality entities for each time unit in said time units of said particular time period in said range from said entity index of said entity index of said particular time period; and

circuitry configured to provide said term count information and said entity count information in said information provided to generate said display.

71. The system of claim 70 wherein said circuitry configured to retrieve said term count information comprises:

circuitry configured to determine said term index for said particular time period.

72. The system of claim 70 wherein said circuitry configured to retrieve said term count information comprises:

circuitry configured to limit said term count information to said count for each of a pre-determined number of said plurality of terms.

73. The system of claim 70 wherein said circuitry configured to retrieve said entity count information comprises:

circuitry configured to determine said entity index for said particular time period.

74. The system of claim 70 wherein said circuitry configured to retrieve said term count information comprises:

circuitry configured to limit said entity count information to said count for each of a pre-determined number of said plurality of entities.

75. The system of claim 66 further comprising:

circuitry configured to determine whether said input query includes a request for terms associated with a particular one of said plurality of entities;

circuitry configured to retrieve a term list that includes each of said plurality of terms associated with one of said plurality of entities from said search matrix; and

circuitry configured to provide said term list in said information for generating said display.

76. The system of claim 75 wherein said circuitry configured to retrieve said term list comprises:

circuitry configured to read entry of said one of said plurality of entities from said entity document map;

circuitry configured to retrieve an entry from each of said plurality of documents in said entry of said one of said plurality of entities from said document term map; and

circuitry configured to add terms from said entry of each of said plurality of documents in said entry of said one of said plurality of entities from said entity document map to said term list.

77. The system of claim 75 wherein said circuitry configured to retrieve said term list comprises:

circuitry configured to limit said term list to a predetermined number of said plurality of terms.

78. The system of claim 66 further comprising:

circuitry configured to determine whether said input query includes a request for entities associated with a particular one of said plurality of terms;

5 circuitry configured to retrieve an entity list that includes each of said plurality of entities associated with one of said plurality of terms from said search matrix; and

circuitry configured to provide said entity list in said information for generating said display.

10 79. The system of claim 78 wherein said circuitry configured to retrieve said entity list comprises:

circuitry configured to read an entry of said one of said plurality of terms from said term document map;

15 circuitry configured to determine each of entry of said entity-document map that includes a document from said entry read from said term-document map; and

circuitry configured to add each entity associated with each entry of said entity-document map determined to include a document from said entry read from said term document map.

20 80. The system of claim 78 wherein said circuitry configured to retrieve said entity list comprises:

circuitry configured to limit said entity list to a predetermined number of said plurality of entities.

25 81. A method for providing information about data in a plurality of documents generated over a time period to generate a display comprising:

maintaining a search matrix for information about said data in said plurality of documents that includes term index and an entity index where said term index stores a count for each of a plurality of terms and each of a plurality of time units in said time period
30 wherein each count represents a number of documents of said plurality of documents generated during a particular time unit that include said term and said entity index stores a count for each of a plurality of entities and each of said plurality of time units of said time period wherein each count represents a number of documents of said plurality of documents generated during a particular time unit that include said entity further wherein a term is a
35 word or a phrase and an entity is a name of at least one selected from a group consisting of a person, a place, a business, and a thing,

retrieving said count for each of a portion of said plurality of terms and each of a portion of said plurality of time units from said term index and said count for each of a portion of said plurality of entities and each of said portion of said plurality of time units to generate count information; and

- 5 providing said count information to provide a display to show a comparison of occurrences of said portion of plurality of terms and said portion of said plurality of entities over said portion of said plurality of time units.

82. The method of claim 81 wherein said step of maintaining said search matrix
10 comprises:
 adding information about a one of said plurality of documents to said search matrix.

83. The method of claim 81 wherein said search matrix further includes a document index that stores information about each of said plurality of documents and said
15 step of adding about said one of said plurality of documents comprises:
 receiving said one of said plurality of documents from a source;
 generating a record for said one of said plurality of documents in a document index;
 pre-processing data in said one of said plurality of documents;
 identifying each term in said data of said one of said plurality of documents;
20 identifying each entity in said data of said one of said plurality of documents; and
 updating said search matrix.

84. The method of claim 83 wherein said step of generating said record in said document index comprises:
25 generating said document record in said document index for said one of said plurality of documents;
 assigning an identifier to said one of said plurality of documents;
 storing said document in said record for said one of said plurality of documents in said identifier; and
30 storing a pointer to a location storing said one of said plurality of documents in said record.

85. The method of claim 83 wherein said step of pre-processing said data of said one of said plurality of documents comprises:
35 converting all text in said data to lower case letters;
 parsing said text; and
 removing each stop word from said parsed text.

86. The method of claim 85 wherein said step of identifying each term in said data comprises:

applying a term extractor to strings to identify each term in said data.

5

87. The method of claim 86 wherein said step of identifying each entity in said data comprises:

comparing each said term identified by said term extractor to entries in a semantic database to identify each entity in said terms.

10

88. The method of claim 83 wherein said search matrix further includes an entity-document map which stores a record for each of said plurality of entities indicating each of said plurality of documents that include said entity, a term-document map which stores a record for each of said plurality of terms indicating each of said plurality of documents that include said term, a document-term map which includes a record for each of said plurality of documents indicating each of said plurality of terms included in each of said plurality of documents, and a time unit-document map which includes a record for each of said plurality of time units indicating each of said plurality of documents generated during each of said plurality of time units and said step of updating said search matrix comprises:

20

updating said time unit-document map;

updating said term index;

updating said entity index;

updating said entity-document map;

updating said term-document map;

25

updating said document-entity map; and

updating said document-term map.

89. The method of claim 88 wherein said step of updating said time unit-document map comprises:

30

updating a field for a last document included in a time unit with said identifier of said one of said plurality of documents; and

incrementing a counter indicating a number of said plurality of documents generated during said time unit.

35

90. The method of claim 89 wherein said step of updating said time unit-document map further comprises:

searching for said record of said time unit in said time-unit document map,

creating an entry for said time unit in said time unit-document map responsive to a determination that said record is not included in said time unit-document map; and

inserting said identifier of said one of said plurality of documents into a field for a beginning document.

5

91. The method of 88 wherein said step of updating said term index comprises:

determining a number of occurrences of each identified term in said one of said plurality of documents;

10 determining whether said number of occurrences of each identified term is greater than or equal to a threshold; and

updating a frequency count of an entry of each identified term in said term index having said number of occurrences greater than or equal to said threshold.

15 92. The method of claim 91 wherein said step of updating said term index comprises:

determining whether said term index includes an entry for a particular term in response to said particular term having a number of occurrences greater than said threshold; and

20 adding an entry for said particular term to said term index in response to a determination that said term index does not include an entry for said particular term.

93. The method of claim 91 wherein said step of updating said term index comprises:

25 determining whether said term index includes an entry for a particular time unit of said one of said plurality of documents; and

adding an entry for said particular time unit to said term index in response to a determination that said term index does not include an entry for said particular time unit.

94. The method of 88 wherein said step of updating said entity index comprises:

30 determining a number of occurrences of each identified entity in said one of said plurality of documents;

determining whether said number of occurrences of each identified entity is greater than or equal to a threshold; and

35 updating a frequency count of an entry of each identified entity in said entity index having said number of occurrences greater than or equal to said threshold.

95. The method of claim 94 wherein said step of updating said entity index comprises:

determining whether said entity index includes an entry for a particular entity in response to said particular entity having a number of occurrences greater than said
5 threshold; and

adding an entry for said particular entity to said entity index in response to a determination that said entity index does not include an entry for said particular entity.

96. The method of claim 94 wherein said step of updating said entity index
10 comprises:

determining whether said entity index includes an entry for a particular time unit of said one of said plurality of documents; and

adding an entry for said particular time unit to said entity index in response to a determination that said entity index does not include an entry for said particular time unit.
15

97. The method of claim 88 wherein said step of updating said entity-document comprises:

inserting said document identifier of said one of said plurality of documents into an entry of each entity identified in said one of said plurality of documents.
20

98. The method of claim 97 wherein said step of updating said entity-document further comprises:

determining whether said entity-document map has an entry for each of entity identified in said one of said plurality of documents; and

25 creating an entry for each said entity identified in said one of said plurality of documents that is determined to not have an entry in said entity-document map.

99. The method of claim 88 wherein said step of updating said term-document map comprises:

30 inserting said document identifier of said one of said plurality of documents into an entry of each term identified in said one of said plurality of documents.

100. The method of claim 99 wherein said step of updating said term-document map further comprises:

35 determining whether said term-document map has an entry for each of term identified in said one of said plurality of documents; and

creating an entry for each said term identified in said one of said plurality of documents that is determined to not have an entry in said term-document map.

5 101. The method of claim 88 wherein said step of updating said document-term map comprises:

generating an entry for said one of said plurality of documents that includes said document identifier and a list of each term identified in said one of said plurality of documents.

10 102. The method of claim 88 wherein said step of updating said document-entity map comprises:

generating an entry for said one of said plurality of documents that includes said document identifier and a list of each entity identified in said one of said plurality of documents.

15

103. The method of claim 88 wherein said search matrix further includes a term index and a search index for a second period of time wherein said second period of time is made up of time units of said period time and said method further comprises:

20 updating said term index for said second time period with information about said plurality of documents generated during said period of time; and

updating said entity index for said second time period with information about said plurality of documents generated during said period of time.

25 104. The method of claim 103 wherein said step of updating said term index of said second period of time comprises:

determining said period of time has expired;

generating an entry for said period of time in said term index for said second period of time;

30 reading entries for each of said plurality of terms in said term index for said period of time;

summing said count from each of said plurality of time units for each of said plurality of terms in said term index of said time period;

determining whether each of said plurality of terms has an entry in said term index for said second period of time;

35 creating an entry in said term index of said second time period for each of said plurality of terms from said term index of said time period determined not to have an entry in said term index of said second time period; and

inserting said sum of counts for each particular term of said plurality of terms into said entry for said particular term and said time period in said term index of said second period of time.

5 105. The method of claim 103 wherein said step of update said entity index of said second period of time comprises:

 determining said period of time has expired;

 generating an entry for said period of time in said entity index for said second period of time;

10 reading entries for each of said plurality of entities in said term index for said period of time;

 summing said counts from each of said plurality of time units for each of said plurality of entities in said entity index of said time period;

 determining whether each of said plurality of entities has an entry in said entity index for said second period of time;

15 creating an entry in said entity index of said second time period for each of said plurality of entities from said entity index of said time period determined not to have an entry in said entity index of said second time period; and

 inserting said sum of said counts for each particular entity of said plurality of entities into said entry for said particular entity and said time period in said entity index of said second period of time.

 106. The method of claim 88 further comprising:

 receiving an input query;

25 determining whether said query is for information about a specific entity, retrieve information about said specific entity from said search matrix; and

 providing said information to generate a display.

30 107. The method of claim 106 wherein said step of retrieving said information comprises:

 reading frequency information for said specific entity for each said time unit from said entity index;

 generating a list of frequencies for time units from said frequency information;

 reading entries for each time unit in said list from said time-unit document map;

35 reading entry for said specific entity from said entity-document map;

comparing documents listed in said entry from said entity-document map to said entries for each time unit to determine each document within said time units and add each document determined to be in range to a list of documents; and
returning said list of frequencies and said list of documents for said information.

5

108. The method of claim 106 further comprising:
determining whether said query is for information about a specific term;
retrieving information about said specific term from said search matrix; and
providing said information to generate a display.

10

109. The method of claim 108 wherein said step of retrieving said information about said specific term comprises:

reading frequency information for said specific term for each said time unit from said term index;

15

generating a list of frequencies for each said time unit from said frequency information;

reading entries for each time unit in said list from said time-unit document map;

reading an entry for said specific term from said term-document map;

20

comparing documents listed in said entry from said term-document map to said entries for each time unit to determine each document within said time units and add each document determined to be in range to a list of documents; and

returning said list of frequencies and said list of documents for said information.

110. The method of claim 106 further comprising:

25

determining whether said input query is for information about a range of time units for a particular time period;

retrieving term count information including said count for each of said plurality terms for each time unit in said time units of said particular time period in said range from said term index for said particular time period;

30

retrieving entity count information including said count for each of said plurality entities for each time unit in said time units of said particular time period in said range from said entity index of said entity index of said particular time period; and

providing said term count information and said entity count information in said information provided to generate said display.

35

111. The method of claim 110 wherein said step of retrieving said term count information comprises:

determining said term index for said particular time period.

112. The method of claim 110 wherein said step of retrieving said term count information comprises:

5 limiting said term count information to said count for each of a pre-determined number of said plurality of terms.

113. The method of claim 110 wherein said step of retrieving said entity count information comprises:

10 determining said entity index for said particular time period.

114. The method of claim 110 wherein said step of retrieve said term count information comprises:

15 limiting said entity count information to said count for each of a pre-determined number of said plurality of entities.

115. The method of claim 106 further comprising:

determining whether said input query includes a request for terms associated with a particular one of said plurality of entities;

20 retrieving a term list that includes each of said plurality of terms associated with one of said plurality of entities from said search matrix; and

providing said term list in said information for generating said display.

116. The method of claim 115 wherein said step of retrieving said term list comprises:

25 reading an entry of said one of said plurality of entities from said entity document map;

retrieving an entry from each of said plurality of documents in said entry of said one of said plurality of entities from said document term map; and

30 adding terms from said entry of each of said plurality of documents in said entry of said one of said plurality of entities from said entity document map to said term list.

117. The method of claim 115 wherein said step of retrieving said term list comprises:

35 limiting said term list to a predetermined number of said plurality of terms.

118. The method of claim 106 further comprising:

determining whether said input query includes a request for entities associated with a particular one of said plurality of terms;

retrieving an entity list that includes each of said plurality of entities associated with one of said plurality of terms from said search matrix; and

5 providing said entity list in said information for generating said display.

119. The method of claim 118 wherein said step of retrieving said entity list comprises:

reading an entry of said one of said plurality of terms from said term document map;

10 determining each of entry of said entity-document map that includes a document from said entry read from said term-document map; and

adding each entity associated with each entry of said entity-document map determined to include a document from said entry read from said term document map.

15 120. The method of claim 118 wherein said step of retrieving said entity list comprises:

limiting said entity list to a predetermined number of said plurality of entities.