United States Patent **6,484,168**

Pennock , et al. **November 19, 2002**

System for information discovery

Abstract

A sequence of word filters are used to eliminate terms in the database which do not discriminate document content, resulting in a filtered word set and a topic word set whose members are highly predictive of content. These two word sets are then formed into a two dimensional matrix with matrix entries calculated as the conditional probability that a document will contain a word in a row given that it contains the word in a column. The matrix representation allows the resultant vectors to be utilized to interpret document contents.

Inventors: **Pennock; Kelly A.** (Richland, WA)**, Miller; Nancy E.** (Kennewick, WA)

Assignee: **Battelle Memorial Institute** (Richland, WA)

Appl. No.: **09/455,849**

Filed: **December 7, 1999**

Related U.S. Patent Documents

Application Number	Filing Date	Patent Number	Issue Date<TD
191004	Nov., 1998		<TD
713313	Sep., 1996		<TD

Current U.S. Class: **1/1** ; 707/999.006; 707/E17.07

Current International Class: G06F 17/30 (20060101); G06F 017/30 ()

Field of Search: 340/172.5 707/3,100,2,6,7,102 377/39 395/701,800

References Cited [Referenced By]

U.S. Patent Documents

3609703	September 1971	Peacock
5333313	July 1994	Heising
5369577	November 1994	Kadashevich et al.
5381454	January 1995	Nusinov et al.
5499382	March 1996	Nusinov et al.
5576954	November 1996	Driscoll
5659730	August 1997	Kelley et al.
5689716	November 1997	Chen
5950189	September 1999	Cohen et al.

Other References

Kirac et al., "Results on Lattice Vector Quantization with Dithering", IEEE, 1996, pp. 811-825..

Primary Examiner: Rones; Charles L.
Assistant Examiner: Pardo; Thuy
Attorney, Agent or Firm: Milde & Hoffberg, LLP

Government Interests

SYSTEM FOR INFORMATION DISCOVERY

This invention was made with Government support under Contract DE-AC06-76RLO 1830 awarded by the U.S. Department of Energy. The Government has certain rights in the invention.

Parent Case Text

REFERENCE TO RELATED APPLICATIONS

This application is a continuation of application Ser. No. 09/191,004 filed on Nov. 12, 1998, a continuation of application Ser. No. 08,713,313 filed on Sep. 13, 1996, both abandoned, hereby incorporated by reference in their entirety.

Claims

We claim:

1. A method for analyzing and characterizing a database of electronically formatted natural language documents comprising the steps of: a) subjecting the database to a sequence of word filters to eliminate terms in the database which do not discriminate document content, resulting in a filtered word set whose members are highly predictive of content; b) defining a subset of the filtered word set as the topic set, said topic set being characterized as the set of filtered words which best discriminate the content of the documents which contain them; c) forming a two dimensional matrix with the words contained within the topic set defining one dimension of said matrix and the words contained within the filtered word set comprising the other dimension of said matrix; d) calculating matrix entries as the conditional probability that a document in the database will contain each word in the topic set given that it contains each word in the filtered word set; and e) providing said matrix entries as vectors to interpret the document contents of said database, wherein one of said sequence of filters comprises a frequency filter, a topicality filter and an overlap filter.

2. The method of claim 1 wherein one of said topicality filter comprises the steps of: a) calculating the expected distribution of each word contained in said database, b) measuring the actual distribution of each word contained in said database, c) expressing the ratio of said actual distribution to said expected distribution, and d) defining said set of topic words as those which fall below a predetermined value of said ratio.

3. The method of claim 1 wherein one of said frequency filter comprises the steps of: a) defining a predetermined upper and lower limit for the frequency of said words in the database, b) determining the frequency of occurrence of each word contained in said database, c) further defining said set of topic words as those words whose frequency of occurrence in the database are above said predetermined lower limit and below said predetermined upper limit.

4. The method of claim 1 wherein one of said overlap filter comprises the steps of: a) defining a preset limit for joint distribution of word pairs occurring within said database, b) calculating the joint distribution of word pairs occurring within said database, c) defining the set of word pairs whose joint distribution falls above said preset limit c) further defining said set of topic words as not containing one of those words for each set of word pairs whose joint distribution falls above said preset upper limit.

5. The method of claim 1, wherein said defining step further comprises the substep of ranking the words containing the filtered word set according to predetermined criteria.

6. The method of claim 5, wherein the predetermined criteria of said ranking substep comprises weighting words according to their inherent indicia of content.

7. The method of claim 1, wherein said defining step further comprises the substep of ranking the words containing the filtered word set according to their frequency in the database.

8. The method of claim 1, wherein said defining further comprises the substep of ranking the words containing the filtered word set according to their rank in the topicality filter.

9. The method of claim 1, wherein said defining step further comprises the substep of ranking the words containing the filtered word set, and truncating words having a relatively lower rank.

10. A computer readable software medium containing computer software for controlling a general purpose computer to perform the method of claim 1.

11. A method for analyzing and characterizing an informational file system of documents, comprising the steps of: (a) compressing vocabulary of the informational file system, said compressing step resulting in a filtered word set indicative of topicality; (b) creating a reduced subset of the filtered word set to produce a topic set; (c) creating a two-dimensional matrix wherein the topic set represents rows of the matrix and the filtered word set represents columns of the matrix; and (d) evaluating the conditional probability of each member of the topic set being present in a document after being provided with an indication of presence of each member of the filtered word set in the matrix.

12. The method of claim 11, further comprising the step of: (e) ranking the filtered word set according to a predetermined criteria.

13. The method of claim 11, further comprising the step of: (e) manipulating the matrix to characterize documents within the informational file system according to the context of the documents.

14. The method of claim 13, wherein said manipulating step comprises, for each document in the informational file system, the steps of: (i) summing at least one vector of each word in the document located in the topic set, said summing step resulting in a unique vector for the document which measures the relationships between the document and other documents of the informational file system across all parameters expressed in the topic set; and (ii) comparing said at least one vector to at least one other vector to determine a relationship between content of the document and content of said other documents.

15. The method of claim 13, wherein said manipulating step further comprises the step of: utilizing a vector space model for content characterization.

16. The method of claim 15, wherein said utilizing step comprises the steps of: (i) measuring a degree of match between documents and a query; and (ii) executing a dot product to rank the documents from top to bottom for natural language-based queries.

17. The method of claim 13, wherein said manipulating step further comprises the step of: constructing at least one document vector for each document in the topic set.

18. The method of claim 17, wherein said constructing step comprises, for each document in the topic set, the steps of: (i) determining words of interest in the document; (ii) extracting at least one vector for each word of interest in the document from a conditional probability matrix; (iii) summing the at least one vector for each word of interest; and (iv) normalizing said summing step to produce a summation of all component magnitude of one.

19. The method of claim 11, wherein said compressing step comprises the step of: filtering the informational file system.

20. The method of claim 19, wherein said filtering step comprises the steps of: (i) measuring an absolute number of occurrences of each word in the informational file system and eliminating words which are outside of a predetermined upper and 3lower frequency range to produce a first set; (ii) comparing actual placement of each word within the informational file system with expected placement of each word within the informational file system; (iii) discarding words from the first set having a certain ratio that exceeds a predefined limit to produce a second set; and (iv) comparing words remaining from the second set to determine words whose placement in the informational file system are correlated with each other.

21. The method of claim 20, wherein said discarding step comprises the steps of: (1) assigning a ratio to words in the informational file system according to randomness of the words in the informational file system; and (2) discarding each word whose said ratio exceeds a certain predefined limit.

22. The method of claim 20, wherein said discarding step comprises the steps of: (1) expressing a ratio between a value representing actual placement of a given word within the informational file system and a value representing expected placement of the given word within the informational file system; and (2) discarding each word whose said ratio exceeds a certain predefined limit.

23. The method of claim 20, wherein said measuring step comprises the steps of: (1) defining a predetermined upper and lower limit for frequency of occurrence of each word in the informational file system; (2) determining frequency of occurrence of each word in the informational file system; and (3) further defining words in the topic set as those words whose frequency of occurrence in the informational file system are above the predetermined lower limit and below the predetermined upper limit.

24. The method of claim 11, wherein said comparing step comprises the steps of: (1) calculating an expected distribution of each word in the informational file system; (2) measuring actual distribution of each word in the informational file system; (3) expressing a ratio of the actual distribution to the expected distribution; and (4) defining the topic set as a set with words which are below a predetermined value of the ratio.

25. The method of claim 20, wherein said discarding step comprises the steps of: (1) defining a preset limit for joint distribution of word pairs occurring within the informational file system; (2) calculating the joint distribution of the word pairs occurring within the informational file system; (3) refining the set of word pairs to include only the word pairs whose joint distribution falls above the preset limit; and (4) further refining the topic set as a set which can contain only a word for each word pair in the refined word pairs.

26. A computer readable software medium containing computer software for controlling a general purpose computer to perform the method of claim 11.

Description

FIELD OF THE INVENTION

The present invention relates generally to a method for automatically analyzing and characterizing the content of electronically formatted natural language based documents.

BACKGROUND OF THE INVENTION

As the global economy has become increasingly driven by the skillful synthesis of information across all disciplines, be they scientific, economic, or otherwise, the sheer volume of information available for use in such a synthesis has rapidly expanded. This has resulted in an ever increasing value for systems or methods which are able to analyze information and separate information relevant to a particular problem or useful in a particular inquiry from information that is not relevant or useful. The vast majority of information available for such synthesis, 95% according to estimates by the National Institute or Science and Technology (NIST), is in the form of written natural language. The traditional method of analyzing and characterizing information in the form of written natural language is to simply read it. However, this approach is increasingly unsatisfactory as the sheer volume of information outpaces the time available for manual review. Thus, several methodologies for automating the analysis and characterization of such information have arisen. Typical for such schemes is the requirement that the information is presented, or converted, to an electronic form or database, thereby allowing the database to be manipulated by a computer system according to a variety of algorithms designed to analyze and/or characterize the information available in the database. For example, vector based systems using first order statistics have been developed which attempt to define relationships between documents based upon simple characteristics of the documents, such as word counts.

The simplest of these methodologies is a simple search wherein a word or a word form is entered into the computer as a query and the computer compares the query to words contained in the documents in the database to determine if matches exist. If there are matches, the computer then returns a list of those documents within the database which contain a word or word form which matches the query. This simple search methodology may be expanded by the addition of other Boolean operators into the query. For example, the computer may be asked to search for documents which contain both a first query and a second query, or a second query within a predetermined number of words from the first query, or for documents containing a query, which consist of a series of terms, of for documents which contain a particular query but not another query. Whatever the particular parameters, the computer searches the database for documents which fit the required parameters, and those documents are then returned to the user.

Among the drawbacks of such schemes is the possibility that in a large database, even a very specific query may match a number of documents that is too large to be effectively reviewed by the user. Additionally, given any particular query, there exists the possibility that documents which would be relevant to the user may be overlooked because the documents do not contain the specific query term identified by the user; in other words, these systems often ignore word to

word relationships, and thus require exacting queries to insure meaningful search results. Because these systems tend to require exacting queries, these methods suffer from the drawback that the user must have some concept of the contents of the documents in order to draft a query which will generate the desired results. This presents the users of such systems with a fundamental paradox: In order to become familiar with a database, the user must ask the right questions or enter relevant queries; however, to ask the right questions or enter relevant queries, the user must already be familiar with the database.

To overcome these and other drawbacks, a number of methods have arisen which are intended to compare the contents of documents in an electronic database and thereby determine relationships between the documents. In this manner, documents that address similar subject matter but do not share common key words may be linked, and queries to the database are able to generate resulting relevant documents without requiring exacting specificity in the query parameters. For example, vector based systems using higher order statistics may be characterized by the generation of vectors which can be used to compare documents. By measuring conditional probabilities between and among words contained within the database, different terms may be linked together. However, these systems suffer from the drawback that they are unable to discern words which provide insight into the meaning of the documents which contain them. Other systems have sought to overcome this limitation by utilizing neural networks or other methods to capture the higher order statistics required to compress the vector space. These systems suffer from considerable computational lag due to the large amount of information that they are processing. Thus, there exists a need for an automated system which will analyze and characterize a database of electronically formatted natural language based documents in a manner wherein the system output correlates documents within the database according to the meaning of the documents and required system resources are minimized.

SUMMARY OF THE INVENTION

Accordingly, the present invention is a System for Information Discovery (SID). The overriding intent of the present invention is to provide a system for analyzing and characterizing a database of electronically formatted natural language based documents wherein the output is information concerning the content and structure of the underlying database in a form that correlates the meaning of the individual documents within the database. A sequence of word filters are used to eliminate terms in the database which do not discriminate document content, resulting in a filtered word set whose members are highly predictive of content. The filtered word set is then further reduced to determine a subset of topic words which are characterized as the set of filtered words which best discriminate the content of the documents which contain them. These two word sets, the filtered word set and the topic set, are then formed into a two dimensional matrix. Matrix entries are then calculated as the conditional probability that a document will contain a word in a row given that it contains the word in the column of the matrix. The number of word correlations which is computed is thus significantly reduced because each word in the filtered set is only related to the topic words, with the topic word set being smaller than the filtered word set. The matrix representation thus captures the context of the filtered words and allows the resultant vectors to be utilized to interpret document contents with a wide variety of querying schemes. Surprisingly, while computational efficiency gains are realized by utilizing the reduced topic word set (as compared with creating a matrix with only the filtered word set forming both the

columns and the rows), the ability of the resultant vectors to predict content is comparable or superior to approaches which consider word sets which have not been reduced either in the number of terms considered or by the number of correlations between terms.

The first step of the system is to compress the vocabulary of the database through a series of filters. Three filters are employed, the frequency filter, the topicality filter and the overlap filter. The frequency filter first measures the absolute number of occurrences of each of the words in the database and eliminates those which fall outside of a predetermined upper and lower frequency range.

The topicality filter then compares the placement of each word within the database with the expected placement assuming the word was randomly distributed throughout the database. By expressing the ratio between a value representing the actual placement of a given word (A) and a value representing the expected placement of the word assuming random placement (E), a cutoff value may be established wherein words whose ratio A/E is above a certain predefined limit are discarded. In this manner, words which do not rise to a certain level of nonrandomness, and thus do not represent topics, are discarded.

The overlap filter then uses second order statistics to compare the remaining words to determine words whose placement in the database are highly correlated with one and another. Measures of joint distribution are calculated for word pairs remaining in the database using standard second order statistical methodologies, and for word pairs which exhibit correlation coefficients above a preset value, one of the words of the word pair is then discarded as its content is assumed to be captured by its remaining word pair member.

At the conclusion of these three filtering steps, the number of words in the database is typically reduced to approximately ten percent of the original number. In addition, the filters have discriminated and removed words which are not highly related to the topicality of the documents which contain them, or words which are redundant to words which reveal the topicality of the documents which contain them. The remaining words, which are thus highly indicative of topicality and non-redundant, are then ranked according to some predetermined criteria designed to weight them according to their inherent indicia of content. For example, they may be ranked in descending order of their frequency in the database, or according to ascending order according to their rank in the topicality filter.

The filtered words thus ranked are then cut off at either a predetermined limit or a limit generated by some parameter relevant to the database or its characteristics to create a reduced subset of the total population of filtered words. This subset is referred to as a topic set, and may be utilized as both an index and/or as a table of contents. Because the words contained in the topic set have been carefully screened to include those words which are the most representative of the contents of the documents contained within the database, the topic set allows the end user the ability to quickly surmise both the primary contents and the primary characteristics of the database.

This topic set is then utilized as rows and the filtered words are utilized as columns in a matrix wherein each of the elements of the columns and the rows are evaluated according to their

conditional probability as word pairs. The resultant matrix evaluates the conditional probability of each member of the topic set being present in a document, or a predetermined segment of the database which can represent a document, given the presence of each member of the filtered word set. The resultant matrix can be manipulated to characterize documents within the database according to their context. For example, by summing the vectors of each word in a document also present in the topic set, a unique vector for each document which measures the relationships between the document and the remainder of the database across all the parameters expressed in the topic set may be generated. By comparing vectors so generated for any set of documents contained within the data set, the documents be compared for the similarity of the resultant vectors to determine the relationship between the contents of the documents. In this manner, all of the documents contained within the database may be compared to one and another based upon their content as measured across a wide spectrum of indicating words so that documents describing similar topics are correlated by their resultant vectors.

OBJECTS

The subject matter of the present invention is particularly pointed out and distinctly claimed in the concluding portion of this specification. However, both the organization and method of operation, together with further advantages and objects thereof, may best be understood by reference to the following description.

DESCRIPTION OF THE PREFERRED EMBODIMENT(S)

In a preferred embodiment of the present invention, the System for Information Discovery (SID) is provided as a series of instructions provided as a code for a programmable computer. The particular programming language selected is not critical, any language which will accomplish the required instructions necessary to practice the method is suitable. As an example, a prototype program of SID was written in C++. Similarly, the particular computer platform selected for running the code which performs the series of instructions is not critical. Any computer platform with sufficient system resources such as memory to run the resultant program is suitable. The prototype program of SID described above has been run on a variety of systems including a Silicon Graphics Indy workstation, a Sun Sparc 2 system, a Sun Sparc 10 system, and a Sun Sparc 20 system. The database that is to be interrogated is a series of documents written in some natural language. While the prototype system has primarily been tested with documents written in English, the methodology of SID will work for any language, particularly those which utilize a Roman alphabet, which may be represented in a written form. The documents must be converted into an electronic form to be loaded into the database. As will be apparent to those skilled in the art, this may be accomplished by a variety of methods.

In the first step of the preferred embodiment, the individual words contained within the database are examined to create a filtered word set comprising about 15% of the words contained in the database. The filtered word set is then further reduced to produce a topic set comprising about 15% of the words in the filtered word set. The filtered word set is produced by sending the database through a series of three filters, the frequency filter, the topicality filter and the overlap filter.

The frequency filter first measures the absolute number of occurrences of each of the words in the database and eliminates those which fall outside of a predetermined upper and lower frequency range. In the preferred embodiment, the upper range was set to exclude words which occur in more than 20% of the documents, and the lower range was set to exclude words which occur in fewer than 1% of the documents, for documents sets of one hundred or more. Those words that are outside of these ranges remain in the database, but are discarded from the filtered word set.

The remaining words are then sent through the topicality filter which compares the placement of each word within the database with the expected placement assuming the word was randomly distributed throughout the database. Algorithmically, a rapid method to identify important topics at the proper "sampling rate" is essential. The approach followed in the preferred embodiment for topic filtering is based on the serial clustering work described in Detecting Content-Bearing Words by Serial Clustering, Bookstein, A., Klein, S. T., Raita, T. (1995) Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval: 319-327. and incorporated herein by reference. The preferred method greatly simplifies the serial clustering as described in Bookstein et al by approximating the size of the text unit with the average size of the document, and then assuming uniform distribution of each word within the document so that word counts in documents which are larger than average are scaled down and word counts in documents which are smaller than average are scaled up. For example, the count for a particular word for a document which contains m * the average number of total words, and a count n of a particular word, is scaled by a factor of 1/m. This approximation avoids the computationally expensive text unit divisions identified in Bookstein et al.

The concept can be understood by considering the placement of points within a grid of cells. Given m points randomly distributed on n cells, some cells can be expected to contain zero points, others one, etc. Numerically, condensation clustering is the ratio between the actual number of occurrences of a term within a text unit (document or subdocument unit) of arbitrary size, to the expected number of occurrences, and is given by:

with

and with U=# documents in the corpus, and T=# occurrences of token t.sub.a in the database.

Thus, topic words, or words that represent subject, are characterized by their condensation clustering value. In the preferred embodiment, words having a condensation clustering value of less than 1.0 are selected for inclusion in the filtered word set.

In the preferred embodiment, the remaining words are then sent through the overlap filter. The overlap filter uses second order statistics to compare the remaining words to determine words whose placement in the database are highly correlated with one and another. Many measures of joint distribution are known to those skilled in the art, and each is suitable for determining values which are then used by the overlap filter. In the preferred embodiment, conditional probabilities are utilized to represent the relationship between words, so that the relationship between term a and term b is given by:

P(t.sub.j /t.sub.i)=the conditional probability of t.sub.i given tj

Word pairs which are closely correlated may have one of their members discarded as only the remaining member is necessary to signify the content of the word pair. Thus, for word pairs having a correlation above a preset value, 0.4 for the preferred embodiment, the overlap filter will discard the lower topicality word of the pair, as its content is assumed to be captured by its remaining word pair member.

After the overlap filter has eliminated redundant words, the final set of filtered words is complete. The preferred embodiment then ranks these filtered words in descending order according to their frequency in the database. The words with a topicality value below a predefined threshold, for the preferred embodiment 0.8, are then selected to comprise the topic set. The words that comprise the topic set and the filtered word set may be displayed to the user as they are extremely useful in and of themselves for communicating the general content of the dataset. In short, at this juncture, the preferred embodiment has provided a listing of key terms which are readily interpreted by humans and which are highly representative of the underlying topicality of the dataset. This topic set is then utilized as rows and the filtered words are utilized as columns in a matrix wherein each of the elements of the columns and the rows are evaluated according to their conditional probability as word pairs. The matrix is described as two sets of words: the topic words (n) and the filtered words, or (m). An n by m matrix is then computed, with the entries in the matrix being the conditional probabilities of occupance, modified by the independent probability of occurrence, or

with M.sub.ij =the ith row, jth column of the conditional probability matrix P(t.sub.j /t.sub.i)=the conditional probability of t.sub.i given tj P(t.sub.j)=the probability of tj Beta=parameter constant to ensure strong correlations

The end product is a list of associations among discriminating words in the data set using only second order statistics. An advantage derived from the exclusive use of second order statistics and smaller word sets in the preferred embodiment is greatly reduced computational time, memory, and the capacity for incremental updates.

The preferred embodiment then utilizes a vector space model for content characterization for two primary reasons. First, to measure the degree of match between documents and a query, the vector space model can be very efficient. For natural language-based queries or extended Boolean queries, the vector-space model allows documents to be ranked from top to bottom using a simple dot product. Queries, in this model, are simply vectors in the vector space, the same as any unit of text (from a single word to a document or even multiple documents). The second primary reason for the vector space model is that it provides a spatial representation for information. The representation conveys significant structural information which is important to many operations such as grouping or clustering or projecting.

A number of functions such as advanced query operations and visualization depend on meaningfully distinguishing documents in vector space due to content. The preferred embodiment capitalizes on the tendency that the greater the differences in content, the more

distinct the vector characterization of the documents. In the preferred embodiment, topics words serve as dimensions in the vector-space model Given that the major topics of a data set have been defined using the described filtering techniques, and that the essential vocabulary of the data set have been probabilistically linked to the topics, the general goal of document content characterization is to map the specific document contents to varying values for each of the topics in the canonical set. Just as functions can be defined by combinations of sinusoids, so SID documents are defined by combinations of topic values. The contents of each document are then judged strongly related to those topics for which relatively large values are calculated. An important advantage of this method is that topics of interest can easily be enhanced or diminished through linear transforms on the topic magnitudes across the document set. This permits users to define spatial relationships among documents based on their interests, instead of a single predefined representation. The limitless combinations of topics and values thus allows a rich method of content characterization in the preferred embodiment.

The construction of the document vectors in the preferred embodiment is a relatively straight-forward process. For each document in the data set, four steps are followed to produce the vector: 1) words of interest are determined (the topic words contained in the document) 2) a vector for each word of interest in the document is extracted from the modified conditional probability matrix (e.g. if the first word of interest is entry n in the conditional probability matrix, the corresponding vector is the nth column of the matrix, with each row of that vector the modified conditional probability associating the word of interest with each topic) 3) the vectors for each word of interest are summed, and 4) the final vector summation is normalized so that the summation of all component magnitudes is one.

In the preferred embodiment, efficient indexing or hashing makes the vector construction process rapid. Interim representations of documents that were originally created when the documents were initially parsed to provide an index of terms, contain words and their associated counts in the document, minus stop words and after stemming. The interim structure is much smaller and therefore more efficient to search for topics. The vector construction is also extremely flexible in the scale of its application. Short and long documents, document components such as chapters, or paragraphs, or even single words, or user queries ranging from strict Boolean queries to natural language queries, each have vectors constructed in the same manner. Document discrimination is thus a derived property of the vector-space representation. Representations that proximally locate documents of similar theme and separate documents which are significantly dissimilar in theme are capable of appropriate discrimination based on the resultant vectors generated.

While a preferred embodiment of the present invention has been shown and described, it will be apparent to those skilled in the art that many changes and modifications may be made without departing from the invention in its broader aspects. The appended claims are therefore intended to cover all such changes and modifications as fall within the true spirit and scope of the invention.

* * * * *

Images

View Cart Add to Cart

Top

Home Quick Advanced Pat Num Help