



- (51) **International Patent Classification:**
G06F 17/00 (2006.01) *G06F 17/30* (2006.01)
- (21) **International Application Number:**
PCT/US2014/047803
- (22) **International Filing Date:**
23 July 2014 (23.07.2014)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P. [US/US]; 11445 Compaq Center Drive W, Houston, Texas 77070 (US).
- (72) **Inventors:** KAFAL, Mehran; 1501 Page Mill Road, Palo Alto 94304-1100 (GB). YEO, Wen; 1501 Page Mill Road, Palo CA, California 94304-1100 (US).
- (74) **Agent:** DAS, Manav; 3404 E Harmony Road, Fort Collins, Colorado 80528-9599 (US).
- (81) **Designated States** (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT,

HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

- (84) **Designated States** (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- as to the identity of the inventor (Rule 4.17(i))
- as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))

Published:

- with international search report (Art. 21(3))

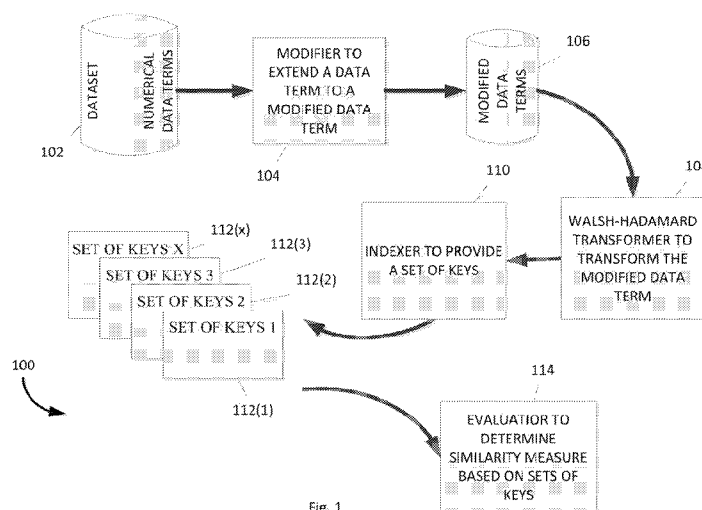
(54) **Title:** PROXIMITY OF DATA TERMS BASED ON WALSH-HADAMARD TRANSFORMS

Fig. 1

(57) **Abstract:** Determining proximity of data terms based on Walsh-Hadamard transforms is disclosed. One example is a system including a modifier, a Walsh-Hadamard transformer, an indexer, and an evaluator. A dataset, including a plurality of numerical data terms, is received via a processing system. The modifier extends a given data term of the plurality of data terms, the extension based on multiple concatenations of the given data term with itself. The Walsh-Hadamard transformer provides coefficients of the Walsh-Hadamard transform of the modified given data term. The indexer provides a set of keys based on the coefficients, and associates the set of keys with the given data term. The evaluator determines a similarity measure for a pair of data terms of the plurality of data terms, the similarity measure based on a number of overlaps between respective sets of keys, and indicative of proximity of the pair of data terms.



PROXIMITY OF DATA TERMS BASED ON WALSH-HADAMARD TRANSFORMS

Background

[0001] A dataset is a collection of data terms. Datasets are analyzed to determine proximity of the data terms. Such proximity may be utilized in finding a data term that is proximate to a received query term.

Brief Description of the Drawings

[0002] Figure 1 is a functional block diagram illustrating one example of a system for determining proximity of data terms based on Walsh-Hadamard transforms.

[0003] Figure 2 is a block diagram illustrating one example of a processing system for implementing the system for determining proximity of data terms based on Walsh-Hadamard transforms.

[0004] Figure 3 is a block diagram illustrating one example of a computer readable medium for determining proximity of data terms based on Walsh-Hadamard transforms.

[0005] Figure 4 is a flow diagram illustrating one example of a method for determining proximity of data terms based on Walsh-Hadamard transforms.

Detailed Description

[0006] A dataset is a collection of data terms. Datasets are analyzed to detect proximity of the data terms. Such proximity may be utilized for an approximate nearest neighbor ("ANN") search.

[0007] As described in various examples herein, proximity of data terms is determined based on Walsh-Hadamard transforms ("WHTs"). Such an approach may be utilized to perform probabilistic dimensionality reduction for the purpose of ANN search and indexing. ANN search via WHT indexing may

be utilized for computer systems that perform in-memory big-data analysis and retrieval. A WHT is an orthogonal, non-sinusoidal transform that takes a signal as input and outputs a set of basis functions. The output functions are known as Walsh functions. A Walsh function takes two values: +1 and -1. Performing a WHT on an input signal provides a set of coefficients associated with the input signal.

[0008] As described herein, the ANN search based on a WHT takes a data term (e.g. a numerical N -dimensional vector) in a dataset and maps it to a set of H keys. Each key is an integer from the set $\{1, 2, \dots, U\}$, where U is generally much larger than N , and N is much larger than H . Generally, U may be a power of 2. For example, we may have $H = 100$, $N = 6000$, and $U = 2^{16}$. The H keys may be based on largest H coefficients provided by the WHT. So we obtain a projection of an N -dimensional object onto a lower H -dimensional object. A similarity measure between two data terms in the dataset may be determined based on a number of common keys. This provides an approximate measure of nearest neighbors in the dataset. An ANN search may then be performed. For example, a received query term may be mapped to a set of keys, and this set of keys may be utilized to search for a nearest neighbor in the dataset based on the similarity measure.

[0009] As described in various examples herein, determining proximity of data terms based on Walsh-Hadamard transforms is disclosed. One example is a system including a modifier, a Walsh-Hadamard transformer, an indexer, and an evaluator. A dataset is received via a processing system, the dataset including a plurality of numerical data terms. A numerical data term is data that may be represented numerically. In one example, a numerical data may be a vector with numerical components. As another example, a numerical data term may be a matrix with numerical entries. In one example, a data term may be represented numerically. For example, the term "True" may be represented by the number "1" and the term "False" may be represented by the number "0". The modifier extends a given data term of the plurality of data terms, the extension based on multiple concatenations of the given data term with itself. The Walsh-Hadamard transformer applies a Walsh-Hadamard transform to the

modified given data term to provide coefficients of the Walsh-Hadamard transform. The indexer provides a set of keys based on the coefficients of the Walsh-Hadamard transform, and associates the set of keys with the given data term. The evaluator determines, via the processing system, a similarity measure for a pair of data terms of the plurality of data terms, the similarity measure based on a number of overlaps between respective sets of keys, and indicative of proximity of the pair of data terms.

[0010] In the following detailed description, reference is made to the accompanying drawings which form a part hereof, and in which is shown by way of illustration specific examples in which the disclosure may be practiced. It is to be understood that other examples may be utilized, and structural or logical changes may be made without departing from the scope of the present disclosure. The following detailed description, therefore, is not to be taken in a limiting sense, and the scope of the present disclosure is defined by the appended claims. It is to be understood that features of the various examples described herein may be combined, in part or whole, with each other, unless specifically noted otherwise.

[0011] Figure 1 is a functional block diagram illustrating one example of a system 100 for determining proximity of data terms based on Walsh-Hadamard transforms. The system 100 receives a dataset, including a plurality of numerical data terms. The system 100 extends a given data term of the plurality of data terms, the extension based on multiple concatenations of the given data term with itself. In one example, the system 100 extends each data term of the plurality of data terms into an extended data term, the extension based on concatenating each data term with itself d times. The system 100 applies a Walsh-Hadamard transform to the modified given data term to provide coefficients of the Walsh-Hadamard transform. The system 100 determines a similarity measure for a pair of data terms of the plurality of data terms, the similarity measure being based on a number of overlaps between respective associated sets of keys, and being indicative of proximity of the pair of data terms. In one example, the given data term is a vector with N components, and the modified given data term is a modified vector with U components, and the

indexer associates the set of U integers with the given data term, each given integer of the set of U integers associated with the given data term if the given integer appears in the set of keys associated with the given data term.

Accordingly, the plurality of data terms in the dataset may be indexed based on the Walsh-Hadamard transforms.

[0012] Such indexing represents each data term with multiple keys to increase the overall probability of overlaps. The multiple keys may correspond to selected WHT coefficient indices (e.g. the largest H indices), thereby representing a high-dimensional data term in low dimensional space. Applying the WHT is computationally more efficient than other comparable transforms. The indexing disclosed herein may be applicable to a data set with numerical data terms.

[0013] System 100 includes a dataset 102 with a plurality of numerical data terms, a modifier 104, a collection of modified data terms 106, a Walsh-Hadamard transformer 108, an indexer 110, sets of keys 112(1), 112(2), ..., 112(x), each set of keys associated with a data term, and an evaluator 114. In one example, the dataset 102 may include a plurality of vectors with numerical, real-valued components. System 100 may be provided with values for H and U . The integer H may be experimentally determined based on the type and number of data terms in the dataset. Generally, U is a very large integer relative to H . In one example, U is a power of 2. The elements of system 100 may be implemented, for example, in software.

[0014] Modifier 104 extends a given data term of the plurality of data terms, the extension being based on multiple concatenations of the given data term with itself. In one example, the extension is based on concatenating each data term with itself d times. In one example, a vector with N numerical components may be extended by concatenating it with itself d times, where d may be selected as a floor(U/N). In one example, the extension includes adding zeros so that the modified vector has U components. For example, if $d = \text{floor}(U/N)$, then the number of additional zeros may be $U \bmod N$. The floor of a real number is the largest integer that is smaller than the real number. For example, the floor of

2.999 is 2, the floor of 10.001 is 10, and so forth. In one example, N may be 6000, and U may be 2^{18} . Accordingly, $d = \text{floor}(2^{18}/6000)$.

[0015] As another illustrative example, N may be 10, and U may be 2^5 . Accordingly, $d = \text{floor}(2^5/10) = \text{floor}(32/10) = \text{floor}(3.2) = 3$, and $U \bmod N = 32 \bmod 10 = 2$. A vector $A = \langle a_1, a_2, \dots, a_{10} \rangle$ may be concatenated $d = 3$ times with itself to obtain a vector: $A' = \langle a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10} \rangle$ of length 30. Two additional zeros may be added to the vector A' to obtain a modified vector: $\langle a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, 0, 0 \rangle$ of length $U = 32$.

[0016] In one example, the modifier may randomly permute components of the extended data term. For example, components of the extended vector may be permuted. In one example, the integers $\{1, 2, \dots, U\}$ may be permuted, and the corresponding permutation may be applied to the modified vector with U components. For example, when $U = 32$, the integers $\{1, 2, \dots, 32\}$ may be permuted to obtain the set $\{32, 1, 2, \dots, 31\}$. Accordingly, the modified vector $\langle a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, 0, 0 \rangle$ may also be permuted to obtain the vector: $\langle 0, a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, 0 \rangle$. In general, an extension followed by a random permutation increases a likelihood of finding similarities between two data terms.

[0017] Dataset 102 is modified via modifier 104 to provide modified data terms 106. System 100 includes a Walsh-Hadamard transformer 108 to apply a Walsh-Hadamard transform to the modified given data term to provide coefficients of the Walsh-Hadamard transform. For example, after application of the Walsh-Hadamard transform to the modified vector $\langle 0, a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, a_1, a_2, \dots, a_{10}, 0 \rangle$, the Walsh-Hadamard transformer may provide a collection of coefficients $c_1, c_2, \dots, c_k$.

[0018] System 100 includes an indexer 110 to provide a set of keys based on coefficients of the Walsh-Hadamard transform, and to associate the set of keys with the given data term. In one example, the highest H coefficients of the Walsh-Hadamard transform of the modified data term may be selected as the set of keys. In general, H may be much smaller than U . In one example, H may be 100, N may be 6000, and U may be 2^{18} .

[0019] Indexer 110 provides sets of keys 112(1), 112(2), ..., 112(x), each set corresponding to a data term, may be provided by the Walsh-Hadamard transformer 108. In one example, each 6000-dimensional vector may be associated with 100 integers selected from the set $\{1, 2, 3, \dots, 2^{18}\}$.

Accordingly, a higher dimensional data object (e.g. with 6000 dimensions) is associated with a lower dimensional index (e.g. with 100 dimensions).

[0020] In one example, the set of keys comprises coefficients of the Walsh-Hadamard transform of the modified data term. As described herein, the Walsh-Hadamard transform for a given modified data term may provide a collection of coefficients $c_1, c_2, \dots, c_k$. The H largest coefficients, $c_{n_1}, c_{n_2}, \dots, c_{n_H}$ may be selected as the set of keys associated with the data term A.

[0021] Table 1 illustrates an example association of data terms A, B, and C, with sets of keys:

Data Term	Set of Keys
A	{1, 5, 9, 13, 16}
B	{2, 3, 4, 7, 13}
C	{1, 5, 7, 8, 11}

Table 1

[0022] In one example, the given data term may be a vector with N components, and the modified given data term may be a modified vector with U components, and the indexer 110 may associate the set of U integers with the plurality of data terms, each given integer of the set of U integers being associated with the given data term if the given integer appears in the set of keys associated with the given data term. Table 2 illustrates an example association of $U = 2^4$ integers $\{1, 2, \dots, 16\}$ to data terms A, B, and C, based on sets of $H = 5$ keys in Table 1:

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A				A				A				A			A
	B	B	B			B						B			
C				C		C	C			C					

Table 2

[0023] As illustrated, integers 1 and 5 are associated with A and C since these integers appear in the set of keys associated with A (see Table 1) and the set of keys associated with C (see Table 1). Likewise, integer 13 is associated with A and B since this integer appears in the set of keys associated with A (see Table 1) and the set of keys associated with B (see Table 1).

[0024] System 100 includes an evaluator 114 to determine a similarity measure for a pair of data terms of the plurality of data terms, the similarity measure based on a number of overlaps between respective sets of keys, and indicative of proximity of the pair of data terms. Table 3 illustrates an example determination of similarity measures for pairs formed from the data terms A, B, and C:

Data Term Pair: (X,Y)	Similarity Measure: S(X,Y)
(A,B)	$S(A,B) = 1$
(A,C)	$S(A,C) = 2$
(B,C)	$S(B,C) = 1$

Table 3

[0032] As illustrated in Table 2, the data terms A and B have index 13 in common in their respective sets of keys. Accordingly, the similarity measure for the pair (A,B), denoted as $S(A,B)$ may be determined to be 1. Also, for example, as illustrated in Table 2, the data terms A and C have indices 1 and 5 in common in their respective sets of keys. Accordingly, the similarity measure for the pair (A,C), denoted as $S(A,C)$ may be determined to be 2. As another example, as illustrated in Table 2, the data terms B and C have index 7 in common in their respective sets of keys. Accordingly, the similarity measure for the pair (B,C), denoted as $S(B,C)$ may be determined to be 1.

[0033] In one example, system 100 may further include a receiver (not illustrated in Figure 1) to receive a query term. In one example, the query term may be a vector with numerical components. The modifier 104 may extend the query term, and the Walsh-Hadamard transformer 108 may apply a Walsh-Hadamard transform to the modified query term to provide coefficients for the modified query term. As described herein, the indexer 110 may associate the query term with a set of keys, the set of keys based on the coefficients for the modified

query term. Table 4 illustrates an example query term Q associated with a set of keys:

Query Term	Set of Keys
Q	{1, 5, 6, 7, 9, 10, 13}

Table 4

[0034] In one example, system 100 may include a classifier (not illustrated in Figure 1) to generate a list of data terms of the plurality of data terms, the list generated based on the set of keys associated with the modified query term. In one example, the classifier may rank the list of data terms based on a similarity measure of the query term with each data term in the list of data terms. Table 5 illustrates an example list of terms associated with the query term Q illustrated in Table 4, and the corresponding similarity measures.

1	5	6	7	9	10	13	Similarity Measures
A	A			A		A	$S(Q,A) = 4$
			B			B	$S(Q,B) = 2$
C	C		C				$S(Q,C) = 3$

Table 5

[0035] As illustrated, the set of keys associated with the query term Q (see Table 4) may be compared with the indexed data terms illustrated in Table 2. Based on such comparison, index 1 appears in the set of keys associated with Q and index 1 is also associated with data terms A and C. Also, for example, index 5 appears in the set of keys associated with Q and index 5 is also associated with data terms A and C. As another example, index 13 appears in the set of keys associated with Q and index 13 is also associated with data terms A and B. The frequency of occurrence of A is 4. This is also the similarity measure for the pair Q and A. The frequency of occurrence of B is 2. This is also the similarity measure for the pair Q and B. The frequency of occurrence of C is 3. This is also the similarity measure for the pair Q and C.

[0036] In one example, the classifier may rank the list of data terms based on a similarity measure of the query term with each data term in the list of data terms. Based on the example illustrated in Table 5, the classifier may rank the list of data terms as A, C, and B.

[0037] In one example, the classifier provides, in response to the query term, at least one data term from the list of data terms based on the ranking. In the example illustrated in Table 5, based on the ranking, the at least one data term may be selected as A, and the classifier may provide A in response to the query term Q. Accordingly, A may be determined as a nearest neighbor for the query term Q.

[0038] In one example, the ranking may not provide an unambiguous candidate for the at least one data term. In such instances, in one example, more than one data term may be provided in response to the query term. Also, for example, if data terms D and E are determined to have the same ranking, then additional measures of similarity may be utilized to determine if D or E may be provided in response to the query term. For example, cosine similarities may be determined for the pairs (D,Q) and (E,Q), and D or E may be selected based on the respective cosine similarities.

[0039] Figure 2 is a block diagram illustrating one example of a processing system 200 for implementing the system 100 for determining proximity of data terms based on Walsh-Hadamard transforms. Processing system 200 includes a processor 202, a memory 204, input devices 216, and output devices 218. Processor 202, memory 204, input devices 216, and output devices 218 are coupled to each other through communication link (e.g., a bus).

[0040] Processor 202 includes a Central Processing Unit (CPU) or another suitable processor. In one example, memory 204 stores machine readable instructions executed by processor 202 for operating processing system 200. Memory 204 includes any suitable combination of volatile and/or non-volatile memory, such as combinations of Random Access Memory (RAM), Read-Only Memory (ROM), flash memory, and/or other suitable memory.

[0041] Memory 204 stores dataset 206, including a plurality of data terms, for processing by processing system 200. Memory 204 also stores instructions to be executed by processor 202 including instructions for a modifier 208, a Walsh-Hadamard transformer 210, an indexer 212, and an evaluator 214. In one example, modifier 208, Walsh-Hadamard transformer 210, indexer 212, and evaluator 214, include modifier 104, Walsh-Hadamard transformer 108, indexer

110, and evaluator 114, respectively, as previously described and illustrated with reference to Figure 1.

[0042] In one example, processor 202 executes instructions of modifier 208 to modify dataset 206 to extend a given data term of the plurality of data terms, the extension being based on multiple concatenations of the given data term with itself. In one example, processor 202 executes instructions of modifier 208 to extend a vector with N numerical components by concatenating it with itself d times, where d may be selected as the floor(U/N). In one example, processor 202 executes instructions of modifier 208 to randomly permute components of the extended given data term.

[0025] Processor 202 executes instructions of Walsh-Hadamard transformer 210 to apply a Walsh-Hadamard transform to the modified given data term to provide coefficients of the Walsh-Hadamard transform.

[0026] Processor 202 executes instructions of an indexer 212 to provide a set of keys based on coefficients of the Walsh-Hadamard transform, and to associate the set of keys with the given data term. In one example, highest H coefficients of the Walsh-Hadamard transform of the modified given data term may be selected as the set of keys. In one example, the given data term may be a vector with N components, and the modified given data term may be a modified vector with U components, and processor 202 executes instructions of an indexer 212 to associate the set of U integers with the plurality of data terms, each given integer of the set of U integers being associated with the given data term if the given integer appears in the set of keys associated with the given data term.

[0027] Processor 202 executes instructions of an evaluator 214 to determine a similarity measure for a pair of data terms of the plurality of data terms, the similarity measure based on a number of overlaps between respective sets of keys, and indicative of proximity of the pair of data terms.

[0028] In one example, processor 202 executes instructions of a receiver (not illustrated in Figure 2) to receive a query term. In one example, processor 202 executes instructions of modifier 208 to extend the query term. In one example, processor 202 executes instructions of Walsh-Hadamard transformer 210 to

apply a Walsh-Hadamard transform to the modified query term to provide coefficients for the modified query term. In one example, processor 202 executes instructions of indexer 212 to provide a set of keys based on coefficients of the Walsh-Hadamard transform, and to associate the set of keys with the query term. In one example, processor 202 executes instructions of a classifier (not illustrated in Figure 2) to generate a list of data terms of the plurality of data terms, the list generated based on the set of keys associated with the modified query term. In one example, processor 202 executes instructions of a classifier to rank the list of data terms based on a similarity measure of the query term with each data term in the list of data terms. In one example, processor 202 executes instructions of a classifier to provide, in response to the query term, at least one data term from the list of data terms based on the ranking.

[0029] Input devices 216 include a keyboard, mouse, data ports, and/or other suitable devices for inputting information into processing system 200. In one example, input devices 216 are used to input a query term. Output devices 218 include a monitor, speakers, data ports, and/or other suitable devices for outputting information from processing system 200. In one example, output devices 218 are used to provide responses to the query term. For example, output devices 218 may provide the at least one data term.

[0030] Figure 3 is a block diagram illustrating one example of a computer readable medium for determining proximity of data terms based on Walsh-Hadamard transforms. Processing system 300 includes a processor 302, a computer readable medium 306, and a Walsh-Hadamard transformer 304. Processor 302, computer readable medium 306, and the Walsh-Hadamard transformer 304 are coupled to each other through communication link (e.g., a bus).

[0031] Processor 302 executes instructions included in the computer readable medium 306. Computer readable medium 306 includes dataset receipt instructions 308 to receive a dataset. The dataset receipt instructions 308 include instructions to receive a plurality of plurality of vectors with numerical components. Computer readable medium 306 includes modification instructions

310 of a modifier to modify a given vector of the plurality of vectors into a modified given vector. The modification instructions 310 comprising further extend instructions 312 to extend the given vector by concatenating it with itself multiple times. The modification instructions 310 comprising further permute instructions 314 to randomly permute the components of the extended given vector.

[0032] Computer readable medium 306 includes Walsh-Hadamard transform instructions 316 of the Walsh-Hadamard transformer 304 to apply a Walsh-Hadamard transform to the modified given vector to provide coefficients of the Walsh-Hadamard transform. Computer readable medium 306 includes indexing instructions of an indexer 318 to associate a set of keys with the given vector, the set of keys based on the coefficients of the Walsh-Hadamard transform. In one example, highest H coefficients of the Walsh-Hadamard transform of the modified given vector may be selected as the set of keys. In one example, the given vector may have N components, and the modified given vector may have U components, and computer readable medium 306 includes indexing instructions of an indexer 318 to associate the set of U integers with the given vector, each given integer of the set of U integers being associated with the given vector if the given integer appears in the set of keys associated with the given vector. Computer readable medium 306 includes similarity measure determination instructions 320 of an evaluator to determine a similarity measure for a pair of vectors of the plurality of vectors, the similarity measure based on a number of overlaps between respective sets of keys, and indicative of proximity of the pair of vectors.

[0033] In one example, computer readable medium 306 includes instructions to receive a query vector, associate the query vector with a set of keys, and provide at least one vector of the plurality of vectors based on the set of keys associated with the query vector.

[0034] Figure 4 is a flow diagram illustrating one example of a method for determining proximity of data terms based on Walsh-Hadamard transforms. At 400, a query term is received. At 402, the query term is modified by concatenating the query term with itself multiple times. At 404, a Walsh-

Hadamard transform is applied to the modified query term to provide coefficients of the Walsh-Hadamard transform. At 406, the query term is associated with a set of keys, the set of keys based on the coefficients of the Walsh-Hadamard transform. At 408, at least one data term is retrieved from a plurality of data terms, the at least one data term being retrieved based on the set of keys associated with the query term. At 410, the at least one data term is provided in response to the query term.

[0035] In one example, modifying the query term may include randomly permuting the components of the concatenated query term.

[0036] In one example, the associated set of keys may include indices of the Walsh-Hadamard transform of the modified query term.

[0037] In one example, the query term is a vector with N components, and the modified query term is a modified vector with U components, and the indexer associates the set of U integers with the vector, each given integer of the set of U integers associated with the vector if the given integer appears in the set of keys associated with the vector.

[0038] In one example, the database may include an association of the set of U integers with the plurality of data terms, each given integer of the set of U integers being associated with a given data term if the given integer appears in the set of keys associated with the given data term.

[0039] Examples of the disclosure provide a generalized system for determining proximity of data terms based on Walsh-Hadamard transforms. The generalized system provides an automatable approach to perform probabilistic dimensionality reduction for the purpose of ANN search and indexing. ANN search via WHT indexing may be utilized for computer systems that perform in-memory big-data analysis and retrieval.

[0040] Although specific examples have been illustrated and described herein, especially as related to healthcare data, the examples illustrate applications to any structured data. Accordingly, there may be a variety of alternate and/or equivalent implementations that may be substituted for the specific examples shown and described without departing from the scope of the present disclosure. This application is intended to cover any adaptations or variations of

the specific examples discussed herein. Therefore, it is intended that this disclosure be limited only by the claims and the equivalents thereof.

CLAIMS

1. A system comprising:
 - a dataset received via a processing system, the dataset including a plurality of numerical data terms;
 - a modifier to extend a given data term of the plurality of data terms, the extension based on multiple concatenations of the given data term with itself;
 - a Walsh-Hadamard transformer to apply a Walsh-Hadamard transform to the modified given data term to provide coefficients of the Walsh-Hadamard transform;
 - an indexer to provide a set of keys based on the coefficients of the Walsh-Hadamard transform, and to associate the set of keys with the given data term; and
 - an evaluator to determine, via the processing system, a similarity measure for a pair of data terms of the plurality of data terms, the similarity measure based on a number of overlaps between respective sets of keys, and indicative of proximity of the pair of data terms.
2. The system of claim 1, wherein the modifier randomly permutes components of the extended data term.
3. The system of claim 1, wherein the given data term is a vector with N components, and the modified given data term is a modified vector with U components, wherein U is considerably larger than N , and the indexer associates the set of U integers with the given data term, each given integer of the set of U integers associated with the given data term if the given integer appears in the set of keys associated with the given data term.

4. The system of claim 3, wherein the set of keys comprises H largest coefficients of the Walsh-Hadamard transform, wherein H is considerably smaller than N .
5. The system of claim 1, further comprising a receiver to receive a query term, and wherein:
 - the modifier extends the query term;
 - the Walsh-Hadamard transformer applies the Walsh-Hadamard transform to the modified query term to provide coefficients for the modified query term; and
 - the indexer associates the query term with a set of keys, the set of keys based on the coefficients for the modified query term.
6. The system of claim 5, further comprising:
 - a classifier to generate a list of data terms of the plurality of data terms, the list generated based on the set of keys associated with the modified query term.
7. The system of claim 6, wherein the classifier ranks the list of data terms based on a similarity measure of the query term with each data term in the list of data terms.
8. The system of claim 7, wherein the classifier provides, in response to the query term, at least one data term from the list of data terms based on the ranking.
9. A method to find an approximate nearest neighbor in a database, the method comprising:
 - receiving, via a processor, a query term;
 - modifying the query term by concatenating the query term with itself multiple times;
 - applying a Walsh-Hadamard transform to the modified query term to provide coefficients of the Walsh-Hadamard transform;

associating the query term with a set of keys, the set of keys based on the coefficients of the Walsh-Hadamard transform;
retrieving, from the database, at least one data term from a plurality of data terms, the at least one data term retrieved based on the set of keys associated with the query term; and
providing, in response to the query term, the at least one data term.

10. The method of claim 9, wherein modifying the query term further comprises randomly permuting the components of the concatenated query term.
11. The method of claim 9, wherein the query term is a vector with N components, and the modified query term is a modified vector with U components, and the indexer associates the set of U integers with the vector, each given integer of the set of U integers associated with the vector if the given integer appears in the set of keys associated with the vector.
12. The method of claim 9, wherein the database comprises an association of the set of U integers with the plurality of data terms, each given integer of the set of U integers associated with a given data term if the given integer appears in the set of keys associated with the given data term.
13. A non-transitory computer readable medium comprising executable instructions to:
 - receive a dataset via a processor, the dataset including a plurality of vectors with numerical components;
 - modify a given vector of the plurality of vectors into a modified given vector, the instructions to modify comprising further instructions to:
 - extend the given vector by concatenating it with itself multiple times, and

randomly permute the components of the extended given vector;

apply a Walsh-Hadamard transform to the modified given vector to provide coefficients of the Walsh-Hadamard transform;

associate a set of keys with the given vector, the set of keys based on the coefficients of the Walsh-Hadamard transform; and

determine, via the processor, a similarity measure for a pair of vectors of the plurality of vectors, the similarity measure based on a number of overlaps between respective sets of keys, and indicative of proximity of the pair of vectors.

14. The non-transitory computer readable medium of claim 13, wherein the given vector has N components, and the modified given vector has U components, wherein U is considerably larger than N , and further including instructions to:

associate the set of U integers with the given vector, each given integer of the set of U integers associated with the given vector if the given integer appears in the set of keys associated with the given vector.

15. The non-transitory computer readable medium of claim 13, further including instructions to:

receive a query vector;

associate the query vector with a set of keys; and

provide at least one vector of the plurality of vectors based on the set of keys associated with the query vector.

1/4

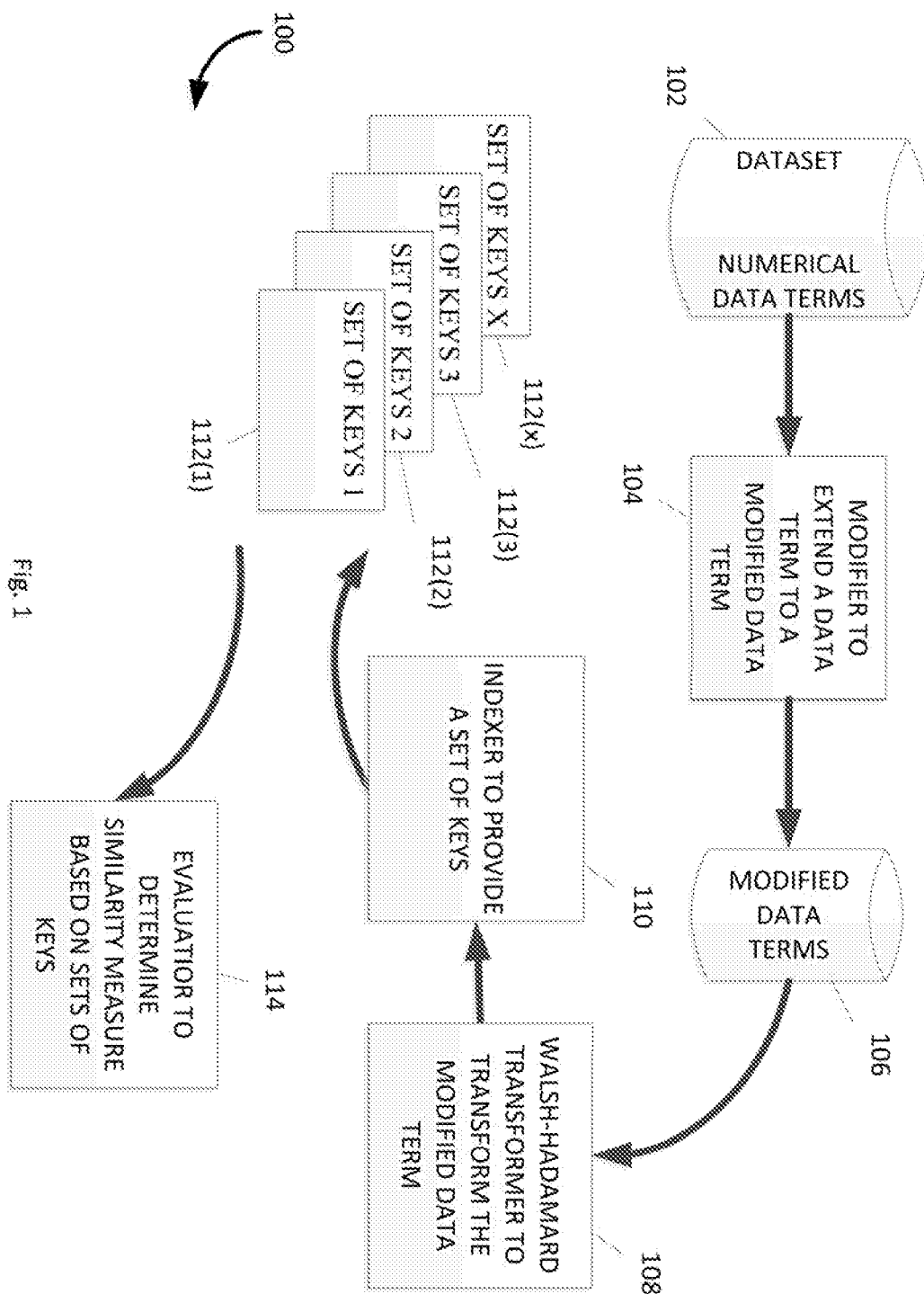


Fig. 1

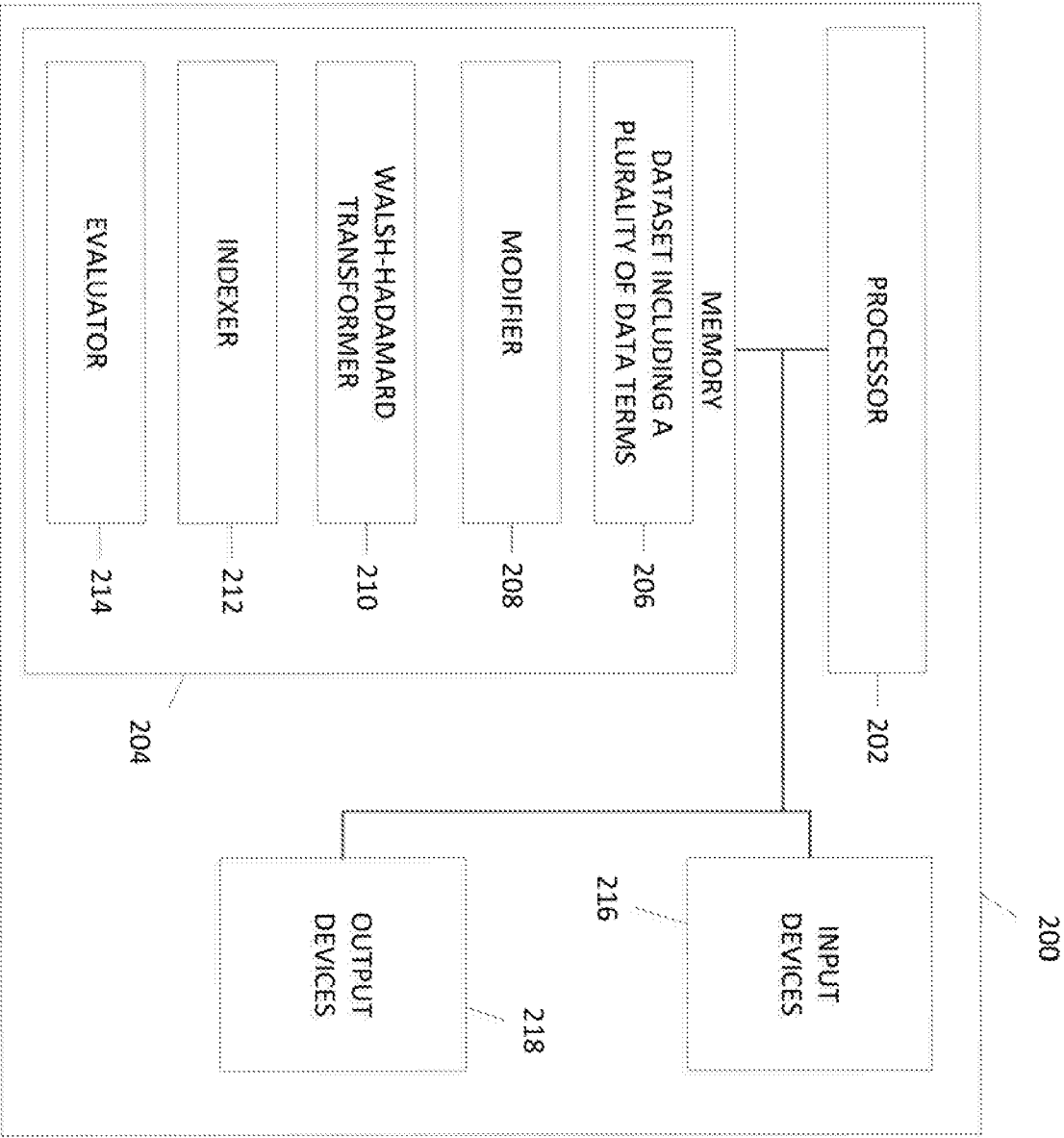


Fig. 2

3/4

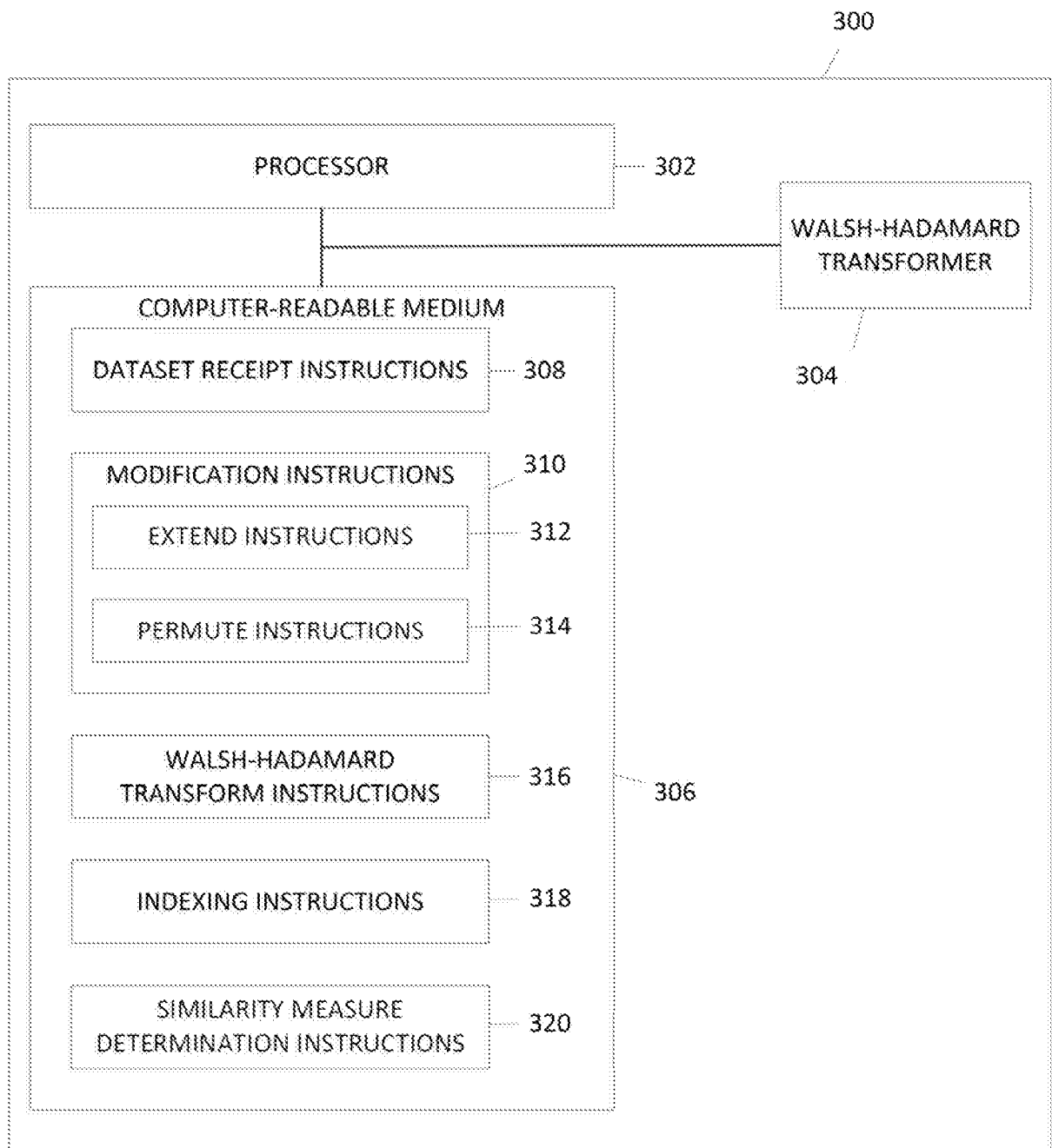


Fig. 3

4 / 4

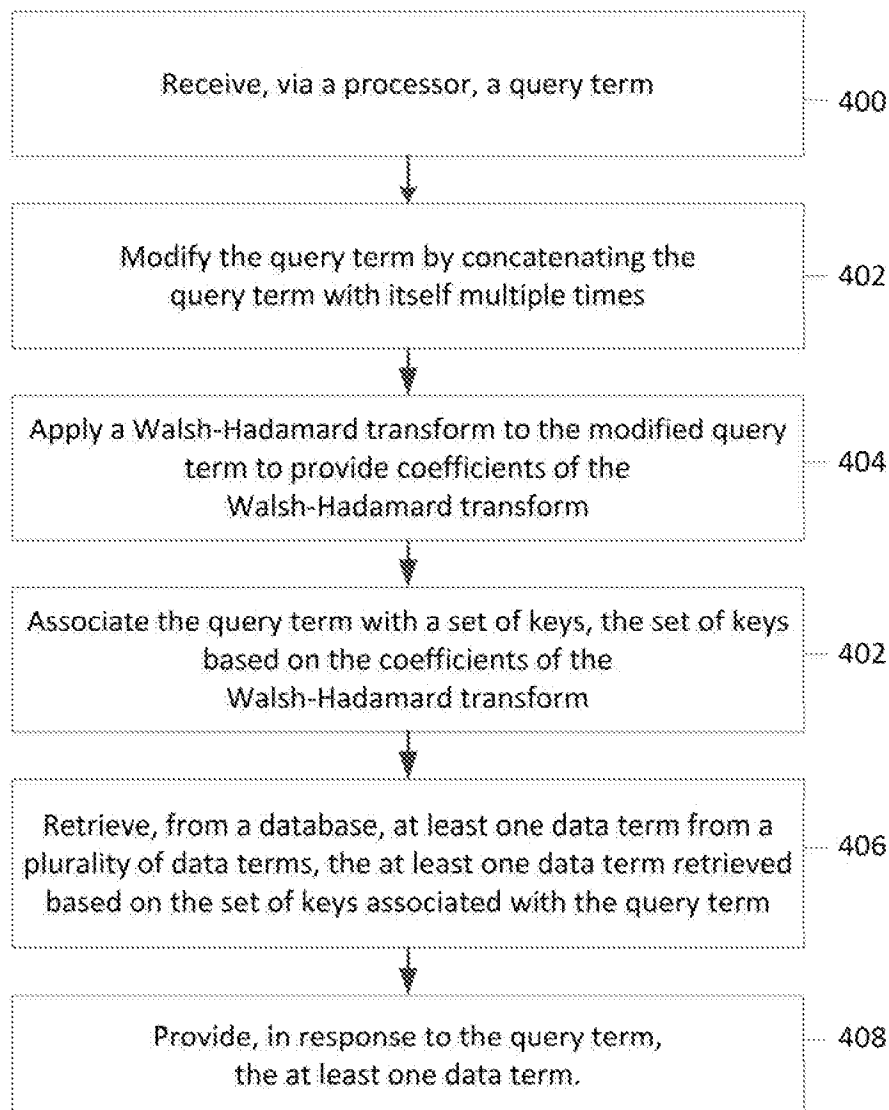


Fig. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/047803**A. CLASSIFICATION OF SUBJECT MATTER****G06F 17/00(2006.01)i, G06F 17/30(2006.01)i**

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 17/00; G06K 9/46; G06F 17/30; G06K 9/62; G06F 7/06

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: proximity, measure, approximate, nearest neighbor, Walsh-Hadamard transform, word set, similarity, coefficient, vector, concatenation, overlap, and similar terms.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2010-0325133 A1 (ROBERT L. ROUNTHWAITE et al.) 23 December 2010 See paragraphs [0043], [0045], and [0058]; and figure 6.	1-15
A	EP 1521210 B1 (SONY CORPORATION) 18 November 2009 See paragraph [0058] and claim 4.	1-15
A	EP 1072982 A2 (MATSUSHITA ELECTRIC INDUSTRIAL CO., LTD.) 31 January 2001 See paragraphs [0009], [0039], and [0043]; and figure 3.	1-15
A	US 2003-0108242 A1 (STEPHEN W. CONANT) 12 June 2003 See paragraph [0012]-[0015] and [0068]; and figure 5.	1-15
A	US 2008-0183682 A1 (CHRISTIAN A. LANG et al.) 31 July 2008 See paragraphs [0008] and [0016]-[0023]; and figures 1-2.	1-15



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

28 February 2015 (28.02.2015)

Date of mailing of the international search report

02 March 2015 (02.03.2015)

Name and mailing address of the ISA/KR

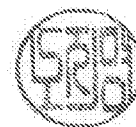
International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701,
Republic of Korea

Facsimile No. ++82 42 472 7140

Authorized officer

NHO, Ji Myong

Telephone No. +82-42-481-8528



INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.

PCT/US2014/047803

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2010-0325133 A1	23/12/2010	US 8606786 B2	10/12/2013
EP 1521210 B1	18/11/2009	CN 1324509 C	04/07/2007
		CN 1552042 A	01/12/2004
		DE 60330147 D1	31/12/2009
		EP 1521210 A1	06/04/2005
		EP 1521210 A4	04/07/2007
		EP 1521210 B9	15/09/2010
		JP 2004-046370 A	12/02/2004
		JP 4623920 B2	02/02/2011
		KR 10-1021044 B1	14/03/2011
		KR 10-2005-0016278 A	21/02/2005
		US 2005-0033523 A1	10/02/2005
		US 7260488 B2	21/08/2007
		WO 2004-006185 A1	15/01/2004
EP 1072982 A2	31/01/2001	CN 1282934 A	07/02/2001
		EP 1072982 A3	26/05/2004
		JP 2001-043236 A	16/02/2001
		KR 10-0408637 B1	06/12/2003
		TW 476034 B	11/02/2002
US 2003-0108242 A1	12/06/2003	None	
US 2008-0183682 A1	31/07/2008	US 2007-0050566 A1	01/03/2007
		US 7512282 B2	31/03/2009
		US 7894627 B2	22/02/2011