(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 December 2015 (03.12.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/183246 A1

(51) International Patent Classification:
G06F 17/00 (2006.01) *G06F 17/30* (2006.01)

(21) International Application Number:
PCT/US2014/039660

(22) International Filing Date:
28 May 2014 (28.05.2014)

(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: **SIMSKE, Steven J**; 3404 E Harmony Road, Fort Collins, Colorado 80528-9544 (US). **VANS, Marie**; 3404 E Harmony Road, Fort Collins, Colorado 80528-9544 (US). **STURGILL, Malgorzata M**; 3404 E Harmony Road, Fort Collins, Colorado 80528-9544 (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: DATA EXTRACTION BASED ON MULTIPLE META-ALGORITHMIC PATTERNS

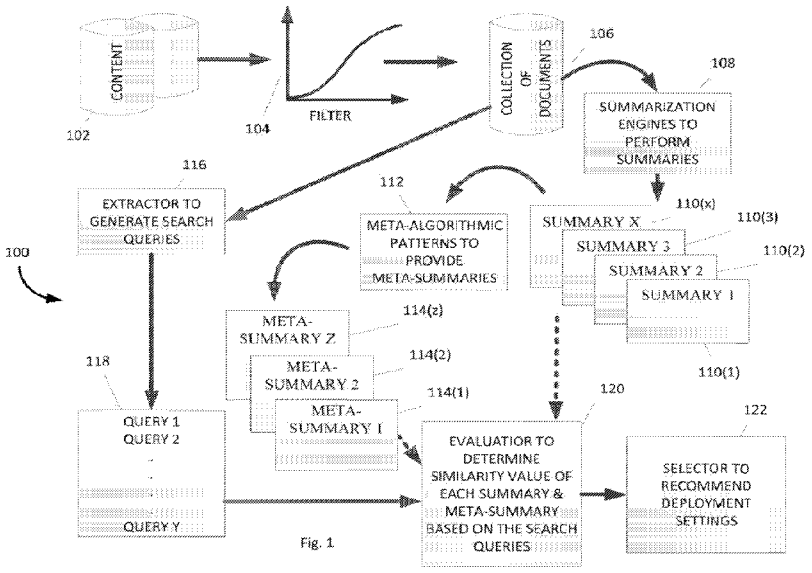

Fig. 1

(57) Abstract: One example is a system including a plurality of combinations of summarization engines and/or meta-algorithmic patterns used to combine a plurality of summarizers, an extractor, an evaluator, and a selector. Each of the plurality of combinations of summarization engines and/or meta-algorithmic patterns receives content to provide a meta-summary of the content. The extractor generates a collection of search queries based on the content. The evaluator determines a similarity value of each combination of summarization engines and/or meta-algorithmic patterns for the collection of search queries. The selector selects an optimal combination of summarization engines and/or meta-algorithmic patterns based on the similarity value.

WO 2015/183246 A1

DATA EXTRACTION BASED ON MULTIPLE META-ALGORITHMIC PATTERNS

Background

[0001] Summarizers are computer-based applications that provide a summary of some type of content, such as text. Meta-algorithms are computer-based applications that may be applied to combine two or more summarizers to yield meta-summaries. Meta-summaries may be used in a variety of applications, including data mining applications.

Brief Description of the Drawings

[0002] Figure 1 is a functional block diagram illustrating one example of a system for data extraction based on multiple meta-algorithmic patterns.

[0003] Figure 2 is a block diagram illustrating one example of a processing system for implementing the system for data extraction based on multiple meta-algorithmic patterns.

[0004] Figure 3 is a block diagram illustrating one example of a computer readable medium for data extraction based on multiple meta-algorithmic patterns.

[0005] Figure 4 is a flow diagram illustrating one example of a method for data extraction based on multiple meta-algorithmic patterns.

Detailed Description

[0006] 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.

[0007] Figure 1 is a functional block diagram illustrating one example of a system 100 for data extraction based on multiple meta-algorithmic patterns. The system receives content, such as a collection of documents, and filters the content. The filtered content is then processed by a plurality of different summarization engines to provide a plurality of summaries. The summaries may be further processed by a plurality of different meta-algorithmic patterns, each meta-algorithmic pattern to be applied to at least two summaries, to provide a collection of meta-summaries, where each meta-summary of the collection of meta-summaries is provided using the at least two summaries. The filtered content is also processed to generate a plurality of search queries based on the collection of documents.

[0008] Summarization may be used as a decision criterion for text analytics, each with its own specific elements. In addition to applications related to text analytics, functional summarization may be used for evaluative purposes. Similarity of document selection during search may be used to evaluate translators. For example, a useful language translator will result in the same search behavior as in the original language for the same collection of queries. In general, similarity of summarization indicates similarity of the documents, and so differences in the behavior of multiple summarization engines rather than the behavior of one engine or the collective behavior of a set of engines is often significant.

[0009] Validation and/or relative assessment of the individual meta-algorithmic approaches is based on the utilization of indices/keywords for search behavior. After the indices/keywords have been extracted, tags on the documents (and their relative order and/or relative weighting) may be utilized for searching on the collection of documents. The optimal overall extraction to represent a large

collection of documents is the one that behaves most closely like the original (unadulterated) collection of documents. That is, the best combination of summarization engines and/or meta-algorithmic patterns is the one that results in search behavior least different from the search behavior for the original collection of documents.

[0010] A similarity score is determined for each combination of meta-algorithmic patterns and summarization engines, the similarity score being indicative of a difference in search behaviors of the collection of documents and the collection of meta-summaries, the search behaviors responsive to the plurality of search queries. The summarization engine and/or meta-algorithmic pattern that provides a summary and/or meta-summary, respectively, that has a minimum difference in the aforesaid search behavior is then recommended for deployment. In this way, a summarization architecture optimized for a search task is determined. As described herein, the summarization architecture may be a combination of a plurality of summarization engines and/or a plurality of meta-algorithmic patterns.

[0011] Meta-summaries are summarizations created by the intelligent combination of two or more standard or primary summaries. The intelligent combination of multiple intelligent algorithms, systems, or engines is termed "meta-algorithmics", and first-order, second-order, and third-order patterns for meta-algorithmics may be defined.

[0012] System 100 includes content 102, a filter 104, filtered collection of documents 106, summarization engines 108, summaries 110(1)-110(x), a plurality of meta-algorithmic patterns 112, meta-summaries 114(1)-114(z), extractor 116, a plurality of search queries 118, an evaluator 120, and a selector 122, where "x", "y", and "z" are any suitable numbers of summaries, queries, and meta-summaries, respectively. Content 102 includes text, meta-data, and/or other computer storable data, including images, sound, and/or video. Content 102 may include a book, an article, a document, or other suitable information. Filter 104 filters content 102 to provide a collection of documents 106 suitable for processing by summarization engines 108. In one example, filter 104 may remove common words (e.g., stop words such as "the", "a", "an",

"for", and "of") from content 102. Filter 104 may also remove images, sound, video and/or other portions of content 102 to provide filtered content 106. In one example, filter 104 is excluded and content 102 is provided directly to summarization engines 108.

[0013] Summarization engines 108 summarize documents in the collection of documents 106 to provide a plurality of summaries 110(1)-110(x). In one example, each of the summarization engines provides a summary including one or more of the following summarization outputs:

(1) a set of key words;

(2) a set of key phrases;

(3) an extractive set of clauses;

(4) an extractive set of sentences;

(5) an extractive set of clustered sentences, paragraphs, and other text chunks; or

(6) an abstractive summarization or semantic.

[0014] In other examples, a summarization engine may provide a summary including another suitable summarization output. Different statistical language processing (SLP) and natural language processing (NLP) techniques may be used to generate the summaries.

[0015] Meta-algorithmic patterns 112 are used to summarize summaries 110(1)-110(x) to provide a plurality of meta-summaries 114(1)-114(z). Each of the meta-algorithmic patterns is applied to two or more summaries to provide a meta-summary. In one example, each of the plurality of meta-algorithmic patterns is based on one or more of the following approaches:

(1) weighted voting;

(2) predictive selection;

(3) tessellation and recombination;

(4) tessellation and recombination with a decisioner;

(5) predictive selection with a secondary engine; or

(6) majority voting

In other examples, a meta-algorithmic pattern may be based on another suitable approach.

[0016] In the (1) weighted voting approach, the output of multiple summarization engines or meta-algorithmic patterns is combined and relatively weighted based on the relative confidence in each summarization engine or meta-algorithmic pattern, and the relative weighting of the terms, phrases, clauses, sentences, and chunks in each summarization. In one example, the output of data mining engines may be given in ranked order (e.g., a matrix R), and the weighted voting results may be given in a weighted output (W matrix). If the output of the data mining engines are given in weighted order (W matrix), then weighted voting results in a weighted matrix (e.g., a matrix W). Thus, this meta-algorithmic pattern allows the combination of both ranked and weighted outputs from the summarization engines or meta-algorithmic patterns. Data mining is the discovery of patterns in large data sets. Summarization engines may be combined to provide a summary for the data extracted. In some examples, data mining may provide an exhaustive description of the text information. In some examples, data mining may provide a gist of a document content (specific data) or content that distinguishes the document from other documents (differential data). For the data mining task, the summaries and meta-summaries may be evaluated to determine the summarization architecture that provides the data mining results that provide a significant recovery of tagged content (e.g., ground truth to-be-mined data). As described herein, the summarization architecture may then be selected and recommended for deployment.

[0017] The (2) predictive selection approach may include looking at the general topic associated with the key terms in a portion of text and applying a specific summarization engine or set of summarization engines based on membership within a class associated with a particular topic or set of topics. In one example, a combination of summarization engines or meta-algorithmic patterns is selected to create an abridged document representation, A, of an original document, D. The document representation A is then used to represent the document. In one example, such a selection of the combination of summarization engines or meta-algorithmic patterns is based on attributes of the document (including document entropy, document author, document language, and document length). In general, different documents will use

different combinations of summarization engines or meta-algorithmic patterns, and the overall result may be compared to the best result from any other meta-algorithmic pattern (such as Weighted Voting) for accuracy.

[0018] In the (3) tessellation and recombination method, two types of tessellations may be utilized: (a) tessellation by commonality across multiple combinations of summarization engines or meta-algorithmic patterns, and (b) tessellation by commonality with other documents. In the case of (a), the common terms are kept initially and then incrementally more are added based on maximum dispersion (that is, maximum weighted differences between abridged documents). In the case of (b), the common terms are discarded initially and then incrementally more are added, again based on maximum dispersion between abridged documents.

[0019] In the (4) tessellation and recombination method with an expert decisioner, the expert decisioner is used to assign additional (presumably optimal) terms (keywords, phrases, etc.) to the tessellated abridged documents described herein. However, rather than adding them based on a generic means (such as maximum difference or "dispersion" as in (3)), here the terms added are guided by the terms that are considered most representative of the documents themselves. For example, assignment of additional terms may be guided by (a) the salient terms of the document class the document belongs to, if this information is available; (b) the overall set of relevant search terms, if these are available, or (c) significant terms when comparing the document to a large set of mixed-class documents.

[0020] In the (5) predictive selection with a secondary engine method, as with the predictive selection described herein, the attributes of document (including document entropy, document author, document language, document length, etc.) are used to choose a particular data mining engine to create the abridged document representation, A, of the original document, D. This A is then used to represent the document thereafter. Different documents will use different data mining engines, and the overall result may be compared to the best result from any other meta-algorithmic pattern (such as Weighted Voting, described herein) for accuracy. In this design pattern, if there is no clear "winner" for predictive

selection, then a secondary meta-algorithmic pattern (one of (1) weighted voting, (3) tessellation and recombination, or (4) tessellation and recombination with an expert decisioner) is selected. This may be utilized when predictive selection has a low confidence level; for example, in systems with only modest training data (or ground truth set).

[0021] In the (6) majority voting method, key terms agreed on by the majority of the summarization engines are selected. Additional terms left over in the majority voting (based on its error rate either overall or within the subclass chosen by the predictive selection pattern, etc.) are added to create the set of terms. This method tends to select more terms than any single summarization engine since it merges two streams of selection, but may also be pruned by selecting only the first few terms from the majority voting. This pattern works well as more summarization engines are added to the system 100, and in particular, when an odd number of engines are integrated.

[0022] Extractor 116 generates a plurality of search queries 118 based on the collection of documents 106. The output of the meta-algorithmic patterns are used for identification of keywords, extraction of salient data and tagging of the documents (e.g. for search, indexing and clustering). Collectively, these are termed "data mining". The original collection of documents may be denoted D{N}, a set of N documents. The collection D{N} is utilized as a bag of words to generate a collection of search queries. This may be achieved by utilizing any of the to-be-deployed combination of summarization engines and/or meta-algorithmic patterns to extract the search queries as the key words, extracted data or tags for the collection D{N}. The output of the summarizers are used to tag documents. For a metadata tagging task, the summaries and meta-summaries are evaluated to determine the summarization architecture that provides the metadata tags (e.g., indices, descriptors, semantic tags) that provide a match to training data. Each summarization architecture is evaluated for its relative value in the search task. The relative value in the search task, (i.e., the relevance or utility for the search task), may be evaluated based on training data, feedback received from users, and/or other suitable criteria applicable to the search task.

[0023] The plurality of search queries is called S{M}. A first action of the plurality of search queries S{M} on the collection of documents D{N} is represented as:

$$S\{M\} \rightarrow D\{N\} : \sum_{i=1}^{M} \sum_{j=1}^{N} R_{ij}(D\{N\})$$

Or, alternatively:

$$S\{M\} \rightarrow D\{N\} : \sum_{i=1}^{M} \sum_{j=1}^{N} W_{ij}(D\{N\})$$

where R is an M x N matrix of ranks, and W is an M x N matrix of weights, depending on what the output of the plurality of search queries is. The R-method is the non-parametric method, while the W-method is the parametric method. As indicated, the R-method is based on a ranking of the plurality of search queries, and the W-method is based on a weighting of the plurality of search queries.

[0024] The meta-algorithmic patterns of two or more data mining engines are used to create a collection of meta-summaries A{N}. A second action of the plurality of search queries S{M} on the collection of meta-summaries A{N} is represented as:

$$S\{M\} \rightarrow A\{N\} : \sum_{i=1}^{M} \sum_{j=1}^{N} R_{ij}(A\{N\})$$

Or, alternatively:

$$S\{M\} \rightarrow A\{N\} : \sum_{i=1}^{M} \sum_{j=1}^{N} W_{ij}(A\{N\})$$

where R is an M x N matrix of ranks, and W is an M x N matrix of weights, depending on what the output of the plurality of search queries is. As indicated, the R-method is based on a ranking of the plurality of search queries, and the W-method is based on a weighting of the plurality of search queries.

[0025] Evaluator 120 determines a similarity score for each combination of meta-algorithmic patterns and summarization engines, the similarity score being indicative of a difference in search behaviors of the collection of documents 106

and the collection of meta-summaries 114(1)-114(z), the search behaviors responsive to the plurality of search queries 118.

[0026] In one example, the similarity score is based on a difference between the first action of the plurality of search queries 118 on the collection of documents 106, and the second action of the plurality of search queries 118 on the collection of meta-summaries 114(1)-114(z). If there are L meta-algorithmic patterns, then the similarity score may be based on a difference between a first action of the plurality of search queries on the collection of documents, and a second action of the plurality of search queries on the collection of meta-summaries, as given by:

$$\left(\sum_{i=1}^{M}\sum_{j=1}^{N}R_{ij}(D\{N\})-\sum_{i=1}^{M}\sum_{j=1}^{N}R_{ij}(A_k\{N\})\right)$$

Or, alternatively:

$$\left(\sum_{i=1}^{M}\sum_{j=1}^{N}W_{ij}(D\{N\})-\sum_{i=1}^{M}\sum_{j=1}^{N}W_{ij}(A_k\{N\})\right)$$

[0027] An optimum pattern is the one satisfying:

$$\min_{k=1\ldots L}\left(\sum_{i=1}^{M}\sum_{j=1}^{N}R_{ij}(D\{N\})-\sum_{i=1}^{M}\sum_{j=1}^{N}R_{ij}(A_k\{N\})\right)$$

Or, alternatively:

$$\min_{k=1\ldots L}\left(\sum_{i=1}^{M}\sum_{j=1}^{N}W_{ij}(D\{N\})-\sum_{i=1}^{M}\sum_{j=1}^{N}W_{ij}(A_k\{N\})\right)$$

[0028] Selector 122 selects for deployment, via the processing system, a combination of the meta-algorithmic patterns and the summarization engines, where the selection is based on the similarity score. In one example, the selector 122 selects for deployment the combination of the meta-algorithmic patterns and the summarization engines that minimize the similarity score. The recommended deployments settings include the summarization engines and/or meta-algorithmic patterns that provide the optimum summarization architecture with respect to the search behaviors. The optimum summarization architecture

may be integrated into a system real-time. The system may be re-configured per preference, schedule, need, or upon the completion of a threshold number of new instances of the tasks.

[0029] In one example, the selector 122 generates a meta-summary of a given document of the collection of documents by applying the selected combination of the meta-algorithmic patterns and summarization engines to the given document. In one example, the selector 122 associates, in a database, the generated meta-summary with the given document.

[0030] In one example, system 100 is fully automatable. As described herein, the summaries and meta-summaries may be evaluated to determine the summarization architecture that provides a document summary that significantly matches the training data. Generally, the larger the training data and the larger the number of summarization engines available, the better the final system performance. System performance is optimized, however, when the training data is much larger than the number of summarization engines. The summarization architecture is then selected and recommended for deployment.

[0031] For example, the number of possible combinations of the meta-algorithmic patterns and the summarization engines is:

$$N_{MP} * (2^{N_{KE}} - 1 - N_{KE})$$

where, N_{MP} is a number of meta-algorithmic patterns used, and N_{KE} is a number of keyword-generating engines used. A plurality of such combinations may be obtained from one of the six meta-algorithmic patterns described herein: (1) weighted voting; (2) predictive selection; (3) tessellation and recombination; (4) tessellation and recombination with a decisioner; (5) predictive selection with a secondary engine; or (6) majority voting. For example, many different combinations may be used with predictive selection. Accordingly, the system behavior may be given an adaptive summarization architecture over time, allowing it to be very general when first deployed and then narrow the number of feasible combinations over time as a number of documents, classes of documents, and/or search terms increases as the system scales and/or evolves. For example, when the system is first deployed, the following may hold:

$$N_{MP} * (2^{N_{KE}} - 1 - N_{KE}) > N_C$$

or

$$N_{MP} * (2^{N_{KE}} - 1 - N_{KE}) > N_{ST}$$

where, N_C is the number of classes of documents and N_{ST} is the number of search terms. This deployment specification may allow many more combinations than the number of classes and/or search terms, which provides design flexibility (e.g., an artificial neural network, genetic algorithm).

[0032] As more documents per class and per search term become part of the system, different combinations of the meta-algorithmic patterns and the summarization engines will be de-selected based on their relative lack of effectiveness. As the system evolves, the following may hold:

$$N_{MP} * (2^{N_{KE}} - 1 - N_{KE}) << N_C$$

and

$$N_{MP} * (2^{N_{KE}} - 1 - N_{KE}) << N_{ST}$$

[0033] Accordingly, the system will have naturally resolved to a smaller set of optimally-combined meta-algorithmic patterns and summarization engines. In one example, as the system is guided by the same combination over a long period (or upon the addition of a substantial amount of documents since the last major system change), meta-algorithmic patterns and summarization engines may be added so that the following may hold:

$$N_{MP} * (2^{N_{KE}} - 1 - N_{KE}) \approx N_C$$

and

$$N_{MP} * (2^{N_{KE}} - 1 - N_{KE}) \approx N_{ST}$$

[0034] This allows some flexibility to changes in the behavior of the collection of documents, while retaining some memory of the former learned combination of meta-algorithmic patterns and/or summarization engines.

[0035] Figure 2 is a block diagram illustrating one example of a processing system 200 for implementing the system 100 for data extraction based on multiple meta-algorithmic patterns. Processing system 200 includes a processor 202, a memory 204, input devices 218, and output devices 220. Processor 202, memory 204, input devices 218, and output devices 220 are coupled to each other through communication link (e.g., a bus).

[0036] 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.

[0037] Memory 204 stores content 206 for processing by processing system 200. Memory 204 also stores instructions to be executed by processor 202 including instructions for summarization engines 208, meta-algorithmic patterns 210, an extractor 212, an evaluator 214, and a selector 216. In one example, summarization engines 208, meta-algorithmic patterns 210, extractor 212, evaluator 214, and selector 216 include summarization engines 108, meta-algorithmic patterns 112, extractor 116, evaluator 120, and selector 122, respectively, as previously described and illustrated with reference to Figure 1.

[0038] In one example, processor 202 executes instructions of filter to filter a collection of documents to provide a filtered collection of documents 206. Processor 202 executes instructions of a plurality of summarization engines 210 to summarize the collection of documents 206 to provide summaries. Processor 202 executes instructions of a plurality of meta-algorithmic patterns 212 to summarize the summaries to provide meta-summaries. Processor 202 executes instructions of extractor 212 to generate a plurality of search queries from the collection of documents 206. Processor 202 executes instructions of evaluator 214 to determine the similarity score for each combination of meta-algorithmic patterns and summarization engines, the similarity score indicative of a difference in search behaviors of the collection of documents and the collection of meta-summaries, the search behaviors responsive to the plurality of search queries. Processor 202 executes instructions of selector 216 to select for deployment a combination of the meta-algorithmic patterns and the summarization engines, the selection based on the similarity score. The selected summarization architecture, i.e. the combination of the meta-algorithmic patterns and the summarization engines, is then recommended for deployment by processing system 200.

[0039] Input devices 218 include a keyboard, mouse, data ports, and/or other suitable devices for inputting information into processing system 200. In one example, input devices 218 are used to input feedback from users for evaluating the summaries and meta-summaries for search queries. Output devices 220 include a monitor, speakers, data ports, and/or other suitable devices for outputting information from processing system 200. In one example, output devices 220 are used to output summaries and meta-summaries to users and to recommend the optimum summarization architecture for data extraction.

[0040] In one example, the selector 216 generates a meta-summary of a given document of the collection of documents by applying the selected combination of the meta-algorithmic patterns and summarization engines to the given document. In one example, the selector 216 associates, in a database, the generated meta-summary with the given document.

[0041] In one example, a search query directed at a document is received via input devices 218. The processor 202 retrieves, from the database, a meta-summary associated with the document; and generates, based on the retrieved meta-summary, search results responsive to the search query. The search results are then provided via output devices 220.

[0042] Figure 3 is a block diagram illustrating one example of a computer readable medium for data extraction based on multiple meta-algorithmic patterns. Processing system 300 includes a processor 302, a computer readable medium 308, a plurality of summarization engines 304, and a plurality of meta-algorithmic patterns 306. Processor 302, computer readable medium 308, the plurality of summarization engines 304, and the plurality of meta-algorithmic patterns 306 are coupled to each other through communication link (e.g., a bus).

[0043] Processor 302 executes instructions included in the computer readable medium 308. Computer readable medium 308 includes document receipt instructions 310 to receive a collection of documents. Computer readable medium 308 includes summarization instructions 312 of a plurality of summarization engines 304 to summarize the received collection of documents to provide summaries. Computer readable medium 308 includes meta-

algorithmic pattern instructions 314 of a plurality of meta-algorithmic patterns 306 to summarize the summaries to provide meta-summaries. Computer readable medium 308 includes query generation instructions 316 of extractor to generate a plurality of search queries from the collection of documents. Computer readable medium 308 includes similarity value determination instructions 318 of evaluator to determine the similarity score for each combination of meta-algorithmic patterns and summarization engines, the similarity score indicative of a difference in search behaviors of the collection of documents and the collection of meta-summaries, where the search behaviors are responsive to the plurality of search queries. Computer readable medium 308 includes deployment instructions 320 of selector to select for deployment a combination of the meta-algorithmic patterns and the summarization engines, the selection based on the similarity score. The selected summarization architecture is then recommended for deployment by processing system 300.

[0044] Figure 4 is a flow diagram illustrating one example of a method for data extraction based on multiple meta-algorithmic patterns. At 400, content is filtered to provide a collection of documents. At 402, a plurality of search queries are generated. At 404, a plurality of combinations of meta-algorithmic patterns and summarization engines are applied to provide a collection of meta-summaries. At 406, the plurality of combinations are evaluated to determine a similarity score of each combination. At 408, a combination of the meta-algorithmic patterns and the summarization engines having the least similarity score is selected.

[0045] In one example, the method may further include generating a meta-summary of a given document of the collection of documents by applying the selected combination of the meta-algorithmic patterns and summarization engines to the given document, and associating, in a database, the generated meta-summary with the given document.

[0046] In one example, the method may further include receiving a search query directed at a document, and retrieving, from the database, a meta-summary associated with the document. The method may further include generating, based on the retrieved meta-summary, search results responsive to the search

query. In one example, the generated search results may be provided via output devices.

[0047] Examples of the disclosure provide a generalized system for using multiple summaries and meta-algorithms to optimize a text-related intelligence generating or machine intelligence system. The generalized system provides a pattern-based, automatable approach to summarization that may learn and improve over time, and is not fixed on a single technology or machine learning approach. In this way, the content used to represent a larger body of text, suitable to a wide range of applications, may be optimized.

[0048] Although specific examples have been illustrated and described herein, a variety of alternate and/or equivalent implementations 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 plurality of summarization engines, each summarization engine to receive, via a processing system, a collection of documents to provide a summary of each document of the collection of documents;

a plurality of meta-algorithmic patterns, each meta-algorithmic pattern to be applied to at least two summaries to provide, via the processing system, a collection of meta-summaries, each meta-summary of the collection of meta-summaries provided using the at least two summaries;

an extractor to generate a plurality of search queries from the collection of documents; and

an evaluator to determine a similarity score for each combination of meta-algorithmic patterns and summarization engines, the similarity score indicative of a difference in search behaviors of the collection of documents and the collection of meta-summaries, the search behaviors responsive to the plurality of search queries.

2. The system of claim 1, further comprising a selector to select for deployment, via the processing system, a combination of the meta-algorithmic patterns and the summarization engines, the selection based on the similarity score.

3. The system of claim 2, wherein the selector generates a meta-summary of a given document of the collection of documents by applying the selected combination of the meta-algorithmic patterns and summarization engines to the given document.

4. The system of claim 1, wherein the evaluation of each combination of meta-algorithmic patterns and summarization engines comprises comparing each

combination of meta-algorithmic patterns and summarization engines to training data.

5. The system of claim 2, wherein the similarity score is based on a difference between a first action of the plurality of search queries on the collection of documents, and a second action of the plurality of search queries on the collection of meta-summaries.

6. The system of claim 5, wherein the first action and the second action are based on a ranking of the plurality of search queries.

7. The system of claim 5, wherein the first action and the second action are based on a weighting of the plurality of search queries.

8. The system of claim 1, wherein the plurality of meta-algorithmic patterns are selected from the group comprising weighted voting, predictive selection, tessellation and recombination, tessellation and recombination with a decisioner, predictive selection with a secondary engine, and majority voting.

9. A method to extract data from documents based on meta-algorithm patterns, the method comprising:

filtering content to provide a collection of documents;

generating a plurality of search queries from the collection of documents;

applying a plurality of combinations of meta-algorithmic patterns and summarization engines, wherein:

each summarization engine provides a summary of each document of the collection of documents,

each meta-algorithmic pattern is applied to at least two summaries to provide, via a processor, a collection of meta-summaries, each meta-summary of the collection of meta-summaries provided using the at least two summaries;

evaluating the plurality of combinations to determine a similarity score of each combination, the similarity score based on a difference between a first action of the plurality of search queries on the collection of documents, and a second action of the plurality of search queries on the collection of meta-summaries; and

selecting a combination of the meta-algorithmic patterns and the summarization engines having the least similarity score.

10. The method of claim 9, further comprising:

generating a meta-summary of a given document of the collection of documents by applying the selected combination of the meta-algorithmic patterns and summarization engines to the given document; and

associating, in a database, the generated meta-summary with the given document.

11. The method of claim 10, further comprising:

receiving a search query directed at a document;

retrieving, from the database, a meta-summary associated with the document; and

generating, based on the retrieved meta-summary, search results responsive to the search query.

12. The method of claim 9, wherein the plurality of meta-algorithmic patterns are selected from the group comprising weighted voting, predictive selection, tessellation and recombination, tessellation and recombination with a decisioner, predictive selection with a secondary engine, and majority voting.

13. A non-transitory computer readable medium comprising executable instructions to:

receive a collection of documents via a processor;

summarize the collection of documents to provide a plurality of summaries via the processor;

summarize the plurality of summaries using a plurality of meta-algorithmic patterns to provide a collection of meta-summaries via the processor;

generate a plurality of search queries from the collection of documents;

determine a similarity score of each combination of a plurality of combinations of meta-algorithmic patterns and summarization engines, the similarity score based on a difference between a first action of the plurality of search queries on the collection of documents, and a second action of the plurality of search queries on the collection of meta-summaries; and

select for deployment, via the processor, a combination of the meta-algorithmic patterns and the summarization engines having the least similarity score.

14. The non-transitory computer readable medium of claim 13, wherein the first action and the second action are based on a ranking of the collection of search queries.

15. The non-transitory computer readable medium of claim 13, wherein the plurality of meta-algorithmic patterns are selected from the group comprising weighted voting, predictive selection, tessellation and recombination, tessellation and recombination with a decisioner, predictive selection with a secondary engine, and majority voting.

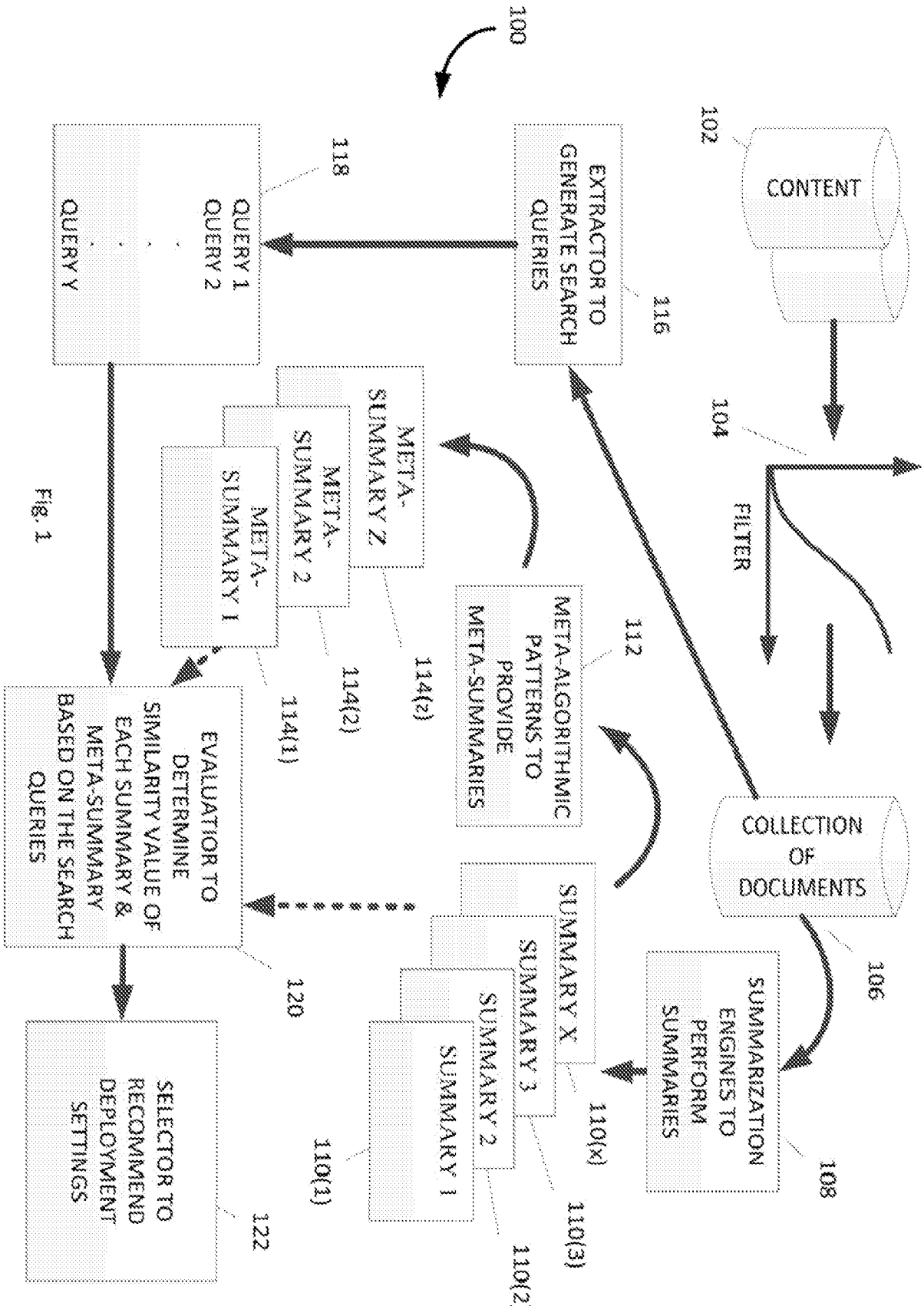
100
102
CONTENT
104
FILTER
106
COLLECTION OF DOCUMENTS
108
SUMMARIZATION ENGINES TO PERFORM SUMMARIES
110(1)
SUMMARY 1
110(2)
SUMMARY 2
110(3)
SUMMARY 3
110(x)
SUMMARY X
112
META-ALGORITHMIC PATTERNS TO PROVIDE META-SUMMARIES
114(1)
META-SUMMARY 1
114(2)
META-SUMMARY 2
114(z)
META-SUMMARY Z
116
EXTRACTOR TO GENERATE SEARCH QUERIES
118
QUERY 1
QUERY 2
QUERY Y
120
EVALUATOR TO DETERMINE SIMILARITY VALUE OF EACH SUMMARY & META-SUMMARY BASED ON THE SEARCH QUERIES
122
SELECTOR TO RECOMMEND DEPLOYMENT SETTINGS

Fig. 1

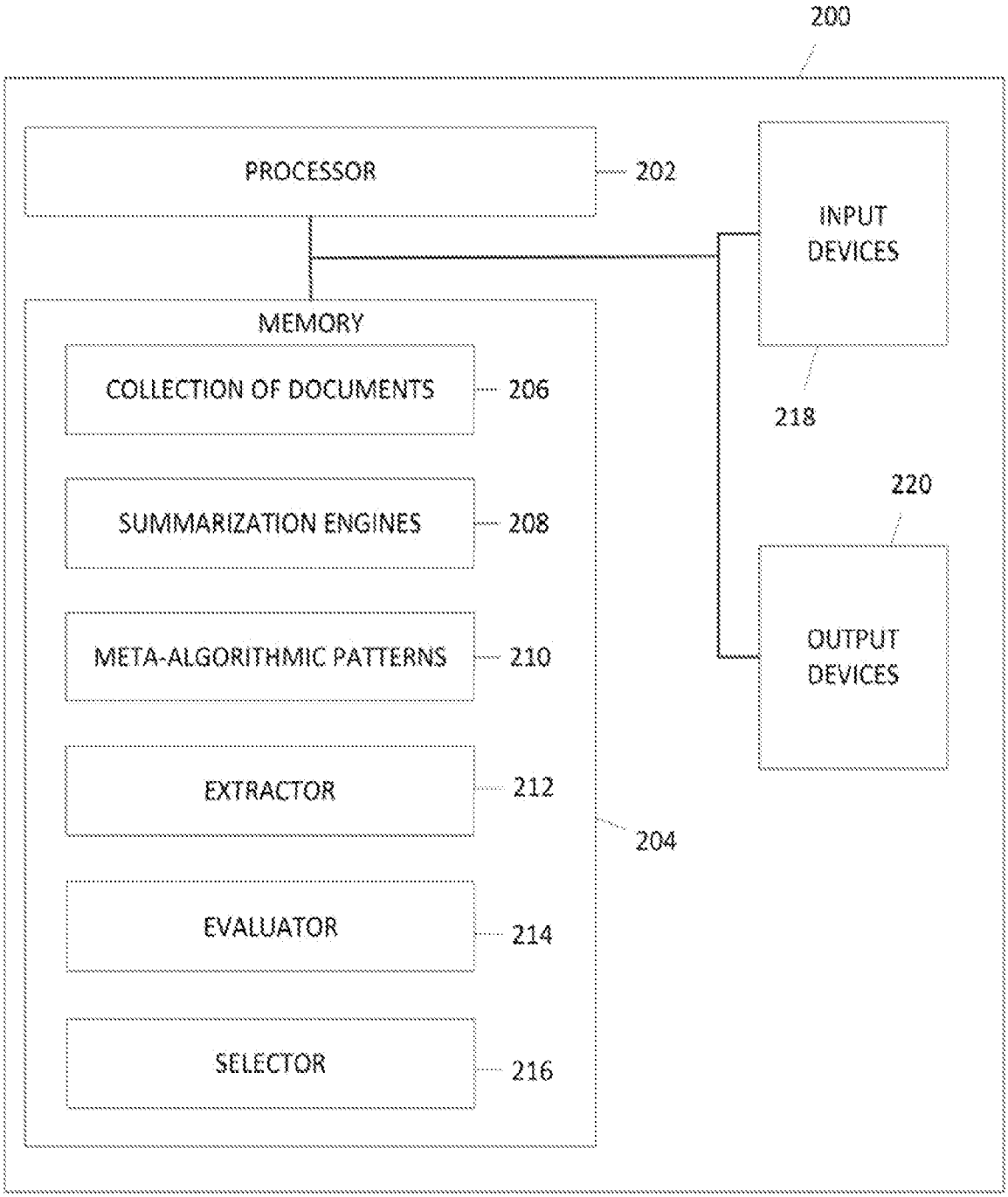
200
PROCESSOR
202
INPUT DEVICES
MEMORY
COLLECTION OF DOCUMENTS
206
218
220
SUMMARIZATION ENGINES
208
META-ALGORITHMIC PATTERNS
210
OUTPUT DEVICES
EXTRACTOR
212
204
EVALUATOR
214
SELECTOR
216

Fig. 2

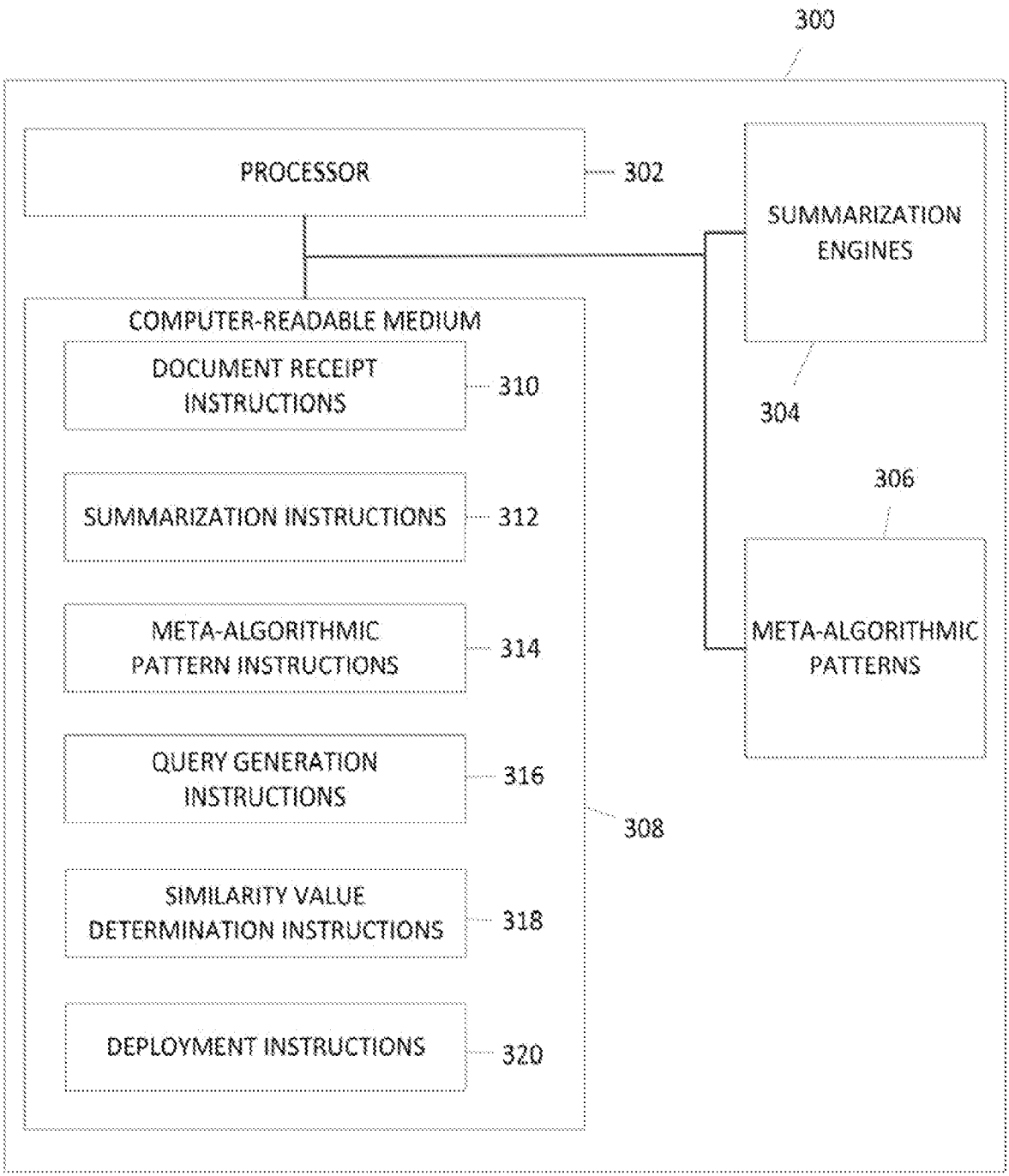
300
PROCESSOR
302
SUMMARIZATION ENGINES
COMPUTER-READABLE MEDIUM
DOCUMENT RECEIPT INSTRUCTIONS
310
304
306
SUMMARIZATION INSTRUCTIONS
312
META-ALGORITHMIC PATTERN INSTRUCTIONS
314
META-ALGORITHMIC PATTERNS
QUERY GENERATION INSTRUCTIONS
316
308
SIMILARITY VALUE DETERMINATION INSTRUCTIONS
318
DEPLOYMENT INSTRUCTIONS
320

Fig. 3

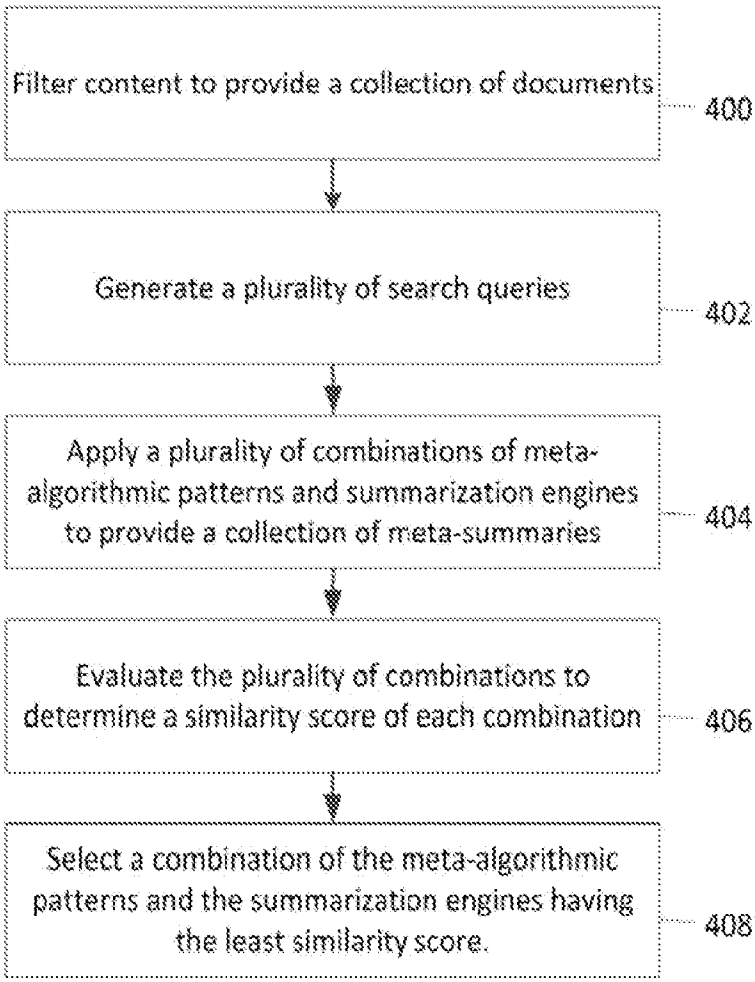
Filter content to provide a collection of documents
400
Generate a plurality of search queries
402
Apply a plurality of combinations of meta-algorithmic patterns and summarization engines to provide a collection of meta-summaries
404
Evaluate the plurality of combinations to determine a similarity score of each combination
406
Select a combination of the meta-algorithmic patterns and the summarization engines having the least similarity score.
408

Fig. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/039660

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; G06F 17/20; G06F 17/30; G06F 17/21; G06F 17/27

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: summarization engine, meta-summary, combination, search query, search behavior, similarity score, 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	WO 2014-058433 A1 (HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.) 17 April 2014 See paragraphs [0007], [0014]-[0015], [0021], [0023], [0030], and [0032]; and figure 1.	1-15
A	US 2008-0281810 A1 (SMYTH, BARRY et al.) 13 November 2008 See paragraphs [0043], [0046]-[0047], and [0052]-[0058]; and figures 1-2.	1-15
A	US 2010-0031142 A1 (NAGATOMO, KENTARO) 04 February 2010 See paragraphs [0141]-[0142], [0146]-[0148], and [0210]-[0213]; and figure 4.	1-15
A	US 7,292,972 B2 (LIN, XIAOFAN et al.) 06 November 2007 See column 2, lines 56-63; column 5, lines 7-15; column 6, lines 8-14, 20-38; and figures 2, 6, and 8A.	1-15
A	US 2012-0240032 A1 (MCKEOWN, KATHLEEN R. et al.) 20 September 2012 See paragraphs [0026] and [0033]; claim 1; and figure 1.	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	Date of mailing of the international search report
02 January 2015 (02.01.2015)	**02 January 2015 (02.01.2015)**

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/US2014/039660

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2014-058433 A1	17/04/2014	None	
US 2008-0281810 A1	13/11/2008	US 7805432 B2	28/09/2010
US 2010-0031142 A1	04/02/2010	CN 101529500 A	09/09/2009
		CN 101529500 B	23/05/2012
		JP 5104762 B2	19/12/2012
		WO 2008-050649 A1	02/05/2008
US 7292972 B2	06/11/2007	US 2004-153309 A1	05/08/2004
US 2012-0240032 A1	20/09/2012	CA 2496567 A1	25/03/2004
		US 2005-203970 A1	15/09/2005
		US 2012-240020 A1	20/09/2012
		US 8176418 B2	08/05/2012
		WO 2004-025490 A1	25/03/2004