



(51) International Patent Classification:
G06F 17/30 (2006.01) **G06F 17/22** (2006.01)

(21) International Application Number:
PCT/IL2009/001218

(22) International Filing Date:
27 December 2009 (27.12.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/193,862 31 December 2008 (31.12.2008) US
12/567,773 27 September 2009 (27.09.2009) US

(71) Applicant (for all designated States except US):
FORNOVA LTD [IL/IL]; Egoz 10/4, 34792 Haifa (IL).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **RUBANOVICH, Michael** [IL/IL]; Egoz 10/4, 34792 Haifa (IL). **BABITSKY, Dmitry** [IL/IL]; Hashikma 6/29, 36812 Nesher (IL).

(74) Agent: **DR. D. GRAESER LTD.**; 13 HaSadna St., P.O. Box 2496, 43650 Raanana (IL).

(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, 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, IS, JP, KE, KG, KM, 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, PE, PG, PH, PL, PT, RO, RS, RU, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (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, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

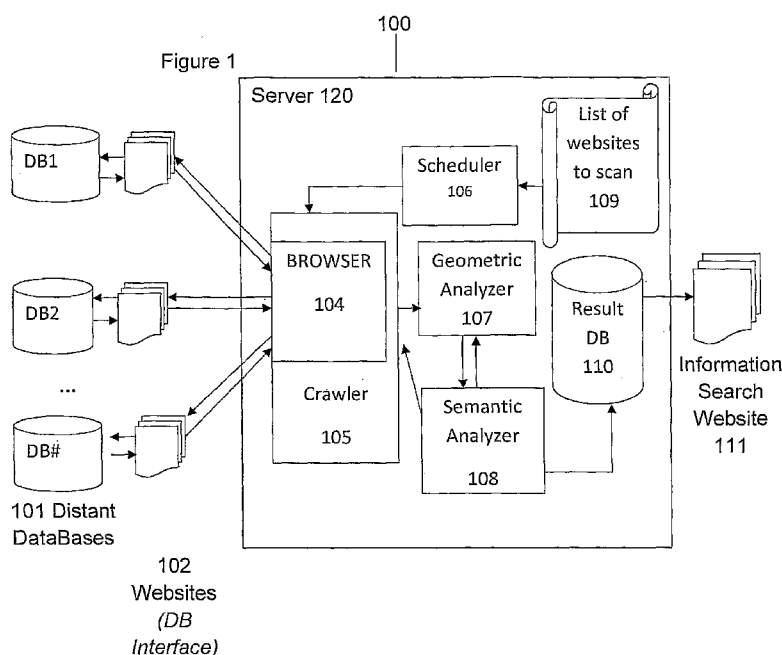
Published:

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

— with amended claims (Art. 19(1))

Date of publication of the amended claims: 7 October 2010

(54) Title: SYSTEM AND METHOD FOR AGGREGATING DATA FROM A PLURALITY OF WEB SITES



(57) Abstract: System and method for collecting information from a plurality of related sites, analyzing the information and storing the relevant information in a data base for future use. According to one embodiment of the present invention, the system uses the provided list of sites, whether obtained automatically or separately, queries them and analyzes the result retrieved from each site. The information may also optionally and preferably be ranked.

AMENDED CLAIMS

received by the International Bureau on 11 AUG 2010 (11.08.2010)

Claims:

1. A method for automatic aggregation of data from a plurality of web sites; comprising:
 - i. Automatically and periodically querying for said data from a plurality of related sites;
 - ii. Analyzing the results from said querying, said results comprising at least one document, said analyzing comprising geometrical analyzing of a page layout of the document, wherein said geometrical analyzing comprises determining one or more geometrical properties of the document; analyzing said one or more geometrical properties to determine a layout of the document; searching for a plurality of record containers within said layout; and determining a relevancy of a record from at least one record container according to a semantic analysis and according to said one or more geometrical properties;
 - iii. Storing the relevant record data in a data base; and
 - iv. Retrieving said data from said data base, upon demand from user.
2. The method of claim 1 wherein said searching for a plurality of record containers within said layout further comprises identifying a plurality of records from each record container; dividing said records into groups, each group having the same geometrical pattern; the method further comprising semantically analyzing a representative from each said group; and wherein if the outcome of said semantic analyzing identifies relevant data, saving said data and said pattern in a data base.
3. The method of claim 2 wherein groups having identical said pattern in other pages are assumed to have the same semantic structure, such that data from said groups is fetched without further semantic analyzing.
4. The method of claim 1, wherein said searching for a plurality of record containers within said layout further comprises ranking the area size of the container and the closeness of the geometric center of the container to the

geometric center of the layout of document; and selecting a record container according to said ranking to form a selected record container, such that said determining said relevancy is performed on said selected record container.

5. The method of claim 4, wherein said determining said relevancy of said record comprises identifying a plurality of records within said selected record container; grouping said plurality of records into groups according to geometrical pattern, such that records having the same geometrical pattern are identified as belonging to the same group; performing semantic analysis on a representative record of each group; and if said representative record is relevant, storing data from said group of records.

6. The method of claim 5, wherein said grouping according to geometrical pattern is performed by identifying geometrical rectangles, or other geometrically defined shapes, within the record container and by ordering the rectangles or other geometrically defined shapes.

7. The method of claim 6, further comprising receiving a query from a user and comparing said query to a plurality of records; and ranking a plurality of records according to said geometrical pattern for said comparing said query.

8. The method of claim 7, further comprising ranking a plurality of records according to one or more of "freshness", ranking of the source website according to reliability and/or popularity, completeness of the record, or prominence of the record within the website.

9. The method of claim 7, further comprising ranking said plurality of records according to a plurality of weighted attributes.

10. The method of claim 7, further comprising dividing said plurality of records into a group of one or more relevant records and a group of one or more non-relevant records before said ranking said plurality of records, such that said ranking said plurality of records is performed only for said group of one or more relevant records, wherein said dividing said plurality of records comprises analyzing said user query to decompose said query to a plurality of items; analyzing each record to decompose said record to a plurality of items; and comparing values of said items for said user query and for said record.

11. The method of claim 10, wherein said comparing said query to a plurality of records further comprise representing each record and said query as a vector of variables, said variables having differential weighting; and comparing said vectors of variables to determine their similarity.
12. A method for geometrical analyzing of a page layout comprising database query results; the method comprising:
- a. Determining at least one record container within said layout by identifying said record container according to said layout;
 - b. If a plurality of record containers is determined, selecting a record container either by using the size relations of the layout records or by deducing the most regular area on a page;
 - c. Dividing the records within said record container into groups, each group having the same geometrical pattern; and
 - d. Analyzing the records according to semantic analysis, said semantic analysis comprising analyzing according to a plurality of keywords.
13. The method of claim 12 wherein rectangles within said chosen record container are identified.
14. The method of claim 13 wherein said identification is done by ordering said records inside said record container and by separating them, using line boundaries.
15. A system for automatic aggregating data from a plurality of web sites; comprising:
- a. A crawler process for fetching data from a provided list of related web sites;
 - b. A geometrical analyzer process for analyzing said data, said data comprising at least one document, said analyzing comprising geometrical analyzing of a page layout of the document, wherein said geometrical analyzing comprises determining one or more geometrical properties of the document; analyzing said one or more geometrical properties to detect a geometric pattern; searching for a plurality of record containers within said layout; and

determining a relevancy of a record from at least one record container according to said geometric pattern;

- c. A semantic layer for textually analyzing said relevant record; and
- d. A data base for storing the information retrieved by said semantic layer.