



- (51) International Patent Classification:
G06F 17/30 (2006.01) *G06F 19/10* (2011.01)
- (21) International Application Number:
PCT/TR2013/000321
- (22) International Filing Date:
14 October 2013 (14.10.2013)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:
2013/02437 28 February 2013 (28.02.2013) TR
- (72) Inventor; and
- (71) Applicant : VATA, Celal Korkut [TR/TR]; Sedat Simavi
Sokak 39/3, Ankara (TR).
- (74) Agent: DESTEK PATENT, INC; Konak Mah. Lefkose
Cad. NM Ofis Park, No: 36/5, 16110 Bursa (TR).
- (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).

Published:

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

(54) Title: COMBINATIONAL DATA MINING

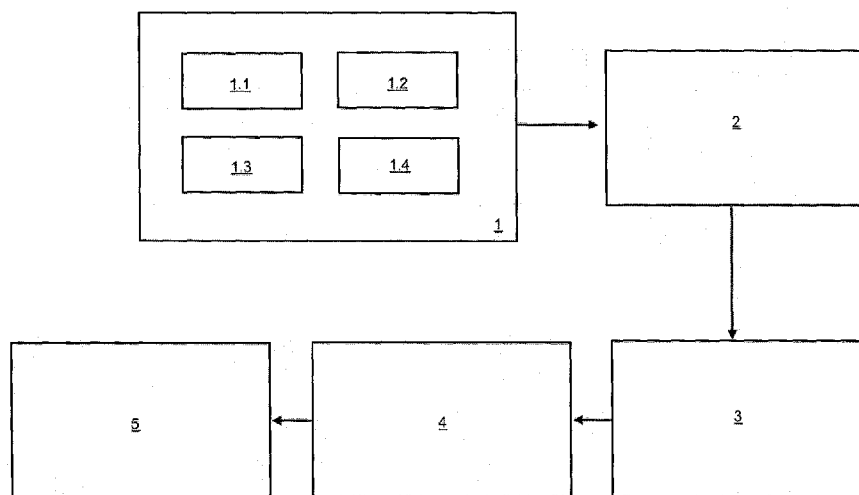


Figure 1

(57) Abstract: A combinatorial data mining system consisting of; a database selection unit (1.1) allowing the user to choose at least one database among others; a unit of term entrance (1.3) under the user choice unit (1) allowing the user to enter terms of interest in to different list (2); a unit of occurrence frequency determination retrieving the occurrence frequencies of the terms of interest separately and co-occurrence frequencies of the terms of different lists in a combinatorial fashion on the database; a unit of data normalization calculating the ratio of term co-occurrence statistics to the separately occurrence statistics using various formula; a data integration unit integrating the normalized numeric results on a matrix and; a data display unit (5) displaying the numerical results graphically in a color code to the user.



DESCRIPTION

COMBINATIONAL DATA MINING

5 TECHNICAL FIELD

The invention of interest is about a data mining system and a data mining method allowing the user to search on a database of interest with the potential of displaying the most relevant and meaningful results of the search terms to the end-user.

10

Prior Art

A classical data mining approach consists of the steps of data cleaning, data integration and data display. International patent applications WO 2001/037072 and WO 2002/005209 are exemplar prior art referring to the steps of data cleaning and data integration steps of data mining.

15

There are efficient methods of data integration and data display. However, the step of background elimination (data cleaning) is usually problematic. The problems can be summarized as the following; same terms in different languages referring to the same concepts are represented by different numerical occurrence statistics across the databases. Therefore, the language barrier can not be overcome. For example, particular investments of a Turkish drug company with a Turkish name can not be effectively searched against the investments of an American drug company with an English name. The second problem is the existence of different terms only as statistical figures with differences of orders of magnitude on huge databases. These statistical figures are mainly raw data and not processed information. For example, a search about the city of Istanbul can not be directly compared with a search on the city of Muş as the city of Istanbul is at least two orders of magnitude more frequently represented than the city of Muş on any public database. The third problem is that the classical data mining systems do not allow the user to search for specific information in a combinatorial fashion.

20

25

30

Although, the steps of data integration and data display of today are quite efficient the inefficiency of the background elimination is the biggest problem of the field. The invention of interest is mainly a system of data normalization before data integration and data display.

35

Therefore there is a great need for an advancement in the technical field to solve the problems mentioned above.

For example, when a user specifically searches for the binary term "data mining" the presence of terms "data" and "mining" separately on the database is the background. The invention of interest efficiently eliminates this problem.

SHORT DESCRIPTION OF THE INVENTION

The invention of interest is aiming to eliminate the problems mentioned above and to potentiate the current data mining technology of today.

The particular work of interest is aiming to eliminate the problem of background information of data mining and to allow the user to retrieve meaningful results regarding the topic of interest.

Another aim of the invention is to allow the user to enter lists of keywords in double or triple combinations.

Another aim of the invention is to allow the user to select among different databases for a combinatorial search of interest.

Another aim of the invention is to display the results of the combinatorial search in a graphical format to the end-user.

Another aim of the invention is to allow the user to compare different search results on different databases with each other to delineate database specific responses.

Another aim of the invention is to allow the user to use terms of different languages on the same platform in a combinatorial fashion.

As mentioned above and further described below the invention of interest is about a combinatorial data mining system with the following specifications;

- A unit for at least one database selection and a unit of keyword lists allowing the user to enter keywords of interest in a combinatorial fashion in different lists,
- A unit of co-occurrence frequency retrieval wherein the unit extracts the co-occurrence and separately occurrence statistics of the terms of interest in a combinatorial fashion from the databases,
- A unit of normalization wherein the ratio of co-occurrence statistics of the terms to the separately occurrence statistics are calculated using various formula,
- A unit of data integration where the normalized data is integrated on a matrix,
- A unit of data display where the data is displayed to the end-user in a graphical format,

The combinatorial data mining system functions on the following bases:

- At least one database is chosen by the user,
- The terms of interest are entered by the user in at least two lists with respect to the order of interest,
- Determination of co-occurrence as well as separately occurrence frequencies for the terms of different lists in a combinatorial fashion,
- Data normalization via ratio calculation of the co-occurrence statistics to the separately occurrence statistics using different ratio formula,
- Background elimination according to the normalization step,
- Graphical display of the results to the end-user,

The invention of interest should be considered along with the items and drawings as below to shed light on the relevant advantages.

BRIEF DESCRIPTIONS OF THE DRAWINGS

Figure 1 is a schematically display of the combinatorial data mining.

REFERENCE NUMBERS

1 User Choice Unit

1.1 Unit of Database Selection

1.2 Unit of Criteria Determination

1.3 Unit of Database Selection

2 Unit of Term Frequency Determination

3 Unit of Data Normalization

4 Unit of Data Integration

5 Unit of Graphical Data display

5

With the option to chose a sub-database under the main database the user can specifically direct his/her search to the database of interest. Furthermore, using the criteria determination unit (1.2) the user can determine whether the terms of interest should be next to each other strictly or else the terms should only be on the same document.

10

The invention of interest allows the user to search for symptoms and diseases and to read and interpret the results in the following fashion:

- The selection of the main database,
- Entrance of the disease and symptom terms into list 1 and list 2 as below using the term entrance unit (1.3),

15

List 1- Name of the Disease	List 2- Symptom
Alzheimer's Disease	Loss of Sleep
Delusional Disorder	Open Eyelids
Bipolar Manic Depressive Disorder	Agitation
	Shaky Hands

- Determination of the occurrence frequencies of terms in the list 1 and list 2 separately on the database,
- Determination of the co-occurrence frequencies of terms in the list 1 and terms in list 2 in a combinatorial fashion,
- Ratio normalization of the term frequencies of list 1 and list 2 in a combinatorial fashion,
- Background elimination with respect to results of the normalization,
- Integration of the cleaned data on a matrix and displaying to the end-user using the color code as below,

20

25

30

	Alzheimer's Disease			
10	Delusional Disorder			
	Bipolar Manic Depressive Disorder			
		Loss of Sleep	Open Eyelids	Agitation
20				Shaky Hands

The matrix displays the relevance of diseases and symptoms using a color code. The relative color intensity reveals the relative correlation of the symptoms to the diseases allowing the user to interpret the results. As seen on the matrix the square of manic depression and agitation is marked with a higher color intensity than that of the square of Alzheimer's disease and agitation. Similarly, the square referring to loss of sleep symptom and Alzheimer's disease is with a higher color intensity than that of bipolar depressive disorder and loss of sleep. Based on these results the user can confidently conclude that loss of sleep is a major symptom of Alzheimer's disease and agitation is a major symptom of bipolar disorder.

The color intensities are a direct function of the numeric results of the normalization procedure.

The invention of interest allows the user to enter terms of different languages into the same list. For example, "Glaxo Smith Klein" the English term, "Sandoz" the German term, "Sanofi" the French term, "Daiichi Sankyo" the Japanese term and the "Abdi Ibrahim" the Turkish term can be entered in to the same list, list one. The terms of chollesterol lowering drugs "Atorvastatin", "Cericastatin", "Fluvastatin" and "Lovastatin" can be entered into the other list, list 2. The results will show the user which company has invested into which drug extensively. The ratio calculation based background elimination allows the user to exclude all the language specific backgrounds for terms internationally. Therefore, the user is able to extract meaning regarding terms in different languages based on the numeric value of the term frequencies of different languages.

Similarly, the Turkish Term of "veri madenciliği", the French Term of "l'exploration de donn'ees", the English term of "data mining" and the Japanese term of "データマイニング" can be entered into the same list. If the English term "data mining" reveals a higher numeric value than the Turkish term "Veri madenciliği" the user can confidently conclude that the concept of data mining is more common in English speaking countries. Therefore, the system has a capacity to dissect the culture specific details in different languages.

10

15

20

25

30

35

CLAIMS

1. A combinatorial data mining system characterized in comprising;

- A unit of database selection allowing the user to choose a database of interest among others (1.1) and a unit of term entrance (1.3) allowing the user to enter terms into different lists on the user selection unit (1),
- A unit of term frequency determination (2) retrieving the database term frequencies separately as well as co-occurrence frequencies of different lists combinatorially,
- A unit of data normalization (3), where the unit calculates the ratio of the co-occurrence statistics to the occurrence statistics of the separation
- A unit of data integration (4), wherein the system integrates the normalized data on a matrix,
- A unit of graphical data display (5), wherein the system displays integrated data graphically to the user,

2. The combinatorial data mining system according to claim 1, wherein the unit allows the user to choose the function of normalization (1.4).

3. The combinatorial data mining system according to claim 1, wherein the unit of criteria determination (1.2) of the user option unit (1) allows the user to determine on the option of term co-occurrence next to each other or on the option of term co-occurrence on the same document.

4. A combinatorial data mining method with the following specifications;

- The user chooses at least database among others,
- The user enters terms of interest in at least two lists in to the system with respect to the order of interest,
- The step of the determination of the co-occurrence statistics of one term of interest with another term of interest on the other list, for each term combination on a row,
- The step of normalization, wherein the statistics of term co-occurrences are ratio normalized to the separately occurrence statistics,
- The step of background elimination with respect to normalization,
- The step of data display in a graphical format,

5. The combinatorial data mining method according to claim 4, wherein the criteria determination unit (1.2) allows the user to choose between the options of term occurrence next to each other and the option of term occurrence separately on the same document.
6. The combinatorial data mining method according to claim 4, wherein the speed of data retrieval is determined by the user via criteria determination unit (1.2).
7. The combinatorial data mining method according to claim 4, wherein the normalization results of numeric values are indicated in quantitative color intensities on a matrix.
8. The combinatorial data mining method according to claim 4, wherein two numeric values of term occurrences and a single value of term co-occurrence are used in the three value normalization-ratio formula.
9. The combinatorial data mining method according to claim 7, wherein three different numerical values are used in different weighted ratio formula.

1/1

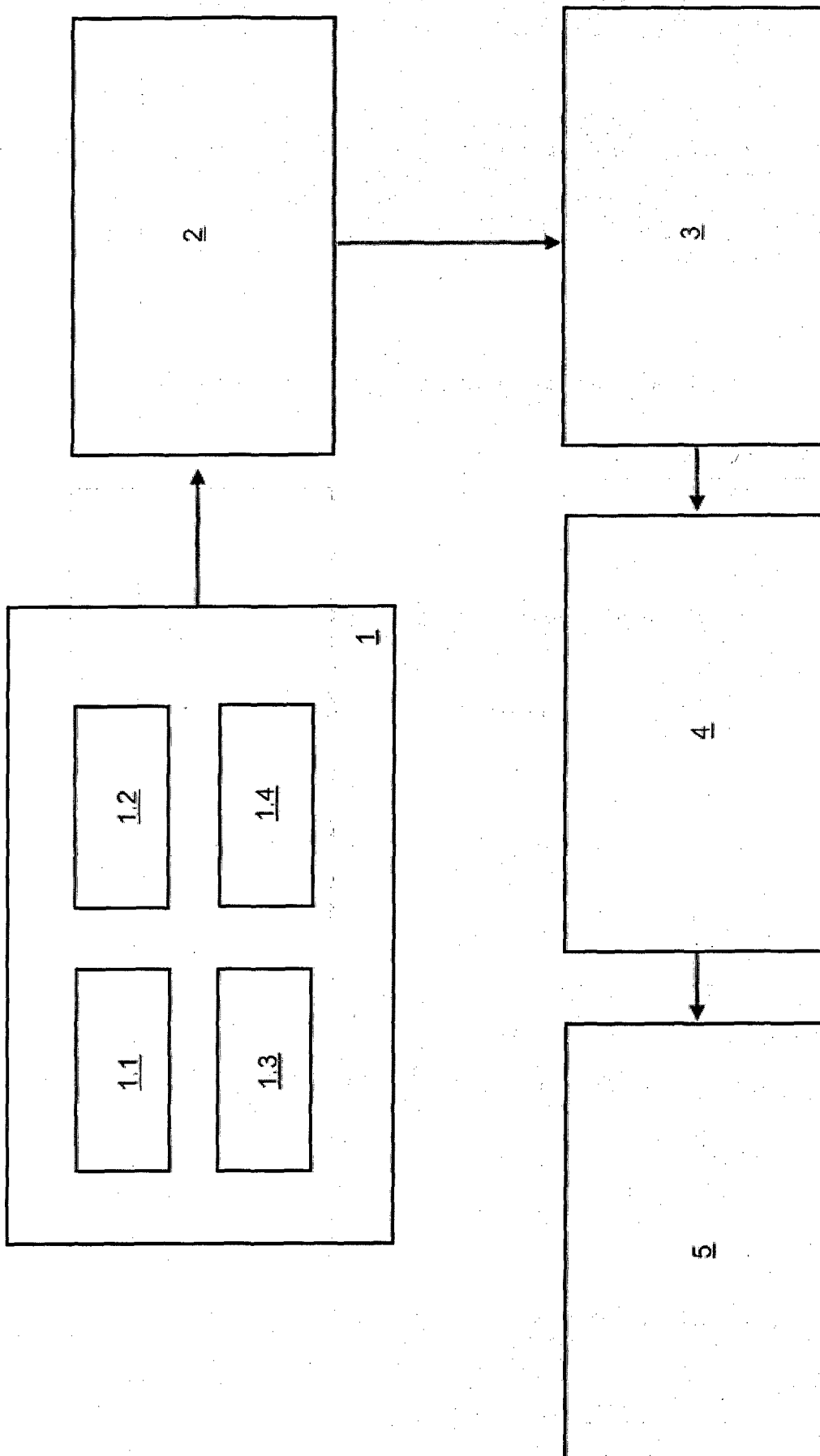


Figure 1

INTERNATIONAL SEARCH REPORT

International application No
PCT/TR2013/000321A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F17/30 G06F19/10
ADD.

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

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

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

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	D Stott Parker ET AL: "Literature Mapping with PubAtlas extending PubMed with a 'BLASTing interface'", 2009 AMIA Summit on Translational Bioinformatics, San Francisco, California, 15 March 2009 (2009-03-15), XP055107246, Retrieved from the Internet: URL:http://summit2009.amia.org/files/symposium2008/S16-Parker.pdf [retrieved on 2014-03-12] the whole document ----- -/--	1-9



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

13 March 2014

Date of mailing of the international search report

20/03/2014

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Nazzaro, Antonio

INTERNATIONAL SEARCH REPORT

International application No

PCT/TR2013/000321

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	BECKER KEVIN G ET AL: "PubMatrix: a tool for multiplex literature mining", BMC BIOINFORMATICS, BIOMED CENTRAL, LONDON, GB, vol. 4, no. 1, 10 December 2003 (2003-12-10), page 61, XP021000471, ISSN: 1471-2105, DOI: 10.1186/1471-2105-4-61 the whole document	1-9
X	FRANCO CAUDA ET AL: "Shared "Core" Areas between the Pain and Other Task-Related Networks", PLOS ONE, vol. 7, no. 8, 10 August 2012 (2012-08-10), page e41929, XP055107250, DOI: 10.1371/journal.pone.0041929 page 2 - page 4; figures 1,3	1-9
X	D S Parker ET AL: "Literature Mapping with PubAtlas - extending PubMed with a 'BLASTing interface' * Consortium for Neuropsychiatric Phenomics, UCLA Finding Associations in PubMed", 1 March 2009 (2009-03-01), XP055107192, Retrieved from the Internet: URL: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041555/pdf/amia-s2009-90.pdf [retrieved on 2014-03-12] the whole document	1-9