Message sequence analysis

Patent LT6685

The invention subject is the technology for identifying statistical links in the sequence of news items, adverts, or other messages. Incoming messages are classified according to several attributes. Selective reclassification is used to account for different trait assessment interpretations. The messages converted into code form an estimator matrix. To detect a pattern in a message sequence on a timescale, it is necessary to compare matrix fragments which follow either before or after messages with the same assessment value according to one or more traits. The correlation dependence with the same data filter on the superimposed time segments is assessed. If the correlation dependence for two or more matrix fragments is high, the data filter becomes narrower. Data on settings and search results are stored in the database as a pattern. The examples discovered are assessed by a person for significance. A new or repeated pattern search starts with settings combining two or more known patterns with similar message codes. The patterns with high significance assessment are more often used to create combined search settings. The data filter is additionally extended using random values. Figuratively speaking, the pattern search criteria evolve by crossing, mutation, and selection. The analysis predictive power is expressed in the assessment of probability with which the new or probable message fits into the previously identified pattern. The past message sequence examples show what typically happens under similar circumstances.

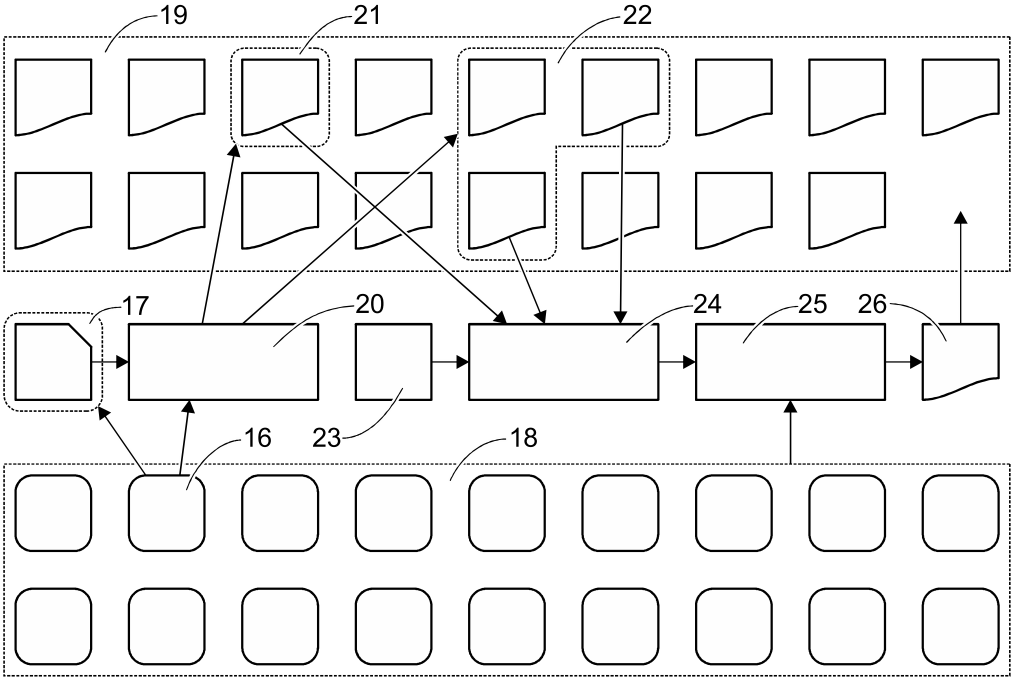


Fig. 11

Field of the invention

The invention is relevant to data processing systems and methods specifically designed for managerial, supervisory, and predictive purposes. The invention is applicable in special sections of business, government, and public services.

Background of the invention

Patent US10181167 describes a method for predicting a politician's behavior, based on unrelated historical data. This example shows an obvious connection between the subject and the circumstances. The circumstances imply the flow of news messages with specific data filtering. The search for interrelations among messages with an indefinite selection of subjects is not applicable.

Description of the invention

The subject matter of this invention is a method to identify statistical relationships in a sequence of news, advertisement, or other messages. Automated analysis of historical and collective experience complements human cognitive abilities with digital intuition.

The analysis is conducted using a computer. The hardware components of a computing system include at least one memory module; at least one processor; at least one data entry interface; at least one data display or transmission device. All data mentioned in this description is stored in the memory module. All computing operations are carried out by the processor.

Incoming messages are classified by several criteria. Data on a message and message occurrence circumstances are stored as formal feature assessments. At the initial stage, a person conducts a formal assessment. A self-learning system uses the accumulated material to automate the classification process. In content, the classifier is similar to the International Patent Classification.

The classifier may be interpreted inaccurately. For the algorithm, it is not consensus in assessment definition that matters, but the consistency of assessment for different messages. Selective reclassification is utilized to take different interpretations of feature assessment into account. When searching for similar messages, the difference in assessments is taken into account. The increase or decrease in differences in people's judgments singles out the probable points of conflict occurrence and resolution.

Messages, converted to code, create an assessment matrix. Such matrices allow building the following diagrams on a time scale: the amount of attention attracted; data density; data volume; the rate of data volume change; diagrams of other derived functions. When plotting diagrams, assessment filters are utilized to single out a targeted data combination. The filters imply exclusion, conjunction, disjunction, inversion, and the use of coefficients. The number of message views or another significance value may serve as a coefficient. The average correlation values of the diagrams with different filters are calculated. Local deviations from the average value show the changes peculiar to specific phenomena. For example, deviations peculiar to holidays or wartime are revealed.

To identify a recurrent pattern among several messages in the matrix, similarities to a specific message are searched for. A similar message is characterized by the same assessment values according to one or more features. Matrix fragments that follow before or after similar messages are superimposed and compared on the timeline. The correlation dependence is assessed, using the same data filter on the time spans compared. If the correlation value is low, the matrix fragments are compared, using a modified data filter. If the correlation dependence is high for two or more matrix fragments, the interconnections are specified in the examples. The length of the matrix fragments compared is selected to find the maximum number of correlation dependence examples. Message features or assessments are excluded from the data filter one by one or inverted to identify the maximum correlation dependence value. The data filter becomes narrower, and the codes of those messages showing a high correlation of message sequences are identified. The data on search settings and message sequence comparison examples are recorded to the database as a “pattern”.

Two or more patterns with similar message codes are used to combine search settings. This results in a wider data filter. The data filter is further expanded by including randomly selected features or assessment values in it. This wider data filter is used to search for new patterns.

The analysis may be conducted for a planned or presumptive message. To show a user what most often happens under similar circumstances, examples of past message sequences are created. Interesting examples attract people's attention measured by the number of views and likes. People's attention becomes feedback for a self-learning algorithm that automatically searches for heuristic combinations.

Description of the drawings

Fig. 1 depicts message classifier, index E. Marked positions: 1 – message abbreviation; 2 – feature assessment formal definition.

Fig. 2 depicts message classifier, index F.

Fig. 3 depicts message classifier, index G.

Fig. 4 depicts message classifier, index H.

Fig. 5 depicts message classifier, index I.

Fig. 6 depicts message classifier, index L.

Fig. 7 depicts message classifier, indexes J, K.

Fig. 8 depicts a message-to-code conversion example. Marked positions: 3 – information on a message; 4 – headline; 5 – a format for writing code, using alphabetic indexes; 6 – message code.

Fig. 9 depicts an example of an assessment matrix. Marked positions: 7 – feature indexes; 8 – message order number; 9 – data classified.

Fig. 10 depicts an example of matrix fragments comparison. Marked positions: 10 – time axis; 11 – message codes; 12 – codes for similar messages; 13 – matrix fragments compared; 14 – an assessment of the correlation among data sequences; 15 – examples with a higher correlation value.

Fig. 11 depicts message sequence pattern identification procedure. Marked positions: 16 – considered message code; 17 – passport number of a person who classified the message; 18 – a matrix with message codes; 19 – database “patterns”; 20 – search for patterns with similar codes; 21 – patterns, the data filter of which is excluded from the settings profile; 22 – patterns, the data filter if which is added to the settings profile; 23 – a randomly selected assessment value or feature; 24 – search settings combination; 25 – patterns search; 26 – data on the pattern identified.

Fig. 12 depicts an example of data on a pattern identified. Marked positions: 27 – a similar message filter; 28 – the number of similar codes at the moment of analysis procedure; 29 – a data filter at the beginning of the analysis procedure; 30 – the length of the matrix fragments compared; 31 – a data filter at the end of the analysis procedure; 32 – the number of pattern examples identified; 33 – numbers of similar messages; 34 – numbers of statistically related messages; 35 – a pattern name; 36 – a significance value.

Fig. 13 depicts an example of a comparison of data volume diagrams. Marked positions: 37 – a time scale; 38 – the number of similar messages; 39 – the number of pattern examples; 40 – diagram correlation coefficient.

Below is an example of an analysis of Russian-language news messages delivered in the first decades of the 21st century. It is assumed that Russia was the source of a large number of misleading messages. Consequently, the patterns identified may stir heightened interest.

Four most noticeable messages are selected from a daily news archive. A person converts event news messages into code (Fig. 8) according to the classifier (Fig. 1-7). The message classifier contains a set of features marked by alphabetic indexes:

A – message date;

B – message position in the daily news list;

C – headline;

D – link to the source;

E – message subject;

F – how the event is described or how it is perceived;

G – at what execution stage the event is;

H – time span between the event and the message about it;

I – data source type;

J – event or consequences location;

K – event cause occurrence location or subject;

L – truth assessment according to other sources;

M – reference to the message number about the same event and a hint for an analyst;

N – message circumstances;

O – calculated message characteristics;

P – passport number of a person who conducted the classification.

The number of combinations based on features E*F*G*H*I*J*K*L is 1,610,612,736 variants.

Data from other sources clarifies the message circumstances (index N). In this example, the oil price is recorded as a circumstance.

Information about a message (index O) is updated each time the algorithm processes the message data. The list of calculated message characteristics: the number of readings of this code throughout the entire matrix history; a list of database patterns referring to this code; significance value of patterns referring to this code. The list of characteristics considered can be expanded.

The message code (6) may contain an error, made by an analyst, or may differ due to ambiguous interpretation of the feature. The messages recorded in the matrix (9, 11) are selectively re-classified by the same or another person. If the new code differs from the previous one, both codes are recorded in the person's passport (17) as a “paradox”. The passport contains the following data: a person identifier; the number of messages classified; the amount of time spent on work; paradoxes; paradox discovery dates. If paradoxes appear frequently, another person checks the work done by the previous person. The list of paradoxes provides different feature assessment options, which are taken into account when searching for codes of similar messages (12). When paradoxes are taken into account, it allows keeping the classifier relatively simple and employing many analysts.

To search for patterns, the settings profile is used (24). The settings profile contains the following data: the length of the matrix fragments compared (30); search direction — before or after a message; a data filter (29). A person creates the first settings profiles.

Pattern discovery procedure:

1) Search for similar codes in the matrix (12). A similar message code contains the same assessment according to one or more features (27).

2) Matrix fragments (13), coming before or after similar codes, are superimposed on the time scale (10).

3) Calculation of the compared sequences correlation (14). A single data filter is used for all matrix fragments (13).

4) Selecting examples with the highest correlation coefficient (15).

5) To identify message codes (34) showing the highest correlation coefficient (15), the criteria in the settings profile are gradually narrowed down. Message assessments are alternately inverted or excluded from the transmission data filter (27), and the compared matrix fragments are shortened (30).

6) The pattern identification procedure with a modified filter is repeated many times while the correlation coefficient increases (15), and the number of examples does not decreases (32-34).

Data on identified pattern examples (Fig. 12, Fig. 13) is published. A person assesses the significance of the pattern identified (36). The number of views, the number of comments, or another indicator of interest is assessed. An analyst can assess significance according to a ten-point scale with grades from “not interesting” to “interesting” (36). A person can name a pattern (35) and classify it as a group of specific phenomena. When published, patterns are sorted by those features, which may cause heightened interest. In particular, those examples are highlighted, which contain messages with the grade "refuted" (Fig. 7). Features of heightened interest are specified based on significance value statistics (36). Patterns with high significance values and with high performance (22) are used for analysis more often. Pattern performance is assessed by the number of examples identified (32), matrix fragments length (30), and a correlation dependence rate (15, 40).

A narrow data filter is common (30, 31) for the pattern saved (Fig. 12). The combination of settings comprised of several patterns (24) returns a wider data filter (27, 29) for a new pattern search procedure (25). The procedure for settings profile creation:

1) The code of the message being analyzed (16) is compared with paradoxes recorded (17). If there is more than one feature assessment value, all assessment values are used to search for patterns with similar messages (20).

2) Search for patterns (21, 22) that contain a data filter (27, 31) that lets through the message analyzed. If the pattern search is repeated for the same message, the previously identified pattern is not taken into account.

3) The combination of data filters (30, 31), using two or more patterns (22). The number of filters combined is limited. Patterns with high significance values (36) are used more often (22). This approach increases the likelihood of identifying a more complex example of a statistical relationship, which may prove interesting for people (26).

4) A data filter (31) of one or more patterns (21) with low significance assessment values (36) is excluded from the settings profile. This approach reduces the likelihood of identifying an example of a statistical relationship uninteresting for people (26).

5) The length (30) of the matrix fragments compared (13) extends to the time span between the event and the news message about it (indexes H, M). This operation can be carried out selectively. The choice may be random.

6) Randomly selected assessment values or features (23) are used to expand a data filter (29).

Figuratively speaking, pattern search criteria evolve through crossing (22), mutation (23), and selection (20). Each time, the matrix (18) is analyzed (25) using modified filters (24). Message analysis can be conducted many times.

The sequence of computing operations, data priority, and threshold values are changed. The computing operations, required to identify pattern examples, are counted. Process optimization is conducted using the most productive settings. For example, the number of combined data filters is specified (22). A data correlation coefficient (15), at which a connection is considered established, is an example of a threshold value. The balance between random values (23) and values from stored patterns (22) is an example of data priority. The sequence and priority of computing operations are changed by a person or are automatically made dependent upon the parameters calculated.

The invention is developed as a universal software product. The classifier and data entry methods are different for different tasks.

The first example of use: the invention may be used by news message rating agencies. A rating is an assessment of the likelihood that a particular message sequence is repeated or may be repeated.

The second example of use: user data and advertising content relationship reconstruction. The matrix contains data on an advertising product consumer and data on the targeted advertising content. Data is collected from many users. Patterns are searched in the matrices of all users. Patterns identified provide users with an explanation of why they get this or that advert.

The third example of use: the study of loyalty conditions. For example, an identified pattern shows how people, to whom special conditions apply, differ from each other. Messages on many users' profiles and their servicing conditions are categorized.

The fourth example of use: message source features assessment. For example, the archive of many users' correspondence is analyzed. Patterns with positive or negative feedback are identified. The user gets a probabilistic assessment of the dialogue partner's ingenuity, sincerity, and commitment.

The fifth example of use: an analyst's features assessment. For example, a person or artificial intelligence is assigned with a task to categorize messages. Different people categorize the same message. One person's assessments make up a fixed-length matrix fragment on a conventional time scale. Patterns identified show groups of people with different views.

Claims

1. A message sequence analysis, designed to search for patterns, using formal assessment of features and conditions, which a value matrix, recorded in one or more memory blocks, is compiled of, is *characterized in that* a processor identifies similarities in the matrix for each message based on one or more features; superimposes matrix fragments that follow before or after similar messages on the timeline; the same data filter is applied to all matrix fragments; when a high correlation among the compared matrix fragments is identified, the comparison data, called a pattern, is recorded in a memory block.

2. A message sequence analysis as defined in claim 1, *characterized in that* the event message classification is carried out taking into account the following features: message date and time; assessment of attention attracted; message subject; event or consequence location ; event cause location or a subject; the emotion the event is described with; what stage of execution the event is at; the time span between the event and the message about it; type of data source; assessment of truth according to other sources; event circumstances according to other data sources; relation to other messages about the same event.

3. A message sequence analysis as defined in claim 1, *characterized in that* the message is re-classified; dissimilar assessments of the same message are recorded to a memory block as a paradox; the processor searches for similar messages, using paradox entries.

4. A message sequence analysis as defined in claim 1, *characterized in that* the processor calculates the correlation among the compared matrix fragments on a time scale, using the following diagrams: a diagram of the attention attracted; a data volume diagram; a data amount change rate diagram; a diagram of other derived functions.

5. A message sequence analysis as defined in claim 1, *characterized in that* the processor utilizes the following data filters: exclusion, conjunction, disjunction, inversion, and a coefficient.

6. A message sequence analysis as defined in claim 1, *characterized in that* the length of the matrix fragments compared includes the time span between the event and the message about it.

7. A message sequence analysis as defined in claim 1, *characterized in that* the processor shortens the matrix fragments compared until the correlation value increases.

8. A message sequence analysis as defined in claim 1, *characterized in that* the processor randomly removes message assessments from the data filter or inverts them, and then it records the pattern with the highest correlation coefficient to a memory block.

9. A message sequence analysis as defined in claim 1, *characterized in that* the processor generates a wider data filter, using two or more patterns recorded in a memory block.

10. A message sequence analysis as defined in claim 1 and claim 9, *characterized in that* human assessment of pattern significance is recorded in a memory block; the processor more often uses patterns with a high significance value to extend the data filter.

11. A message sequence analysis as defined in claim 1, *characterized in that* the processor extends the data filter by including randomly selected assessment values or features into it.

12. A message sequence analysis as defined in claim 1, *characterized in that* the processor changes the sequence of computational operations, data priority, and threshold values, calculates the number of operations required to identify the pattern and records the value in a memory block.

13. A message sequence analysis as defined in claim 1, *characterized in that* the search for patterns is carried out for a planned or fictional message.

E – Phenomenon

oth	Other
fue	Fuel, energy
fos	Fossils
tra	Transport
pur	Food, purveyance
med	Health, medicine
ecl	Ecology
ecn	Economy
bld	Building, infrastructure
tec	Technics and techology
sci	Science, knowledge acquistion
eso	Esoterica, mythology
rel	Religion, mores
edu	Education
spo	Competition, sport
cul	Culture, game, spectacle
inf	Spread of information
dom	Domestic incident
tch	Technogenic
fir	Fire
ear	Earthquake, volcano
flo	Flood
atm	Atmospheric
spa	Space
abn	Abnormal
gov	Government policy
lib	Liberals
soc	Socialists
nac	Nationalists
isl	Islamists
chr	Christians
ele	Elections, appointment
bur	Bureaucracy, regulation
cor	Corruption
ref	Reform
deb	Debt
dip	Dispute
jus	Justice
vio	Violence
pro	Protest
mig	Migration
pol	Policing
cri	Crime
spy	Secret service
mte	Military technology
mtr	Military training
ter	Terror
war	War or incident

1 2

Fig. 1

F – Perception

cel	Celebration
ach	Achievement, find
dem	Demonstration of opportunities
hlp	Giving help
con	Concession
pla	Proposal, plan
ent	Entertainment
was	Story as it was
wil	Notification how it will be
sus	Suspicion
ren	Justification, refutation
req	Prosecution, requirement
per	Persecution
los	Loss, sadness
def	Defeat, mourning
cat	Catastrophe

Fig. 2

G – The state of affairs

wng	A warning
ini	Initiative
prp	Preparation
obs	Observation
act	Action
fin	The final
gen	Generalizing
rem	Reminder

Fig. 3

H – Time gap

yea	Antiquity - years
mon	Up to 13 months before
wek	Last week event
day	Last day event
exp	The expected event happened
pda	Prediction for the coming days
pmo	Prediction for months
pye	Prediction for years

Fig. 4

I – Source

myt	Mythical
ano	Anonymous
opi	Opinion
wit	Witnesses
sta	Statistics
ofi	State services
pob	Professional observers
sev	Several sources

Fig. 5

J – Where are the consequences **K** – Whence the reason

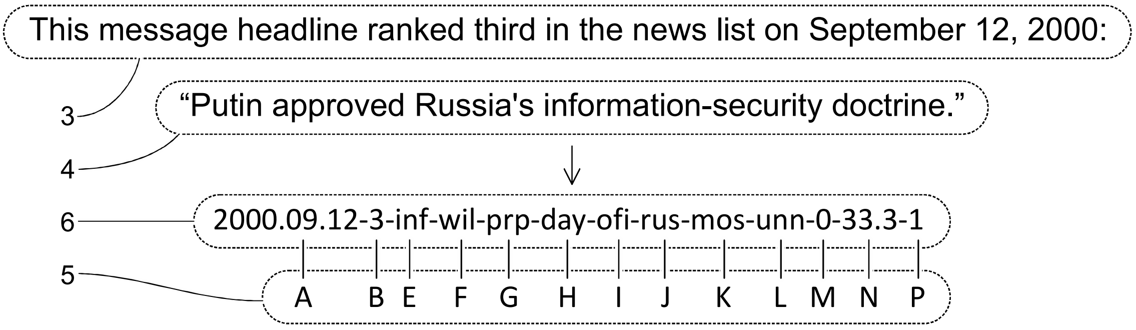
mos	Moscow
che	Chechnya, Dagestan, Ingushetia, Kabardino-Balkaria
rus	Russia
ukr	Ukraine
bel	Belarus
eae	Eastern Europe: Bulgaria, Hungary, Moldova, Poland, Romania, Slovakia, Czech Rep.
bal	Baltic states: Lithuania, Latvia, Estonia
grb	Great Britain
noe	Northern Europe: Norway, Finland, Sweden, Ireland, Iceland, Denmark
wee	Western Europe: Austria, Belgium, Germany, Liechtenstein, Luxembourg, the Netherlands, France, Switzerland
soe	Southern Europe: Albania, Andorra, Bosnia and Herzegovina, Vatican, Greece, Spain, Italy, Cyprus, Macedonia, Malta, Monaco, Portugal, San Marino, Serbia, Slovenia, Croatia, Montenegro, Yugoslavia.
tur	Turkey
afr	Africa
isr	Israel
mie	Middle East: Libya, Egypt, Cyprus, Iran, Iraq, Syria, Palestine, Saudi Arabia, Arab Emirates, Jordan, Kuwait, Lebanon, Oman, Qatar, Bahrain, Yemen.
cau	Caucasus: Georgia, Armenia, Azerbaijan, Abkhazia, Ossetia.
soa	South Asia: India, Pakistan, Afghanistan, Bangladesh, Bhutan, Nepal, Maldives, Sri Lanka.
cea	Central Asia: Mongolia, Kazakhstan, Kyrgyzstan, Uzbekistan, Turkmenistan, Tajikistan, Afghanistan
chi	China PRC, Taiwan
kor	South Korea, DPRK
jap	Japan
sea	Southeast Asia: Indonesia, Philippines, Singapore, Vietnam, Cambodia, Laos, Myanmar, Thailand, Malaysia, Brunei, East Timor
aus	Australia, New Zealand, Papua, Solomon Islands, Fiji, Vanuatu, Samoa
ant	Antarctica
sam	South America: Brazil, Argentina, Chile, Bolivia, Paraguay, Uruguay, Peru, Ecuador, Guyana, Guyana, Suriname, Falkland, South Georgia
car	Caribbean: Mexico, Colombia, Venezuela, Cuba, Guatemala, Dominican Republic, Haiti, Honduras, Nicaragua, Costa Rica, Panama, Jamaica, Puerto Rico, Trinidad and Tobago, Guadeloupe, Belize, Barbados, St. Vincent and the Grenadines Virgin is., Grenada, Cayman, Saint Kitts, Aruba, Anguilla, St. Maarten, Sint Maarten.
usa	USA
arc	Arctic, Canada, Greenland
coa	Coalition
reg	International Regulator: UN, WHO, WTO, OPEC, IMF, IAEA…
glo	Globally
nlo	No location

Fig. 6

L – Verity

unn	Consent or unknown
sov	Suspicious overlay
suc	Suspicious in content
dis	Disproved

Fig. 7



This message headline ranked third in the news list on September 12, 2000:

"Putin approved Russia's information-security doctrine."

2000.09.12-3-inf-wil-prp-day-ofi-rus-mos-unn-0-33.3-1

Fig. 8

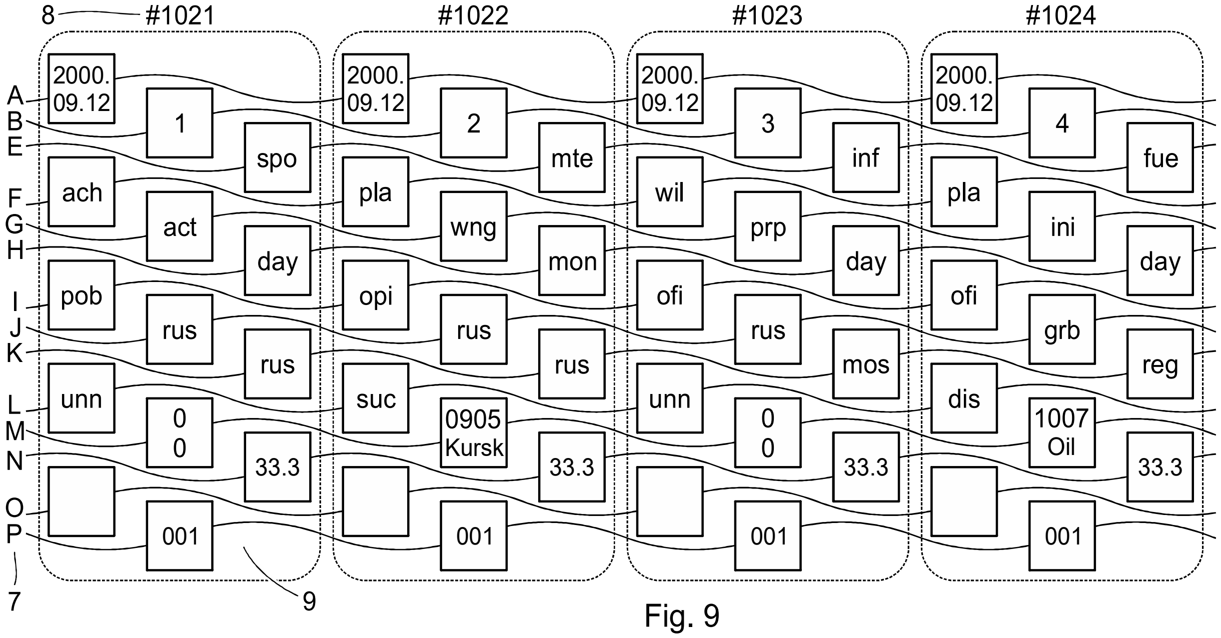
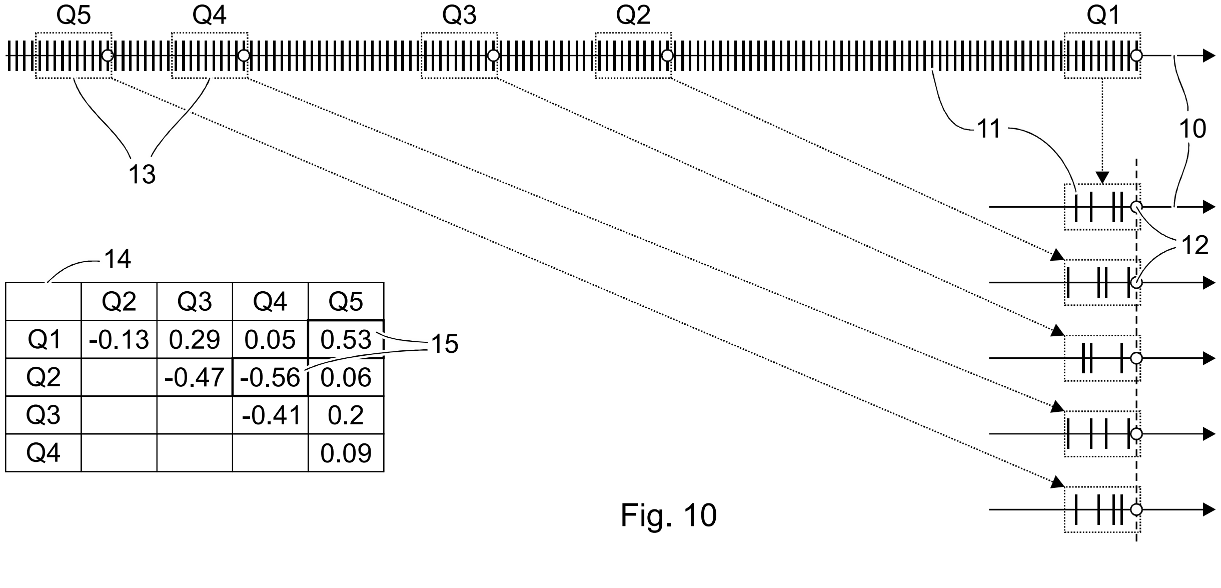


Fig. 9



	Q2	Q3	Q4	Q5
Q1	-0.13	0.29	0.05	0.53
Q2		-0.47	-0.56	0.06
Q3			-0.41	0.2
Q4				0.09

Fig. 10

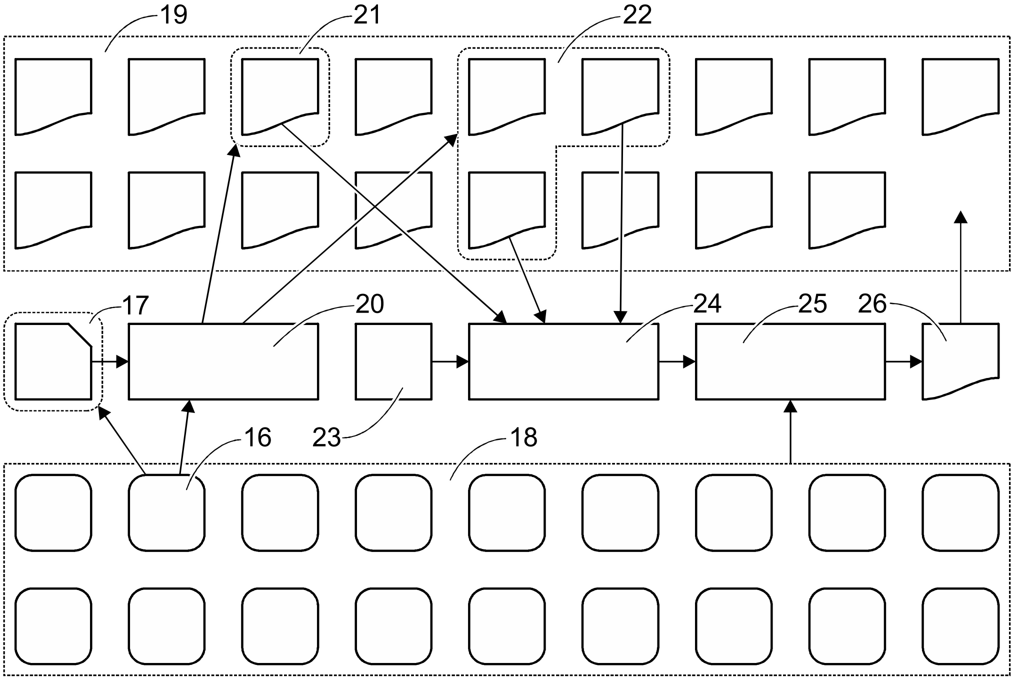
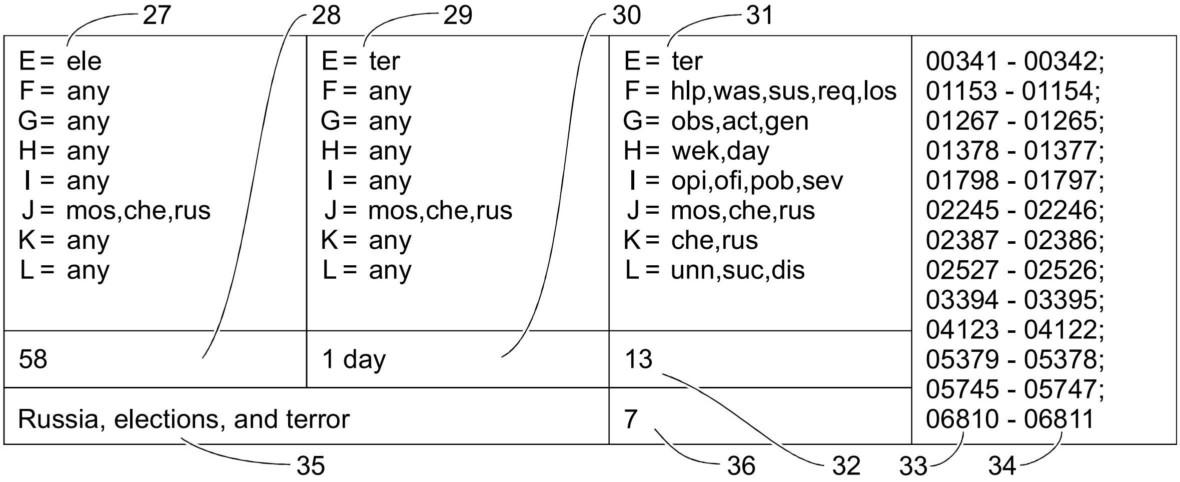


Fig. 11


27
28
29
30
31
35
36
32
33
34

E= ele
F= any
G= any
H= any
I = any
J = mos,che,rus
K= any
L = any
	E= ter
	F= any
	G= any
	H= any
	I = any
	J = mos,che,rus
	K= any
	L = any
		E= ter
		F = hlp,was,sus,req,los
		G= obs,act,gen
		H= wek,day
		I = opi,ofi,pob,sev
		J = mos,che,rus
		K= che,rus
		L = unn,suc,dis
			00341 - 00342;
			01153 - 01154;
			01267 - 01265;
			01378 - 01377;
			01798 - 01797;
			02245 - 02246;
			02387 - 02386;
			02527 - 02526;
			03394 - 03395;
			04123 - 04122;
			05379 - 05378;
			05745 - 05747;
			06810 - 06811

58	1 day	13	
Russia, elections, and terror		7	

Fig. 12

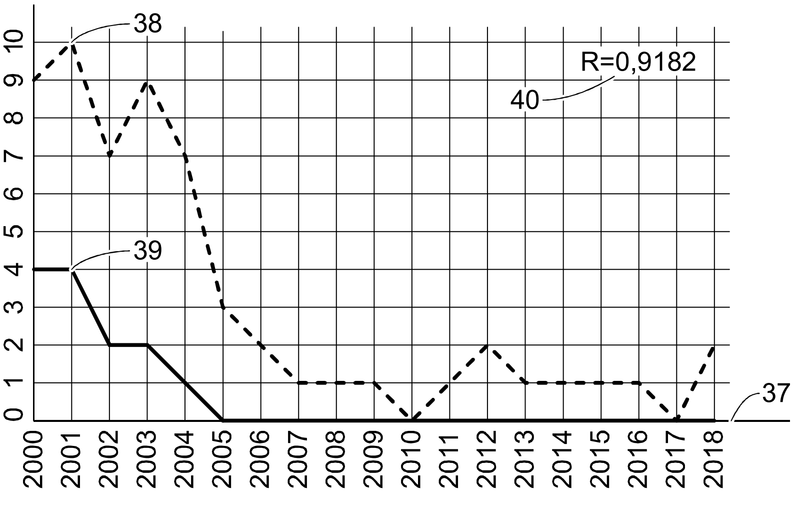


Fig. 13