US 20070033189A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2007/0033189 A1**

Levy et al. (43) **Pub. Date: Feb. 8, 2007**

(54) **METHOD AND SYSTEM FOR EXTRACTING WEB DATA**

(76) Inventors: **Ori Levy**, Ramat-HaSharon (IL); **Jonathan Schler**, Petach-Tikva (IL); **David Tenne**, RaAnana (IL); **Akiva Navot**, Petach-Tikva (IL)

Correspondence Address:
Martin D. Moynihan
PRTSI, Inc.
P.O. Box 16446
Arlington, VA 22215 (US)

(21) Appl. No.: **11/517,418**

(22) Filed: **Sep. 8, 2006**

Related U.S. Application Data

(62) Division of application No. 11/364,169, filed on Mar. 1, 2006.

(60) Provisional application No. 60/705,442, filed on Aug. 5, 2005.

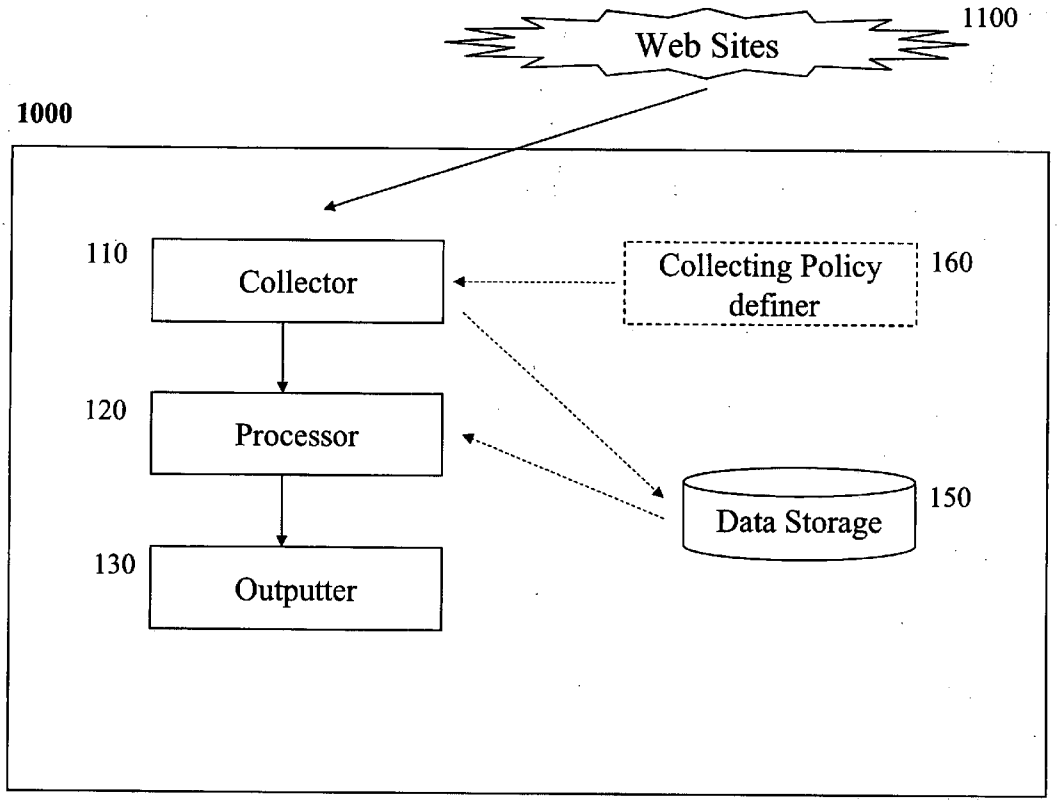
Publication Classification

(51) **Int. Cl.**
G06F 7/00 (2006.01)

(52) **U.S. Cl.** **707/7**

(57) **ABSTRACT**

An apparatus for providing an analysis of attitudes expressed in web content, comprising: a collector for collecting attitude-data in relation to a predetermined subject from one or more pre-selected web site, the attitude-data containing attitudes in relation to the predetermined subject; a processor, associated with the collector, for processing the attitude data so as to generate an attitude analysis; and an outputter, associated with the processor, for outputting the attitude analysis, thereby to provide an indication of attitudes being expressed in the web content in relation to the predetermined subject.



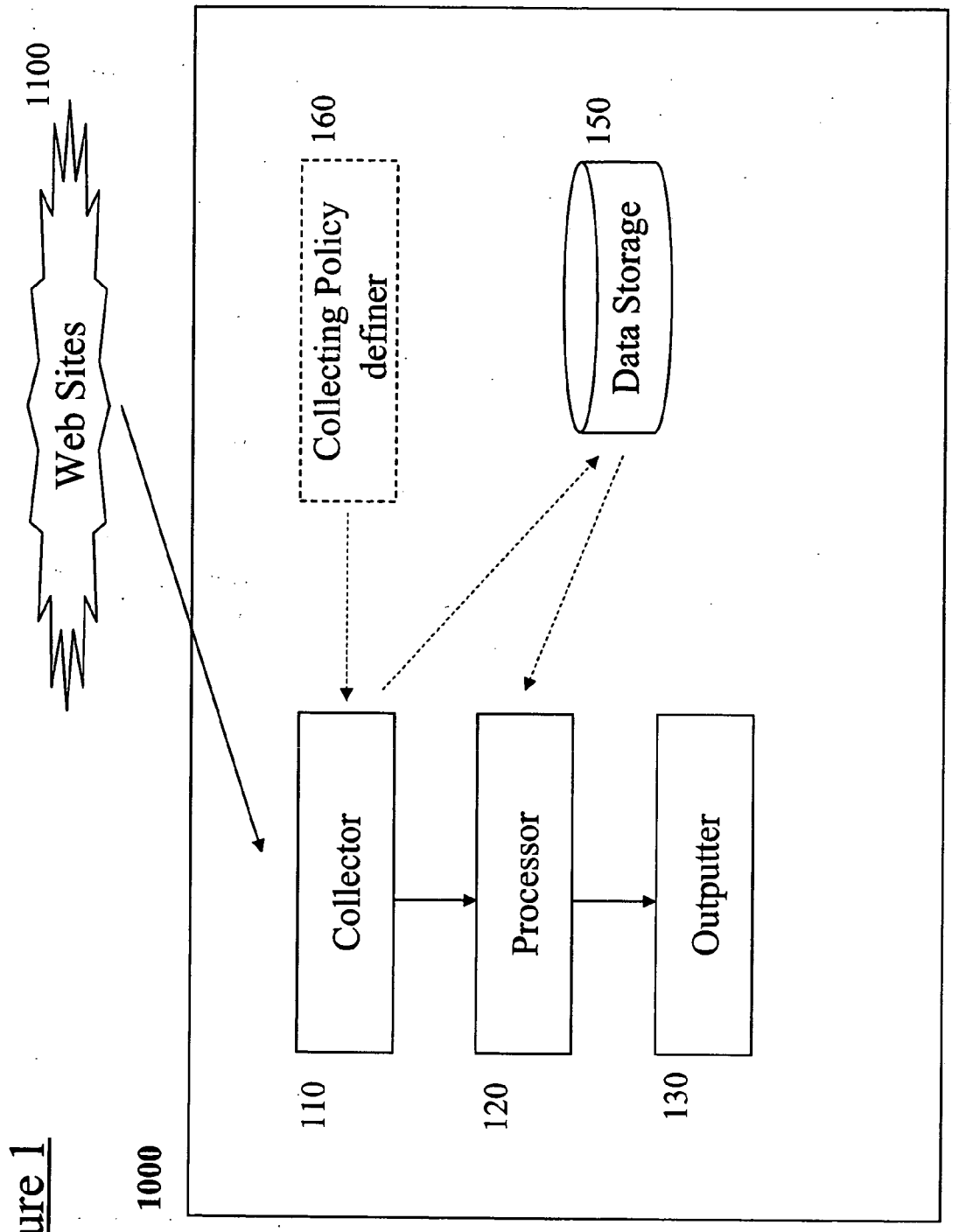

1100
Web Sites
160
Collecting Policy definer
150
Data Storage
Collector
Processor
Outputter
110
120
130
1000

Figure 1

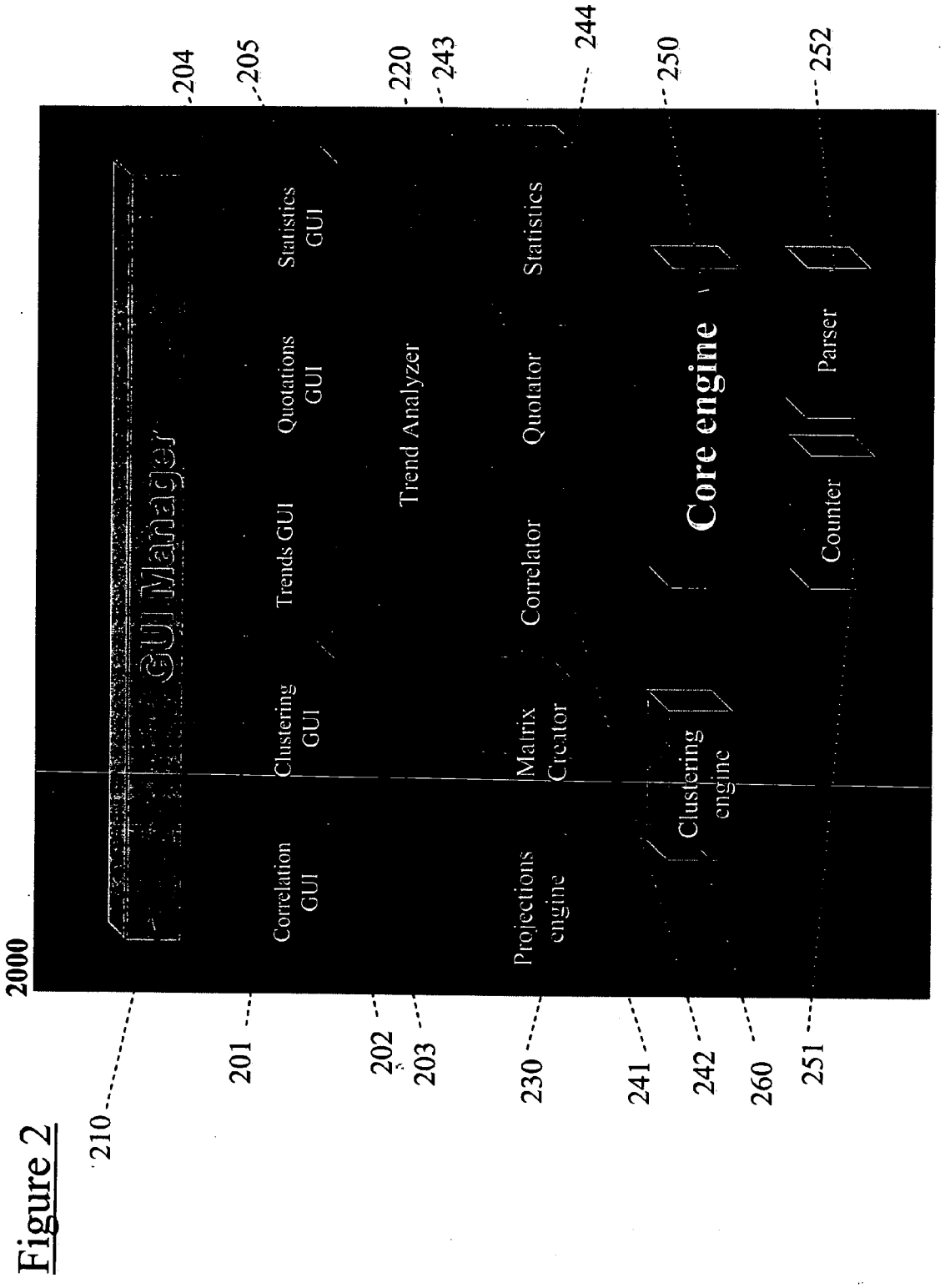
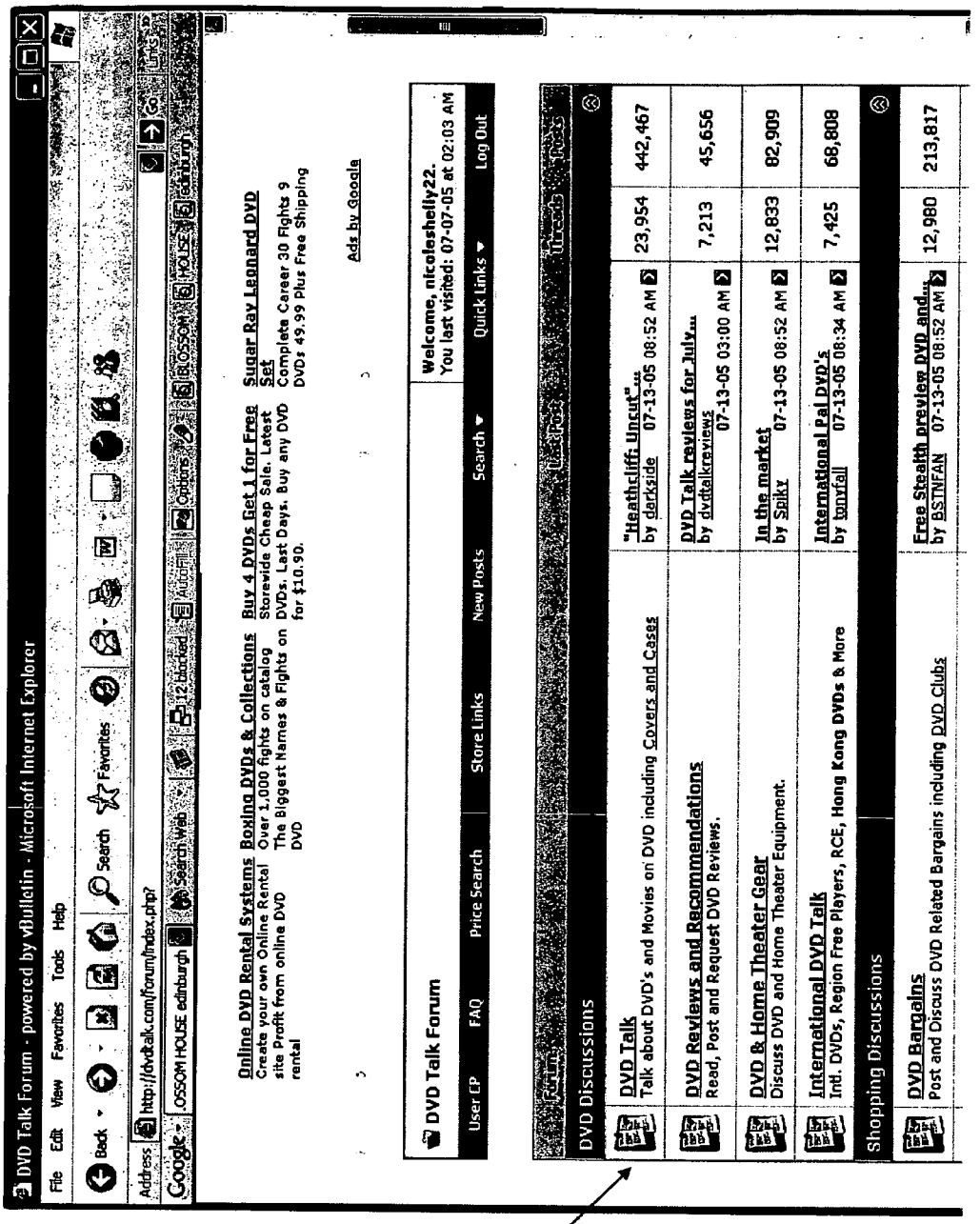

2000
GUI Manager
Correlation GUI
Clustering GUI
Trends GUI
Quotations GUI
Statistics GUI
Projections engine
Matrix Creator
Correlator
Trend Analyzer
Quotator
Statistics
Clustering engine
Core engine
Counter
Parser
210
201
202
203
230
241
242
260
251
204
205
220
243
244
250
252

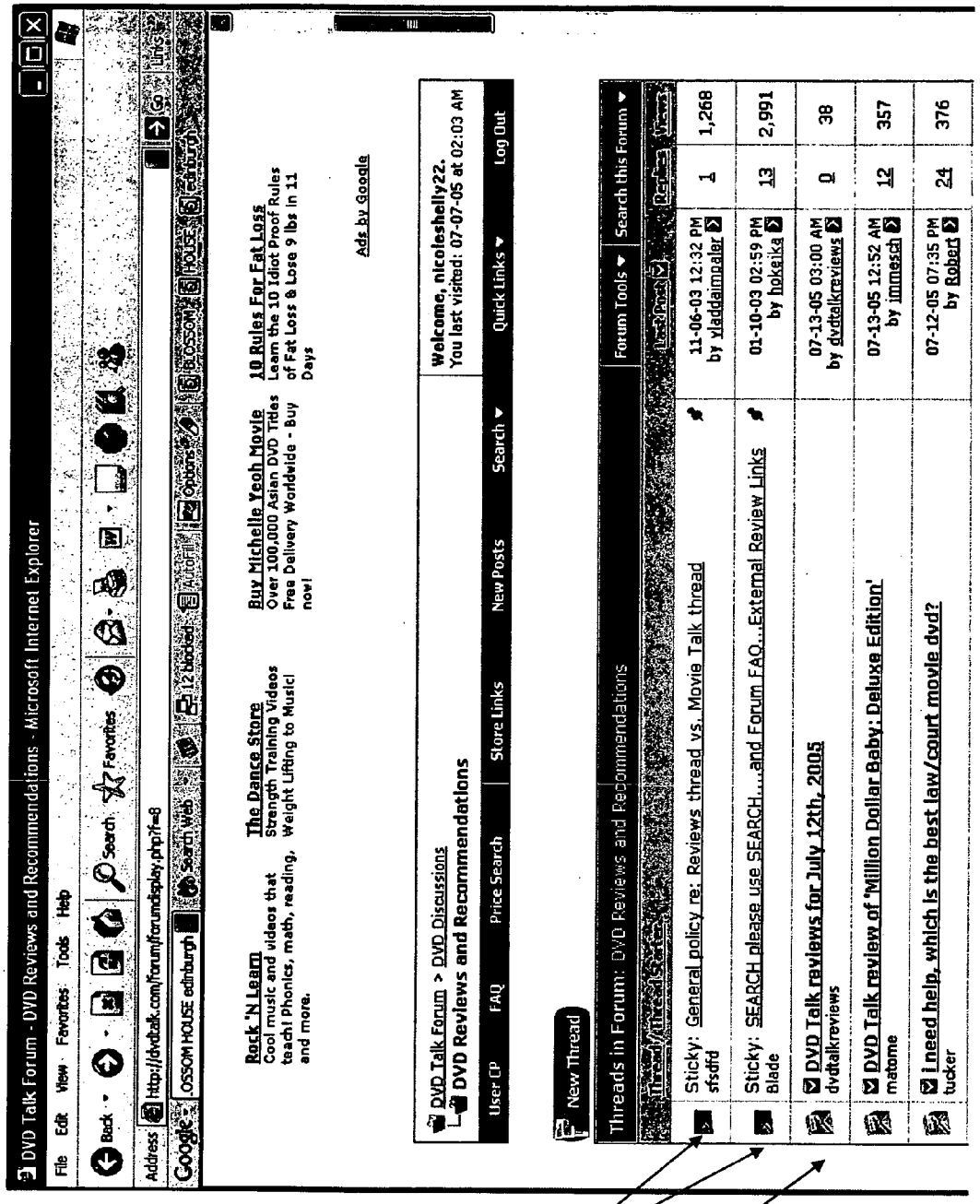
Figure 2

Figure 3


DVD Talk Forum - powered by vBulletin - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://dvdtalk.com/forum/index.php?
Google OSSOM HOUSE edinburgh Search Web 12 blocked AutoFill Options BLOSSOM HOUSE edinburgh
Online DVD Rental Systems
Create your own Online Rental site Profit from online DVD rental
Boxing DVDs & Collections
Over 1,000 fights on catalog The Biggest Names & Fights on DVD
Buy 4 DVDs Get 1 for Free
Storewide Cheap Sale. Latest DVDs. Last Days. Buy any DVD for $10.90.
Sugar Ray Leonard DVD Set
Complete Career 30 Fights 9 DVDs 49.99 Plus Free Shipping
Ads by Google
DVD Talk Forum
Welcome, nicoleshelly22.
You last visited: 07-07-05 at 02:03 AM
User CP FAQ Price Search Store Links New Posts Search Quick Links Log Out
Forum Last Post Threads Posts
DVD Discussions
DVD Talk
Talk about DVD's and Movies on DVD including Covers and Cases
"Heathcliff Uncut"... by derkside 07-13-05 08:52 AM
23,954
442,467
DVD Reviews and Recommendations
Read, Post and Request DVD Reviews.
DVD Talk reviews for July... by dvdtalkreviews 07-13-05 03:00 AM
7,213
45,656
DVD & Home Theater Gear
Discuss DVD and Home Theater Equipment.
In the market by Spikr 07-13-05 08:52 AM
12,833
82,909
International DVD Talk
Intl. DVDs, Region Free Players, RCE, Hong Kong DVDs & More
International Pal DVD's by tonyfall 07-13-05 08:34 AM
7,425
68,808
Shopping Discussions
DVD Bargains
Post and Discuss DVD Related Bargains including DVD Clubs
Free Stealth preview DVD and... by BSTNFAN 07-13-05 08:52 AM
12,980
213,817

310

Figure 4


DVD Talk Forum - DVD Reviews and Recommendations - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://dvdtalk.com/forum/forumdisplay.php?f=8
Rock 'N Learn
Cool music and videos that teach! Phonics, math, reading, and more.
The Dance Store
Strength Training Videos Weight Lifting to Music!
Buy Michelle Yeoh Movie
Over 100,000 Asian DVD Titles Free Delivery Worldwide - Buy now!
10 Rules For Fat Loss
Learn the 10 Idiot Proof Rules of Fat Loss & Lose 9 lbs in 11 Days
Ads by Google
DVD Talk Forum > DVD Discussions
DVD Reviews and Recommendations
Welcome, nicoleshelly22.
You last visited: 07-07-05 at 02:03 AM
User CP FAQ Price Search Store Links New Posts Search Quick Links Log Out
New Thread
Threads in Forum: DVD Reviews and Recommendations
Forum Tools Search this Forum
Thread / Thread Starter
Sticky: General policy re: Reviews thread vs. Movie Talk thread
sfsdfd
11-06-03 12:32 PM by yladdaimpaler
1
1,268
Sticky: SEARCH please use SEARCH....and Forum FAQ...External Review Links
Blade
01-10-03 02:59 PM by hokeika
13
2,991
DVD Talk reviews for July 12th, 2005
dvdtalkreviews
07-13-05 03:00 AM by dvdtalkreviews
0
38
DVD Talk review of 'Million Dollar Baby: Deluxe Edition'
matome
07-13-05 12:52 AM by immesch
12
357
I need help, which is the best law/court movie dvd?
tucker
07-12-05 07:35 PM by Robert
24
376
411
410

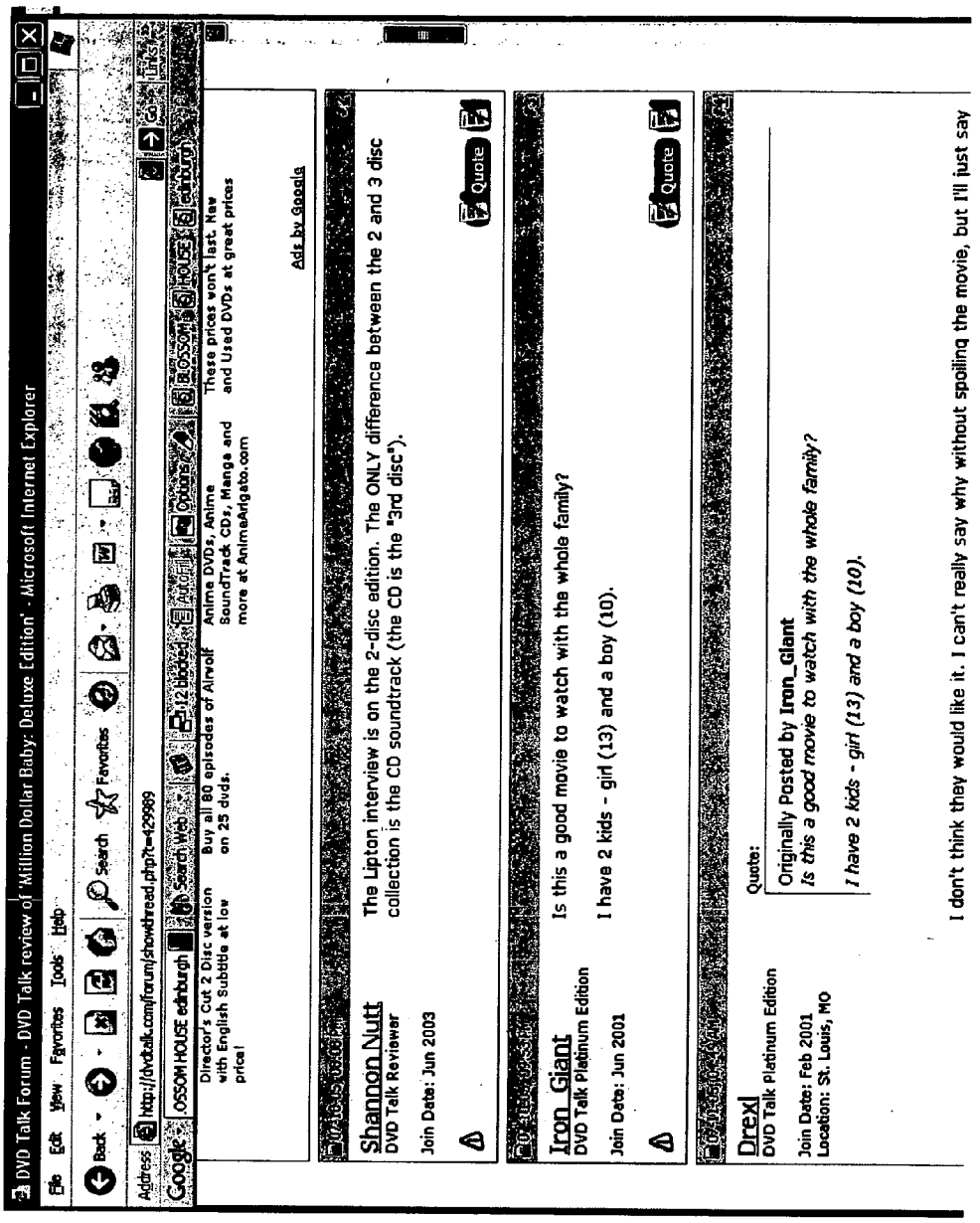

DVD Talk Forum - DVD Talk review of 'Million Dollar Baby: Deluxe Edition' - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://dvdtalk.com/forum/showthread.php?t=429989
Google OSSOM HOUSE edinburgh Search Web 12 blocked AutoFill Options BLOSSOM HOUSE edinburgh
Director's Cut 2 Disc version with English Subtitle at low price!
Buy all 80 episodes of Airwolf on 25 dvds.
Anime DVDs, Anime SoundTrack CDs, Manga and more at AnimeAirgato.com
These prices won't last. New and Used DVDs at great prices
Ads by Google
Shannon Nutt
DVD Talk Reviewer
Join Date: Jun 2003
The Lipton interview is on the 2-disc edition. The ONLY difference between the 2 and 3 disc collection is the CD soundtrack (the CD is the "3rd disc").
Quote
Iron Giant
DVD Talk Platinum Edition
Join Date: Jun 2001
Is this a good movie to watch with the whole family?
I have 2 kids - girl (13) and a boy (10).
Quote
Drexl
DVD Talk Platinum Edition
Join Date: Feb 2001
Location: St. Louis, MO
Quote:
Originally Posted by Iron_Giant
Is this a good movie to watch with the whole family?
I have 2 kids - girl (13) and a boy (10).
I don't think they would like it. I can't really say why without spoiling the movie, but I'll just say

Figure 5

Figure 6

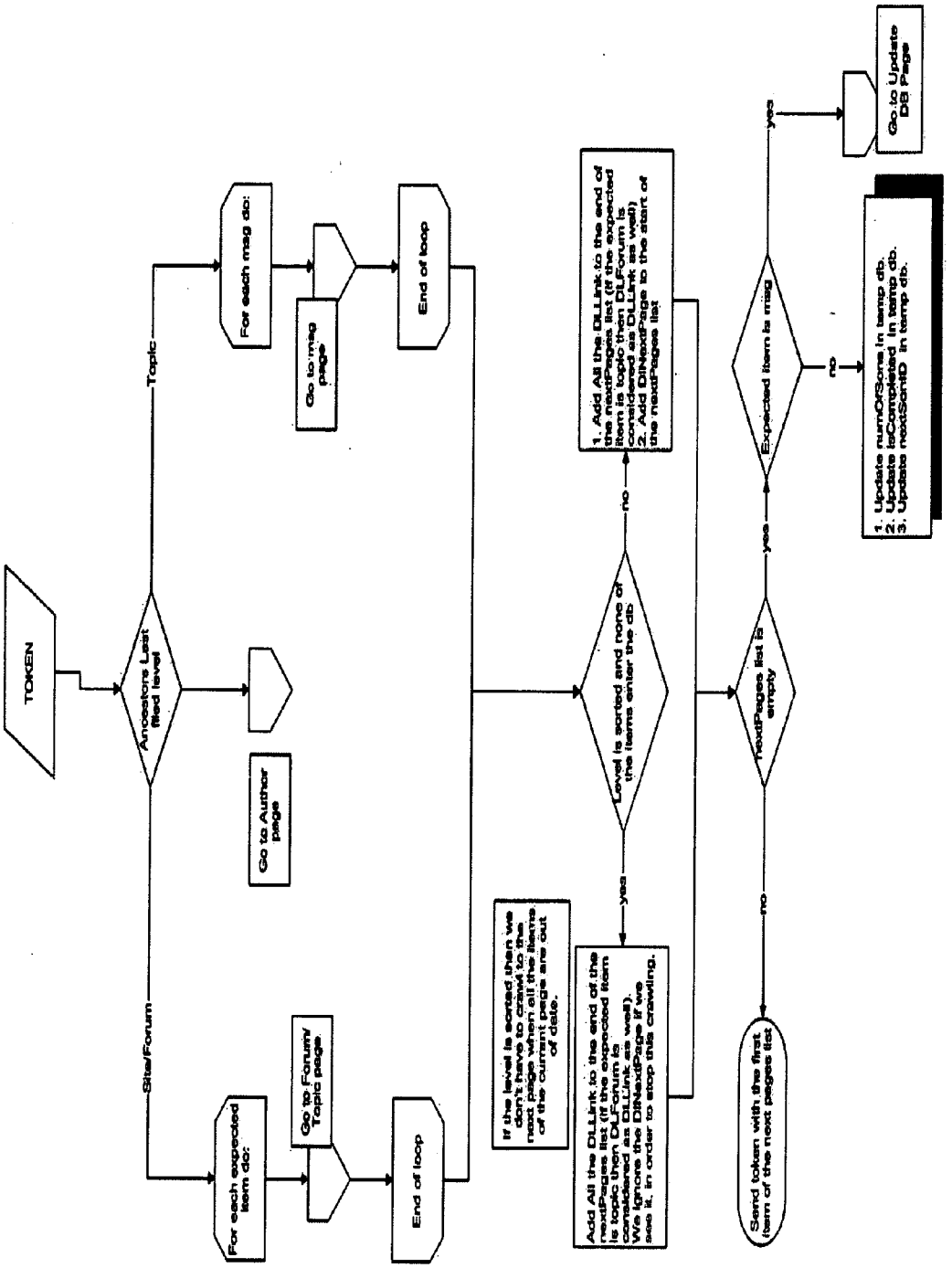
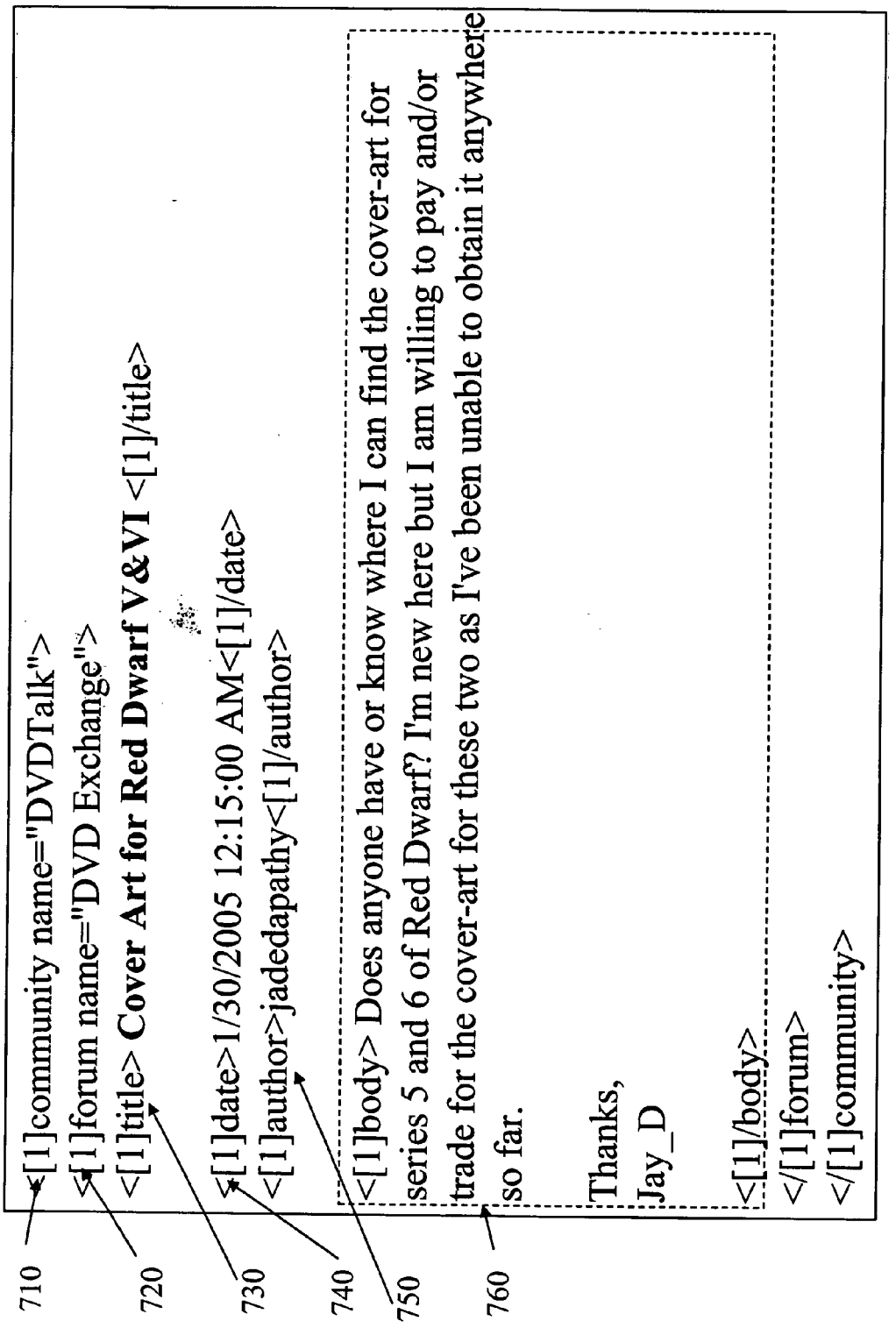

TOKEN
Ancestors Last filed level
Site/Forum
Topic
Go to Author page
For each expected item do:
Go to Forum/ Topic page
End of loop
For each msg do:
Go to msg page
End of loop
If the level is sorted then we don't have to crawl to the next page when all the items of the current page are out of date.
Level is sorted and none of the items enter the db
yes
no
Add All the DLLink to the end of the nextPages list (if the expected item is topic then DLForum is considered as DLLink as well). We ignore the DlNextPage if we see it, in order to stop the crawling.
1. Add All the DLLink to the end of the nextPages list (if the expected item is topic then DLForum is considered as DLLink as well).
2. Add DlNextPage to the start of the nextPages list
NextPages list is empty
yes
no
Send token with the first item of the next pages list
Expected item is msg
yes
no
Go to Update DB Page
1. Update numOfSons in temp db.
2. Update isCompleted in temp db.
3. Update nextSonID in temp db.

Figure 7


710
720
730
740
750
760
<[1]community name="DVDTalk">
<[1]forum name="DVD Exchange">
<[1]title> **Cover Art for Red Dwarf V&VI** <[1]/title>
<[1]date>1/30/2005 12:15:00 AM<[1]/date>
<[1]author>jadedapathy<[1]/author>
<[1]body> Does anyone have or know where I can find the cover-art for series 5 and 6 of Red Dwarf? I'm new here but I am willing to pay and/or trade for the cover-art for these two as I've been unable to obtain it anywhere so far.
Thanks,
Jay_D
<[1]/body>
</[1]forum>
</[1]community>

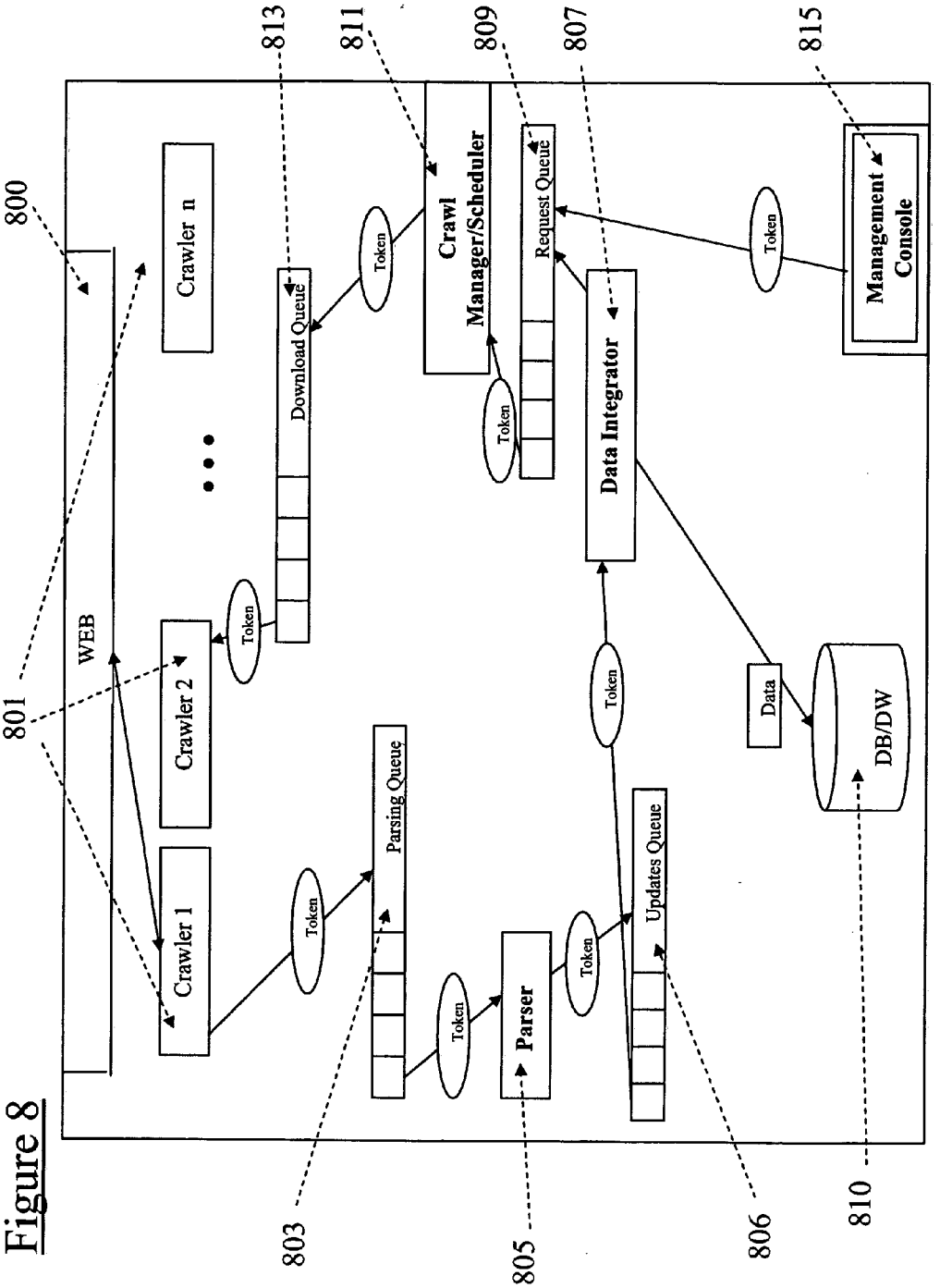

800
801
WEB
Crawler 1
Crawler 2
Crawler n
Token
Download Queue
Crawl Manager/Scheduler
Request Queue
Data Integrator
Management Console
Parsing Queue
Parser
Updates Queue
Data
DB/DW
803
805
806
810
811
813
807
809
815

Figure 8

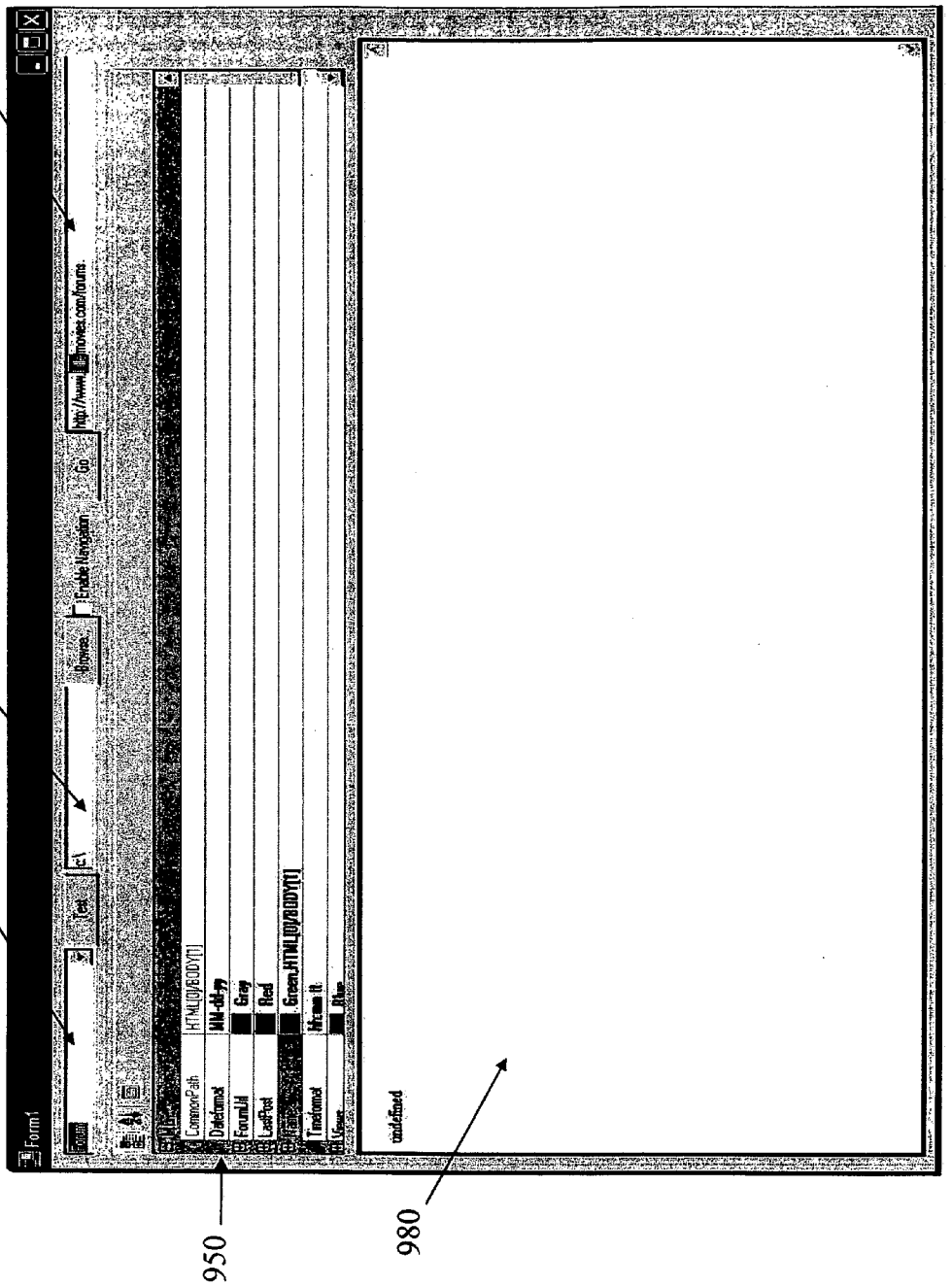

910
920
930
950
980
Form1
Enable Navigation
Go
Movies
Text
CommonPath
DateFormat
ForumUrl
LastPost
TimeFormat
HTMLObjectDOM(1)
MM-dd-yy
Grey
Red
Green,HTMLObjectDOM(1)
Maroon
Blue
undefined

Figure 9

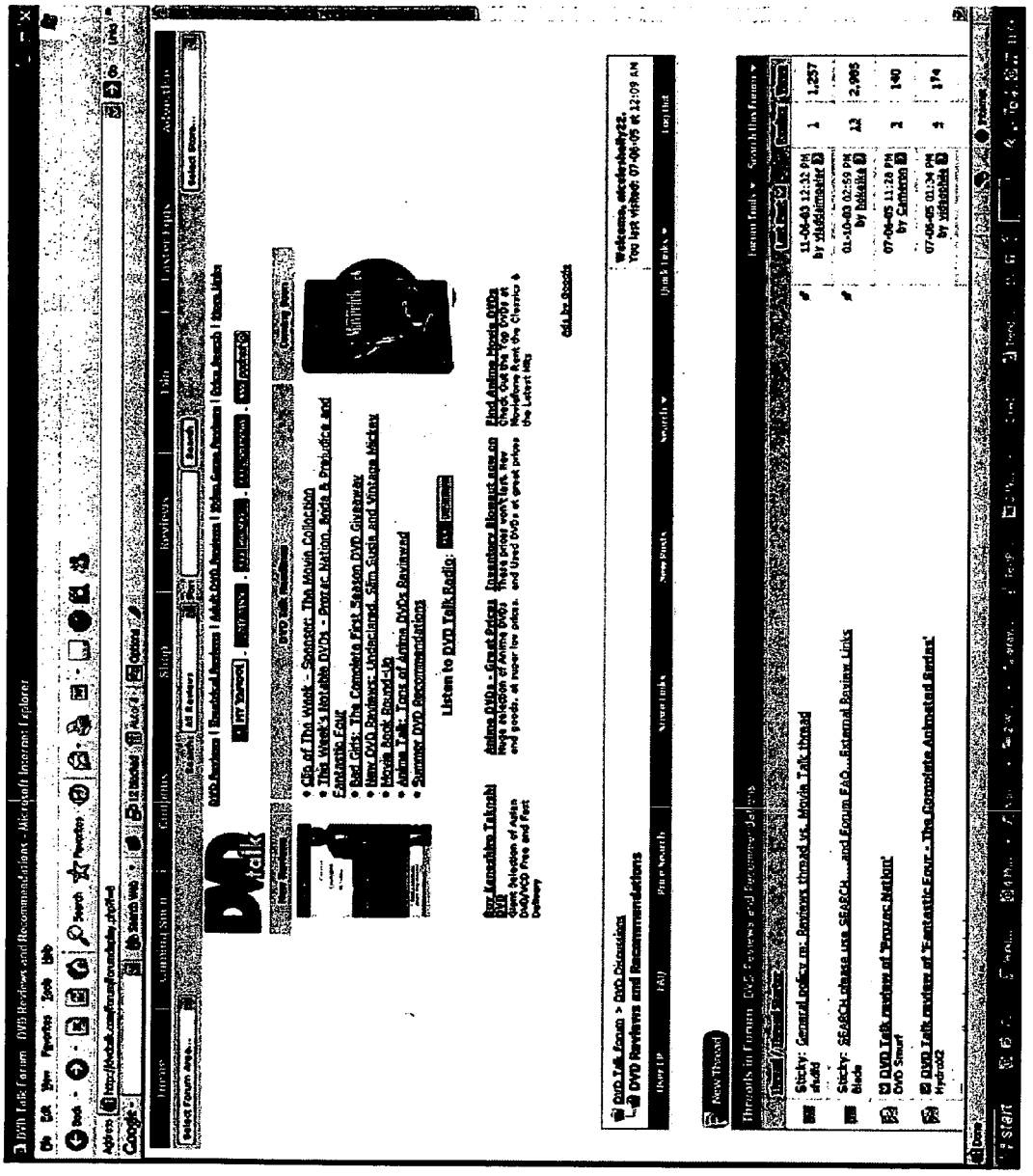


Figure 10

Figure 11

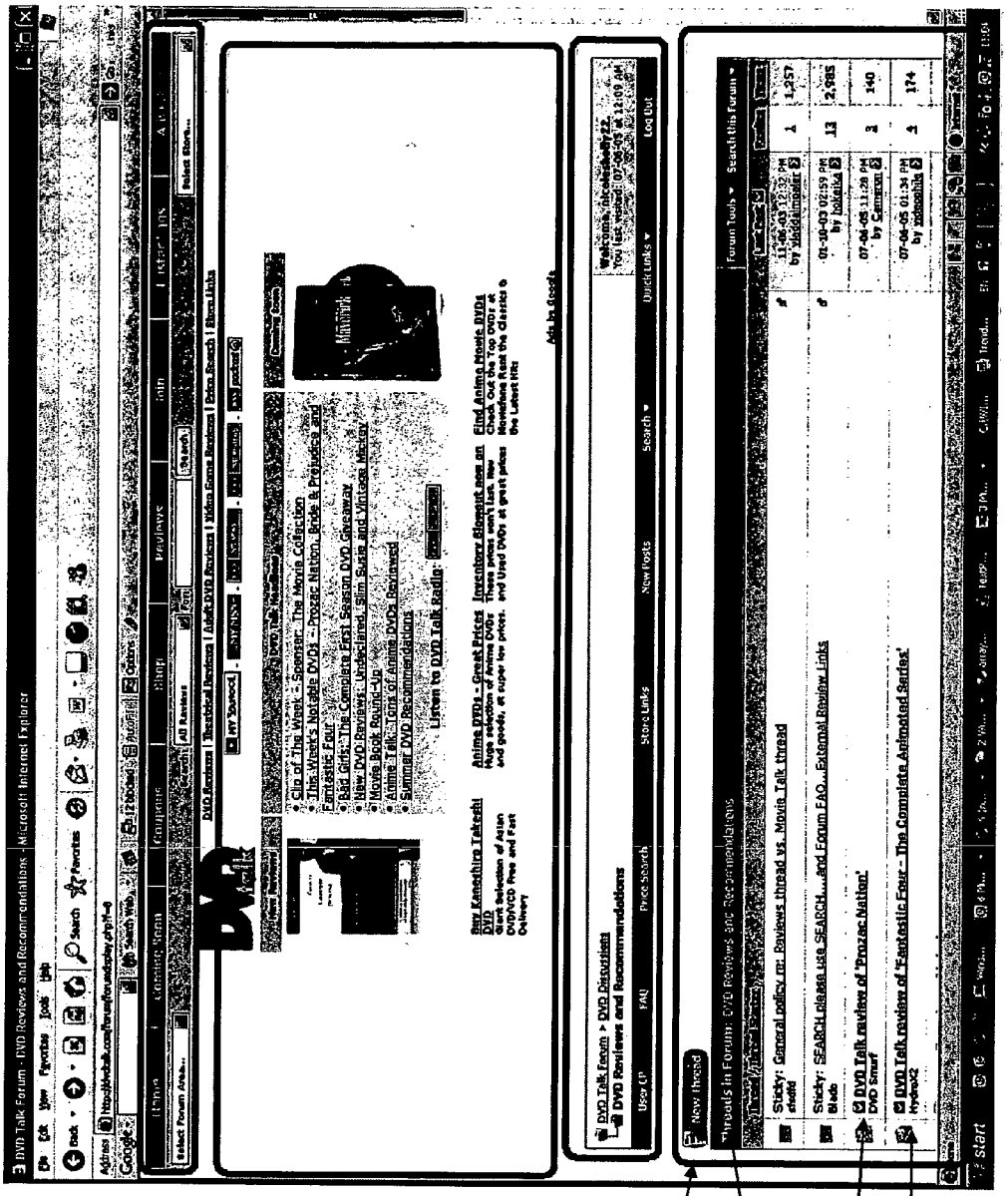


Figure 12

Title: //tr[td[2]/div[1]/a[1]/strong[1]][td[3]/div[1]/div[3]][td[2]/div[1]/a[1]/strong[1]]

Figure 13

Element_paths_collection = get all absolute paths to extracted elements on page

Common_path = first_element_full_path

Foreach element_path in Element_paths_collection do

Common_path = Find_common_path_between(element_path, Common_path)

End foreach

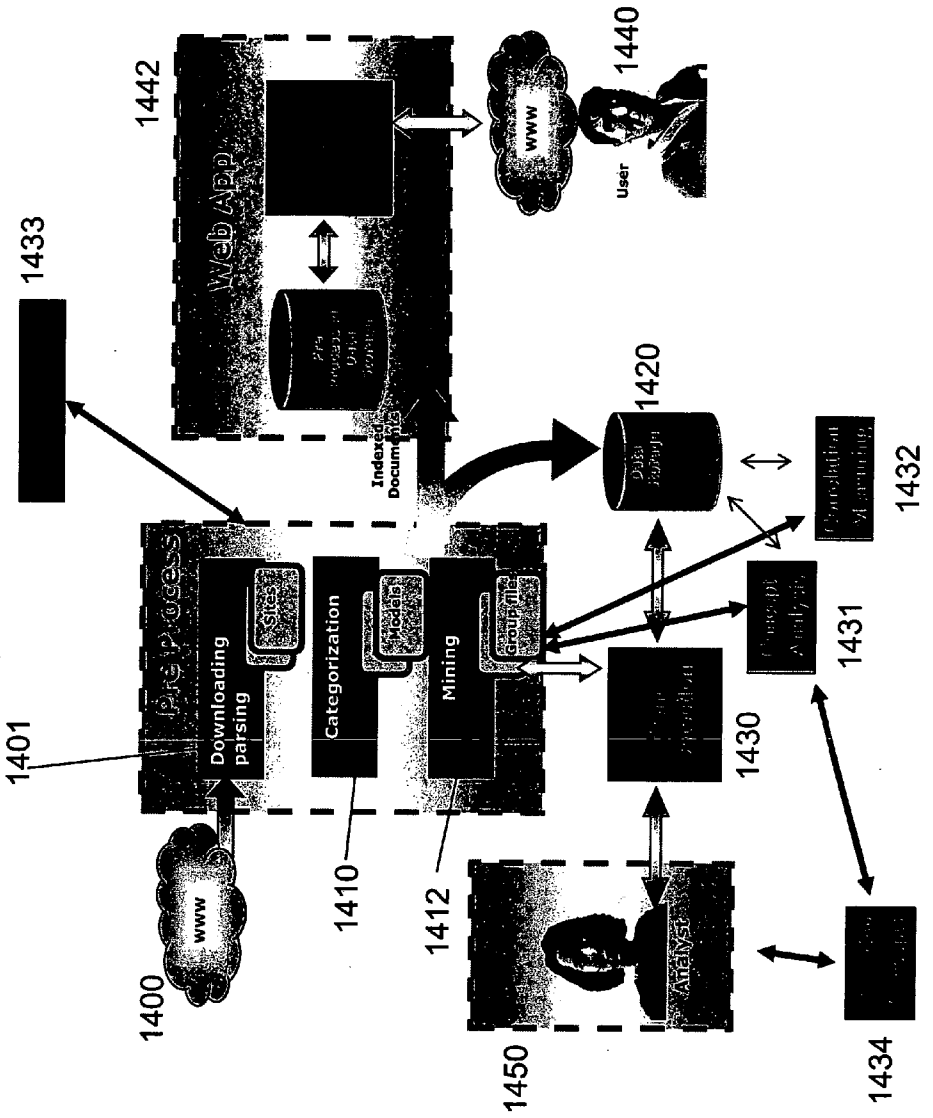

1442
Web APP
1440
WWW
User
1433
1420
Index Documents
1432
Pre-process
Downloading parsing
Sites
Categorization
Models
Mining
Groups
1401
1431
1430
1410
1412
1400
WWW
Analyst
1450
1434

Figure 14

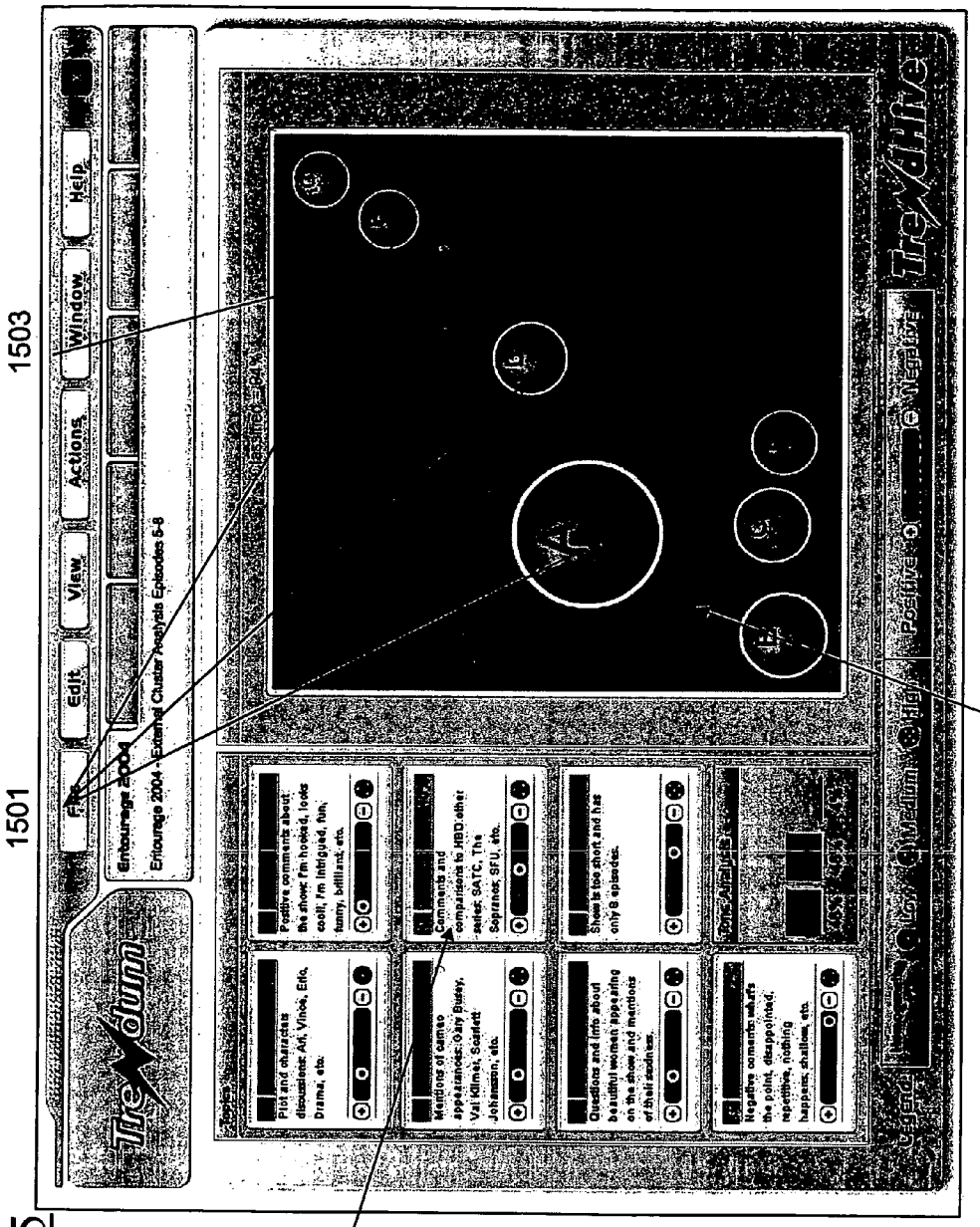

1501
1503
1505
1507
File
Edit
View
Actions
Window
Help
Entourage 2004
Entourage 2004 - External Cluster Analysis Episodes 6-8
Topics
Positive comments about the show: I'm hooked, looks cool!, I'm intrigued, fun, funny, brilliant, etc.
Comments and comparisons to HBO: other series: SATC, The Sopranos, SFU, etc.
Shows is too short and has only 8 episodes.
Plot and characters discussions: Ari, Vince, Eric, Drama, etc.
Mentions of cameo appearances: Gary Busey, Val Kilmer, Scarlett Johansson, etc.
Questions and info about beautiful women appearing on the show and mentions of their sexiness.
Negative comments: sucks, the point, disappointed, repetitive, nothing happens, shallow, etc.
Legend
Low
Medium
High
Positive
Negative
TreeMap

Figure 15

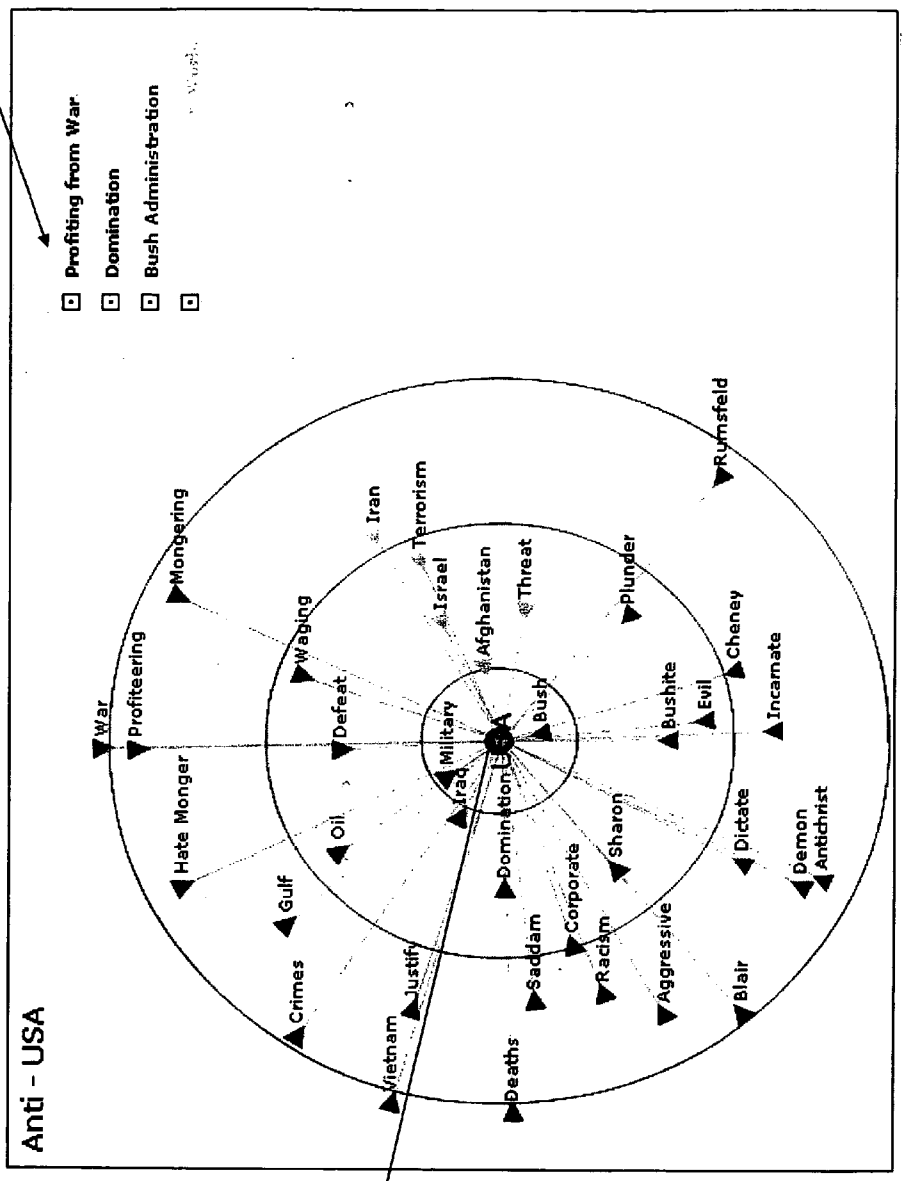

1605
Profiting from War
Domination
Bush Administration
Anti - USA
War
Profiteering
Mongering
Hate Monger
Waging
Defeat
Oil
Gulf
Crimes
Justify
Vietnam
Deaths
Iran
Terrorism
Israel
Afghanistan
Threat
Military
Iraq
Domination
Bush
Saddam
Corporate
Racism
Sharon
Plunder
Bushite
Evil
Cheney
Incarnate
Rumsfeld
Dictate
Demon
Antichrist
Aggressive
Blair
1601

Figure 16

Figure 17

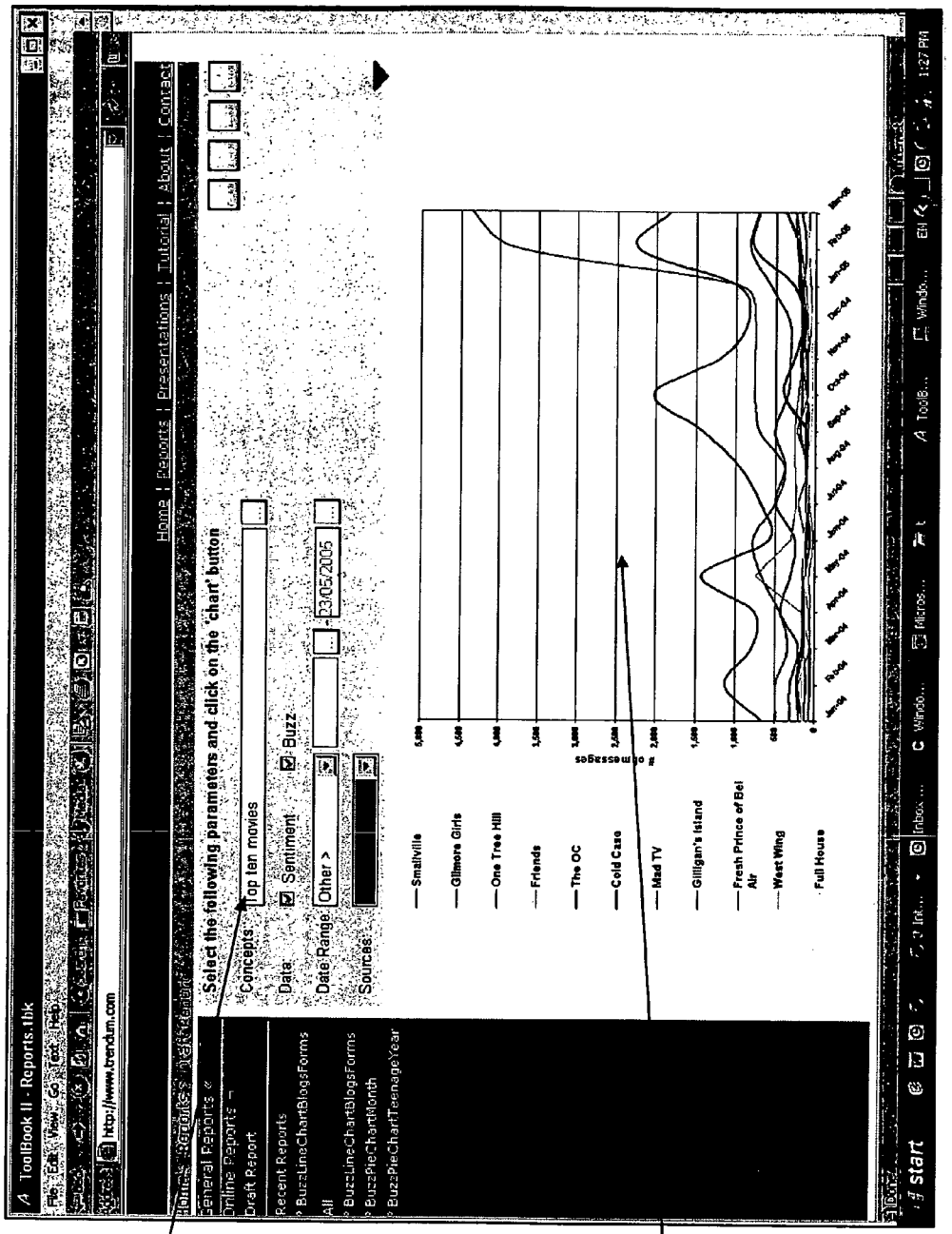
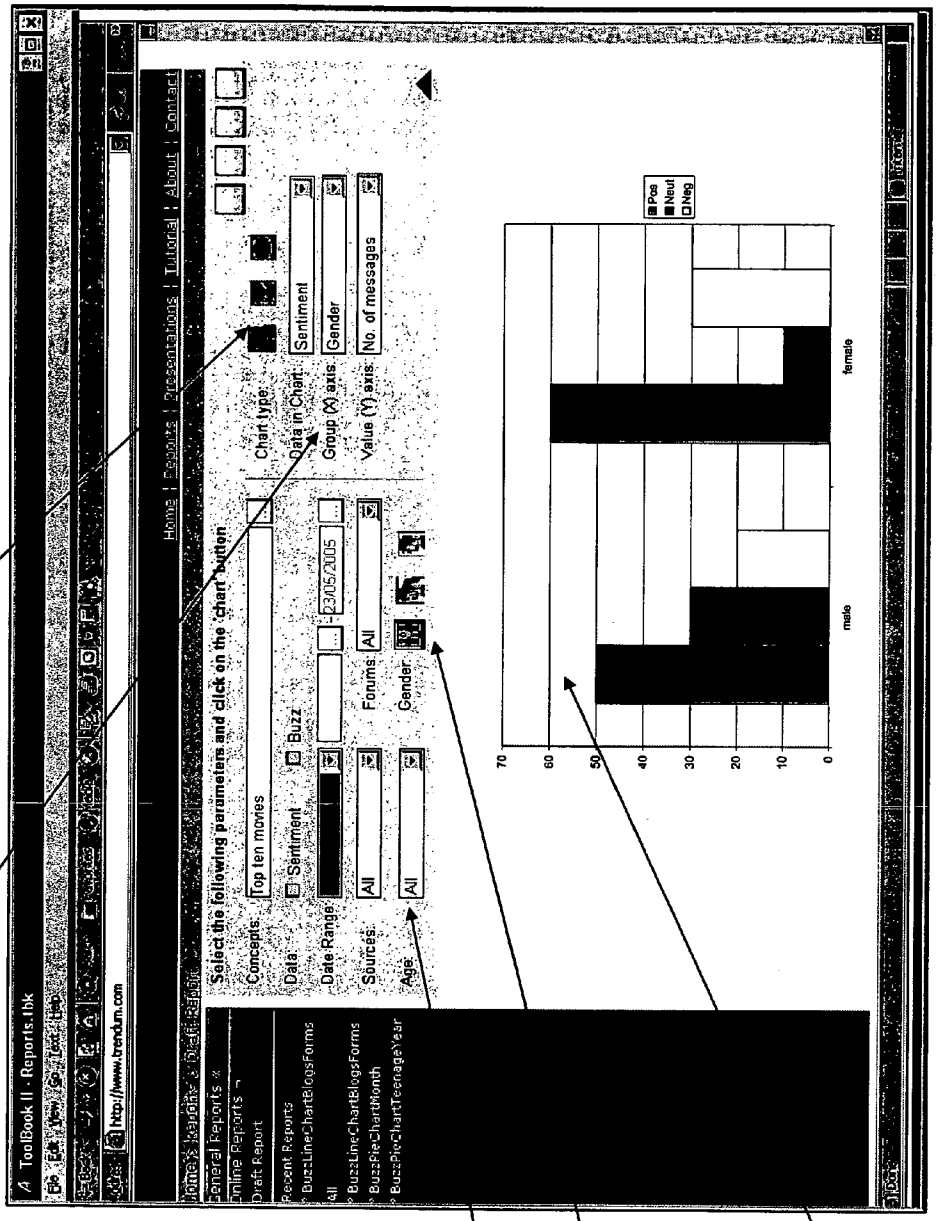

1701
1703
ToolBook II - Reports.tbk
http://www.trendum.com
Home | Reports | Presentations | Tutorial | About | Contact
Select the following parameters and click on the 'chart' button
Concepts:
Top ten movies
Data:
Sentiment
Buzz
Date Range:
Other >
03/05/2005
Source:
General Reports
Online Reports
Draft Report
Recent Reports
BuzzLineChartBlogsForms
All
BuzzLineChartBlogsForms
BuzzPieChartMonth
BuzzPieChartTeenageYear
Smallville
Gilmore Girls
One Tree Hill
Friends
The OC
Cold Case
Mad TV
Gilligan's Island
Fresh Prince of Bel Air
West Wing
Full House
# of messages
start

Figure 18


1805
1807
1803
1801
1810
ToolBook II - Reports.tbk
http://www.trendum.com
Home | Reports | Presentations | Tutorial | About | Contact
General Reports
Inline Reports
Draft Report
Recent Reports
BuzzLineChartBlogsForms
All
BuzzLineChartBlogsForms
BuzzPieChartMonth
BuzzPieChartTeenageYear
Select the following parameters and click on the 'chart' button
Concepts:
Top ten movies
Data:
Sentiment
Buzz
Date Range:
-23/05/2005
Sources:
All
Forums:
All
Age:
All
Gender:
Chart type:
Data in Chart:
Sentiment
Group (X) axis:
Gender
Value (Y) axis:
No. of messages
Pos
Neut
Neg
female
male
70
60
50
40
30
20
10
0

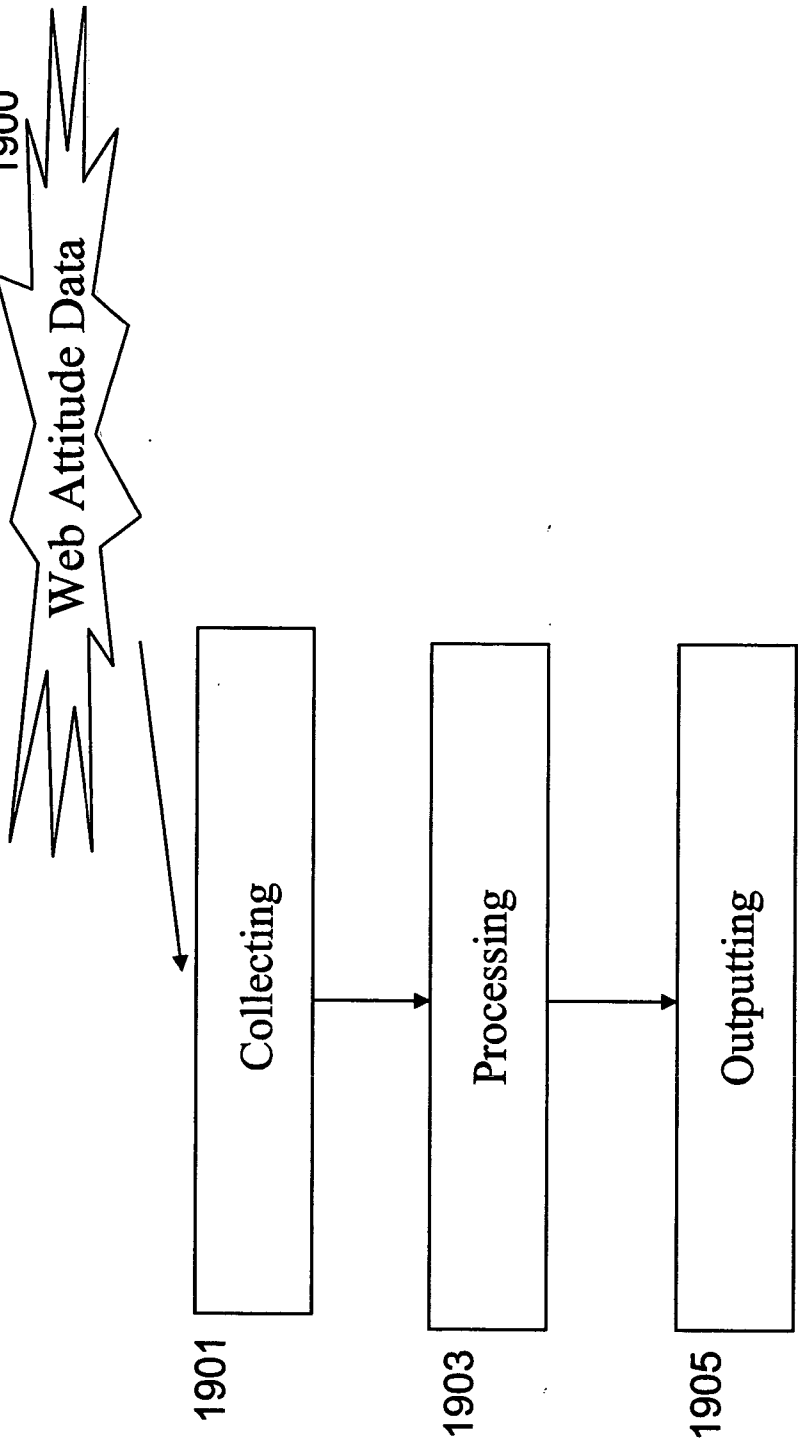

1900
Web Attitude Data
Collecting
Processing
Outputting
1901
1903
1905

Figure 19

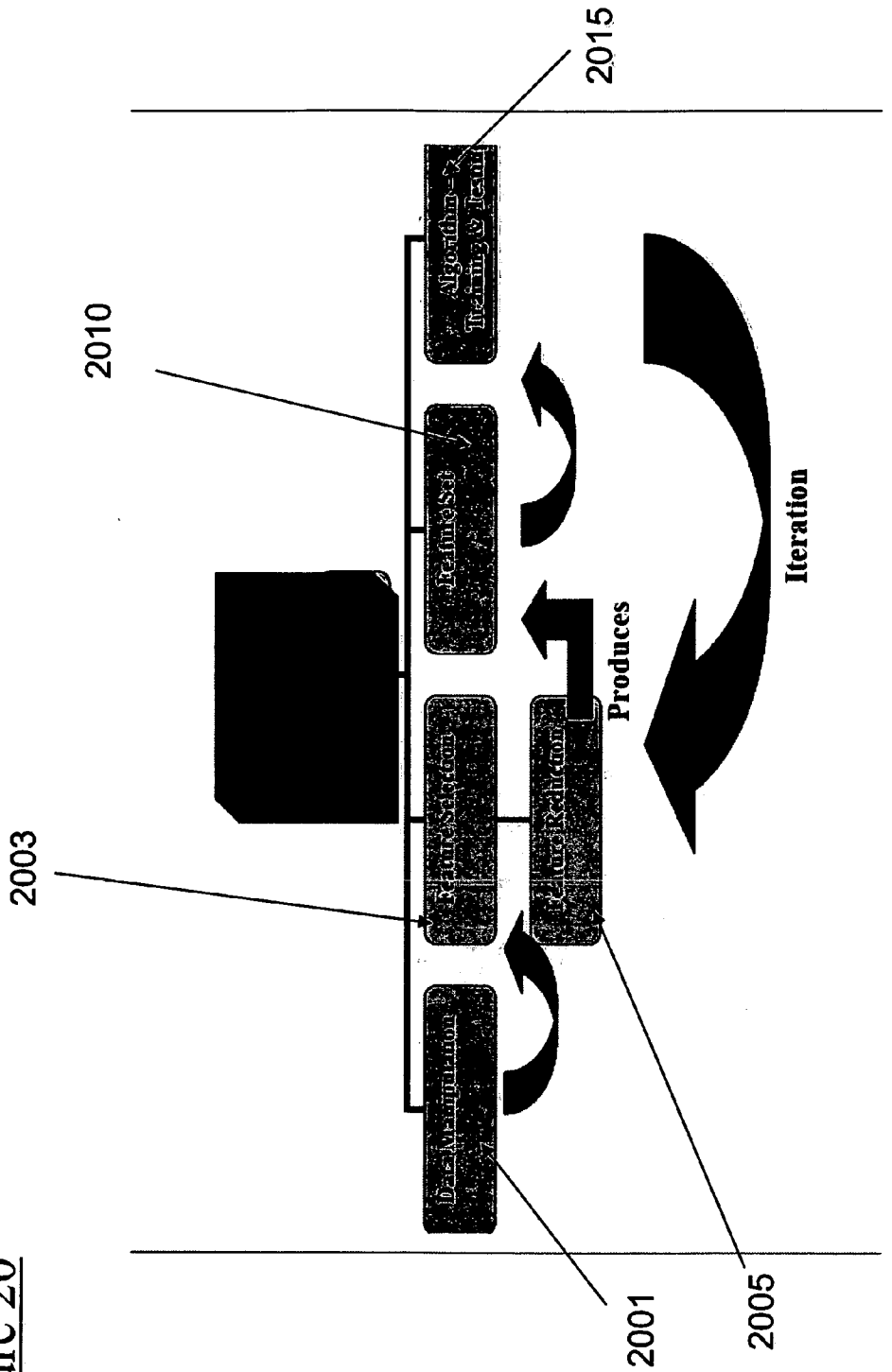


Figure 20

Figure 21

Find k most relevant words // either most frequent, or highest relevance in the //collection (based on distance metric described above

Run density-word clustering algorithm

Build distance matrix for all words

Find most "dense" word //(with most close words)

Merge 2 closest words into one cluster

Repeat until no close word to merge

Fill relevant messages/documents to each word cluster

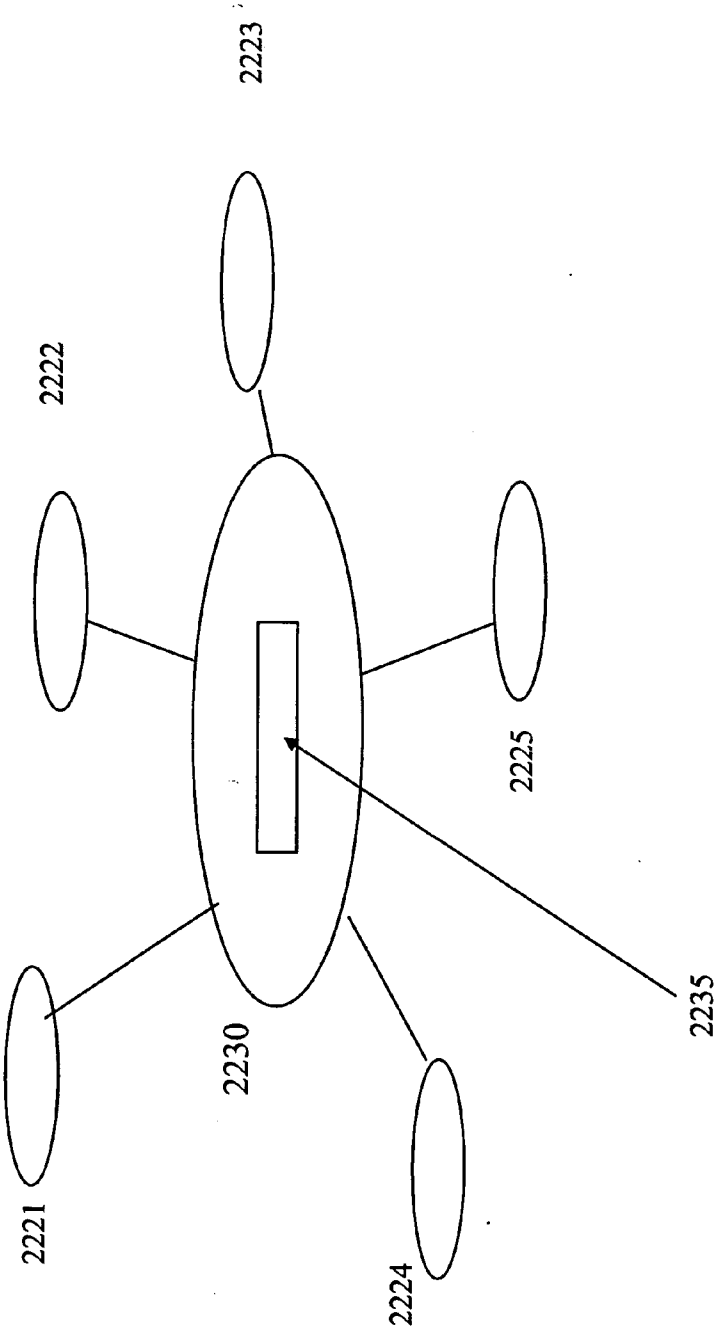


Figure 22

METHOD AND SYSTEM FOR EXTRACTING WEB DATA

RELATED APPLICATIONS

[0001] This Application is a division of U.S. patent application Ser. No. 11/364,169, filed on Mar. 1, 2006, which claims the benefit of U.S. Provisional Patent Application No. 60/705,442, filed on Aug. 5, 2005, the contents of which are hereby incorporated by reference.

FIELD AND BACKGROUND OF THE INVENTION

[0002] The present invention relates generally to an apparatus and method for public attitude analysis. More particularly but not exclusively, the present invention relates to an apparatus and a method for extracting and analyzing public attitude relevant data.

[0003] Modern organizations spend billions of dollars on Public Relations (PR) and advertisement campaigns in order to bring to the public a message, create a positive atmosphere, and influence stakeholders, opinion leaders and customers.

[0004] However, measuring the impact imposed by such campaigns on the public is very difficult.

[0005] Traditional methods for measuring or predicting the impact imposed by public campaigns on the public are inherently limited.

[0006] For example, Consumer marketing research includes both attitudinal and behavioral market research. Consumer marketing research generally refers to the study of consumers and their purchasing habits and activities.

[0007] Attitudinal research generally includes studies that focus on understanding consumers and how consumers make purchasing decisions. Attitudinal research can be defined as research that represents a person's ideas, convictions or liking with respect to a specific object or idea. Opinions are essentially expressions of attitudes. Consequently, attitudes and opinions can be used almost interchangeably to represent a person's ideas, convictions or liking with respect to a specific object or idea. Collecting consumer purchasing information allows, for example, product manufacturers, to drill down to human purchasing dispositions. Attitudinal research may assist in determining the likelihood of product purchase, how future products can be improved, whether product changes are acceptable, etc.

[0008] Behavioral research can be defined as the study of consumer behavior. Behavioral research studies what people do, that is, how people act.

[0009] Behavioral data, reflecting what consumers actually purchase in the marketplace, as opposed to what researchers infer consumers will or will not purchase, has always been available. However, comprehensive behavioral data is not always easy to capture for a variety of reasons.

[0010] The field of consumer marketing research which includes attitudinal and behavioral market research requires gathering data related to, for example, consumer attitudes and consumer behavior, in order to analyze such attitudes and behavior. Consumer data may be gathered through the distribution of incentive items activated via participation in consumer research programs and consumer surveys, such as the ones described in U.S. Patent Publication 20030070338, entitled: "Removable label and incentive item to facilitate collecting consumer data". However, incentive based methods may produce biased results.

[0011] Prior art methods for measuring public attitudes include conducting polls on a presumably representative sample of target audiences. For example, U.S. Pat. No. 3,950,618 entitled: "System for Public Opinion research" describes an automatic system for processing a public opinion poll. However, such methods are based on an assumption that such samples are indeed representative of the target audiences.

[0012] Another popular prior art method for evaluating public attitudes which is very often employed involves focus group techniques. A focus group is a group of people, presumed to be representative of a target population, such as parents or customers, gathered to provide answers to open-ended questions on specific topics and share their opinions.

[0013] Prior Art lacks methods for capturing public attitudes which do not rely on the careful selection of a representative sample or the actual behavior and the availability of comprehensive data pertaining to the actual behavior.

[0014] Prior art has so far failed to incorporate public attitude spread by word of mouth, specifically as far as the Internet is concerned. The web added a new dimension to the media mix—online news groups, discussion groups, forums, chats and blogs—are all forms of communications that did not exist ten years ago, and today they are an inseparable part of the media mix. The public is an inseparable part of the media. The public is fed from the media and feeds the media through its new means of communication.

[0015] There is thus a widely recognized need for, and it would be highly advantageous to have an apparatus and method for extracting and analyzing public attitude data which is devoid of the above limitations.

SUMMARY OF THE INVENTION

[0016] According to one aspect of the present invention there is provided an apparatus for providing an analysis of attitudes expressed in web sites, comprising: a collector for collecting attitude-data in relation to a predetermined subject from at least one pre-selected web site, the attitude-data containing attitudes in relation to the predetermined subject, a processor, associated with the collector, for processing the attitude data so as to generate an attitude analysis, and an outputter, associated with the processor, for outputting the attitude analysis, thereby to provide an indication of attitudes being expressed in the web content in relation to the predetermined subject.

[0017] According to a second aspect of the present invention there is provided an apparatus for crawling web content to provide data for attitude analysis of attitudes expressed in the web content in relation to a predetermined subject, the apparatus comprising a crawler, configured to crawl a plurality of pre-selected web sites, for collecting attitude-data from the web sites, the attitude data comprising attitudes relating to the predetermined subject, the crawler being further configured to provide the attitude data to a predetermined location for the attitude analysis.

[0018] According to a third aspect of the present invention there is provided a method for analyzing attitudes expressed in web content, the attitudes being in relation to a predetermined subject, comprising: automatically collecting attitude data from at least one pre-selected web site, the attitude-data expressing a plurality of attitudes in relation to the predetermined subject, electronically processing the attitude data so as to generate attitude information indicative of the plurality of attitudes, and outputting the attitude-information, thereby to provide an analysis of the attitudes in relation to the predetermined subject.

[0019] According to a fourth aspect of the present invention there is provided a device for interactive setting of a data collection policy using a web page display, comprising a web page displayer, for displaying a web page to a user, operable for defining a data collection policy in relation to the web page. Preferably, the device's web page displayer is further operable to define a specific data collection policy in relation to a respective region of the web page.

[0020] Unless otherwise defined, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. The materials, methods, and examples provided herein are illustrative only and not intended to be limiting.

[0021] Implementation of the method and system of the present invention involves performing or completing certain selected tasks or steps manually, automatically, or a combination thereof. Moreover, according to actual instrumentation and equipment of preferred embodiments of the method and system of the present invention, several selected steps could be implemented by hardware or by software on any operating system of any firmware or a combination thereof. For example, as hardware, selected steps of the invention could be implemented as a chip or a circuit. As software, selected steps of the invention could be implemented as a plurality of software instructions being executed by a computer using any suitable operating system. In any case, selected steps of the method and system of the invention could be described as being performed by a data processor, such as a computing platform for executing a plurality of instructions.

BRIEF DESCRIPTION OF THE DRAWINGS

[0022] The invention is herein described, by way of example only, with reference to the accompanying drawings. With specific reference now to the drawings in detail, it is stressed that the particulars shown are by way of example and for purposes of illustrative discussion of the preferred embodiments of the present invention only, and are presented in order to provide what is believed to be the most useful and readily understood description of the principles and conceptual aspects of the invention. In this regard, no attempt is made to show structural details of the invention in more detail than is necessary for a fundamental understanding of the invention, the description taken with the drawings making apparent to those skilled in the art how the several forms of the invention may be embodied in practice.

[0023] In the drawings:

[0024] FIG. **1** is a block diagram of an apparatus for analyzing attitudes expressed in web sites, according to a preferred embodiment of the present invention;

[0025] FIG. **2** is a detailed block diagram of an apparatus for analyzing attitudes expressed in web sites, according to a preferred embodiment of the present invention;

[0026] FIG. **3** is an exemplary main forum web page;

[0027] FIG. **4** is an exemplary forum header web page;

[0028] FIG. **5** shows an exemplary message header page;

[0029] FIG. **6** is a flow chart illustrating an implementation of a predefined collecting policy for a specific forum web site, according to a preferred embodiment of the present invention;

[0030] FIG. **7** shows an exemplary XML format parsed attitude-data bearing page representation, according to a preferred embodiment of the present invention;

[0031] FIG. **8** is a block diagram illustrating an apparatus for collecting attitude-data from web site(s) according to a preferred embodiment of the present invention;

[0032] FIG. **9** shows an exemplary collecting policy definer graphical user interface (GUI), according to a preferred embodiment of the present invention;

[0033] FIG. **10** shows an exemplary Web site page;

[0034] FIG. **11** shows an exemplary user marked Web site page, according to a preferred embodiment of the present invention;

[0035] FIG. **12** shows an exemplary relative title position encoding in a change query language script according to a preferred embodiment of the present invention;

[0036] FIG. **13** shows an exemplary pseudo-code, for crawling a specific web site page, according to a preferred embodiment of the present invention;

[0037] FIG. **14** is a flowchart illustrating attitude data processing according to a preferred embodiment of the present invention;

[0038] FIG. **15** shows an exemplary graphic representation of the results of clustering, according to a preferred embodiment of the present invention;

[0039] FIG. **16** shows an exemplary graphic representation of the results of correlation measurement according to a preferred embodiment of the present invention;

[0040] FIG. **17** shows a first graphic representation of attitude-data analysis according to a preferred embodiment of the present invention;

[0041] FIG. **18** shows a second exemplary graphic representation of attitude-data analysis according to a preferred embodiment of the present invention;

[0042] FIG. **19** is a flow diagram of an exemplary method for analyzing attitudes expressed in web sites, according to a preferred embodiment of the present invention;

[0043] FIG. **20** is a flow diagram of an exemplary method for categorizing attitude-data text according to a preferred embodiment of the present invention;

[0044] FIG. **21** is an exemplary pseudo-code algorithm for clustering concepts relating to attitude-data, according to a preferred embodiment of the present invention; and

[0045] FIG. **22** is a simplified block diagram of an exemplary architecture of an apparatus for analyzing attitudes expressed in web sites, according to a preferred embodiment of the present invention.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0046] The present embodiments comprise apparatus and method for analyzing public attitudes expressed in web sites or any kind of electronic information found on the web, in a holistic approach.

[0047] The embodiments, according to the present invention, are based on collecting information found on the web, generally considered an influential medium, where authentic attitudes are expressed daily. The websites are in effect, today's word of mouth (WOM) as communicated by millions of Internet users.

[0048] Millions of web users express their views and feelings in online news groups, discussion groups, forums, chat sites, internet blogs etc. All these new means of communication, intensively used by the public today, have become a major part of the media where people are exposed to ideas, products, and messages and where people express their attitudes.

[0049] Embodiments of the present invention aim at collecting the immense amount of high value authentic data pertaining to people's attitudes found in Web sites and holistically analyzing the data, so as to provide with high value attitude information.

[0050] The principles and operation of an apparatus and a method according to the present invention may be better understood with reference to the drawings and accompanying description.

[0051] Before explaining at least one embodiment of the invention in detail, it is to be understood that the invention is not limited in its application to the details of construction and the arrangement of the components set forth in the following description or illustrated in the drawings. The invention is capable of other embodiments or of being practiced or carried out in various ways. Also, it is to be understood that the phraseology and terminology employed herein is for the purpose of description and should not be regarded as limiting.

[0052] Reference is now made to FIG. **1**, which is a block diagram of an apparatus for analyzing attitudes expressed in web sites, according to a preferred embodiment of the present invention.

[0053] An apparatus **1000** according to a preferred embodiment of the present invention comprises a collector **110**. The collector **110** is configured for collecting data, including but not limited to attitude data, containing attitude expression, from pre-selected web site(s) **1100** the attitude data relating to a predefined subject. Preferably, the number of the pre-selected web sites **1100** may reach hundreds of thousands of web sites.

[0054] The pre-selected web sites typically include Chat sites, Interactive news groups, Discussion groups, Forums, Blogs and the like where people express their views and feelings. For example: Internet users may express their views regarding a proposed tax reform, to be discussed by a government, regarding a new product etc.

[0055] According to a preferred embodiment, the collector **110** is programmed as a crawler in a spider network, arranged to detect new attitude data in the pre-selected web sites. For tracking the new attitude data added to a pre-selected web site, the collector **110** utilizes a script, written in a change detection language, as described in greater detail herein below.

[0056] For example, the script may define which parts of a specific page of a pre-selected web site bear a fixed content such as a logo of a firm operating the site, and which parts contain dynamic content, bearing attitude data, such as a continuous flow of user's messages in a web site's chat room.

[0057] In another example, the script may define a comparison to be made by the collector **110** between current content of a page or a part of a page and attitude data previously downloaded from the same page or part of the page.

[0058] The script may be generated using a collecting policy definer **160**, as described in greater detail herein below, and illustrated using FIGS. **9** and **10**.

[0059] The apparatus **1000**, according to a preferred embodiment, further comprises a processor **120**, associated with the collector **110**, used for processing the attitude-data. The processing of the attitude-data may typically include parsing the attitude-data, content analysis techniques, data mining, and other data analysis techniques. These techniques may implement any one of a variety of algorithms, which includes but is not limited to: neural networks, rule reduction, decision trees, pattern analysis, text and linguistic analysis techniques, or any relevant known in the art algorithm.

[0060] The apparatus **1000** according to a preferred embodiment, further comprises an outputter **130**, associated with the processor, for outputting resultant attitude information based on the processed attitude-data. Preferably, the output information is presented to a user utilizing a set of graphical tools, as described in greater detail herein below. The graphical tools may be implemented as a stand alone desktop application, as a web browser based application, as a client application in a client-server architecture, etc.

[0061] Preferably, the apparatus further comprises a data-storage **150** where the attitude-data is stored.

[0062] More preferably the data storage **150** is a data warehouse, provided with a storage area and, preferably, with advanced means for analysis of the attitude-data. In a preferred embodiment, the data warehouse is provided with a set of graphical tools aimed at enabling a user to navigate the processed attitude-data, explore it, and easily find the information the user is interested in.

[0063] The graphical tools may be implemented in as a desktop application, a web application, or any other known in the art alternative.

[0064] In a preferred embodiment, the collector **110** may continuously monitor the pre-selected web site(s) **1100**, on a 24 hours a day and seven days a week basis.

[0065] Optionally, a specific schedule for collecting the attitude-data may be set with respect to specific web site(s).

[0066] According to a preferred embodiment of the present invention, the collector **110** works in a continuous mode. Preferably, the collector utilizes a change detection language or mechanism, and downloads relevant pages of the pre-selected web site(s), according to a predefined collecting policy.

[0067] According to a preferred embodiment of the present invention, the collector **110** further includes a crawler.

[0068] The crawler is responsible for crawling the pre-selected web pages for new data, and for downloading relevant web pages there from. Preferably, the crawler is an open system that has capabilities to download all kind of data on the network including, but not limited to: Web pages, Forums, Discussion boards, and Blogs.

[0069] Reference is now made to FIG. **2** which a detailed block diagram of an apparatus for analyzing attitudes expressed in web sites, according to a preferred embodiment of the present invention.

[0070] An apparatus **2000**, according to a preferred embodiment of the present invention comprises a GUI Manager **210** which manages the interaction with a user of the apparatus **2000**.

[0071] The GUI Manager **210** includes a Correlation GUI component **201** which is configured to present correlation data pertaining to correlations among phrases having relevance-relationships with a common concept relating to the predefined subject, as found in the attitude data, as described in greater derail herein below.

[0072] The Correlation GUI component is connected with a correlator **242** which is configured to measure correlation between one or more phrases and a respective common concept relating to the predefined subject, as described in greater detail herein below. The concept may describe an attitude towards the subject such as but not limited to negative, a positive, or a neutral attitude, including any words that do not express a sentiment directly but may be conceptually related in people's minds.

[0073] The Correlation GUI is further connected to a Matrix Creator **241** which is configured to a create and populate a N×N Matrix with values indicating distances between correlated phrase, as described in greater detail herein below.

[0074] The GUI Manager **210** further includes a Clustering GUI component **202** which presents clusters relating to concepts in the attitude data to the user, as described in greater detail hereinabove. The concept may describe any information regarding the subject, such as: attitude towards the subject such as a negative, a positive, or a neutral attitude, as described hereinabove, or other related concepts, people, products, emotions etc.

[0075] Clustering the concepts may be carried out by a clustering engine **260**, utilizing clustering methods, as described in greater detail hereinabove.

[0076] Preferably, the Correlation GUI **201** component and the Clustering GUI component **202** are further connected with a Projections engine **230** which graphically positions items representing clusters and correlated phrases on the GUI's screens, such as the screen presented in FIG. **15** herein below.

[0077] The GUI Manager **210** further includes a Trend GUI component **203** which presents the user with trend data. The trend data is generated by a Trend Analyzer **220** which is configured to detect trends in the attitude data. The trend analyzer is fed by a Statistics component **244** which generates statistical data pertaining to the appearing of attitude expressing phrases in the attitude data, as described in greater detail herein below. For example, trend GUI may facilitate the detection of a shift in public discussion of a specific concept, or expression of specific attitudes.

[0078] The GUI Manager **210** may further include a Statistics GUI component, **205** connected to Statistics component **244**, for presenting the statistics data generated by the Statistics component **244** to the user.

[0079] The GUI Manager **210** further includes a Quotations GUI component **204** which presents the user with quotations relating to the concepts presented to the user by the correlator GUI **201**, as described hereinabove. The Quotations GUI component **204** is fed by a Quatator **243** which is configured to extract relevant attitude expressing quotations from the attitude data.

[0080] The Statistics component **244**, the Quatator **243**, the Correlator **242**, and the Matrix creator **241** are connected to a core engine **250** which includes a parser **252**, for parsing the attitude data that is downloaded from crawled pages and a counter **251** for counting the appearances of concepts etc. in the attitude data, as described in greater detail herein below.

[0081] Reference is now made to FIG. **3** which shows an exemplary main forum web page.

[0082] The exemplary main forum web page is a DVD Talk forum main web page. In this example, the crawler is preconfigured for crawling the web site, downloading all messages that appear in the threads (topics) of the site forums. In this example, the crawler first crawls the links in the exemplary main forum web page, to all forum header pages **310** available in the pre-selected web site. Preferably, the crawler is further preconfigured to filter out non-relevant links so as to avoid downloading or attempted downloading of irrelevant pages.

[0083] Reference is now made to FIG. **4** which shows an exemplary forum header web page.

[0084] After the crawler gets the links to the forum header pages from the exemplary main forum page, the crawler crawls relevant threads **410** appearing in each of the header web pages, according to the links **310**. Preferably, the crawler is pre-configured for filtering out non-relevant threads like the general policy and search threads appearing in this example **411**.

[0085] Reference is now made to FIG. **5** which shows an exemplary message header page.

[0086] According to a preferred embodiment of the present invention, for each of the relevant threads **410**, the crawler extracts relevant attitude-data, which contain attitude expressions. As shown in the example page on FIG. **5**, each message optionally comprises a date, a title, an author, and a message body.

[0087] Optionally, each message also contains a list of quotes (quotations from other cited messages), and signature. The quotes are marked in the message so that, during his analysis procedure, the user can choose if he wants his analysis to be performed on the messages including the quotes or not. The message signatures (when present) are filtered during the crawling process, in order to avoid skewing the results, as described in greater detail herein below.

[0088] The data has to be extracted from the page, while omitting all irrelevant information. It is important to remember that there are many types of irrelevant information that may be found on such message pages. The irrelevant information includes but is not limited to: other messages, signatures, html tags, ads etc. and those vary from one site to another.

[0089] Preferably, the collector **110** implements a predefined collecting policy. The collecting policy may include specific guidelines with respect to specific ones of the pre-selected web sites. These guidelines may define which parts of the pre-selected web site(s) to crawl, in what order, etc.

[0090] For example, reference is now made to FIG. **6** which is a flow chart illustrating an implementation of a predefined collecting policy for a specific forum web site, according to a preferred embodiment of the present invention.

[0091] In a preferred embodiment, the collector **110** uses a HTTP request for downloading the relevant page(s) of the pre-selected web site(s), according to URL addresses.

[0092] Preferably, the crawler may be further configured for handling relevant aspects of the crawling such as—session objects, login information, cookies, etc.

[0093] According to a preferred embodiment, the crawler is further responsible for scheduling downloading processes of relevant pages of the pre-selected web site(s) (i.e. request per time quantum per site).

[0094] In addition, the crawler may be also configured for determining in what order the pages are downloaded. Preferably, network traffic is also carefully monitored by the crawler, with respect to the pre-selected web sites, so as to avoid generating excessive traffic on the web site(s), by carefully scheduling the downloading process.

[0095] Optionally, the crawler verifies that a pre-defined time interval is kept between one access to a certain web site and another access, so as to try avoiding creation of network overload on the web site.

[0096] According to a preferred embodiment, several crawlers are employed in parallel in the downloading process and each of the crawlers is configured for downloading respective web site(s).

[0097] According to a preferred embodiment of the present invention, the collector **110** further includes a parser. Once a relevant page is downloaded by the crawler, it may be forwarded to the parser.

[0098] The parser is configured for parsing the relevant page and for extracting relevant attitude data from the relevant page or links top pages that contain this relevant data.

[0099] Relevant data sections may be found on the message text, message title, date, author and other places on the page. The parser is further configured for filtering out irrelevant information on the page, like html tags, adds, header, footer etc.

[0100] In a preferred embodiment of the present invention, the parser may apply a script, customized specifically by a user for each web site, to extract relevant attitude-data from the web site, while filtering out non-relevant or corrupted data. The non relevant data may include but is not limited to: hidden data such as html tags and scripts that are mainly used for page definition and page control, and non relevant content data like texts that are presented on the web page but are not relevant with respect to the attitude-data, such as a page number, a commercial footer, a banner etc.

[0101] In a preferred embodiment of the present invention, after irrelevant data is removed and only the relevant attitude-data remains in the web page, the parser converts the web page into a mark-up language format representation. Optionally, the mark-up language is XML. In the mark-up language format representation, relevant data and metadata may be encoded in a searchable and indexable format.

[0102] According to a preferred embodiment, specific types of data, found on the web page, are handled by the parser in a specific manner, in accordance with a predefined policy.

[0103] For example, in message boards, very often a user issues a new message, citing a message previously posted by another user. The cited message appears in the new message. However, for analysis purposes it may be ignored, as it may skew the statistics of the results if it is counted twice in spite of the fact it is not a new unique message. In another example, message signatures may also skew the results, as they are identical for all messages a specific user issues. During analysis, the words appearing in the signatures may skew the statistics.

[0104] Thus the parser may be configured to recognize and ignore parts of messages such as quotations from other messages or signatures.

[0105] Reference is now made to FIG. **7** which shows an exemplary semi-XML format parsed attitude-data bearing page representation, according to a preferred embodiment of the present invention. The provided example illustrates the encoding of page including a community name (DVD Talk in the example) **710**, a forum name (DVD Exchange in the example) **720**, a message title **730**, a date **740**, an author **750** and the body of the message **760** are encoded in a searchable and indexable XML language format representation.

[0106] In a preferred embodiment of the present invention, the collector **110** further comprises a data integrator (updater).

[0107] The data integrator is responsible for verifying that only relevant pages documents crawled from the internet are stored in the data storage **150**. The data integrator checks that a current document does not already exist in the data storage **150**. The data integrator is also responsible for checking the completeness of the download, i.e that no errors are found, the parsing is carried out successfully, etc.

[0108] When the Updater identifies that all pages are downloaded (For example, according to the expected num-

ber of pages that should be downloaded), it crawls all the user profiles, and then folds the whole downloaded data set to the data storage **150** or calls another component, say a utility of a data base management system (DBMS), for folding the data to the data storage **150**.

[**0109**] The data integrator may be configured for integrating the attitude data into a complete and non-redundant attitude data. The integration of attitude data by the data integrator may include but is not limited to: handling redundancy of data, preventing keeping duplicate page etc. Integrating the data may further include ensuring complete download of all relevant data bearing pages of the web site(s). i.e.—that the attitude-data is error free, that the parsing is successfully completed etc. The data integrator may be further configured for indicating when and if all relevant pages are downloaded.

[**0110**] In a preferred embodiment, the data integrator is further configured for deciding if an apparent error, detected when downloading a web page, is recoverable or should the web page be regarded as corrupted and be accordingly ignored.

[**0111**] The data integrator may be further configured for updating the data storage **150** with the attitude-data, while carrying out the integration of the attitude-data as described hereinabove.

[**0112**] According to a preferred embodiment of the present invention, the apparatus **1000** further comprises a collecting policy definer **160** which is associated with the collector **110** and is used for defining the collecting policy.

[**0113**] Preferably, the collecting policy may address various aspects of the collecting process. The collecting policy may define which web site(s) or what kind of web sites the collector collects attitude-data from. The collecting policy may provide specific guidelines for crawling through a specific web site. The specific guidelines may define which kinds of data that are found on pages of the web site are to be ignored, in what order should the web site be crawled, how the pages are parsed, how different types of data are marked up, etc.

[**0114**] Reference is now made to FIG. **8** which is a block diagram illustrating an apparatus for collecting attitude-data from web site(s) according to a preferred embodiment of the present invention.

[**0115**] An apparatus according to a preferred embodiment of the present invention comprises one or more crawler(s) **801**. Each crawler may be assigned to respective pre-selected web site(s) **800** for crawling, to locate and download relevant web pages carrying attitude-data therefrom.

[**0116**] Each of the downloaded web pages is then put in a parsing queue **803** where from, in its turn, the page is parsed by a parser **805**. Preferably, the parser **805** is configured to parse a web page and create a mark-up language base representation of the web page. Preferably, the mark-up language is XML. The parser may be further configured for forwarding the parsed page(s) to an update queue **806**.

[**0117**] An apparatus according to a preferred embodiment also includes a data integrator (updater) **807**. Preferably, the data integrator **807** is configured for fetching the parsed pages data from the updates queue **806**, integrating the data by handling redundancy of data, preventing from keeping duplicate pages etc. Integrating the data may further include ensuring complete download of all relevant data bearing pages of the web site(s)—such as next pages, navigation from forum to topic and then to the message itself etc.—utilizing a request queue **809**, ensuring that the attitude-data is error free, verifying that the parsing is successfully completed etc. Finally, the data integrator updates a database (DB) or a data warehouse (DW) **810** with the parsed pages carrying the attitude-data.

[**0118**] A preferred embodiment of the present invention may further include a crawl manager/scheduler **811** which manages the crawler(s) **801** and schedules the crawling of pre-selected web page(s) according to the request queue **809**, utilizing a downloads queue **813** to be used by the crawler(s) **801**. Preferably, the request queue is managed by a collecting policy definer **160**, preferably implemented as a management console **815**.

[**0119**] This crawl manager/scheduler **811** is responsible for scheduling the download process (i.e. request per time quantum per site); in addition it is responsible for the order of pages being downloaded.

[**0120**] Network traffic is carefully monitored by the various web sites, and trying to avoid generating over traffic on the downloaded web sites, a carefully schedule may be implemented for the download process. The Crawl Manager **811** is responsible for the scheduling and verifies that a pre-defined time interval is kept between one access to a certain web site and the other, in that way the generation of network overload on the crawled web sites by the crawling may be avoided.

[**0121**] In addition, employing several crawlers **801** together allows parallelism in the downloading process, downloading many web sites in parallel while accessing each one only once in a while.

[**0122**] Also, as the ratio of new user post pages (documents) to exiting pages is not very high, an updated list of the new post pages may be maintained and used for further reducing crawling activities on the crawled web sites.

[**0123**] Reference is now made to FIG. **9** which shows an exemplary collecting policy definer graphical user interface (GUI), according to a preferred embodiment of the present invention.

[**0124**] According to a preferred embodiment of the present invention, the collecting policy definer **160** includes a graphical user interface (GUI), which graphically facilitates the definition of a collecting policy by a user of the apparatus **1000**.

[**0125**] In the exemplary collecting policy GUI of FIG. **9**, on the top of the screen there is a settings bar where the user inputs the address of the web site page **910** for collecting attitude-data there from, the output file **920** to save the results in, and the page type **930**.

[**0126**] Below the setting bar there is a window **950** where the user may provide other definitions. For example: The color coding for each of the marked fields, the date format being used in this particular site (e.g. European, American, or other) and optionally other relevant definitions.

[**0127**] Below the window **950** is the main working area of the application **980**. The main working area **980** has the

behavior of a browser and loads a web site page so as to allow the user to define the specific collecting policy with regards to the specific web site page.

[0128] When the page is loaded the user may mark the relevant parts on the page, indicating what section reflects what part of information to be crawled, or optionally, to be ignored. This operation is preferably repeated for each part of a web site (i.e.: forums list, topic list, message pages, author profiles, etc.).

[0129] According to a preferred embodiment of the present invention, the collecting policy definer **160** is configured to use the definitions made using the GUI, as described hereinabove, for generating a script encoded collecting policy. The collecting policy may specifically define how each element on the page is crawled or parsed.

[0130] Reference is now made to FIG. **10** which shows an exemplary Web site page.

[0131] The exemplary page (http://dvdtalk.com/forum/forumdisplay.php?f=8) is a Forum web site page. The exemplary page has several main parts: a header, headlines, banners, a quick launch area for starting frequently used forums, and a list of forums.

[0132] Using the GUI of the collecting policy definer **160** the user may graphically select part(s) of the web site page and define a collecting policy for the part(s) as well as for the whole page.

[0133] Reference is now made to FIG. **11** which shows an exemplary user marked web site page, according to a preferred embodiment of the present invention.

[0134] The web site page of FIG. **10** is now presented having its main parts graphically selected and marks by the user.

[0135] With regards to the exemplary page, the user may define that only the elements of the list of forums **11-10** are to be crawled and parsed.

[0136] According to a preferred embodiment, each element is regarded by the collecting policy definer **160** as having a position relative to a parent element.

[0137] In the example of FIG. **11**, each element **1111-1112** of the list of forums **1110** has a relative position with respect to the header of the list **1120**. Consequently, when the absolute position of the header is changed, say when a new advertisement banner is positioned by an operator of the web site, just above the list of forums, the relative position of each element on the list remains the same.

[0138] Reference is now made to FIG. **12** which shows an exemplary relative title position encoding in a change query language script according to a preferred embodiment of the present invention.

[0139] The provided exemplary position is relative to a header of an html table. The table header has fixed position on the page, and the shown exemplary relative title position is encoded in relation to the fixed position. The parser **805** uses the definition provided by the user as illustrated in FIG. **10-11** and explained herein above, to correctly encode a generic title position relative to the fixed position of the table header.

[0140] Reference is now made to FIG. **13** which shows an exemplary pseudo-code, for finding the specific element definition in the collecting policy GUI (FIG. **9**) on a specific web site page, according to a preferred embodiment of the present invention.

[0141] The provided exemplary pseudo-code describes a sequence of steps for collecting policy GUI (FIG. **9**) to extract from the exemplary web site page that may be a part of the collecting policy, encoded in a script, based on user provided definitions, as illustrated and explained using FIGS. **10-11** hereinabove.

[0142] According to a preferred embodiment of the present invention, the collector **110** is configured for carrying out several steps of processing with regards to the collected attitude-data.

[0143] Reference is now made to FIG. **14** which is a flowchart illustrating attitude data processing according to a preferred embodiment of the present invention.

[0144] According to a preferred embodiment of the present invention, the attitude data is processed in a pipeline mode, wherein each document/message in the crawled web pages undergoes a series of steps that are applied to it in a row.

[0145] According to a preferred embodiment, Internet sites **1400** are crawled for new messages, bearing attitude-data, based on a script in a change query language as described herein above.

[0146] Preferably, any given web page may be downloaded using HTTP protocol. However, the page has to be parsed in order to extract its information. This is already the role of the above described parser.

[0147] The parser may represent the downloaded web page as a XML tree, and apply a change query language script, specifically customized for each web site, to extract the relevant information from it, skipping all the non-relevant info.

[0148] For example, the change query language may be an Extensible Style sheet Language Transformation XSLT language, which is a language for transforming XML documents into other XML documents.

[0149] The XSLT script may have the ability to ignore all kind of non relevant data, based on an ad-hoc customization, as described in greater detail herein below, for the collecting policy definer.

[0150] The relevant pages are downloaded and parsed **1401** to identify their relevant text section and the metadata relating to the new attitude-data such as: title, author, or date, is extracted from the collected attitude-data.

[0151] The processor **120** may further include a runtime environment which may be further configured for labeling each message/document with relevant metadata.

[0152] Then, the processor **120**, using an on-line interface, categorizes **1410** relevant texts of the collected attitude-date using supervised approaches.

[0153] Next, the processor **120** carries out classical text categorization by content, which involves assigning each message/document a list of topics being discussed in it, based on the identification and analysis of issues discussed

in the collected attitude data. In addition, processor **120** carries out text categorization by sentiment, which involves assigning each message/document its polarity label (positive, negative or neutral).

[**0154**] According to a preferred embodiment of the present invention, the content based categorization of the collected attitude-data may be based on an output generated by a training/testing environment which may be a part of the processor **120**, and may be used to form the model for categorizing the attitude expression, i.e. the logic of how to identify titles, topics, age groups, gender etc, as described in greater detail herein above.

[**0155**] Optionally, the processor **120** may utilize one of the text categorization techniques in a range which includes but is not limited to: Feature Selection, Feature filtering, and Training as described in greater detail herein below.

[**0156**] In a preferred embodiment, the processor **120** is also configured to carry out text categorization by style technologies. Such technologies may add and categorize vital data about the document author, like his age or his/her gender, without having any direct background knowledge about the author.

[**0157**] Categorization by style technologies are based on the idea of analyzing the writing style, the language used by the author, the use of foreign language words etc. to indirectly learn about the author. Learning about the author, the attitude data may be categorized according an age group, gender, etc.

[**0158**] Style text categorization may enrich the queries and analysis the end user can perform on the data. Since this style derived information is static, it may be generated in a metadata pre-processing stage as well.

[**0159**] According to a preferred embodiment of the present invention the processor **120** may include a Statistics Generator for generating various statistics relating to the collected attitude-data.

[**0160**] Preferably, the processor **120** includes data mining tools for mining **1412** the collected and processed attitude-data, so as to provide a user with means for carrying out pattern analysis and trend detection **1430** in the attitude data.

[**0161**] According to a preferred embodiment of the present invention, the processor may implement any of the methods described hereinbelow for categorizing the texts of the attitude-data and for further analyzing the attitude-data, say for providing statistics relating to the attitude-data or for mining the attitude-data.

[**0162**] The results of the categorization and data mining steps are output and stored in a data storage (a database or a data warehouse) **1420**.

[**0163**] Preferably, the processor **120** may further include a concept analyzer, operable by an analyst/user **1450** for concept analyzing **1431** the attitude data, for finding in the attitude-data relevance-relationship(s) between a phrase, comprising one or several words, and a respective concept, as described in greater detail hereinbelow.

[**0164**] More preferably the processor **120** may also include a correlation measurer, configured for measuring **1432**, in the attitude-data, correlations among phrases having relevance-relationships with a common concept, and for measuring correlation between one or more of these phrases and the common concept, as described in greater detail herein below.

[**0165**] According to a preferred embodiment of the present invention, the processor **120** may further include a quotation extractor, for extracting **1433** from the attitude-data key quotations which are found to be descriptive of a relevance-relationship existing in the attitude-data between a concept and respective phrases (comprising one or more words), as described in grater detail herein below.

[**0166**] According to a preferred embodiment, the processor may further include a clusterer. The clusterer may be operable by a user/analyst **1450** for clustering concepts **1434** relating to the attitude-data, as described in greater detail herein below.

[**0167**] According to a preferred embodiment of the present invention, the outputter **130** provides a user **1440** or an analyst **1450** with various graphical tools for examining, exploring, and analyzing attitude-information, generated by collecting and processing the attitude-data. Optionally, the graphical tools may be provided as a web application **1442**, so as to allow the user to examine and explore the attitude data remotely via the web.

[**0168**] Reference is now made to FIG. **15** which shows an exemplary graphic representation of the results of clustering concepts in the attitude data, according to a preferred embodiment of the present invention, as described hereinabove.

[**0169**] With clustering, individual messages are analyzed for a central attitude and then added a corresponding cluster of attitudes.

[**0170**] In the central part of the screen the user can see the generated clusters as circles **1501**, clusters with more messages/documents are denoted as bigger circles, their distance is displayed by their visual layout. Clusters that are in the red-region **1503** are clusters of negative attitude, while positive attitude ones are in the green part **1505**.

[**0171**] On the left screen side, the user can see the topic of each cluster **1507**. Clicking on one of the clusters displays to the user a set of relevant message/document citations for each of the clusters.

[**0172**] Reference is now made to FIG. **16** which shows an exemplary graphic representation of the results of correlation measurement according to a preferred embodiment of the present invention.

[**0173**] The correlation measurer, discussed hereinabove, measures correlations of relevant phrases for a central concept as well their cross-relationships. An exemplary visualization of results of the measurement is shown in FIG. **16**.

[**0174**] In the center is the main concept ("USA") **1601** surrounded by words indicating anti-American attitude expression in the web. The colors describe the various phrase types that are related to the central term, and their cross relations, according to a provided legend **1605**. Optionally, the layout algorithm may be based on a SVD (factor analysis) formula combined with MDS (multi dimensional scaling), wherein an n×n matrix is used to measure the

distances between each pair among the relevant phrases, n denoting the number of phrases.

[0175] Reference is now made to FIG. **17** which shows a first exemplary graphic representation of attitude-data analysis according to a preferred embodiment of the present invention.

[0176] According to a preferred embodiment the outputter **130** includes a user friendly graphical front end environment for defining and viewing attitude-information. Preferably, there are two types of front end: a desktop application and a web based client.

[0177] For example, the front end environment may provide a user with means for tracking trends, buzz, and sentiment, which are preferably based on the data warehouse **150** capabilities such a multidimensional data analysis.

[0178] Users may analyze their company's\product's word of mouth over time according to the different markets and vertical markets. Such analysis may prove very beneficial for the users.

[0179] In addition the user has the ability to compare his company\product to other products or companies in his vertical market or to a benchmark, set according to an industry standard. For example, as illustrated in FIG. **17**, the user may investigate the concept of the top ten movies **1701**, as depict in a chart showing the trend among the ten most popular in a monthly basis **1703**.

[0180] Reference is now made to FIG. **18** which shows a second exemplary graphic representation of attitude-data analysis according to a preferred embodiment of the present invention.

[0181] Preferably, more advanced capabilities then the ones presented in FIG. **17** are available for the advanced user. For example, as shown in FIG. **18**—analysis according to gender **1801**, analysis according to age range **1803**, selection of chart types **1805**, selection of axis data **1807**, etc. are further available more advanced capabilities.

[0182] For example, when the user chooses to analyze the sentiment with regards to his product according to gender **1801**, with respect to all age groups (combined) **1803**, he may be presented a bar chart **1810** depicting the positive vs. negative vs. natural attitudes towards his produce.

[0183] Reference is now made to FIG. **19** which is a flow diagram of an exemplary method for analyzing attitudes expressed in web sites, according to a preferred embodiment of the present invention.

[0184] According to a preferred embodiment, attitude data **1900** relating to a subject which is predetermined by a user, say using the apparatus **1000**, is collected **1901** from pre-selected web site(s), say by a collector **110**, as described hereinabove.

[0185] The pre-selected web sites may include, but are not limited to: Chat sites, Interactive news groups, Discussion groups, Forums, blogs and the like where people express their views and feelings. For example: Internet uses may express their views regarding a proposed tax reform, to be discussed by a government, regarding a new product etc.

[0186] Optionally, the collecting may include any number of web sites.

[0187] Next, the collected attitude-data is processed **1903**, say by a processor as described hereinabove.

[0188] The processing **1903** of the attitude-data may typically include content analysis techniques, data mining, and other data analysis techniques. These techniques may implement any one a variety of algorithms, which includes but is not limited to: neuronal networks, rule reduction, decision trees, pattern analysis, text and linguistic analysis techniques, or any relevant known in the art algorithm. Detailed exemplary algorithms, usable for processing of the attitude-data are provided herein below.

[0189] Finally, the processed attitude-data is used for outputting **1905** attitude-information to a user, say by an outputter **130**, as described hereinabove.

[0190] The outputting **1905** may be carried out utilizing graphical tools for presenting and analyzing attitude-information, as described in greater detail hereinabove.

[0191] According to a preferred embodiment of the present invention, the collecting **1901** may include crawling the web sites according to a predefined policy. Preferably, the collecting further includes parsing relevant downloaded pages of the pre-selected web sites, as described in greater detail hereinabove.

[0192] Preferably, the crawling is carried out according to a policy defined by a user, say by a collecting policy definer **160**, as described hereinabove.

[0193] According to a preferred embodiment the processing **1903** is carried out in an initial pre-processing step, where metadata relating to the collected attitude-data is processed in advance.

[0194] According to a preferred embodiment, the processing **1903** includes categorizing relevant text of the collected attitude-date using supervised approaches.

[0195] Preferably, in addition to classical text categorization by content, which involves assigning each message/document a list of topics being discussed in it, a preferred embodiment may include using text categorization by style technologies. Such technologies may add and categorize vital data about the document author, like his age or his/her gender, without having any direct background knowledge about the author.

[0196] As described herein above, categorization by style technologies are based on the idea of analyzing the writing style, the language used by the author, the use of foreign language words etc. to indirectly learn about the author.

[0197] Style text categorization may enrich the queries and analysis the end user can perform on the data. Since this style derived information is static, it can be generated in a metadata pre-processing stage.

[0198] Reference is now made to FIG. **20** which is a flow diagram of an exemplary method for categorizing attitude-data text according to a preferred embodiment of the present invention.

[0199] The general flow of the exemplary categorization process includes: data manipulation **2001**, and then feature selection **2003** and feature reduction **2005**, applied, as described in greater detail hereinabove, for yielding a feature set/cluster **2010**. The example further includes train\test

2015 procedures for forming a model which best represents the attitude-information in the collected attitude-data.

[0200] Data Manipulation

[0201] Texts cannot be directly interpreted by a classification system. Because of this, an indexing procedure that maps a text into a compact representation of its content is preferably uniformly applied to training, validation, and testing of messages/documents, for successfully carrying out the categorization and mining of the attitude data.

[0202] The choice of a representation for text depends on what one regards as the meaningful units of text (the problem of lexical semantics) and the meaningful natural language rules for the combination of these units. Similarly to what happens in IR (Information Retrieval), in TC (Text Categorization) a text may be represented as a vector of pairs of terms and their weights. Each of the document terms (sometimes called features) occur at least once (in at least one message/document). There are different ways to understand what a term is and different ways to compute term weights.

[0203] A typical way for understanding a term is to identify the term using a word. The way is often referred to as either the set of words or the bag of words approach to document representation, because a bag or set of words is available from which to select the meaning of the term. With the bag of word approach, a list of words and word combinations is weighted according to the number of appearances of each word or word combination in the document. Predefined stop words/combinations are then excluded from the list, and the term is understood in light of the weights of the remaining words/combinations.

[0204] Feature Selection

[0205] Feature selection may relate to various types of features ranging from textual ones, like words, dictionary based words and also some more grammatical features like part-of-speech tags and their combination. Preferably, Feature selection further includes combinations of phrases, represented as N-grams. N-grams are phrases combining a number (n) of words.

[0206] Feature Filtering

[0207] Unlike in text retrieval, in TC the high dimensionality of the term space may be problematic, as the objective of TC is to extract an attitude from a mass of words rather than to search for a given phrase. In fact, while typical algorithms used in text retrieval can scale up to high values of terms, the same does not hold of many sophisticated learning algorithms used for TC which is about extracting the general attitude rather then its detailed expression.

[0208] Preferably, because of this problem, a Feature filter is also implemented. The effect of the filtering is to reduce the size of the term space. The filtering may apply methods for feature reduction that include but are not limited to: dictionary based reduction, term frequency reduction, and information-gain filtering.

[0209] With dictionary based reduction, a limitation is made to a certain group of words that appears in a predefined dictionary words list (like function words).

[0210] Term frequency reduction is based on filtering out features that appear in too many messages/documents, such as "I" and "The", or in too few messages/documents. That is to say, terms that appear in too many messages are regarded as too general whereas terms that appear in too few messages are regarded as too specific. Information gain filtering measures the decrease in entropy as a result of the presence of a certain term in the text. This is useful to identify the features that are best distinguishing between groups in the space of documents/messages.

[0211] For example, entropy may be formally defined as:

$$IG(t) = \sum_{1}^{m} P(C_i)\log P(C_i) + P(t)\cdot\left[\sum_{1}^{m} P(C_i \mid t)\log P(C_i \mid t)\right] + P(\bar{t})\cdot\left[\sum_{1}^{m} P(C_i \mid \bar{t})\log P(C_i \mid \bar{t})\right]$$

Where:

C denotes a category.

$$P(C_i) = \frac{\#\,docs \text{ in category } C}{\#\,docs \text{ in all categories}}$$

$$P(t) = \frac{\#\,docs \text{ where } t \text{ appears}}{\#\,\text{all } docs}$$

$$P(C_i \mid t) = \frac{\#\,docs \text{ where } t \text{ appears in } C_i}{\#\,\text{all } docs \text{ where } t \text{ appears}}$$

$$P(\bar{t}) = 1 - P(t)$$

$$P(C_i \mid \bar{t}) = \frac{\#\,docs \text{ where } t \text{ does not appears in } C_i}{\#\,\text{all } docs \text{ where } t \text{ does not appear}}$$

[0212] Train\Test Procedure

[0213] Preferably, one or more machine learning algorithms is applied on the data set to find a model which best extracts attitude data from the messages/document downloaded from the crawled web sites.

[0214] For example, given a collection of messages/documents discussing "sports" and "non-sports", the model learns how to distinguish sport messages/documents from non-sport ones.

[0215] In order to do this several models of text categorization may be applied in including but not limited to: Decision Tree (J48), Naïve Bayes, and SVM.

[0216] Decision Tree—a decision tree (DT) for text categorization is a tree in which internal nodes are labeled by terms, branches departing from them are labeled by the weight that the term has in the test document, and leafs are labeled by categories.

[0217] Such a tree categorizes a test document by recursively testing the weights that the terms labeling the internal nodes have in a vector, until a leaf node is reached. The label of this node is then assigned to the document. Most such trees use binary document representations, and are thus binary trees.

[0218] There are a number of standard packages for DT learning, and most DT approaches to TC have made use of such packages. Among the most popular ones are ID3 (used

by Fuhr et al. [1991]), C4.5 (used by Cohen and Hirsh [1998], Cohen and Singer [1999], Joachims [1998], and Lewis and Catlett [1994]), and C5 (used by Li and Jain [1998]).

[0219] Naïve Bayes—Let X be the data record (case) whose class label is unknown. Let H be some hypothesis, such as "data record X belongs to a specified class C." For classification, we want to determine P (H|X)—the probability that the hypothesis H holds, given the observed data record X.

[0220] P (H|X) is the posterior probability of H conditioned on X. For example, the probability that a fruit is an apple, given the condition that it is red and round. In contrast, P(H) is the prior probability, or a priori probability, of H.

[0221] In this example P(H) is the probability that any given data record is an apple, regardless of how the data record looks. The posterior probability, P (H|X), is based on more information (such as background knowledge) than the prior probability, P(H), which is independent of X.

[0222] Similarly, P (X|H) is posterior probability of X conditioned on H. That is to say, it is the probability that X is red and round given that we know that it is true that X is an apple. P(X) is the prior probability of X, i.e. it is the probability that a data record from our set of fruits is red and round.

[0223] Bayes theorem is useful in that it provides a way of calculating the posterior probability, P(H|X), from P(H), P(X), and P(X|H). Bayes theorem may be formally defined by the equation:

$$P(H|X) = P(X|H)P(H)/P(X).$$

[0224] SVM—The support vector machine (SVM) method has been introduced in TC by Joachims [1998, 1999] and subsequently used by Drucker et al. [1999], Dumais et al. [1998], Dumais and Chen [2000], Klinkenberg and Joachims [2000], Taira and Haruno [1999], and Yang and Liu [1999].

[0225] In geometrical terms, it may be seen as an attempt to find, among all the surfaces _1, _2, : : : in j. T j-dimensional space that separate the positive from the negative training examples (decision surfaces), the _i that separates the positives from the negatives by the widest possible margin. That is to say, such that the separation property is invariant with respect to the widest possible translation of _i.

[0226] This idea is best understood in a case where the positives and the negatives are linearly separable, in which the decision surfaces are (jT j−1)-hyper planes.

[0227] The SVM method chooses the middle element from the "widest" set of parallel lines, that is to say, from the set in which the maximum distance between two elements in the set is highest. It is noteworthy that this "best" decision surface is determined by only a small set of training examples, called the support vectors. The method described is applicable also to a case where the positives and the negatives are not linearly separable.

[0228] As argued by Joachims [1998], SVM offers two important advantages for TC: One being that term selection is often not needed, as SVM tends to resistant to overfitting—that is to producing a too complex statistical model compared with the amount of data, and can handle large dimensionality, and the other being that no human and computer processing effort in parameter tuning on a validation set is needed, as there is a theoretically motivated default choice of parameter settings which has also been shown to provide the best effectiveness.

[0229] The above described methods and algorithms are usually implemented in an on-line supervised manner, involving an analyst/user. A preferred embodiment of the present invention further implements unsupervised approaches. Preferably the unsupervised approaches facilitate processing relatively large volumes of textual attitude-data.

[0230] A preferred embodiment of the present invention involves unsupervised approaches that are based on data mining techniques.

[0231] A preferred embodiment of the present invention may utilize a two layers approach. One layer is an application layer and the other is an open query layer where the user may define relevant queries.

[0232] The application layer may use, but is not limited to using:

[0233] Data representation—a data representation component may be used for internally representing text of the attitude-data.

[0234] Memory and performance efficient data-structures are essential for performing the complex online analysis tasks. The data representation component translates the text to a compact binary representation, enabling faster analysis, for example using following steps.

[0235] Frequency analysis—a frequency analyzer may be used to provide the user with various statistics on different parameters, like: most frequent words, phrases, number of authors, unique authors, or distribution over time frame. The frequency analyzer may utilize a counter for counting words, phrases, etc. The counter provides raw data that is then processed by the frequency analyzer, to generate various statistics data.

[0236] Concept Analysis—a concept analyzer may be employed for finding the most interesting and relevant phrases relating to a certain concept, in the attitude-data.

[0237] The analysis handles single word phrases as well as relevant multiple word phrases. The concept analyzer may scan all the words or phrases in the collection, and assign a relevance score to each of them, to indicate relevance of the word or phrase to the researched concept.

[0238] Preferably, the relevance is measured by the ratio between a frequency for the word/phrase for co-occurring with a "leading concept/word" (i.e. the concept/word currently being analyzed) to the frequency of the co-occurrence not with the "leading concept/word". The higher this ratio is, the more relevant is this word/phrase.

[0239] In order to extract phrases (longer than one word), the analysis may include examining the top K (usually 100) words, and then look for phrases containing at least one of

the top K words. Those phrases whose relevance score (as being calculated for single words) is higher than a certain threshold are considered relevant.

[0240] Correlator measurement—according to a preferred embodiment, a correlation measurer may be used to reveal interesting relationships between phrases and concepts in the attitude-data.

[0241] When trying to analyze a concept, one of the important information is what is mentioned\related to this concept, and how these are issue-related. This is done by measuring correlation.

[0242] According to a preferred embodiment of the present invention, the relevant phrases that were identified in the concept analysis stage are populated in a matrix where the distances between all the pair of phrases are calculated, as described ion greater detail herein below.

[0243] Then, the matrix may be populated into a visual interface, with the analyzed concept/phrase in the middle, and the relevant phrases surrounding it, as illustrated in FIG. **14** and discussed hereinabove.

[0244] The distance from the central concept measures the relevance to it, and the distances among the other phrases themselves represents their closeness. These metrics are directly derived from the distances in the distance matrix, populated as described below.

[0245] Preferably, in order to calculate the distance between two phrases, two parameters are taken into consideration: the significance of the co-occurrence of these phrases and the frequency of this occurrence.

[0246] According to a preferred embodiment, the distance between phrases a and b is calculated according to the formula:

$$D(a,b)=\frac{freq(a,b)^{1.5}}{freq(a,\tilde{b})}$$

[0247] freq(a,b)=Frequency for a to co-appear with b, for some measure of togetherness

[0248] freq(a,$\tilde{b}$)=Frequency for a to appear where b does not appear

[0249] Note that D(a,b) is not symmetric with D(b,a).

[0250] In a preferred embodiment, a distance between the two phrases, as put in the matrix, is the maximum of the two:

$$DV(a,b)=\mathrm{Max}(D(a,b),D(b,a))$$

[0251] Quotation extraction—a quotation extractor is preferably employed for extracting key quotations from a data file that contains a given list of concepts, in order to provide a user with the relevant text citations best describing a relationship, existing in the attitude-data between a concept and its neighbor (relevant) phrases.

[0252] The challenge in the above case is identifying ad-hoc the most relevant documents, finding in them, the most relevant phrases and then displaying the phrases to the end user. The relevance in this case is measured by the frequency of the searched phrases in the text, in coordination with their distance in the message/document itself.

[0253] Clustering—according to a preferred embodiment, the concepts relating to the attitude-data may be clustered, say by a clusterer, as discussed hereinabove.

[0254] Clustering may include aimed clustering which includes clustering the concepts that strongly relate to a given topic. Clustering may also include free clustering where a given attitude-data set is clustered into distinctive groups which strongly relate to one another. This functionality is useful when analyzing new domains where the analyst doesn't have any prior knowledge on it.

[0255] Reference is now made to FIG. **21** which is an exemplary pseudo-code algorithm for clustering concepts relating to attitude-data, according to a preferred embodiment of the present invention.

[0256] Free clustering may be implemented using a clustering algorithm as exemplified using FIG. **21**, to provide the user with the list of most relevant document clusters in the collection, along with cluster names and list (and view) of the documents belonging to each cluster.

[0257] The algorithm of FIG. **21** has the following advantages over the classical clustering algorithms: no predefined fixed number of clusters as in the classical clustering algorithms, ability to control the words that build the different clusters, and ability to merge and split clusters.

[0258] A general well known problem of traditional clustering algorithms pertains to relevance of the generated clusters to needs of the end-user, and that the traditional algorithms are based on the end-user's previous knowledge of well known world facts.

[0259] The example algorithm enables the user to control the output and quality of the final clusters, thus overcoming these shortcomings.

[0260] According to a preferred embodiment of the present invention, the processing of the attitude-data further includes data mining techniques.

[0261] Preferably, the data mining techniques may include, but are not limited to Pattern analysis and Trend analysis.

[0262] With Pattern analysis the processing includes searching for patterns in the statistics that may be provided by a statistics generator as described hereinabove.

[0263] The process may reveal relationships that are not obvious or sift out meaningful data from noise, exploiting favorable patterns and avoiding bad ones. Pattern analysis is a traditional part of data mining algorithms as applied on data stored in relational databases. However, in a preferred embodiment, Pattern analysis is further applied to unstructured textual data.

[0264] With Trend analysis, the processing further includes detecting emerging trends in the attitude-data, like new emerging products, consumer habits and more.

[0265] Optionally, trend analysis may be done by applying linear regression principles on the data set results. Once a list

of related phrases is discovered, an analysis of correlation trends over time using linear regression is carried out

[0266] If a strong positive (or negative) correlation trend (by having a high absolute value of the correlation derivative) is discovered, it is checked for consistency over time, by measuring the mean squared error.

[0267] The phrases that have the strongest trend derivative, and the least error, are regarded as those with the higher trends, and are displayed to the user along with their trend graph, and regression equation.

[0268] Platform Architecture

[0269] Reference is now made to FIG. **22** which is a simplified block diagram of an exemplary architecture of an apparatus for analyzing attitudes expressed in web sites, according to a preferred embodiment of the present invention.

[0270] An architecture according to a preferred embodiment may be a distributed environment architecture having loosely coupled components **2221-5**, communicating through one central fault tolerant management and data center **2230**.

[0271] High availability of the data center **2230** is ensured by running the data center in a computer server cluster with redundant machines.

[0272] The central data center **2230** preferably runs on top of a central data storage (data base/data warehouse) **2235**, secured with redundant machines ensuring high-availability. The data-center **2230** stores the current system status and configuration (along with data to be analyzed) as well as the communication messages between the various system components.

[0273] Having a message based communication system enables full distribution of the various run time components, thus having full scaling capability. This architecture also enables real time configuration changes, affecting immediately all the running components without requiring a restart of the whole system or waiting for long update time.

[0274] Preferably, all the components communicate in an asynchronous mode, using messages. All the messages are posed to queues waiting for processing by each of the components. Each component owns one input message queue, one output queue and one management (commands) queue. The input queue contains the processing requests waiting for a component to be processed, upon completion, the processed document is posted to an output queue (which is actually the input for the next component in the pipeline).

[0275] An apparatus according to a preferred embodiment of the present invention may provide means for proper storage for any volume of data with fast access capabilities.

[0276] It is expected that during the life of this patent many relevant devices and systems will be developed and the scope of the terms herein, particularly of the terms "Collcetor", "Processor", "Outputter", "Database" and "data Warehouse", is intended to include all such new technologies a priori.

[0277] Additional objects, advantages, and novel features of the present invention will become apparent to one ordinarily skilled in the art upon examination of the following examples, which are not intended to be limiting. Additionally, each of the various embodiments and aspects of the present invention as delineated hereinabove and as claimed in the claims section below finds experimental support in the following examples.

[0278] It is appreciated that certain features of the invention, which are, for clarity, described in the context of separate embodiments, may also be provided in combination in a single embodiment. Conversely, various features of the invention, which are, for brevity, described in the context of a single embodiment, may also be provided separately or in any suitable subcombination.

[0279] Although the invention has been described in conjunction with specific embodiments thereof, it is evident that many alternatives, modifications and variations will be apparent to those skilled in the art. Accordingly, it is intended to embrace all such alternatives, modifications and variations that fall within the spirit and broad scope of the appended claims. All publications, patents and patent applications mentioned in this specification are herein incorporated in their entirety by reference into the specification, to the same extent as if each individual publication, patent or patent application was specifically and individually indicated to be incorporated herein by reference. In addition, citation or identification of any reference in this application shall not be construed as an admission that such reference is available as prior art to the present invention.

What is claimed is:

1. Apparatus for crawling web content to provide data for attitude analysis of attitudes expressed in said web content in relation to a predetermined subject, the apparatus comprising a crawler, configured to crawl a plurality of pre-selected web sites, for collecting attitude-data from said web sites, said attitude data comprising attitudes relating to said predetermined subject, said crawler being further configured to provide said attitude data to a predetermined location for said attitude analysis.

2. The apparatus of claim 1, further comprising a policy definer, operable for defining specific guidelines with respect to crawling said pre-selected web sites.

3. The apparatus of claim 1, wherein said crawler is further configured for downloading of relevant pages of the pre-selected web site according to a predetermined schedule.

4. The apparatus of claim 1, wherein said crawler is further configured to use a script in a change query language, for said collecting.

* * * * *