From news to facts: an hadoop-based social graphs analysis

Piera Laura Puglisi[†], Daniele Montanari[*], Alessandro Petrella[†], Marco Picelli[‡] and Daniela Rossetti[*]
[*]ICT eni - Semantic Technologies
Via Arcoveggio 74/2, Bologna 40129, Italy
Email: {daniele.montanari, daniela.rossetti}@eni.com
[†]GESP - Geographic Information System
Via Marconi 71, Bologna 40122, Italy
Email: {pieralaura.puglisi, alessandro.petrella}@external.eni.com
[‡]Overit Gruppo Engineering
Via Bassi 81 33080 Fiume Veneto (PN), Italy
Email: marco.picelli@overit.it

***Abstract*—This paper describes a system combining a distributed setup based on Hadoop, MapReduce, Impala and a general semantic model focusing on common entities (people, organizations, places) and their connections as co-occurrences and facts offering the analysts the opportunity to do mining in social networks. Emphasis is given to recall rather than precision, suggesting the analyst many possible relations and connections to be explored. Early studies have shown interesting results, and further explorations are planned to extend the reach and abilities of the system.**

I. Introduction

The ability to extract relevant facts from any text has always been one of the ultimate goals of any textual analysis [25]. Such a task is made increasingly difficult when the domains involved increase in number and similarity, or the entities have ambiguous features, or the relations are difficult to describe formally. Following the diffusion of technologies like the Hadoop distributed file system [15], the MapReduce framework [16], and the related frameworks supporting the so-called Hadoop ecosystem [26], a new impetus has been offered to handle larger datasets, provide faster analysis, with improved accuracy of the outcome.

From an application point of view, our effort is motivated by the requests from users for timely and relevant information supporting multiple activities like due diligence, business analysis, security, and more. Several forms of Business Intelligence are employed; one in particular is the collection and analysis of data, reports, analysis, and news about both macro topics (like world economies and politics), and micro topics (like individuals, companies, local groups, infrastructures, brands, etc), and specific domain interests like intellectual property rights, security, and safety related issues.

At ENI a system has already been developed and deployed to support the acquisition and processing of news and other specific types of content. These documents, acquired from internet sources and provider-syndicated information streams, are processed using a semantic technology for indexing and content categorization using standard taxonomies; based on this collection the analysts can perform a semantic search through the content. However, the outcome of such a search does not return a presentation of the interesting entities and their relations with other entities involved in events, to represent a story spanning much more than the single contribution of a document; this has become the goal of the project presented in this paper.

The extraction of the entities and the relations will result in a social graph which can be abstractly analyzed to facilitate future querying by the users. This results in a mix of asynchronous acquisition of new content, its analysis and positioning on the distributed storage, e.g. further analysis at the graph level, and concurrent querying by the users, which may result in a temporarily inconsistent data system.

This paper reports on the preliminary architecture being developed to build *facts* (about entities) and *relations* (between entities) extracted from the original documents, and provide a resulting view of possibly many entities combined together to form a social network of relevant entities. The crucial idea is to use a weaker form of relations, called co-occurrence, where two entities are connected (just like for a full relation) if they appear in the same sentence which can be characterized by a certain classification, and then apply Semantic Text Mining techniques and social graph analysis techniques to discover and efficiently represent the relations binding the entities recognized in text.

This approach still presents issues like the need to disambiguate the entities involved (for homonymy or synonymy) and the modeling of the domains, hence the use of semantic technologies is necessary. The use of a distributed system can be exploited to enhance and better scale the impact of the semantic capabilities.

The experiment has a Big Data connotation because of the volume, due to the great amount of data being processed, and also for the velocity required, since the documents are pushed into the system continuously and the analyst should be able to query them immediately after the acquisition. The early outcome tests seem interesting, and the risk of having some noise has been considered manageable, to present the analyst with an acceptable number of options in the social graph.

It should be noted that the visualization of the results is

978-1-4799-5313-4/14/$31.00 ©2014 IEEE

crucial for a proper management of a possibly very large network of connected entities. While this part is also included in this study, however it is not reported here.

The remainder of this paper is organized as follows. In section II, we report related works focusing on text mining techniques and on different hadoop-based architectures. Section III shows the current architecture of the solution. Section IV introduces the semantic definitions of co-occurrence and facts and the resulting model. The final section reports the results of our work together with conclusions and further extensions.

II. Related Work

In recent years, the terms Text Mining [20][21], Hadoop [15][9], MapReduce [16] and Big Data [18][19] have been widely used and can often be found in the same context. Particularly, the concept of Big Data has become more important and has inspired a large number of works. Many recent papers focus on storage and analysis of large data sets using Hadoop as distributed framework. In this section, we will briefly describe some related works that employe emergent technologies to solve Big Data problems. First, we will focus on papers that perform text mining and fact mining following a semantic approach similar to our semantic model; then, we will report some works that use Hadoop and MapReduce to support deep analysis of large data sets.

Text Mining and Fact Mining. In the context of semantic analysis, we use the Cogito semantic technology [2]. Specifically, we run the *text mining module* of Cogito API to extract co-occurrences and the *fact mining module* to derive facts from texts, as described in section II. Follows some related approaches that support similar semantic annotations.

Gloss (Global Semantic System) [7] is an integrated system for the analysis and retrieval of data in the environmental and public security domain. The goal of this system is to access and analyze large quantities of distributed multimedia data that may be relevant to the prevention of emergency situations. Following the approach of Kyoto [23], a system that allows to derive facts from text, Gloss has developed an architecture that processes and annotates data with enriched linguistic, semantic and geospatial information using NLP tools, semantic resources (WordNets) and a geographic database [7].

Another approach in the field of *fact mining*, is described in [24]. The authors present an approach for extracting ontological facts from texts using YAGO [27], a large ontology automatically derived from Wikipedia and WordNet [28]. It comprises entities and relations, and currently contains more than 1.7 million entities and 15 million facts. To extract ontological facts from a text, the system performs first a *semantic entities extraction* from text and then matches these entities with Yago facts.

Hadoop and MapReduce. Big data is a keyword to think about and to understand what are the right steps to do. Many works explored different solutions to handle big data issues and most of them converged to the adoption of Hadoop and MapReduce as emerging technologies.

IBM Watson Research has been working for some time on the topic of Big Data and already in 2008 presented DisCo (Distributed Co_clustering) [8], a pragmatic data mining process that involves data gathering, pre-processing, analysis, and presentation. Disco was developed using Hadoop with MapReduce. The paper introduces practical approaches for distributed data pre-processing, able to scale and process data on the order of petabytes.

Another big name as Facebook has worked with these technologies. In [6] the authors characterize the structure of the Facebook social graph confirming the six degrees of separation phenomenon on a global scale. The calculations were performed on a Hadoop cluster with 2,250 machines, using the Hadoop/Hive data analysis framework.

A recent work proposes BC-PDM [22], a system hadoop-based that supports parallel ETL process, statistical analysis, data mining, text mining and social network analysis. The authors present a three-levels architecture based on HDFS [31], Hbase [30] and Hive [29] as distributed file system. The results show good performances and the effectiveness of this approach on real mobile communication data.

Novelty of our approach. Although we adopt the same solutions of other recent works (e.g. a semantic technology), what differentiates our project from the others relies on the main, distinguishing targets of our system. First thing, we aim to analyze tens of millions of news, acquired from internet sources and provider based information streams, in order to convert them into *facts* and *co-occurrences* using a performing engine such as Cogito. This configures our project as dealing with Big Data issues. Second thing, we aim to discover hidden information about entities of interest involved in unusual events or find out unexpected connections, using inference rules and mining techniques. Finally, there is the need to give the user a quick response to support his/her analysis, so that fast processing time is another important requirement.

III. The architecture of the solution

The architecture of the solution is based on the Hadoop distributed file system and the MapReduce framework, which were the initial elements of the design. Following the very early experimental setup, a basic development system was configured, based on virtual Linux SLES11 SP1 machines and Cloudera 4.5 [32]. The reason for adopting a virtual environment was mostly due to the homogeneity with other development systems and the fact that for the time being we were not so much interested in performance, but rather in the correctness of the algorithms and the overall system. The performance evaluation remains as a relevant element for our upcoming investigations.

Figure 1 illustrates the current elements in the system that are briefly explained below.

A. Legacy Information Management System

Historically, the system was organized as shown in figure 2. Acquisition of structured and unstructured data comes from both intranet and external sources and processed via an ETL and a semantic pipeline. Results are placed in a

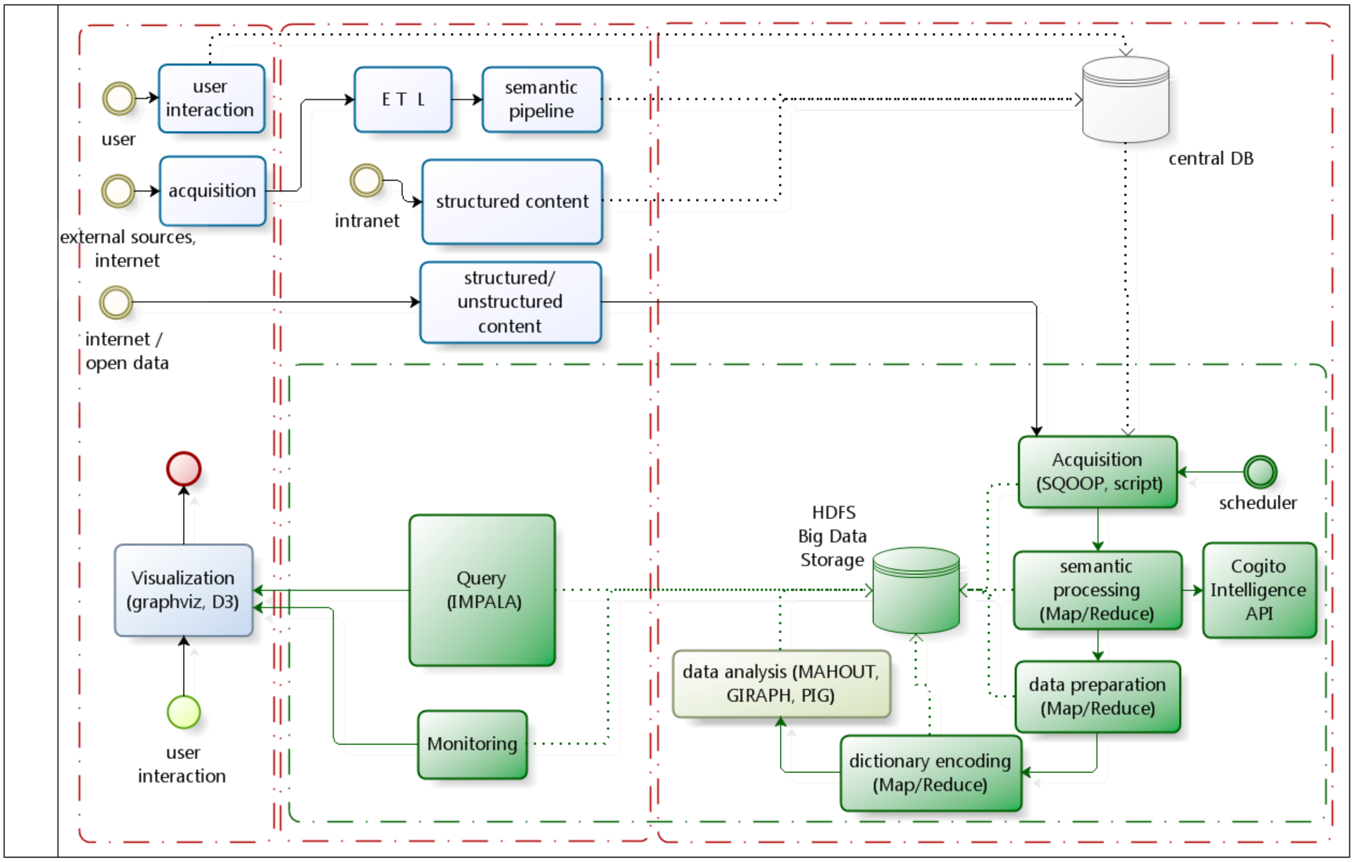

user interaction
user
ET L
semantic pipeline
central DB
acquisition
external sources, internet
intranet
structured content
structured/ unstructured content
internet / open data
Acquisition (SQOOP, script)
scheduler
HDFS Big Data Storage
semantic processing (Map/Reduce)
Cogito Intelligence API
Visualization (graphviz, D3)
Query (IMPALA)
data analysis (MAHOUT, GIRAPH, PIG)
data preparation (Map/Reduce)
Monitoring
dictionary encoding (Map/Reduce)
user interaction

Fig. 1. High level architecture

central database, accessed by the application front-end for user interaction and by the Hadoop system, to load data in the distributed file system in an asynchronous fashion. Future data flows could, if necessary, be directly loaded in HDFS.

B. Unstructured Analysis System

The architecture of the system has been enriched with new elements as shown in figure 3. Apache Sqoop [33] is a tool designed for efficiently transferring bulk data between Apache Hadoop and structured datastores such as relational databases. It is used to fetch data from the central databases collecting the content of interest (documents, news articles, etc.) which in previous ETL were downloaded and tagged as generically remarkable (at a minimum) under a preliminary evaluation.
To enrich the knowledge base in future version of the system, open data sources could be imported in Hadoop.

Via the Cogito Intelligence API component [3], deployed on all the nodes in the cluster to leverage parallelism, the unstructured documents are processed. By calling this component from within a map function, the analysis of the document is performed and the result stored in HDFS (for more details on the API refer to section IV).
This is the most CPU intensive step in the pipeline, where Hadoop plays a crucial role in the architecture: when the volume of data increases, the system can be scaled up by adding more nodes to the cluster.

The output of the previous step is then processed by custom MapReduce functions, in order to reorganize the data in different files, each representing a relational table.
Finally the data is compressed using dictionary encoding [34].

At this point, it is possible to apply different tools to the data with respect to the needs: for example, Giraph [35] has been used to create a graph with entities as nodes to compute shortest paths, following the bulk synchronous parallel paradigm [36].
Pig [37] is used to apply batch transformation and was previously used at query time, successively replaced by Impala [38] to provide an interactive, sub-second response time.
Finally, machine learning algorithms can be applied for experimenting using Apache Mahout [39].

C. Visualization

Visualization is currently done using GraphViz [40] while D3 [41] is under consideration for a more flexible solution (see figure 4).

IV. Semantic Analysis

A. Cogito Semantic Technology

Cogito is a technology developed by Expert System [1] to perform deep semantic analysis. It is based on a rich semantic

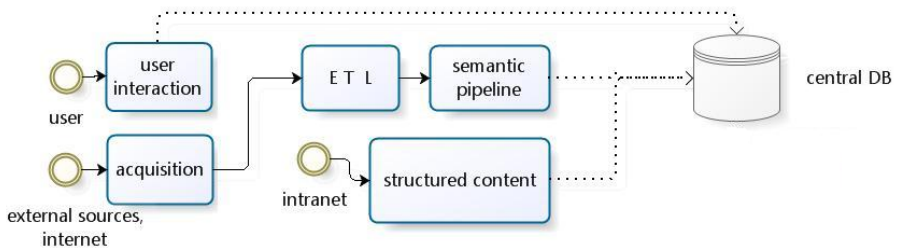


Fig. 2. Legacy system

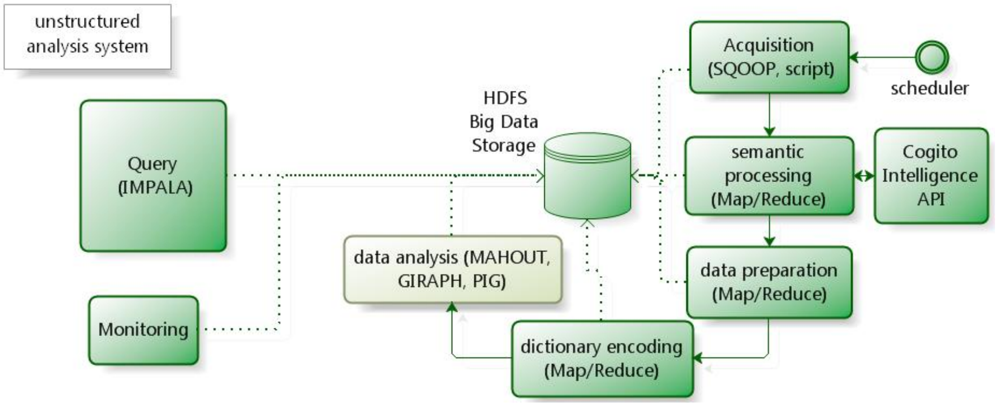

unstructured analysis system
Query (IMPALA)
Monitoring
HDFS Big Data Storage
data analysis (MAHOUT, GIRAPH, PIG)
dictionary encoding (Map/Reduce)
Acquisition (SQOOP, script)
scheduler
semantic processing (Map/Reduce)
Cogito Intelligence API
data preparation (Map/Reduce)

Fig. 3. Unstructured Analysis System

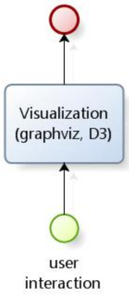


Fig. 4. Visualization

network that allows a complete understanding of a text, finding hidden relationships, trends and events inside the data. Thanks to the rich information discovered from texts, Cogito allows the transformation from unstructured information to structured data [2].

The heart of Cogito is the *Sensigrafo*, a semantic network available in different languages, which enables the disambiguation of terms. Sensigrafo contains more than 1 million concepts and more than 4 million relationships for the English language.

The Cogito semantic network includes common words, which comprise 90% of all content, and rich vertical domain dictionaries including Corporate & Homeland Security, Finance, Media & Publishing, Oil & Gas, Life Sciences & Pharma, Government and Telecommunications.

Cogito technology is employed in a family of tools and systems, the more recent of which is Cogito Intelligence API [3]. It is a semantic-based system with a knowledge domain dedicated to crime and intelligence including specialized classification for Intelligence, Geography, Entity and Relationships Extraction, Name Recognition and more.

COGITO Intelligence API is a classical Web Service that may be accessed by means of modular requests in standard XML format; the response is typically structured in different features, depending on the kind of the analysis performed on text: *text mining*, *categorization*, *semantic tagging*, *fact mining* and *extraction of entity and relationships*. These features can be integrated into platforms or applications in order to perform semantic analysis on documents, social data, web pages and other kind of unstructured information.

In this paper, we will focus on two of these services which are particularly interesting for our analysis: the *Text Mining* feature, for its capability to extract entities and their co-occurrences from the text, and the *Fact Mining* feature, which gives the possibility to identify events and entities related to them.

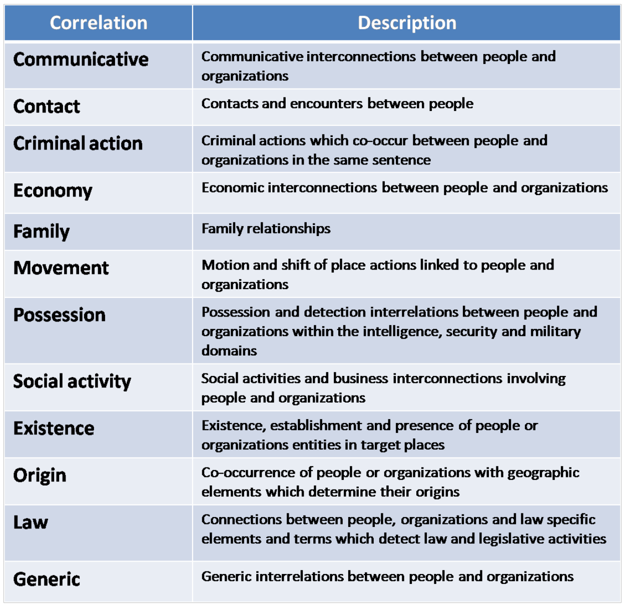
1) Text Mining: Text Mining feature performs the extraction of entities together with their associated semantic co-occurrences.

Entity extraction. The typical entities extracted by text mining service are the *standard entities* and the *domain entities*. Standard entities include *people*, *organizations*, *places*, *dates*, *addresses* etc. The semantic engine is also able to identify a proper name in a text and correlate it to its correct context. For example, "Arthur Andersen" may be categorized as "People" or "Organization" depending on the context [3]. Domain-specific text mining extracts specialized knowledge of several types of entities (e.g. Terrorist Organizations, Biological Agents, World Leaders, US Senators). Beyond the typology of entity, Cogito API extracts also related information such as sex, alias, human role etc. We categorize these additional information as 'attributes'. Formally, for each entity e, Cogito Intelligence associates a list of n attributes $a_1, ..., a_n$ denoted as Attributes(e).

Example. Given the text: "*After working his way through college with the help of scholarships and student loans, President Obama moved to Chicago*", Cogito API extracts the entities *Barack Obama* (people) and *Chicago* (place). Figure 5 shows the entity Barack Obama and his attributes (sex: M, human role: president etc.).

Co-occurrences Extraction. In addition to the extraction of entities, Cogito Intelligence API discovers connections between semantic entities. The focus of *Co-occurrences Extraction* is to detect and extract the occurrence of pairs of entities which are linked by a verb that identifies an action or activity. Cogito Intelligence API offers coherent suggestions and contextualization of acquired text with reference to specific co-occurrences, based on a set of 12 different types of interrelations summarized in figure 6.

In this paper, we denote a co-occurrence as a triple $(e_i, correlation_t, e_j)$ such that e_i and e_j are extracted from text t using the entity extraction feature, while $correlation_t$ belongs to the set of correlations listed in figure 6.



Correlation	Description
Communicative	Communicative interconnections between people and organizations
Contact	Contacts and encounters between people
Criminal action	Criminal actions which co-occur between people and organizations in the same sentence
Economy	Economic interconnections between people and organizations
Family	Family relationships
Movement	Motion and shift of place actions linked to people and organizations
Possession	Possession and detection interrelations between people and organizations within the intelligence, security and military domains
Social activity	Social activities and business interconnections involving people and organizations
Existence	Existence, establishment and presence of people or organizations entities in target places
Origin	Co-occurrence of people or organizations with geographic elements which determine their origins
Law	Connections between people, organizations and law specific elements and terms which detect law and legislative activities
Generic	Generic interrelations between people and organizations

Fig. 6. Typology of correlations.

Example. Following this notation, the co-occurrence in figure 5 will be denoted as $(Obama, Movement, Chicago)$. The correlation *Movement*, as reported in figure 6, represents a motion and shift of place action linked to an entity. In the example above, this correlation is given by the verb *moved* in the sentence: "President Obama *moved* to Chicago".

2) Fact Mining: Fact Mining feature allows identification of facts as well as the entities (people, organizations, places) and tags (URLs, phone numbers, emails, etc.) related to them. A fact is described through the domain and the topic representing the category it lies in, according to a specific facts taxonomy.

The Fact Mining service focuses on single sentences, assigning a category to them and providing extraction of entities, tags and domain-specific entities recurring in the sentence. In other words, the Fact Mining service performs Text Mining of Entities and Tags within specific text sections, thus identifying connections between relevant entities and specific contexts/domains.

Example. Figure 7 shows an example of a fact discovered from text: *"Peshraw Majid Agha, the Chairman of the Empire World and Falcon Group, describes this burgeoning real estate climate and its realtor with the lucrative oil sector: The KRG has invested heavily in their natural resources"*. Here, the set of entities (Peshraw Majid Agha - People, Falcon Group - Organization, KRG - Organization) are linked together by the expression *real estate* in a fact of type *Construction and Property* (domain).

Follows a formalization of the concept of *fact*.

Definition IV-A.1: (**Fact**). A fact f_i is defined as a finite set of elements (d,[t]; e_1,...,e_n) such that d is the domain of the fact, t is the topic (it is a more specific category of the domain and is optional) and e_1,...,e_n are the entities that co-occur in the fact.

Following this notation, the fact reported in figure 7 will be represented as:

(*Construction and Property*; Peshraw Majid Agha, FalconGroup, KRG).

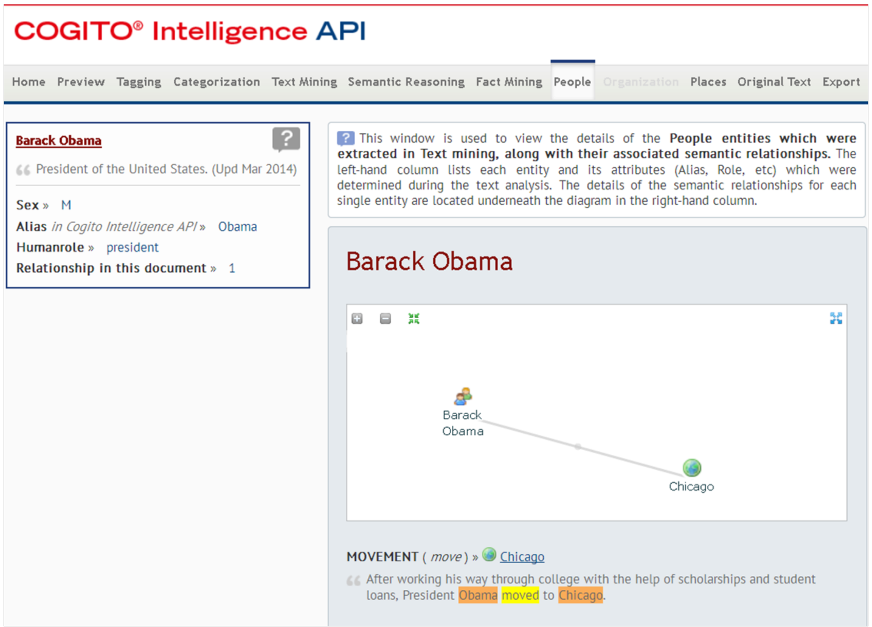
B. Semantic model

The text mining and fact mining features of Cogito API, described in the previous sections, have been integrated in our hadoop-based application. Since Cogito technology allows us to extract a set of relevant co-occurrences and facts, the goal of our system is to discover hidden relations or patterns from millions of news coming from our information sources (news, web pages etc.).

For the purpose of our work, we introduce a semantic model using an graph representation of *co-occurrences* and *facts*.

Suppose we have run Cogito API on a large corpus of documents from which we have extracted co-occurrences and facts. Let be e_i the entity that is the object of our analysis. We generate a social graph following these steps:

- For each entity e_j involved in at least one co-occurrence with e_i, we add an *entity node* n_{e_j} in the graph that maps the entity e_j.


COGITO® Intelligence API
Home Preview Tagging Categorization Text Mining Semantic Reasoning Fact Mining People Organization Places Original Text Export
Barack Obama
Chicago

Barack Obama

President of the United States. (Upd Mar 2014)

Sex » M

Alias *in Cogito Intelligence API* » Obama

Humanrole » president

Relationship in this document » 1

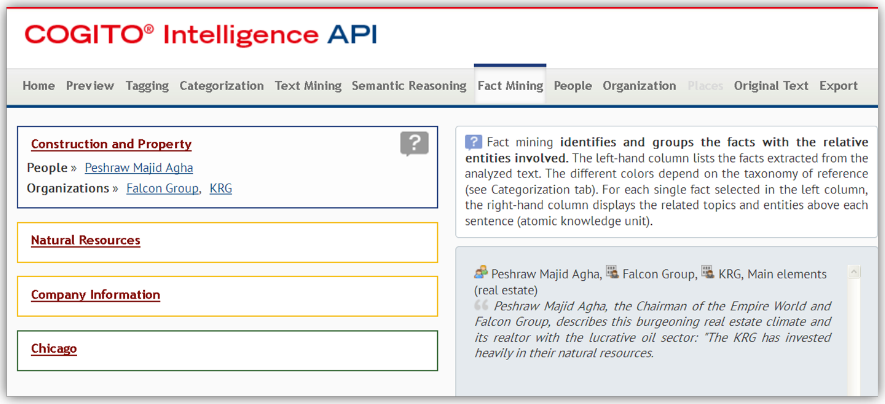
This window is used to view the details of the **People entities which were extracted in Text mining, along with their associated semantic relationships.** The left-hand column lists each entity and its attributes (Alias, Role, etc) which were determined during the text analysis. The details of the semantic relationships for each single entity are located underneath the diagram in the right-hand column.

Barack Obama

MOVEMENT (*move*) » Chicago

After working his way through college with the help of scholarships and student loans, President Obama moved to Chicago.

Fig. 5. Text Mining feature of Cogito API focused on People entity.


COGITO® Intelligence API
Home Preview Tagging Categorization Text Mining Semantic Reasoning Fact Mining People Organization Places Original Text Export

Construction and Property

People » Peshraw Majid Agha

Organizations » Falcon Group, KRG

Natural Resources

Company Information

Chicago

Fact mining **identifies and groups the facts with the relative entities involved.** The left-hand column lists the facts extracted from the analyzed text. The different colors depend on the taxonomy of reference (see Categorization tab). For each single fact selected in the left column, the right-hand column displays the related topics and entities above each sentence (atomic knowledge unit).

Peshraw Majid Agha, Falcon Group, KRG, Main elements (real estate)

Peshraw Majid Agha, the Chairman of the Empire World and Falcon Group, describes this burgeoning real estate climate and its realtor with the lucrative oil sector: "The KRG has invested heavily in their natural resources.

Fig. 7. Fact mining feature of Cogito API.

- For each co-occurrence $(e_i, correlation_t, e_j)$, we add an undirected edge between n_{e_i} and n_{e_j} labeled with $correlation_t$. We also report the cardinality of the co-occurrence, aggregating all correlations of the same type between the same pair of entities.
- For each fact f_j $(d, [t]; e_1, ..., e_i, ..., e_n)$ involving the entity e_i: we add a *fact node* n_{f_j} that maps the set of entities $e_1, ..., e_{i-1}, e_{i+1}, ..., e_n$ that co-occur with e_i in at least one fact of domain d and topic t;
we add a directed edge having n_{e_i} as source and n_{f_j} as target.

Figure 8 shows a social graph focused on entity *Naftomar* generated from a set of documents where the organization is reported. *Naftomar* is involved in:

- one co-occurrence of type *Behaviour* with the place *Athens*;
- three co-occurrences of type *Generic* with the place *Greece*;
- one co-occurrence of type *Generic* with the organization *Platts*;
- one fact of domain *Shipping Service* with the place *Egypt*.

As a further analysis, a user can navigate the graph and visualize more detailed information. For example, clicking on an edge between two *entity nodes*, the system will show the full texts from which the co-occurrences where extracted by the semantic engine. Clicking on a *fact node*, represented as a rectangle in figure 8, the system will report the full texts containing the extracted facts. Finally, it is possible to change

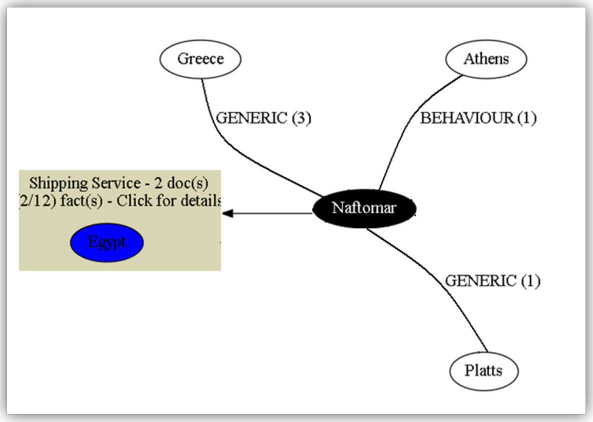


Fig. 8. Social graph focused on entity *Naftomar*.

the object of the analysis just clicking on another *entity node*: the system will generate a new social graph focused on the new entity of interest.

V. Conclusions and Future Work

This paper introduces a system that performs a semantic analysis on millions of news using Cogito Intelligence API. Cogito makes use of a text mining module and a fact mining module to extract co-occurrences and facts from texts. A semantic model is introduced to represent all extracted information as social graphs that can be navigated and queried by the user. The architecture of the solution is based on the Hadoop distributed file system and the MapReduce framework.

The system described in this paper is still quite experimental and exploratory, and we expect that further features will be introduced and tested. In particular, the architecture might be extended to include Apache Spark to exploit its graph handling capabilities. Apache Mahout or similar machine learning environment may also be employed to develop derived information from the original document intake. Further extensions may also be envisioned to support e.g. streaming operations.

References

[1] Expert System, *http://www.expertsystem.net/*

[2] Cogito Semantic Technology, *http://www.expertsystem.net/products-technology/cogito-technology*

[3] Cogito Intelligence API, *http://www.intelligenceapi.com*

[4] Gregory Mone, *Beyond Hadoop*, Communications of the ACM, 2013.

[5] Lipika Dey, Muhammad Abulaish, Jahiruddin and Gaurav Sharma, *Text Mining through Entity-Relationship Based Information Extraction*, IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, Pages 177-180, 2007.

[6] Johan Ugander, Brian Karrer, Lars Backstrom and Cameron Marlow, *The Anatomy of the Facebook Social Graph*, arXiv preprint arXiv:1111.4503, 2011.

[7] Francesca Frontini, Carlo Aliprandi, Clara Bacciu, Roberto Bartolini, Andrea Marchetti, Enrico Parenti, Fulvio Piccinonno and Tiziana Soru, *GLOSS, an infrastructure for the semantic annotation and mining of documents in the public security domain*, EEOP2012: Exploring and Exploiting Official Publications Workshop Programme. p. 21, 2012.

[8] Spiros Papadimitriou and Jimeng Sun, *DisCo: Distributed Co-clustering with Map-Reduce*, ICDM'08, Eighth IEEE International Conference on. IEEE, 2008.

[9] Tom White, *The definitive guide of Hadoop*, Media, 2009.

[10] Paul Zikopoulos and Chris Eaton, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*, McGraw-Hill Osborne Media, 2011.

[11] James Manyika et al., *Big data: The next frontier for innovation, competition, and productivity*, 2011.

[12] Andrew McAfee and Erik Brynjolfsson, *Big Data: The Management Revolution*, Harvard business review, 90.10: 60-68, 2012.

[13] Alvin W. Wolfe, *Social network analysis: Methods and applications*, Cambridge university press, 1994.

[14] Andrew McAfee and Erik Brynjolfsson, *Social Network Analysis For Organizations*, Academy of management review, 4.4: 507-519, 1979.

[15] Apache Hadoop, *http://hadoop.apache.org/*

[16] Dean, Jeffrey, and Sanjay Ghemawat, *MapReduce: simplified data processing on large clusters*, Communications of the ACM 51.1 (2008): 107-113.

[17] Abedini Farhad and Seyedeh Masoumeh Mirhashem, *From Text to Facts: Recognizing Ontological Facts for a New Application*, International Journal of Machine Learning and Computing, Vol.2, No.3, June 2012

[18] James Manyika et al., *Big data: The next frontier for innovation, competition, and productivity*, (2011).

[19] Paul Zikopoulos and Chris Eaton, *Understanding big data: Analytics for enterprise class hadoop and streaming data*, McGraw-Hill Osborne Media, 2011.

[20] Michael W. Berry and Malu Castellanos, *Survey of text mining*, New York: Springer, 2004.

[21] Ronen Feldman and James Sanger, *The text mining handbook: advanced approaches in analyzing unstructured data*, Cambridge University Press, 2007.

[22] Le Yu et al., *Bc-pdm: Data mining, social network analysis and text mining system based on cloud computing*, Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012.

[23] Kyoto, *http://kyoto-project.eu/xmlgroup.iit.cnr.it/kyoto/index.html*

[24] Farhad Abedini and Seyedeh Masoumeh Mirhashem, *From Text to Facts: Recognizing Ontological Facts for a New Application*, International Journal of Machine Learning and Computing, Vol.2,No.3,June 2012

[25] Norman Fairclough, *Analysing discourse: Textual analysis for social research*, London: Routledge, 2003.

[26] Monteith, J. Yates, John D. McGregor, and J. Ingram, *Hadoop and its evolving ecosystem*, Proceedings of the Fifth International Workshop on Software Ecosystems. 2013.

[27] Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum, *Yago: A large ontology from wikipedia and wordnet*, Web Semantics: Science, Services and Agents on the World Wide Web 6.3 (2008): 203-217.

[28] Christiane Fellbaum, *WordNet*, Blackwell Publishing Ltd, 1999

[29] Ashish Thusoo et al., *Hive-a petabyte scale data warehouse using hadoop*, In Data Engineering (ICDE), 2010 IEEE 26th International Conference on (pp. 996-1005). IEEE.

[30] GEORGE Lars, *HBase: the definitive guide*, O'Reilly Media, Inc., 2011

[31] Dhruba BORTHAKUR, *HDFS architecture guide*, Hadoop Apache Project, http://hadoop.apache.org/common/docs/current/hdfs design. pdf, 2008

[32] Cloudera, *http://www.cloudera.com*

[33] Apache Sqoop, *http://sqoop.apache.org*

[34] Jacopo Urbani, Spyros Kotoulas, Eyal Oren, and Frank van Harmele, *Scalable Distributed Reasoning using MapReduce*

[35] Apache Giraph, *http://giraph.apache.org*

[36] BSP, *http://en.wikipedia.org/wiki/Bulk_synchronous_parallel*

[37] Apache Pig, *http://pig.apache.org*

[38] Impala, *http://impala.io*

[39] Apache Mahout, *http://mahout.apache.org*

[40] GraphViz, *http://www.graphviz.org*

[41] D3, *http://d3js.org/*