Interactive Visualization of Non-formalized Data Extracted from News Feeds: Approaches and Perspectives

D.A. Kormalev, E.P. Kurshev, A.N. Vinogradov, S.A. Belov and S.V. Paramonov

Abstract In this article, we consider problems related to the task of interactive visualization of data extracted from news feeds, along with possible approaches to its solution. We describe the general concept of visualization taking into account the most recent developments in the field and provide a general description of our approach and the experimental implementation of a visualization system.

Keywords Data visualization · Information extraction · Natural language processing · News feeds

Introduction

The task of information visualization is a long-standing one, however when the information necessary for decision making is presented as non-formalized data (text documents, images, and multimedia), wide usage of such data poses a problem, as most of the existing visualization technologies are oriented at well-structured and normalized information, with an exclusion for some visualization techniques with niche applications.

D.A. Kormalev (✉) · E.P. Kurshev · A.N. Vinogradov
Ailamazyan Program Systems Institute of RAS, Pereslavl-Zalessky, Russia
e-mail: dk@conrad.botik.ru

E.P. Kurshev
e-mail: epk@epk.botik.ru

A.N. Vinogradov
e-mail: andrew@andrew.botik.ru

S.A. Belov · S.V. Paramonov
JSC "CTT Group", Moscow, Russia
e-mail: s.belov@cttgroup.ru

S.V. Paramonov
e-mail: s.paramonov@cttgroup.ru

K.V. Anisimov et al. (eds.), *Proceedings of the Scientific-Practical Conference "Research and Development - 2016"*, https://doi.org/10.1007/978-3-319-62870-7_10

Today's proliferation of digital documents and web expansion prioritize the development of a technology that will enable the use of information from large text collections in the process of decision making. According to various estimates, over 80% of commercial companies' information is presented in the form of text documents. In order to enable the visualization of the information from textual collections, text-to-data technologies are being developed, generally referred to as text mining technologies. These technologies rely on a family of techniques: information retrieval, information extraction, text classification and clustering, topic detection, data unification, etc.

Text-to-data techniques have a number of limitations that constrain the use of visualization methods. In particular, one key limitation is the level of granularity (structuring) of the output information. The granularity level must meet the requirements of the visualization task; however, the existing methods of text mining sometimes fail to provide the necessary level of information granularity (at least without damage to precision and recall). In this case, the degree of structuring is an indicator of how close we get to the concept of "data" (vs. text-based fragments). Another important limitation is the precision of the acquired data. Wrong input information may lead to wrong decisions; therefore the use of text mining technology requires a mechanism to protect against processing errors.

The focus of our study was the effective visualization of data coming from unstructured natural language text collections and feeds. The goal of the project was to develop a technology for interactive visualization of non-formalized heterogeneous data. The intended application of this technology is in decision support systems based on news feed monitoring and analysis. The project scope covered the following subtasks:

(a) develop functional models and algorithms for information object (IO) and link extraction from textual news feeds;
(b) develop functional models and algorithms for preprocessing and visualization of extracted IOs and links taking into consideration their multidimensional nature and heterogeneity;
(c) create an experimental software implementation of the above algorithms;
(d) develop specifications for future development of decision support systems taking advantage of the developed interactive data visualization technology;
(e) develop interface solutions to integrate the developed interactive data visualization technology with other software systems.

A variety of visualization technologies have been developed recently; however, there is no universal visualization method suitable for all decision-making tasks. The vast majority of existing software decision support systems follow some kind of problem oriented approach to visualization. Therefore, the relevance of our study is based on the fact that the task implies the research of visualization methods covering a maximum range of applications in the field of decision support.

The novelty of our research is based on a two-pronged approach to visualization of non-formalized heterogeneous data. The first part is to improve text mining

technology, and information extraction technology in particular. The second part is to search for new approaches to semi-structured information visualization, i.e., to develop visualization methods making it possible to work around the flaws of the text mining technology. Both parts are covered in our study. Such a two-way approach facilitates the search for new scientific and engineering solutions and will enhance the functional capabilities of modern decision support systems.

Visualization Concept

The practical aspects of intelligent text analysis (e.g., in application to web feed processing) should be considered in the context of a general applied problem from a certain field: business, administration, research, security, etc. An integral part of the system under consideration is the human expert who is able to synthesize findings from analysis outcomes and make decisions based on them. In this context, a text analysis system becomes an augmented intelligence tool rather than an artificial intelligence (AI) component. The difference is that the former does not replace the human intelligence but enhances its capability to perceive and assess large volumes of data. It is the critical difference between systems that enhance and scale human expertise rather than those that attempt to replicate all of human intelligence.

In the light of intensive development of AI systems, the scope of their application is being considered. The US government published a Request for Information (RFI) in June 2016 [4], asking the leading IT market players to present their views on the use of AI technologies in various social areas.

In response IBM proposed its vision [5] making a case for the critical role of human expertise to enable decision making in complex systems. The company highlights the need to focus on augmented intelligence technology within its own "cognitive computing" concept. Apart from the used technology stack, the key features of the technology are:

- the support for Big Data operations;
- a cognitive interface making it possible to include the human operator "in the loop".

The visual channel is the most natural way of human perception. Even after the IOs and links are extracted from the raw textual data, the resulting dataset is too abstract and incomprehensible for the operator. It is necessary to present this data in a visual form (view) and provide the tools for view interaction and manipulation.

Information visualization is the task of processing abstract data regardless of its nature or source. Input information goes through several stages of transformation producing a set of visual images to be displayed to the user. The visualization system can be described with a universal model covering general data stages and data transformations. In particular, when developing a domain- and application-oriented visualization system, such an approach makes it possible to describe specific system elements detailing their functions and requirements.

A widely used example of such description model is the Information Visualization Data State Reference Model [2]. This model includes the description of data stages and data transformations within an abstract technology allowing for visualization of arbitrary data.

The general structure of the Reference Model is provided in Fig. 1. The Data State Model breaks down each technique into four Data Stages, three types of Data Transformation and four types of Within Stage operators. The visualization data pipeline is broken into four distinct Data Stages: Value, Analytical Abstraction, Visualization Abstraction, and View. Transforming data from one stage to another requires one of the three types of Data Transformation operators: Data Transformation, Visualization Transformation, and Visual Mapping Transformation.

The main output of the natural language text processing is a set of extracted IOs and links. A natural way to present this information is to use graphs. The advanced theoretical base for graph processing is complemented by a broad range of practical

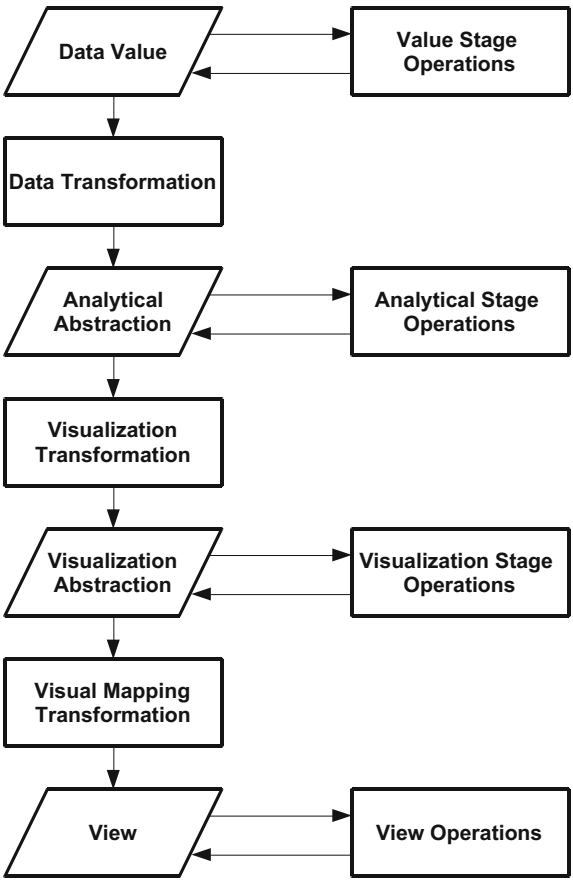

Fig. 1 Information visualization data state reference model

algorithms supporting graphs of different kinds. We can securely assume that natural language processing systems (and other systems with similar information nature and complexity) will use graph representation for analytical abstraction.

Further visual representation of this data may use the "natural" form (direct graph visualization) or other data representation techniques (charts, trees, heat maps, etc.). The choice of visual data representation depends primarily on the type of objects being displayed and on the nature of existing relations between them (e.g., hierarchical vs. non-hierarchical). Therefore, several models of visual representation are to be provided.

The exact combination of object layout and display (rendering) is chosen depending on the number of entities (graph nodes) being visualized, the level of detail, and the volume of metadata.

The characteristics of data representation models for each level, therefore, will be driven by design decisions regarding technology choices.

To make major technology and architecture decisions of a visualization system, one should consider the following questions.

Data access mode. Static data access implies that the abstract analytical data representation can be obtained right after the extraction from the semantic collection, and that it is immutable during further operations with the collection. So, all user queries regarding the visualization of a data set (subgraph) are routed to a static persistent data representation. Dynamic data access implies preparing a visual representation "on the fly," i.e., the queries are routed to the raw data that goes through transformations before being displayed.

The choice of visual representation means. While the attribute set of the IOs in the semantic collection can be extremely wide, we can classify the types of data queries using a certain taxonomy. The choice of visual representation means used in the system is made with regard to the resulting classification.

Requirements to user-level tools. In the vast majority of cases, user actions are associated with dynamic scenarios involving access to additional objects and attributes that are not displayed at the moment.

The above considerations define the set of data transformations that are required to execute the scenario involving either static or dynamic data access, along with UI level tools, graph algorithm set requirements and the functional service architecture to process user queries.

Our Approach and Experimental Implementation

The whole process of interactive visualization of non-formalized heterogeneous data naturally breaks down into two tasks: (1) news feed monitoring and text processing and (2) interactive visualization of extracted IOs and links.

Our experimental software implementation is modular, making it possible to use only the necessary component set depending on the task at hand. The text processing and visualization subsystems may be used independently. They

communicate only indirectly, through a fact database supporting arbitrary types of IOs and links, which are modeled indirectly (rather than through a traditional direct schema definition for specific types of entities and relations). There are no predefined IOs, links and attribute—the taxonomy can be modified online to support new types of objects.

1. The system connects to a set of news feeds and automatically processes the text of incoming documents according to the extraction task. It should be noted that the extraction task can only be solved when the problem specification and the input text meet certain requirements, i.e., the technology does not involve "real" text understanding, but rather it is a "pattern matcher on steroids." Let us consider this process in more detail.

First, the incoming messages are preprocessed in order to extract the metadata and to obtain the message text for further processing (the incoming messages are not necessarily plain text).

The following linguistic analysis provides domain-independent linguistic information for the text under consideration. Linguistic analysis tools include such means as tokenization, POS tagging, morphology, and syntax analysis. These form the basis for application of information extraction methods.

After the linguistic analysis is done, domain-specific analysis is performed involving pattern matching over linguistic structures and providing additional information about identified fragments of interest. The patterns describe target information and possible contexts in the terms of linguistic features, vocabulary, and semantic categories. Our approach to IO and link extraction mostly relies on deterministic pattern-based algorithms employing finite transducers that operate over interval markup. It has common features with CPSL [1] and later JAPE [3] implementation with a number of enhancements, such as support for new data types and operations, variables, look-ahead constructions, non-greedy quantifiers, positional functions, and a number of implementation-specific optimizations.

Another distinctive feature of the extraction model is that it is supported by a knowledge resource containing both linguistic and domain knowledge used to improve the analysis. The knowledge resource is a specialized shared subsystem that is available to all the extraction modules. It contains domain background information (entity classes, possible relations between them, their attributes, etc.) along with extracted factual information (when the document processing is complete the newly extracted content is used to populate the fact database). The knowledge resource also contains domain dictionary information.

When the text processing task is complete, the extracted IOs and links are stored in the fact database for later interactive visualization.

The processing scenario provided above is rather generic and lists only the most notable modules. In fact, the processing pipeline is fully configurable and supports the use of various processing resources through a common API.

2. The visualization subsystem supports a number of visualization scenarios that allow for various ways of graphical data representation. The scenarios include

visualization of IO link structure (using graph layout algorithms), providing aggregated information about IO features (using bar charts) and interactive refining of the view (subgraph) presented to the user. After the subgraph is ready, it can be processed to create a visual image using a range of view options.

The main representation is a graph (including tree-like and network-like structures) with links between OIs. Additional tools that are used to represent aggregated information and interactively refine the view include bar charts and time lines.

The visualization subsystem can be controlled via either a web-based UI (for end users) or a REST API (for integration with external systems). A number of auxiliary technical solutions are provided, including news feed downloader, input document format convertor, and visualization output exporter into common graphic formats.

Conclusions

Building upon theoretical findings (functional models and algorithms) we have developed a suite of scientific and engineering solutions for processing and visualization of data coming from heterogeneous news feeds, and implemented the developed models and algorithms in the form of an experimental software prototype. The prototype performs pattern-based extraction of IOs and links from the input textual data using general linguistic and specific domain knowledge. The extraction results are stored in a fact base for further creation, processing and interactive visualization of a multidimensional link matrix. An API is provided for access from the external systems.

The developed technology and solutions will make a basis for development of decision support systems in the new domains, where previously there were no effective ways to employ the vast arrays of textual information in the decision-making process. Various decision support systems will benefit from the developed technology facilitating the perception of large volumes of complex heterogeneous data and providing a comprehensive picture of the controlled object state. The technology, in general, has significant potential in various industries as it is aimed at improving management quality.

The outcomes of the project can be applied in software systems and solutions for decision support in the public and private sectors. A number of proposed methods and algorithms may be used for a wide range of applications related to automated document analysis, e.g., in the areas of state and corporate governance, business intelligence, defense analytics, marketing, library services and publishing, etc. The results of the study can be used to support further work on such topics as war room information support, open information data mining and social process modeling.

The industrial partner of the study, JSC “CTT Group,” and its business partners express interest in the practical application of the results.

Acknowledgments This research was performed under financial support from the state, represented by the Ministry of Education and Science of the Russian Federation. Agreement (contract) no. 14.604.21.0138 dated 06 Nov 2014. Project unique identifier: RFMEFI60414X0138.

References

1. Appelt, D.E.: The common pattern specification language: Technical report/SRI International, Artificial Intelligence Center. (1996).
2. Chi, E.H.: A taxonomy of visualization techniques using the data state reference model. In: Proceedings of the IEEE Symposium on Information Vizualization 2000, INFOVIS '00, Washington, DC, USA, 69–75 2000.
3. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: a framework and graphical development environment for robust NLP tools and applications. In: Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics. Philadelphia, 850–854 (2002)
4. Request for information: preparing for the future of artificial intelligence [Electronic resource] URL. https://www.whitehouse.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf
5. Response to request for information preparing for the future of artificial intelligence [Electronic resource] URL. https://www.research.ibm.com/cognitive-computing/ostp/rfi-response.shtml