CNRS – INSHS/INS2I * GDR 3359 MoDyS * UMR 3495 MAP * Réseau ISA
École thématique MoDyS 2012 [MOVE_REAL]

Modélisation et visualisation des dynamiques spatiales : Raisonner sur le temps long et ses incertitudes

Visual Analysis of Relatedness in Dynamically Changing Repositories

Coupling Visualization with Machine Processing for Gaining Insights into Massive Data

Vedran SABOL
Know-Center GmbH
vsabol@know-center.at

1 – Visual Analytics: Definition, Goals and Challenges

Users are increasingly confronted with massive amounts of information containing potentially useful knowledge, which is implicitly present in the data and needs to be unveiled. Knowledge discovery (KD) and data mining [5] are research fields addressing the analysis of large data sets with automatic methods. However, human powerful visual system remains largely unchallenged in recognising complex patterns in large data. With the aim of getting the best of both worlds, the novel field of visual analytics (VA) combines automatic analysis with visual methods [2, 4], tightly integrating humans, with their knowledge and intuition, in the analytical process. VA holds the promise of dealing with large, complex data sets, which classical information visualisation (IV) cannot directly handle. VA builds upon IV by including automated analysis in the process of deriving new knowledge: the user is empowered to see patterns, pose a hypothesis, and then apply further automatic and interactive analysis steps to drill down in data and confirm or reject the hypothesis. The balance between interactive and automatic analysis, design of VA workflows, issues of scalability and data quality remain major challenges of the field [1, 3].

2 – Relatedness Analysis in Dynamic Text Repositories and Semantic Knowledge Bases

Despite the increasing use of various modalities, such as audio, video, 3D models, or sensory data, text remains an essential data type in many domains such as media, research, patent management etc. Analysts dealing with large text corpora need to gain insight into topical structures of the corpus, discover arising trends and important events, and identify relationships between entities such as organisations and persons.

The presented visual analysis approach provides methods for identifying patterns in data sets based on relatedness between the data items, in this case the topical similarity between text documents. A scalable, incremental (i.e. capable of accommodating data changes) projection algorithm [7], hierarchically clusters and projects the documents into the 2D plane so that topical similarity is represented by spatial proximity. The resulting layout is visualised by an information landscape - a visual metaphor (up in Fig. 1) allowing the user to immediately spot clusters of topically related documents (represented by hills), see their sizes,

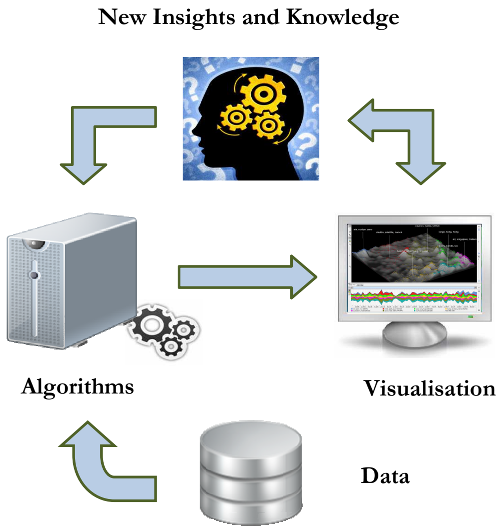

Fig. 1. Visual Analytics combines automatic analysis with visual methods.

***Abstract**: We are facing massive amounts of data, which contain hidden but potentially useful knowledge and facts. Visual analytics combines automatic data analysis with interactive visualisation, promising to provide the advantages of both.*

Besides their sheer size, real-world data sets are permanently undergoing changes. As a consequence, visual and algorithmic methods are required, which are capable of handling dynamic data. The presented visual analytics approach focuses on methods for identifying patterns in data sets based on (some kind) of relatedness between the data items. In the example at hand, large growing document corpora are analysed to identify major topical clusters and relationships between them, as well as to discover topical trends, events and temporal correlations.

Additionally, to demonstrate the applicability of the methods on data other than text, selected techniques are used for visual exploration and navigation of semantic knowledge bases (ontologies).

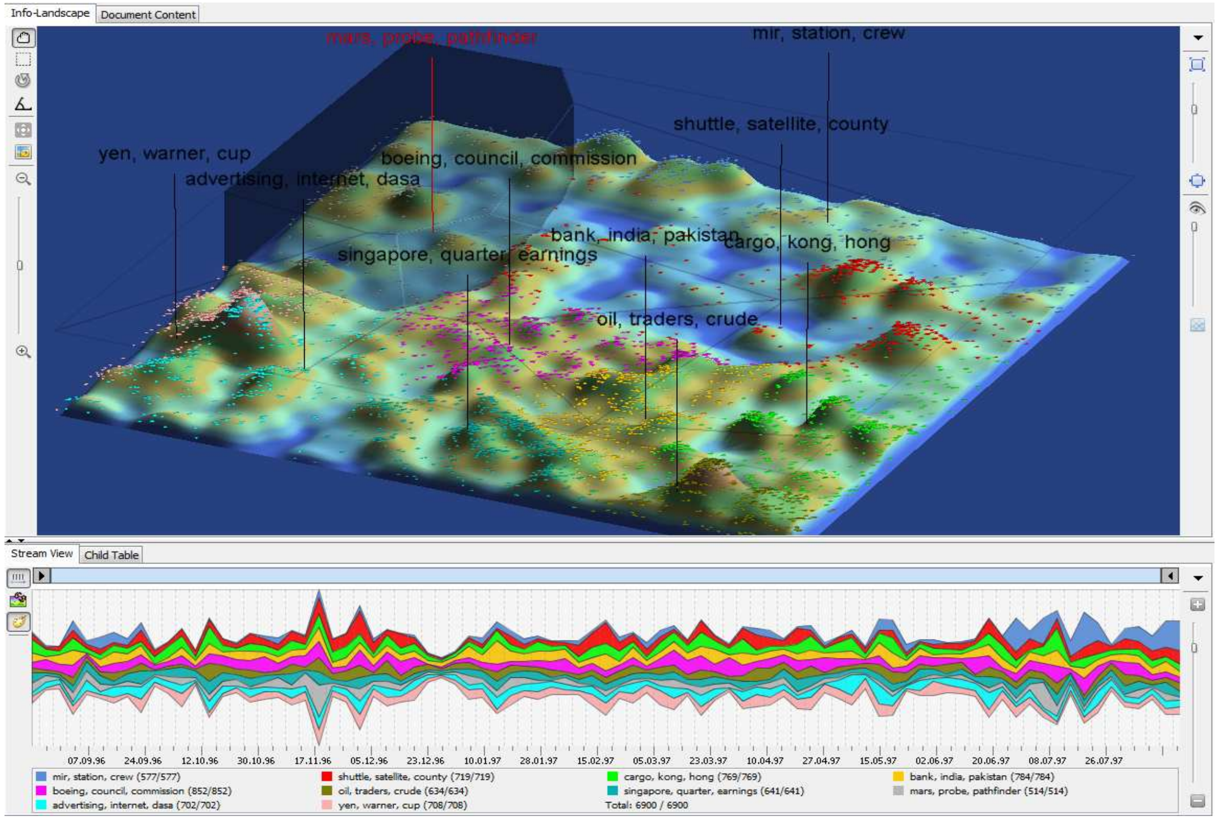
Info-Landscape
Document Content
mars, probe, pathfinder
mir, station, crew
shuttle, satellite, county
yen, warner, cup
advertising, internet, dasa
boeing, council, commission
bank, india, pakistan
cargo, kong, hong
singapore, quarter, earnings
oil, traders, crude
Stream View
Child Table
07.09.96
24.09.96
12.10.96
30.10.96
17.11.96
05.12.96
23.12.96
10.01.97
28.01.97
15.02.97
05.03.97
23.03.97
10.04.97
27.04.97
15.05.97
02.06.97
20.06.97
08.07.97
26.07.97
mir, station, crew (577/577)
shuttle, satellite, county (719/719)
cargo, kong, hong (769/769)
bank, india, pakistan (784/784)
boeing, council, commission (852/852)
oil, traders, crude (634/634)
singapore, quarter, earnings (641/641)
mars, probe, pathfinder (514/514)
advertising, internet, dasa (702/702)
yen, warner, cup (708/708)
Total: 6900 / 6900

Fig. 2. Topical-temporal visualisation of 6900 news articles.

estimate their topical relatedness, and identify outliers. At the same time temporal developments of the topical clusters are shown in the stream view (down in Fig. 1). Visualisations are integrated by a coordinated multiple view framework, so that interactions performed in one view are reflected in the other(s) enabling interactive **topical-temporal analysis of text data** [8]. Topical clusters and/or events of interest can be analysed into depth by retriggering the projection algorithm on the selected data subset. Entities extracted from the text, such as persons, organizations or locations, can be highlighted and correlated with the topical clusters. Fast filtering is supported by advanced retrieval techniques.

Finally, to demonstrate the applicability of the described methods on semantic data, which is (in contrast to text) highly structured with relationships being the central feature of the data type, selected techniques are employed to realise visual interfaces for **explorative navigation** and alignment **of ontologies** [6].

[1] Keim, D., Mansmann, F., Thomas, Jim., *Visual analytics: how much visualization and how much analytics?*, in: SIGKDD Explorations, 11 (2), pp. 5-8, 2009.

[2] Keim, D.A., Mansmann, F., Oelke, D., Ziegler, H., *Visual analytics: Combining automated discovery with interactive visualizations.* In: Discovery Science, LNAI, pages 2–14, 2008.

[3] Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H, *Visual analytics: Scope and challenges*, Visual Data Mining, pages 76–90, 2008.

[4] Thomas, J.J., Cook, K.A. (eds.), *Illuminating the Path: The Research and Development Agenda for Visual Analytics,* IEEE Computer Society, 2005.

[5] Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., *From data mining to knowledge discovery in databases.* AI Magazine 17, pages 37–54, 1996.

[6] Kienreich, W., Wozelka, R., Sabol, V., Seifert, C., *Graph Visualization Using Hierarchical Edge Routing and Bundling*, in Proc. of the 3rd international Eurovis workshop on visual analytics (EuroVA), 2012.

[7] Muhr, M., Sabol, V., Granitzer, M., *Scalable recursive top-down hierarchical clustering approach with implicit model selection for textual data sets*, Proc. of the 2010 Workshops on Database and Expert Systems Applications at DEXA '10, pp. 15 - 19, 2010.

[8] Sabol, V., Kienreich, W., Muhr, M., Klieber, W., Granitzer, M., *Visual knowledge discovery in dynamic enterprise text repositories*, Int. Conf. on Information Visualisation (IV'09), pp. 361–368, 2009.

[9] Granitzer, M., Kienreich, W., Sabol, V., Andrews, K., Klieber, W., *Evaluating a System for Interactive Exploration of Large, Hierarchically Structured Document Repositories*, InfoVis '04, the tenth IEEE Symposium on Information Visualization, 2004.