Visualization of aggregated events from news reports

Thesis Design Master Information Studies
Track: Human Centered Multimedia

Submission date: 14 March 2013

Author: Chun Fei Lung (10409343)
Graduate School of Informatics
Faculty of Science
University of Amsterdam

Supervisor: dr. Lora Aroyo
Department of Computer Science
Faculty of Sciences
VU University Amsterdam

University of Amsterdam

Visualization of aggregated events from news reports

Information Sciences master project research plan

Chun Fei Lung
Faculty of Sciences
VU University Amsterdam
Amsterdam, The Netherlands
2528780@student.vu.nl

ABSTRACT

Using techniques such as named-entity recognition it is possible to extract important data from unstructured documents, such as events. This results in datasets which – while very useful to information professionals – can be hard to analyze for other stakeholders of the dataset. In order to make the data more accessible to users who are less well-versed in constructing queries, a graphical user interface for querying and analyzing the dataset could be offered.
This research focuses on the issue of allowing users to formulate queries on an event dataset in an intuitive way, and visualizing the results in a way that satisfies users' information needs both effectively and effciently.

Categories and Subject Descriptors

H.3 [**Information storage and retrieval**]: Information Search and Retrieval – Search process; H.3 [**Information storage and retrieval**]: Information Search and Retrieval – Relevance feedback; H.5 [**Information interfaces and presentations**]: User interfaces – Graphical user interfaces

General Terms

Design, Experimentation, Human Factors

Keywords

Information visualization, information retrieval, user interfaces, event models, Semantic Web

1. INTRODUCTION

Events play an important role in our lives: when we recall parts of our life using our episodic memory this is always about events, news reports nearly always cover interesting events, and many schedules list future events. When historic events such as verbalized accounts of people's memories or reports of events in newspapers are analysed, they are often stored in unstructured textual form.

Using natural language processing, it is possible to extract events from the unstructured text and store them in a structured format that makes it possible to easily infer information. so that new insights can be gained. In order to present such information in a clear way, visualizations are often used. For instance, one can use maps to show at a glance where certain events have happened or will happen, or line charts to show how events unfold over time.

In this master project research plan, we propose and construct a system that allows users to easily visualize events and interact with those events and related concepts, so that they can get answers on questions that they may have about those events.

The research plan is structured as follows. In Section 2 we present some related work on the relevant topics of this project. We present a number of problem statements in section 3, and make explicit the corresponding research questions that address these problems in section 4. The methodology that will be followed for answering the research questions is outlined in section 5. Finally, an indication of the overall time allocation and planning is given in section 6.

2. RELATED WORK

In this section, we give a brief overview of some related work that has been performed on the topics central to this project.

2.1 Information retrieval

Technically, the proposed system is an information retrieval system. As the name already implies, such systems are used for the retrieval of information from a dataset in a way that satisfies a user's information need as effectively and efficiently as possible.

Note that a user's information need and a user's search query need not be the same – in fact, when presented with a free most users do not know exactly how they should search for a certain item, and need to refine or complete revise their search queries to find what they are actually looking for [6].

Information retrieval (IR) systems that allow searching in documents, usually handle input using simple keyword-based queries. For searching in structured data, keyword search could be used as well, although often special query languages such as SQL are required. For retrieving information from multiple data sources from the Semantic Web, languages such as SPARQL can be used.

Many IR systems, recommender systems in particular, make use of implicit input as well. This input may include information about the user (e.g. age, background), but also some degree of understanding of the users' goals.

2.2 Visualization

Results from IR systems can be presented in various ways, such as lists, tables, or visualizations.

Information visualization is a relatively new field which focuses on the generation of interactive graphical representations of information [2]. This is usually done by providing a user with a bird's-eye graphical view of a large dataset, which optionally can be viewed at various levels of granularity. This allows him/her to easily gain new insights into whatever is visualized.

Information visualization distinguishes itself from related concepts such as scientific visualization and data visualization, in that in the case of the latter concepts, the data to be visualized is very often spatial or quantitative in nature, and has certain "predefined" methods to visualize it, whereas this is often not instantly clear in the case of information visualization [2].

In practice, when we want to know something, we are not only interested in information however. The *data-information-knowledge-wisdom (DIKW) hierarchy* (also known as the *DIKW pyramid*) is a model often used when one wants to contextualize data, information, knowledge, and wisdom [8]. In short, the model suggests that data can be combined to create information, information can be combined to create knowledge, and knowledge can be used to achieve wisdom.

In visualizations, a distinction between data and information can be made as well: Chen for instance lists a number of visualizations for different types of data, information and knowledge [3].

Along another dimension, we can also differentiate between multiple levels of granularity in showing data and information. One could for example distinguish between three types of views: *singular* views, in which the focus is on one specific information object only, *summative* views, in which for each of a number of information objects a summary is displayed, and finally *aggregate* views, which show information about multiple information objects – but not the actual information objects in question.

When combining these viewpoints on the visualization of data, information and knowledge, we find that there are many interesting contributions still to be made.

2.2.1 Examples

We now present some examples of papers which showcase interesting methods of visualization or interaction, or which have had a profound impact on subsequent systems either within or outside of academia. Many important developments in information visualization have been described in the work of C. Chen [2], Shneiderman [9], and as mentioned before, M. Chen [3] gives a good overview of the various visualization techniques in existence.

Buschbeck et al. describe an application that allows users to browse hierarchically structured events with an arbitrary number of granularity levels [1]. Elmqvist and Fekete present a more general model for visualizations using hierarchical aggregation [4]. Hirsch et al. on the other hand take a different approach, by visualizing information extracted from Wikipedia and Freebase as interactive force-directed layout graphs [5].

An interesting observation is made by Van De Moere, who argues that for a visualization tool to be viable in a non-academic setting, it should not only be functional, but be aesthetically pleasing as well [14]. This does not seem to be the case with many of the tools described above.

2.3 MONA Project

An example of a currently active research project which utilizes visualizations is the MONA Project. This project "aims at developing a tool that gathers source material (e.g. news articles, blog posts), extracts events and their properties (e.g. actors, locations, timestamps) from these materials, analyses links between events, and visualizes the results".[1]

In the MONA Project, social scientists collaborate with computer scientists in order to find answers on questions related to activist organizations, e.g. how their roles change over time. Much work has already been done by [7], who are working on extracting information from texts using natural language processing tools, and saving this information using the Simple Event Model, an event

[1] `http://monaproject.wordpress.com/about/`

model which can be used for many different types of events due to its minimal semantic commitment [12]. What remains then, is the analysis of this dataset and visualizing it.

3. PROBLEM STATEMENT

In the previous section, a few examples were given of systems that visualize information and allow users to browse through data and information sets to satisfy their information needs or to gain insights. While this overview is by no means complete, to our knowledge these systems currently are state of the art in academic research. We find that existing systems cope with some limitations which make them unsuitable for the MONA Project. We discuss these limitations in this section.

3.1 Information reliability

For many systems it is assumed that the information shown is correct. This need not necessarily hold however, especially if the information was extracted using a technique such as named-entity extraction. Sometimes information that is extracted is simply wrong, but not seldom are errors caused by disambiguation issues. Problems relating to precision and recall of named entities are described by many authors, including [5] and [10].

Because this has consequences for the usefulness of information that is extracted using such a method – and ultimately the user's trust in the system that presents the information – it is necessary to make clear to a user how likely it is that a certain piece of information is correct, and why the system believes information to be correct [11].

3.2 Translating user's information needs

As stated in section 2.1, many users have difficulty translating information needs into queries: especially when searching specialized datasets, novices have difficulty to write constructed language queries. Machines on the other hand, still have considerable difficulty understanding (often ambiguous) natural language queries.

A good interactive visualization system should therefore be able to offer a method of phrasing questions that is easily understandable for both human users and the system's underlying information retrieval component.

4. RESEARCH QUESTIONS

Based on the problem statements in the previous section, we can see that amongst users who are not information scientists, there is a need for a system that allows them to easily find an answer to questions that they may have on a particular event-related subject. We propose a system that does exactly that, and pose the main research question that drives the design and development of such a system: *'How can event datasets be browsed such that a user's information needs are satisfied effectively and efficiently?'*. From this we can formulate two sub-questions which guide our research.

4.1 Information needs

An obvious question would be why users would want to browse through an event dataset. The existence of both academic and commercial systems which allow for events to be browsed, suggests that users have plenty of reasons to do so. We do not know the underlying reasons for browsing through datasets however. The first sub-question deals with this: *'What are prototypical information needs with regards to event datasets?'*. Answering this question allows us to not only understand what users want to know, but also serves as an indication of what kinds of interaction should be supported by the system, and how answers should be presented to them.

4.2 Visualization of events

The second sub-question then is: *'Which techniques for visualization and interaction with visualizations exist, and for which data and information types that occur in events are they most suited?'*. By identifying which techniques exist, what characteristics they have in common, and which characteristics set them apart from other techniques, we can create an overview that makes clear which visualizations can be used for what types of event data or information. By mapping these to the specific data and information types that are available to us from the MONA Project, we can generate useful visualizations from the data and demonstrate whether this is successful.

5. METHODOLOGY

The goals of the project are to understand which characteristics an interactive visualization system should have and demonstrate how interaction with visualized event data and information can help satisfy information needs. We have previously stated the problem statements and the research questions that guide us to a solution to those problems. This section describes the phases that are necessary to complete the project.

5.1 Literature study

The first phase consists of a literature study. More specifically, we look into two topics: visualization techniques and interacting with visualizations.

5.1.1 Visualization techniques

Many visualization techniques have been devised. Some of these techniques are especially suitable for visualizing simple data (e.g. pie charts, maps), while for visualizing n-dimensional data other techniques might be needed (e.g. graphs, heat maps).

First, the visualization techniques are identified from descriptions in academic literature. Then these techniques are classified according to typical data usage scenarios.

5.1.2 Interacting with visualizations

Because we want to know how people can browse through visualized data, we also need to look at the methods that existing tools have applied to allow interaction with visualizations. Furthermore, we try to identify the design patterns that commonly occur in such tools.

5.2 Making use of events

In one of the later project phases, we want to evaluate which visualizations and methods of interaction work best. For this, we need not only a set of events which we can visualize, but also a set of prototypical information needs that we can use to evaluate the system.

To our knowledge, no widely-used event dataset for evaluation purposes exist. Therefore we make use of events extracted as part of the MONA Project. As described in section 2.3, this project focuses on capturing and extracting events in which political activists were involved.

Many of the different information needs that users could have for an event dataset, actually seem to be similar to each other. In the case of the MONA Project, examples of questions that have been asked were *'What are the most used tactics?'*, *'Who are the most active actors?'*, *'Which actors use labour strike as a tactic?'* and *'What are the most effective tactics used by Colombian actors?'*. While the intention behind each question is clearly different, we already see a pattern: questions often seem to start with one of the Six Ws (who, what, when, where, why, how), with some other variable terms which a user wants to know more about.

While we could get an exhaustive list of questions from (potential) users, this would be very laborious. Due to the similarity of the questions that users may have on events, we hypothesize that this could to some degree be automated by generating questions based on the information that is available in a dataset according to a simple grammar.

This can be done by gathering a large number of questions that people might want to ask about the MONA Project's activist events, and finding patterns in the information needs that drive those questions. This should make it possible to easily construct parameterized queries that can answer questions for virtually any set of events.

5.3 Prototyping

The prototype should allow a user to explore events and thus satisfy his/her information need or gain new insights into a topic related to an event. This prototype should therefore fulfil the functional requirements listed in Table 1. The second column indicates how important it is that the requirement is fulfilled using the MoSCoW prioritization [13].

Table 1: Functional requirements

Functional requirement	Priority
1. Load events which are stored according to a specific event model into the system.	Must
2. Allow users to browse events in various methods.	Must
3. Clearly indicate the relevancy of a certain information object	Must
4. Possibility for users to provide feedback on incorrectly tagged or linked information	Should
5. Incorporate user feedback to correct data about a particular event	Should

There are also some non-functional requirements that could be taken into account, e.g. accessibility, maintainability, performance, reliability, and usability. Because performance and usability directly affect how well users will be able to use the prototype during the evaluations, these are required. Because the project is not about developing a production-grade system however, most of the other non-functional requirements logically have a lower priority.

The prototyping phase takes place over the course of a number of weeks (see section 6), in which the system is iteratively developed and evaluated (see section 5.4) using a rapid software development methodology. In the first few weeks, throwaway prototyping will be used – this allows us to focus purely on determining what works and what does not. Later on, an evolutionary approach is taken, in which the prototype is fully developed.

5.4 Evaluation

During and after development, the prototype is evaluated using two different methods. Depending on the stage of development, these methods are used either in a formative or summative way.

The first is concerned with determining how well the prototype satisfies user's information needs. For this, an evaluation method from the field of information retrieval is used. This method involves asking participants to execute a number of informational queries according to some information need that is explained to participants, and letting them rate the relevance of the results returned by the prototype and from the event dataset. This allows us to calculate the precision and recall of the prototype [6]. Alternatively, as queries can be constructed by means of concatenation, we could count the number of steps between the output that the system provides, and the output that it should give.

The second method is less structured, and involves user tests in which participants are given the opportunity to freely use the prototype. By gathering feedback either during the tests using a concurrent think-aloud protocol, or only after the tests using a retrospective think-aloud protocol, we can gather insights in what visualization techniques and interaction methods work well.

6. PLANNING

The workload for the project is spread over a period of approximately eighteen weeks. An overview of the projected workload for each period is given in Table 2. The hours mentioned in this table are meant to be taken as averages over the entire period: in some weeks it might not be possible to work on the master project due to coursework or other occupations, while in slower weeks it might be possible to spend more time on the project.

Furthermore, coursework might partially overlap with the project, in the sense that certain parts of the project may benefit from work done for other courses, and are thus best scheduled either parallel with or after those courses. These other courses and the respective background knowledge they provide are listed in Table 3.

A more detailed overview of the planning is given in Table 4. The first column shows when the activity in the second column should be completed. While the weekly activities and deadlines are somewhat subject to change due to the reasons mentioned above, this has largely been take into account already by not planning (time-consuming) activities in busy weeks. The final deadline in the first week of July is hard and shall under no circumstance be moved.

The end result of the master project is a research paper of at most ten pages. This includes references, but not appendices.

Table 2: Overall time allocation

Period	Allocated time	Other activities
4	10 hrs/week	60 hrs/week (courses) 20 hrs/week (other)
5	30 hrs/week	40 hrs/week (courses) 20 hrs/week (other)
6	50 hrs/week	30 hrs/week (other)

Table 3: Courses

#	Course	Subjects
2	Research Methods	Research methodologies
	Knowledge & Media	Linked data, events
4	The Social Web	Social network analysis
	Internet Information	Information retrieval
	Mobile Systems	Prototyping
5	Visual Analytics	Visualizations

Table 4: Project milestones

Date	Activity
03/03	Project start
10/03	
17/03	Identify prototypical information needs
24/03	
31/03	
07/04	Literature study of visualizations completed
14/04	First mock-ups/wireframes/scenarios
21/04	Refined mock-ups/wireframes/scenarios
28/04	
05/05	
12/05	
19/05	Presentation of preliminary results
26/05	Prototyping phase finished
02/06	Thesis defense date fixed by this week
09/06	Prototype evaluations analyzed
16/06	
23/06	Finalized master thesis
30/06	
07/07	Public thesis defense + submit thesis

7. ACKNOWLEDGEMENTS

This master project is performed at VU University Amsterdam (VUA), as part of the joint Information Sciences master project programme with the University of Amsterdam (UvA). All work is done in the context of the MONA Project under the supervision of dr. Lora Aroyo, associate professor at VU University Amsterdam.

8. REFERENCES

[1] S. Buschbeck, A. Jameson, and T. Schneeberger. New forms of interaction with hierarchically structured events, 2011.

[2] C. Chen. Information visualization. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):387–403, 2010.

[3] M. Chen, D. Ebert, H. Hagen, R. S. Laramee, R. van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, information, and knowledge in visualization. *IEEE Comput. Graph. Appl.*, 29(1):12–19, jan 2009.

[4] N. Elmqvist and J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *Visualization and Computer Graphics, IEEE Transactions on*, 16(3):439–454, May–June 2010.

[5] C. Hirsch, J. Hosking, and J. Grundy. Interactive visualization tools for exploring the semantic graph of large knowledge spaces. In *Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW2009)*, volume 443, 2009.

[6] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[7] T. Ploeger, B. Armenta, L. Aroyo, F. de Bakker, and I. Hellsten. Making sense of the Arab revolution and Occupy: Visual analytics to understand events. In *Proceedings of the Workhop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, ISWC 2012, pages 61–70, 2012.

[8] J. Rowley. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, 33(2):163–180, 2007.

[9] B. Shneiderman, C. Dunne, P. Sharma, and P. Wang. Innovation trajectories for information visualizations: Comparing treemaps, cone trees, and hyperbolic trees. *Information Visualization*, 11(2):87–105, 2012.

[10] R. Sipoš, A. Bhole, B. Fortuna, M. Grobelnik, and D. Mladenić. Demo: Historyviz – visualizing events and relations extracted from Wikipedia. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *The Semantic Web: Research and Applications*, volume 5554 of *Lecture Notes in Computer Science*, pages 903–907. Springer Berlin Heidelberg, 2009.

[11] M. Skeels, B. Lee, G. Smith, and G. G. Robertson. Revealing uncertainty for information visualization. *Information Visualization*, 9(1):70–81, 2010.

[12] W. R. van Hage, V. Malaisé, R. Segers, L. Hollink, and G. Schreiber. Design and use of the Simple Event Model (SEM). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128–136, 2011.

[13] H. van Vliet. *Software Engineering: Principles and Practice*. Wiley Publishing, 3rd edition, 2008.

[14] A. Vande Moere and H. Purchase. On the role of design in information visualization. *Information Visualization*, 10(4):356–371, 2011.