Sensemaking on Wikipedia by Secondary School Students with SynerScope

W.R. van Hage[1,2], F. Núñez Serrano[2,3], T. Ploeger[1], and J.E. Hoeksema[1,2]

[1] SynerScope B.V.
[2] VU University Amsterdam
[3] Universidad Politécnica de Madrid

Abstract. Visual analytics of linked data can be done by secondary school students with minimal preparation. We study the learning curve of students while answering typical Web analytics questions on Wikipedia and DBpedia using SynerScope visual analytics software. We find that after a short tutorial students are able to answer most complex questions in a few minutes, learning by trial and error. Older students are faster on average, but motivation appears to be a stronger factor than age for success. Answering speed doubles within two hours of experience while correctness increases.

1 Introduction

The world will soon face a critical shortage of data scientists, professionals with analytical expertise that can take advantage of (linked) data to answer questions [7]. One strategy to mitigate this problem is to enable non-experts to take over part of the data science tasks. We pose that data science is comprised of many tasks that do not all require expert-level knowledge. In this article we restrict ourselves to a category of data science sensemaking tasks on Web data that is common in data journalism and involves basic analytics operations, search, and Web browsing. We hypothesise that, given the right tools, untrained people can quickly be trained to do such tasks, avoiding a complete data science education.

The goal of this article is to test this hypothesis by doing an experiment to demonstrate the feasibility of having untrained people do prototypical sensemaking tasks given visual analytics tools. Specifically, we look at secondary school students with no analytical experience, and ask them to answer complex questions about Wikipedia content using the SynerScope[4] visual analytics software illustrated in Figure 1. We want to know if users can get to an answer after a minimal amount of training in the tool. We want to know how long it takes them to find an answer and if their time-to-answer decreases as their experience with the tool increases, and what the influence is of their age and corresponding level of education.

The line of reasoning we follow is that the required skills for such sensemaking data science tasks can be rapidly acquired or substituted with appropriate tools. If this is the case and if SynerScope is an appropriate tool for the task, then we should be able to show that unskilled people can accomplish the sensemaking tasks.

[4] `http://www.synerscope.com`

This idea of empowering people by means of augmented reasoning through human-computer interaction is not new [6], but in recent years the development of interactive tools for visual analytics have intensified. Some of these tools are targeted at programmers (e.g., [1, 10, 11]), while other tools target non-programmers (e.g., [13, 12, 14, 9, 2, 8]). For this experiment we need a tool from the latter category that is network centric and allows search and Web browsing. We use SynerScope [13, 5, 4], one of the tools that meets these requirements.

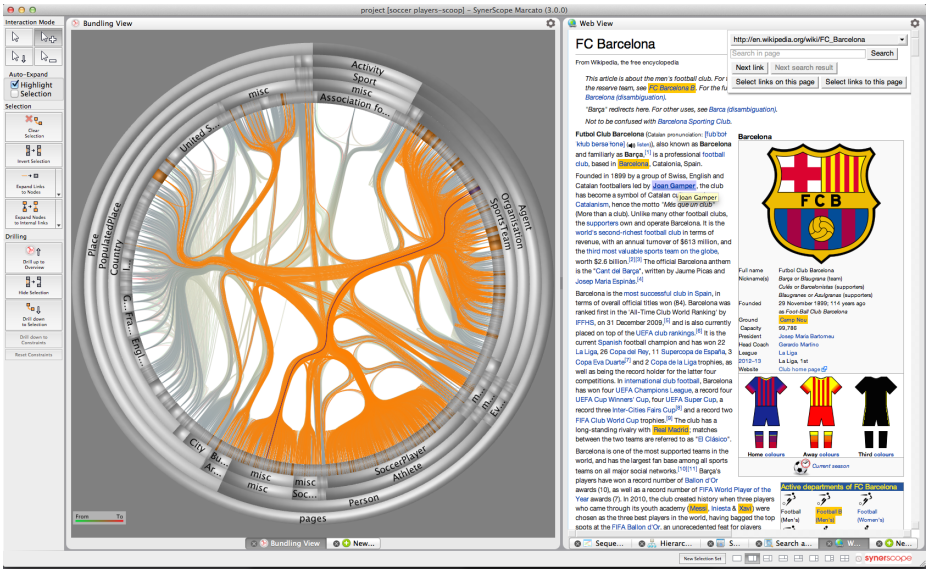

Fig. 1. A screenshot of the SynerScope visual analytics tool showing Wikipedia and DBpedia data. This picture shows two coordinated views: a hierarchical edge bundling network view and a Web browser.

The rest of this paper is organised as follows: Section 2 describes the SynerScope software in more detail. Section 3 outlines the experimental set-up, including the tasks, tooling, and procedure. Section 4 shows our findings. Section 5 discusses our findings, draws conclusions and suggests future work.

2 The SynerScope Software

SynerScope is a visual analytics application that delivers real time interaction with dynamic network-centric data. SynerScope supports simultaneous visualisations and coordinates user interaction, enabling the user to identify causal relationships and to uncover unforeseen connections.

The central interaction paradigm of SynerScope is Multiple and Coordinated Views. SynerScope shows a number of different perspectives on data, for example, relations

and time, and each selection made in either of these views causes an equivalent selection to be made in all other views. This enables the user to explore correlations between different facets of data.

SynerScope is designed to work with a very basic information schema. This schema consists of two object types: Nodes and Links. Links connect two Nodes. Both Nodes and Links can have additional attributes of a number of data types, including integers, floating point numbers, free text, date and time, latitude and longitude.

What follows is a short overview of each visualisation that is offered by SynerScope.

Table View The Table View provides a traditional spreadsheet view on the data. For each type of Node and each type of Link, there is a separate sheet. The Table View shows all the data as a table of values.

Hierarchical Edge Bundling View The Hierarchical Edge Bundling View (HEB) is the primary network view in SynerScope. Each Node is visualised as a point on a circle, and each Link is visualised as a curved line between its source and target Node.

The Nodes are grouped hierarchically, based on one or more of their attributes. The Links between Nodes of the same hierarchical category are bundled together (as if they were tied together with a cable tie).

Massive Sequence View The Massive Sequence View (MSV) is the primary temporal view in SynerScope. Each Node gets a fixed position on the horizontal axis. Nodes are grouped hierarchically in the same fashion as in the HEB. Links between Nodes are represented by a horizontal line between the respective positions of the Nodes. On the vertical axis the user can select a scalar attribute, typically a time or date. This orders the Links temporally.

Map View The Map View is the primary spatial view in SynerScope. The user can select two attributes from any Node or Link data source to interpret as WGS84 latitude and longitude coordinates. These attributes are used to plot the Nodes (not the Links) on a map as points.

Scatter Plot View The Scatter Plot View uses Cartesian coordinates to relate the values of two attributes of either Nodes or Links. Dots are drawn on a two-dimensional chart, the positioning relative to the horizontal and vertical axis being determined by the attribute's values. A third attribute can used to set the size of the dots.

Search and Filter View The Search and Filter View is an interactive view that allows the user to select Nodes or Links by searching by value.

Web View The Web View is an interactive view that allows the user to view any URL's that are an attribute of a node or a link.

The user can interact with SynerScope's views in several ways: By selecting and highlighting data, drilling down to or up from a selection, and expanding selections from nodes to connected links or vice versa. Every interaction method is coordinated across multiple views.

3 Experimental Set-up

Sensemaking Tasks In the experiment we look at 10 exemplar Web analytics questions that each require a combination of at least two of the following operations to answer: network navigation, filtering on categorical and numerical variables, grouping and counting, search, Web browsing within Wikipedia, and zooming in on data selections. Examples of the questions are: *"How many former AFC Ajax soccer players died in Paramaribo and what was the cause of death?"*, or *"Which page about a disease is linked to most from pages about physicists?"*. The complete set of questions can be found on FigShare [15]. Question number 8 is marked as a difficult question, because it is the only question that involves a set intersection between two sets of network patterns.

SynerScope Visual Analytics Tooling The SynerScope tool used by the students is a graphically accelerated visual analytics application that combines a number of views on networked data. It offers real-time interactive exploration using scatter plots, timelines, maps, hierarchical edge bundling network layouts, an integrated Web browser, a search engine, and a spreadsheet table view. The selections made in any of these views are propagated to all the other views. A video illustrating interaction with the Wikipedia data can be found on FigShare [3].

Procedure The experiment consists of five parts: (1) a 30m plenary introduction to the experiment and the data sets used, (2) a 15m plenary tutorial to the SynerScope visual analytics tool, (3 and 4) two 45m sessions where students try to answer questions using SynerScope, (5) a concluding discussion and personal interviews. The students are asked to answer as many as possible of 10 questions about 3 subsets of Wikipedia within 90m. Each set centers around pages on a specific topic.

Data Sets The topics covered in the experiment are: (1) Athletes classified as soccer players and trainers of AFC Ajax, FC Barcelona, and Manchester United, (2) Scientists classified as physicist, (3) Artists in the pop genre. Each of these three sets consist of around 3000 Wikipedia pages about the topic (the seed set), all the pages that are linked to from the seed pages (the "out" context), all pages that link to the seed pages (the "in" context), and all the links between the seed, "out" context, and "in" context pages. This amounts to three sets of around 100k–200k pages and 300k–500k page links. Each page is assigned around 18 attributes with information about the page, such as the page title, the number of words on the page, the in degree and out degree, a three-level hierarchical topic classification of the main subject of the page (e.g. Actor-Artist-Person, or Building-ArchitecturalStructure-Place) derived from the DBpedia rdf:type property of the corresponding DBpedia resource, birth/death date and place, and topic-specific properties such as respectively soccer team, university, or band. An example of the three schemas can be found in the hand-outs for the students [15]. We made a selection of the DBpedia types (downloaded september 2013) that form a hierarchical partitioning of the Wikipedia pages. We only considered types from the DBpedia ontology, ignoring other type hierarchies such as Yago, FreeBase, and Schema.org. The selection process involved dividing the types into three hierarchical layers, and imposing a preferential

ordering onto the types. For example, Amsterdam was assigned City at level 1, PopulatedPlace at level 2, and Place at level 3, discarding types such as Settlement to form a proper partition. When type information is missing, a placeholder type is assigned.

Test Subjects The students involved as test subjects in the experiment are 63 middle school and high school students (9 female, 54 male) from three schools in the Amsterdam area between the ages of 12 and 18, divided into 34 groups of size 1–3. The experiments were performed in two labs of the VU University Amsterdam Network Institute.[5] One running SynerScope in the Amazon cloud accessed through a Web-based client (OTOY), the other running SynerScope natively on gaming PCs with modern NVIDIA GeForce GPUs. The students were paired up and given a hand-out describing the three data sets, listing all the questions, and containing a form to record the answers and the time taken [15]. During the experiment students were assisted by answering specific technical questions, but were given no other guidance that would help them find answers.

4 Results

There was a large variation in the productivity of the various students, as can be seen in Figure 2. This can be expected of students that have no intrinsic motivation to cooperate in the experiment. The motivated students answered all questions, while two groups did nothing and are excluded from the results. In general the total number of 10 questions was too high to answer for most students in two 45 minute sessions. Most students managed to answer the questions of two topics (6 or 7 questions). Of the questions that were answered, about 60% was answered correctly. There was a large variation, depending on the difficulty of the question. This is illustrated in Figure 3. Some questions were answered partially. For example, when asked for a number and explanation only the number or the explanation was answered correctly. We performed significance tests for the differences in duration between all the categories shown in Figure 2 with a Welch's t-test, and similarly for the categories in Figure 4. There was a slight increase in the number of questions that were answered correctly over time. This trend is significant according to a Mann-Kendall test ($p = 0.0318$), even when counting partial answers as false answers. Students performed faster and more consistently for subsequent questions. This is illustrated in Figure 4 (right), specifically with questions 1–7 which were consistently answered before time ran out. This increase in speed is significant between the first and last of the questions in the sequence at a confidence level of 95%. Older students seemed to be faster than younger students, but their answers were of a comparable correctness. Although the difference in mean time taken between the fastest and slowest age groups is a factor 2, a Mann-Kendall test does not show a significant downward trend ($p = 0.178$). This is due to the relatively small number of observations (34 student teams) and a class of particularly talented middle school freshmen that performed on par with 18-year-olds, but with a significantly higher accuracy. The data used to derive these conclusions can be found on Figshare [15].

[5] Network Institute Tech Labs, `http://www.networkinstitute.org/tech-labs/`

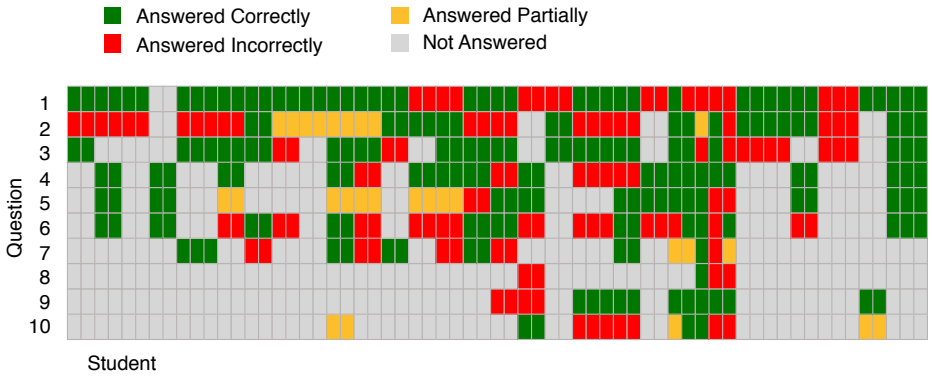

Fig. 2. An overview of the completeness and correctness of the answer to each question by each student. Each column represents the answers given by a student. Each row represents one of the 10 questions. Roughly 55% of the questions were answered, about 60% of the answers were correct.

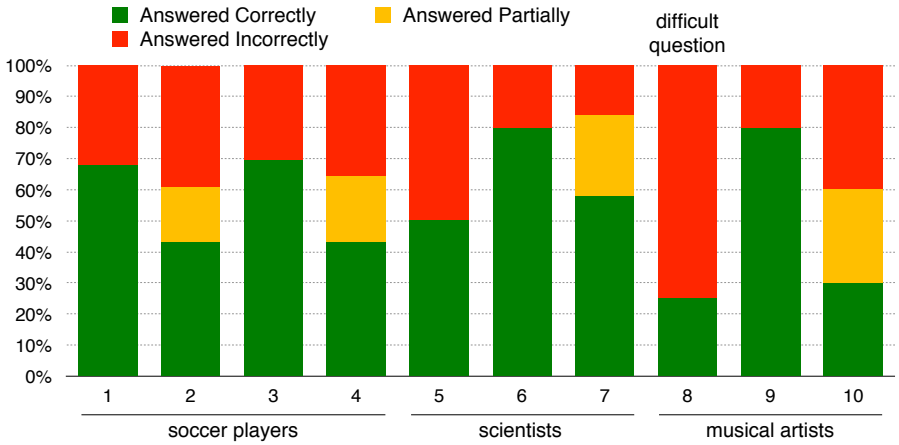

Fig. 3. An aggregation of the correctness of the answers per question. For the questions that were consistently answered (1–7) in the 90m experiment is a rising trend in the quality of the answers. Most students ran out of time before attempting question 8–10.

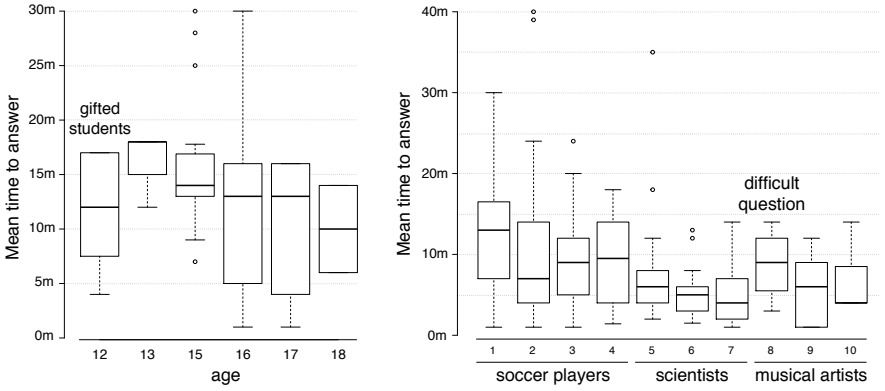

Fig. 4. (left) Time taken to answer a question, aggregated over all questions per age. Older students are faster than younger students with the exception of a group of gifted 12-year-olds; (right) Aggregated time taken to answer each of the 10 questions. There is a non-significant decreasing trend in the time taken per question.

5 Discussion

Given the preliminary nature of these results, we can not draw very strong conclusions yet. If we had more test subjects, we could have repeated the experiment with the topics offered in a randomized order, which would strengthen the conclusions by removing the learning effect and topic preference between the various topics.

We are impressed by what our young test subjects were able to achieve. When given the right tools, visual analytics of linked data really can be done by secondary school students with minimal preparation. We found that after a short tutorial students are able to answer most complex questions in a few minutes, learning by trial and error. Within two hours of experience, answering speed doubles within while correctness increases.

The older test subjects more frequently asked for help when they get stuck than the younger test subjects, who just found their own way through trial and error, and therefore also take longer to get to an anser than the older students (as can be seen in Section 4. Overall, motivation appears to be a stronger success factor than age. This belief is hard to make concrete, but it is reinforced by our observation that students are quick to accept their first findings as a definitive answer to the question they were working on. When students found information they thought was the right answer, they were fairly quick to accept that answer and wanted to move on to the next question as soon as possible. In contrast to professionals, the students did not verify their answers. For instance, when the students had to find out how many AFX Ajax soccer players died in Paramaribo, they typically accepted all the soccer players that died in Paramaribo as an answer, without checking if they played in AFC Ajax. We think this can be explained by the lack of feedback during the experiment. Students were not penalised for wrong answers or rewarded for right answers, and the experiment was a one time encounter with the software. We expect that many of the incorrect or partial answers could have been improved if the students were to have verified their answers.

The experiment reinforced our belief that visual analytics software must be highly interactive and present immediate feedback to the user. During the interviews at the end of the session students were generally positive about the software and tasks and thought the experiment gave them a new perspective on Wikipedia. Their main negative remark was that SynerScope running on Amazon was distractingly slow. In actuality, the software was equally fast on Amazon instances as on local machines, but the lag introduced by network congestion, network latency, and video compression, removed the sensation of true interactivity. In isolated cases, for example, when zooming out to the entire data set of 400k links, students had to wait a few seconds. Delays in interaction like these appeared to interrupt the student's train of thought.

We found that students of all ages are able to effectively use the SynerScope tool to answer the questions. Older students are usually faster, but not significantly more accurate. We would like to further test these findings with older and younger subjects.

Acknowledgements

Thanks go to the Damstede, Pieter Nieuwland College, and Cygnus Gymnasium schools for their participation in this experiment. We thank the VU Network Institute for the use of their facilities, and Samir Naaimi for his assistance during the experiments. This work was done within the context of the SAGAN project supported by ONR Global NICOP grant N62909-14-1-N030, the EU FP7 NewsReader project (316404), and the Dutch COMMIT Data2Semantics project.

References

1. D3.js: D3.js - data-driven documents (2014), http://d3js.org/
2. Gapminder: Gapminder: Unveiling the beauty of statistics for a fact based world view (2014), http://www.gapminder.org/
3. van Hage, W.R.: SynerScope on Wikipedia (movie) (06 2014), http://dx.doi.org/10.6084/m9.figshare.1061499
4. Holten, D., Cornelissen, B., van Wijk, J.: Trace Visualization Using Hierarchical Edge Bundles and Massive Sequence Views. In: Visualizing Software for Understanding and Analysis, 2007. VISSOFT 2007. 4th IEEE International Workshop on. pp. 47–54 (June 2007)
5. Holten, D.: Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. IEEE Transactions on Visualization and Computer Graphics 12(5), 741–748 (Sep 2006), http://dx.doi.org/10.1109/TVCG.2006.147
6. Licklider, J.C.R.: Man-computer symbiosis. Human Factors in Electronics, IRE Transactions on (1), 4–11 (1960)
7. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Hung Byers, A.: Big data: The next frontier for innovation, competition, and productivity (2011), http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation
8. Paulheim, H.: Explain-a-lod: Using linked open data for interpreting statistics. In: Proceedings of the 2012 ACM international conference on Intelligent User Interfaces. pp. 313–314. ACM (2012), http://www.ke.tu-darmstadt.de/resources/explain-a-lod

9. Qlikview: Business Intelligence and Data Visualization Software — Qlik (2014), http://www.qlik.com/
10. R: The R Project for Statistical Computing (2014), http://www.r-project.org/
11. Skjæveland, M.G.: Sgvizler: A javascript wrapper for easy visualization of sparql result sets. In: Extended Semantic Web Conference (2012), http://dev.data2000.no/sgvizler/
12. Spotfire, T.: TIBCO Spotfire - Business Intelligence Analytics Software & Data Visualization (2014), http://spotfire.tibco.com/
13. SynerScope: SynerScope — Connecting the dots (2014), http://www.synerscope.com/
14. Tableau: Business Intelligence and Analytics — Tableau Software (2014), http://www.tableausoftware.com/
15. van Hage, W.R., Ploeger, T., Hoeksema, J., Núñez Serrano, F.: Wikipedia SynerScope experiment (06 2014), http://dx.doi.org/10.6084/m9.figshare.1060254