NARRATOR: SYSTEM FOR REPORT GENERATION IN NATURAL LANGUAGE

Inna Novalija, Marko Grobelnik
Artificial Intelligence Laboratory
Jožef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773144; fax: +386 1 4251038
e-mail: {inna.koval, marko.grobelnik}@ijs.si

ABSTRACT

This paper presents NARRATOR, a system for report generation in natural language. The idea behind the developed system is based on merging statistical data with report templates predefined by the user. The user has a possibility to generate different kinds of reports with respect to various indicators, time periods and natural language statements.

1 INTRODUCTION

The aim of the narrative reporting is to analyze the data and to present them in a simple and understandable way to the user.

The motivation for automatic developing of stories in natural language can be different – to minimize human efforts and funding, to obtain interesting conclusions from the data, to create compelling entertaining content in short period of time etc.

For instance, the Quill technology of the Narrative Science [1] company allows to merge Artificial Intelligence with Big Data analytics and to transform data into stories, which are similar to stories authored by people.

Nichols [2] describes the method for creating machine-generated content and a system called "News at Seven" – an automatically generated news and entertainment show.

From the business prospective of view a lot of companies are interested in using technologies for automated reporting, since it allows systematically draw inferences from data.

The NARRATOR system, presented in this paper, is being developed for website performance analysis and following report generation in English. The system is based on statictical data coming from Google Analytics services [3]. The NARRATOR system works with general and user specific information – the users can adapt the system according to their needs. Furthermore, in this paper we describe the architecture, principles of work and provide a demonstration link to the NARRATOR system.

The paper is structured as follows: Section 2 provides the system architecture; Section 3 describes the system interface; in Section 4 we provide the inside on the data mining techniques employed in the system; and finally, Section 5 concludes the paper.

2 SYSTEM ARCHITECTURE

The NARRATOR system architecture includes several components.

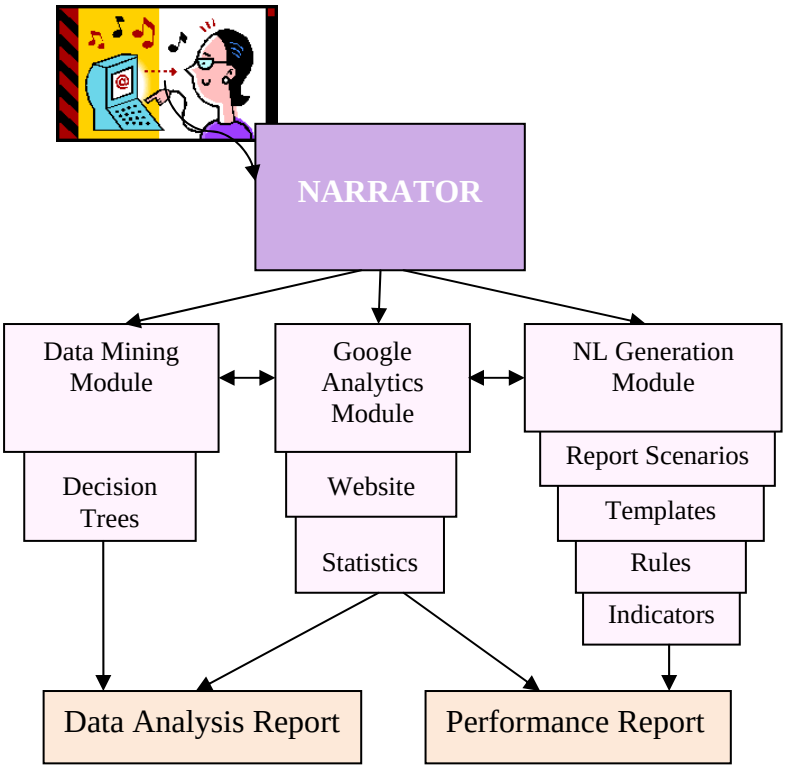

Figure 1: *NARRATOR architecture.*

As it is visible on Figure 1, the main NARRATOR modules are *Google Analytics Module, Natural Language (NL) Generation Module and Data Mining Module*.
The user connects to the NARRATOR system and selects the website she wants to work with in the processes of the performance and data analysis. The connection to the website is performed in the Google Analytics Module with a help of Google Analytics API [4]. The important note is that in the current settings of the NARRATOR system the user is assumed to have access to google analytics tools.
In addition, the user also selects the the report scenario she wants to get. The report scenario contains sets of templates, which connect natural language statements to rules and relevant indicators. The report scenarios, indicators, trules and templates are described below in this section.
Following that, the NARRATOR system obtains a set of statistical performance data for the particular website from the Google Analytics services. The reports are generated in the NL Generation Module based on the report scenario and statistical data. Finally, the user is provided with a natural language report for the selected website.
In addition, inside the NARRATOR system, we are developing a Data Mining Module, which allows to get more inside pictures of the website data. In this module the data mining techniques, such as decision tree generation algorithms, are used to analyze the website statistical data and to discover the rules hidden inside the data.

2.1 Google Analytics Tools

Google Analytics (GA) [3] is a free service from Google that provides detailed statistics about the visitors of the particular website. GA can be used to obtain the information on the mobile analysis, content analysis, conversion analysis, social analysis and advertising analysis. The Google Analytics API [4] can be used to develop custom applications, such as reporting tools.

2.2 Indicators

While automatically generating reports in natural language, it is necessary to connect the textual statements to the performace features. Indicators are the key performace features and the main building blocks for the narrative reports. Currently, inside the NARRATIVE system we have a list of 126 predefined indicators, which can combined together by the users as new indicators. The typical indicators are the following:

- *number of visits last week*
- *number of pageviews last week*
- *most popular keyword last week*
- *exit pages last week*
- *average time on site last week*
- *browsers used by visitors last week*
- *continents of visitors last week*
- *t-test visits last 4 weeks visits previous 4 weeks*
- *number of visits previous week*
- *top departures by visits last week*
- *visits bounce rare last week*
- *top paths by visits last week*
- *top arrivals by visits last week etc.*

2.3 Rules

Rules are used in the templates to trigger the natural language statement from the template to appear in the report. If all rules from the particular template validate, then the NL statement from the template is added to the report.
Rules contain indicators, numerical and logical operators. For instance, for the following rule to be validated, the indicator value *number of visits last week* should be larger than the indicator value *number of visits previous week*:

```
[number of visits last week] > [number of visits previous week]
```

or

```
[number of pageviews per visit last week] == [number of pageviews per visit previous week]
```

2.4 Templates

As stated above, templates combine together indicators, rules and statements in natural language:

```
<text>
        Traffic was down last week from the previous week
</text>
<rule>
        [number of visits last week]
        < [number of visits previous week]
</rule>
```

or

```
<text>
        The 4 week number of visits average is down significantly from the previous 4 week average.
</text>
<rule>
        [number of visits last 4 week average]
         < [number of visits previous 4 week average]
</rule>
<rule>
        [t-test visits last 4 weeks visits previous 4 weeks]
         < [5]
</rule>
```

Templates are merged into template sets. We assume that the text statement from no more than one template (from each template set) appears in the final report.

2.5 Report Scenarios

Finally, report scenarios contain the references to the specific template sets. If the user wants a template set to be included in the report, she should add it to the specific report scenario. For instance, a general weekly report can provide information about traffic (traffic was up, traffic was down, traffic stayed the same) in the last week with comparison to the previous week, about the most popular keyword last week, most popular browser last week etc.

3 SYSTEM INTERFACE

The NARRATOR system is accessible via web at:

http://narrator.ijs.si

Figure 2 presents a NARRATOR interface with connection to user's Google Analytics account. The user provides her Google Analytics username and password and, following that, gets the access to the websites suitable for analysis and report scenarios.

Figure 2: *NARRATOR interface – GA connection.*

Figure 3 demonstrates a performance report generated for the website **videolectures.net** [5] – a website that provides free and open access educational video lectures repository.
The report includes the performance analysis of the visitos to the website in the last week versus previous week. From the report it can be seen that traffic was up last week, as well as the number of pageviews and the number of visitors.

Figure 3: *NARRATOR interface – generated report.*

The NARRATOR website provides the user with possibility to add and modify new indicators (based on the predefined indicators), rules, templates and report scenarios.

4 DATA MINING TECHNIQUES

In order to allow the users to get more insides on their data, the NARRATOR system provides a mechanisms to analyze the website statistical information with using data mining techniques, such as decision trees.
Decision tree learning [6] represents a method commonly used in data mining. As an input, a list of variables is provided and the model that predicts a value of the target variable is built.
The interior tree nodes correspond to the input variables; the edges connect tree nodes and each leaf represents a value of the target variable given the values of the input variables taking to the account the path from the root to the leaf.
We have selected C4.5 algorithm [7] by Quinlan, which is an extension of ID3 algorithm. C4.5 can be used for classification, which is a benefit for mining the website data.
The motivation for mining the website data comes from the idea that we can provide the user not only with website performance reports, but also with interconnections between data attributes and useful attribute characteristics, which would contribute to the decision making process of the user.
In our data mining experiment, we set a task to build an experimental decision tree for the website **videolectures.net** for the attribute *average time on site last week*, which would connect the statistical website performance data (such as *mobile visitors, visitor type, visitor continents, visit day of week* – for details, see Table 1).

Table 1: *Average Time On Site - decision tree attribute values.*

Attribute Name	**Attribute Values**
isMobile	YES, NO
continent	AFRICA, AMERICAS, ASIA, EUROPE, OCEANIA, NOT_SET
visitorType	NEW_VISITOR, RETURNING_VISITOR
dayOfWeek	0 (Sunday),1,2,3,4,5,6
avgTimeOnSite	0_10, 10_100, 100_500, 500_AND_MORE (seconds)

The built decision tree (with 12 leaves) is provided below:

isMobile = YES
| continent = AFRICA
| | dayOfWeek <= 3
| | | dayOfWeek <= 0: 100_TO_500
| | | dayOfWeek > 0: 0_TO_10
| | dayOfWeek > 3: 10_TO_100
| continent = AMERICAS: 10_TO_100
| continent = ASIA
| | dayOfWeek <= 4: 10_TO_100
| | dayOfWeek > 4: 100_TO_500
| continent = EUROPE: 100_TO_500
| continent = OCEANIA: 10_TO_100
| continent = NOT_SET
| | dayOfWeek <= 0: 100_TO_500
| | dayOfWeek > 0: 10_TO_100
isMobile = NO
| visitorType = NEW_VISITOR: 100_TO_500
| visitorType = RETURNING_VISITOR: 500_AND_MORE

From the experimental decision tree, for instance, it is noticeable that more non-mobile returning visitors tend to spend 500 and more seconds at the videolectures.net website, while more non-mobile new visitors tend to spend between 100 and 500 seconds at the website.
In the future work we plan to extend the data mining part of the NARRATOR system and to provide the users with more possibilities to analyze their data.

5 CONCLUSION

NARRATOR is a system providing the possibility to transform the numerical data from Google Analytics services into reports in natural language.
NARRATOR provides a set of report scenarios, based on numerical and textual indicators.
NARRATOR uses Google Analytics data for a specific periods of time.
In addition, inside the NARRATOR system, we are developing a functionalty, which allows to get more detailed analysis of the website data. The data mining techniques, such as decision tree generation algorithms, are used to analyze the website statistical data and to discover the rules hidden in the data.
For the future work we consider the further development of the NARRATOR data mining functionalities, as well as the technology adaptation for other data sources and data streams.

6 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency, the IST Programme of the EC under PASCAL2 (IST-NoE-216886).

References

[1] Narrative Science, http://www.narrativescience.com (accessed in August 2012).
[2] Nichols, N. Machine-Generated Content: Creating Compelling New Content from Existing Online Sources. PhD thesis, Northwestern University, June 2010.
[3] Google Analytics, http://www.google.com/analytics (accessed in August 2012)
[4] Google Analytics API, https://developers.google.com/analytics (accessed in August 2012).
[5] Videolectures.net, http://videolectures.net (accessed in August 2012).
[6] Decision tree learning, http://en.wikipedia.org/wiki/Decision_tree_learning (accessed in August 2012).
[7] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.