Simulating Data Journalism to Communicate Hydrological Information from Sensor Networks

Martin Molina

Department of Artificial Intelligence
Technical University of Madrid, Spain
martin.molina@upm.es

Abstract. Presenting relevant information via web-based user friendly interfaces makes the information more accessible to the general public. This is especially useful for sensor networks that monitor natural environments. Adequately communicating this type of information helps increase awareness about the limited availability of natural resources and promotes their better use with sustainable practices. In this paper, I suggest an approach to communicating this information to wide audiences based on simulating data journalism using artificial intelligence techniques. I analyze this approach by describing a pioneer knowledge-based system called VSAIH, which looks for news in hydrological data from a national sensor network in Spain and creates news stories that general users can understand. VSAIH integrates artificial intelligence techniques, including a model-based data analyzer and a presentation planner. In the paper, I also describe characteristics of the hydrological national sensor network and the technical solutions applied by VSAIH to simulate data journalism.

Keywords: sensor networks, automated data journalism, intelligent knowledge-based system, multimedia-presentation system, computational sustainability.

1 Introduction

Sensor networks for monitoring natural environments usually generate important data that many potential users can use. Monitoring water in natural environments can help different types of users (e.g., municipalities, civil protection, consultants, scientist researchers or educators) make decisions related to agriculture, hydroelectric energy production, flood risk or climate change.

In general, it is important to communicate this type of information appropriately to increase the awareness of the limited availability of natural resources (i.e., water) and promote their better use with sustainable practices. It is thus useful to have web applications that make the information more accessible to the general public. However, these applications are often difficult to use because they present low-level information without adequate data interpretations and explanations.

In this paper, I describe an approach to improve communicating this type of information to wide audiences. This approach is based on data journalism, a professional practice in which journalists look for news in databases (usually online databases) and create a story understandable and useful to the general public.

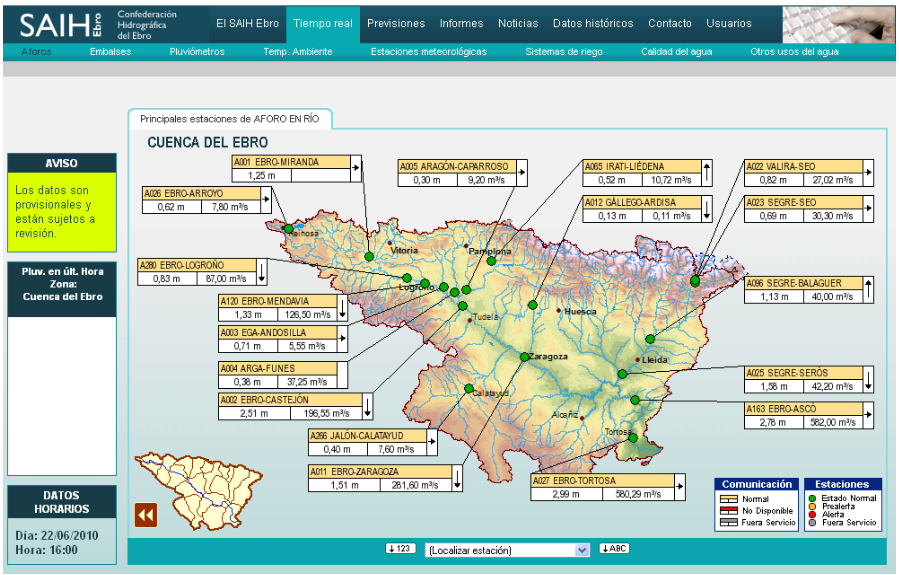
J. Pavón et al. (Eds.): IBERAMIA 2012, LNAI 7637, pp. 722–731, 2012.

I suggest the idea of using artificial intelligence techniques to simulate data journalism and improve sensor data communication. I analyze this approach by describing a pioneer system called VSAIH, which looks for news using hydrological data from a national sensor network in Spain and creates news stories that are potentially useful to different types of users. VSAIH integrates artificial intelligence techniques, including a model-based data analyzer and presentation planner, to generate descriptions.

The remainder of this paper describes a national hydrological sensor network. I then describe how VSAIH simulates data journalism to communicate hydrological information from the sensor network.

2 A National Sensor Network

In Spain, there is a national hydrological sensor network called SAIH (SAIH is the Spanish acronym for Hydrological Automatic Information System) [6]. The SAIH network measures real-time hydrological data using thousands of sensors geographically distributed in rivers and basins in Spain. SAIH is a mature infrastructure that has collected hydrological data for over 20 years.



CUENCA DEL EBRO

AVISO

Los datos son provisionales y están sujetos a revisión.

Pluv. en últ. Hora
Zona:
Cuenca del Ebro

DATOS HORARIOS

Día: 22/06/2010
Hora: 16:00

A001 EBRO-MIRANDA 1,25 m
A026 EBRO-ARROYO 0,62 m 7,80 m³/s
A280 EBRO-LOGROÑO 0,83 m 87,00 m³/s
A120 EBRO-MENDAVIA 1,33 m 126,50 m³/s
A003 EGA-ANDOSILLA 0,71 m 5,55 m³/s
A004 ARGA-FUNES 0,38 m 37,25 m³/s
A002 EBRO-CASTEJÓN 2,51 m 196,55 m³/s
A266 JALÓN-CALATAYUD 0,40 m 7,60 m³/s
A011 EBRO-ZARAGOZA 1,51 m 281,60 m³/s
A005 ARAGÓN-CAPARROSO 0,30 m 9,20 m³/s
A065 IRATI-LIÉDENA 0,52 m 10,72 m³/s
A012 GÁLLEGO-ARDISA 0,13 m 0,11 m³/s
A022 VALIRA-SEO 0,82 m 27,02 m³/s
A023 SEGRE-SEO 0,69 m 30,30 m³/s
A096 SEGRE-BALAGUER 1,13 m 40,00 m³/s
A025 SEGRE-SERÓS 1,58 m 42,20 m³/s
A163 EBRO-ASCÓ 2,78 m 582,00 m³/s
A027 EBRO-TORTOSA 2,99 m 580,29 m³/s

Fig. 1. Web application of the Ebro basin (provided by the *Confederación Hidrográfica del Ebro* in Spain)

The entire SAIH network is divided into nine sub-networks according to the main basins in Spain (e.g., Ebro, Tajo, Júcar). There are nine control centers in Spain, one for each sub-network. The information is recorded periodically and sent to the control centers (e.g., every 30 or 15 minutes). Control centers process and store the data in local databases. The Ministry of Environment of Spain coordinates the control centers and integrates the recorded data into a global database.

The SAIH system includes the following main types of sensors: pluviometers to measure rainfall at a certain geographic place, flow stations to measure water flow in river channels, level stations located at reservoirs or at river channels to measure the water level and volume stations located at reservoirs to measure the volume of stored water. The SAIH network also includes other sensors (e.g., sensors for gates in reservoirs, snow level) and telecommunication devices (e.g., radio emitter-receiver systems, optical fiber networks).

Because many users can use the data recorded by the SAIH network, the Ministry of Environment of Spain and the local governments related to the control centers created open web applications to help users freely consult the data. Figure 1 shows an example of a web application for one of the nine SAIH sub-networks (the Ebro basin). This screen shows a map of the Ebro basin where a user can consult real-time data about water flows and levels in rivers. It uses natural and intuitive presentations with graphics (usually maps and 2D charts for time series).

However, this type of web application presents some limitations when communicating information to general users. One main limitation is that the information is normally presented at sensor level, so it is difficult to have an aggregated view. The user must search sensor-by-sensor, consulting individual time series. It is difficult for the user to see the aggregated behavior of related information (e.g., temporal evolution of rain and related water flows). The web application also assumes that the user is familiar with hydrological measures and operations with graphical user interfaces to consult the data.

3 Simulating Data Journalism

The concept of data journalism [3, 11] identifies a type of professional practice in which human journalists look for news in databases (e.g., analyzing quantitative data from online databases). In general, journalists are experts at looking for news and writing stories using particular communication styles that help communicate with wide audiences.

In this paper, I suggest simulating data journalism with artificial intelligence methods to improve information communication from sensor networks to the general public. I use the term automated data journalism for this approach.

An example of automated data journalism in hydrology is the VSAIH application. VSAIH automatically analyzes sensor data from the SAIH sensor network and creates news stories that summarize relevant information. VSAIH generates news according to three different goals: flood risk management, water management, and sensor validation. These types of news are distributed in geographic areas according to the main river basins in Spain. There are ten areas that correspond to nine areas for sub-networks plus another geographic area for the whole nation. VSAIH thus includes 3 goals x 10 areas = 30 news generators.

Each generated piece of news follows the typical journalistic style used in newspapers. There is a headline that summarizes the main idea in a short sentence, and the body text develops the story to answer the usual journalistic questions (e.g., who, what, where) with evidential facts to support the affirmations. In this domain, it is important to provide adequate evidence presenting the actual sensor measures to help users trust the system's descriptions.

> ***Flow above normal in the Jerte river at El Torno***
>
> *The Jerte River at El Torno has a recorded flow of 38 m3/s, which represents an increase of 8 m3/s compared to the previous hour. The normal flow at this point of the river is 5 m3/s. With respect to this flow, the following hydrological behavior can be highlighted. It has rained in the Tajo Basin over the past 24 hours. Cabezuela del Valle is the location with maximum rainfall over the past 24 hours with a value of 24 mm. The rainfall at Las Becillas was 22 mm, and the rainfall at Los Angeles Casar was 21 mm. The Jerte Reservoir at Plasencia has a recorded volume increase of 1.44 Hm3 over the past 24 hours.*

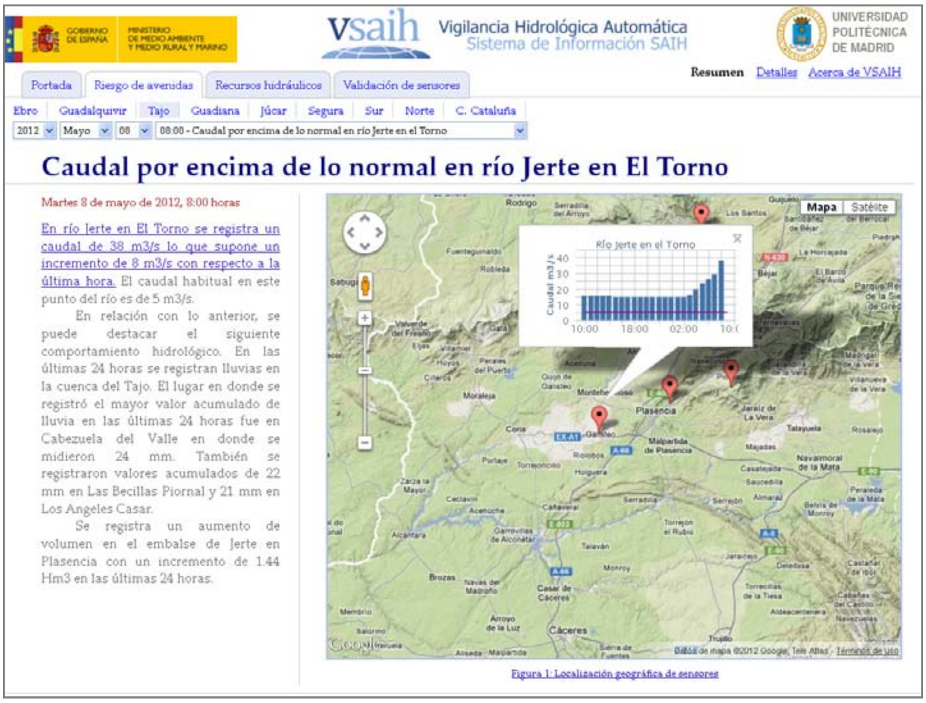
Fig. 2. Generated summary related to flood risk management

Figure 2 shows an example of a generated news story related to flood risk management (it is a translation from the original Spanish text in Fig. 3). The headline describes that there is high flow in a river. The body text follows a discourse that gives answers to questions such as what (high flow), which hydrologic component (the Jerte river), and where (at El Torno). The body text gives information to help explain the causes (places with significant rain in the previous 24 hours).

In general, the text presents a qualitative interpretation of hydrological quantities (e.g., high flows, moderate rain). The description shows quantitative measures recorded by sensors, which are contrasted with other values (previous hour and normal value) to help quantify the relevance of the situation. The text also shows spatial and temporal aggregations, including maximum flow values in rivers and cumulative rain amounts in the previous 24 hours.

The news is presented to the user on a user interface designed as a digital newspaper. Many users are familiar with this type of presentation, so this design helps facilitate communication with a wide audience. The digital newspaper includes a first page that summarizes the most relevant news. The user can select a news story to display on a new page with more details. Figure 3 shows an example page. Tabs at the top of the page allow the user to select different types of news.

Most news includes graphics complementing the text descriptions. The presented graphics depend on the content of the news. There are maps with highlighted geographic points to show, for example, the spatial location of river sections with certain flow or places with significant rain. The points on the map are related to the text descriptions using hyperlinks. This helps the user better understand the text description (especially when the user is not familiar with certain geographic areas). There are also charts to show time series that describe temporal measure behaviors. There are also graphical animations that show information from meteorological radar or image satellites.



Caudal por encima de lo normal en río Jerte en El Torno

Martes 8 de mayo de 2012, 8:00 horas

En río Jerte en El Torno se registra un caudal de 38 m3/s lo que supone un incremento de 8 m3/s con respecto a la última hora. El caudal habitual en este punto del río es de 5 m3/s.

En relación con lo anterior, se puede destacar el siguiente comportamiento hidrológico. En las últimas 24 horas se registran lluvias en la cuenca del Tajo. El lugar en donde se registró el mayor valor acumulado de lluvia en las últimas 24 horas fue en Cabezuela del Valle en donde se midieron 24 mm. También se registraron valores acumulados de 22 mm en Las Becillas Piornal y 21 mm en Los Angeles Casar.

Se registra un aumento de volumen en el embalse de Jerte en Plasencia con un incremento de 1.44 Hm3 en las últimas 24 horas.

Figura 1: Localización geográfica de sensores

Fig. 3. Sample presentation page generated by the VSAIH application (Fig. 2 shows English translation)

4 The VSAIH Architecture

To generate news, VSAIH simulates two main tasks: (1) analyzing hydrological data, and (2) generating presentations to effectively communicate the news to general users. To support these tasks, the VSAIH architecture follows a knowledge-based approach with three knowledge models for the system, abstraction and presentation [12]. The system model represents the set of river basins in Spain. This model follows a component-based approach using single components (e.g., river sections, reservoirs, geographic places where the rain is measured) and aggregated components (e.g., rivers, basins) with quantitative attributes and qualitative states.

The system model is used to analyze sensor data with an abstraction model, which uses logic- and rule-based representations to abstract data. Abstraction is important for generating headlines and concise descriptions in the body text. Abstraction includes the following more specific tasks: interpret quantitative data (i.e., the value of flow 362 m3/s means an above-normal flow in the Segre river), select the most relevant states (i.e., a flow above normal is more relevant than light rain), aggregate information (i.e., the Guadalhorce river and Limonero reservoir are part of the Andalusian basin), and generalize properties (i.e., the qualitative states light rain and heavy rain can be generalized by the qualitative state rain).

VSAIH uses a particular notion of relevance to filter data. The degree of relevance of a state is directly proportional to the distance between the current and normal (or desirable) states. Relevance is thus context-dependent. A low water level in rivers or reservoirs is a desirable state for flood risk, but it is not desirable for water management (e.g., for agriculture or hydroelectric energy production).

PATTERN: *describe_with_causes_and_effects*
GOAL: *inform*(x)
CONDITIONS: $\mathit{causes}(x, c) \wedge \mathit{effects}(y, e)$
DISCOURSE: {*summarize_state*(x), *elaborate_state*(x),
elaborate_causes(c), *elaborate_effects*(e)}

PATTERN: *describe_state_of_single_component*
GOAL: *elaborate_state*(x)
CONDITIONS: $(\mathit{state}(y, s) \in x) \wedge \mathit{single\_component}(y) \wedge \mathit{quantity}(y, q)$
DISCOURSE: {*elaborate_with_quantity*(y, q),
contrast_to_previous (y, q),
contrast_to_average(y, q)}
…

Fig. 4. Example discourse patterns

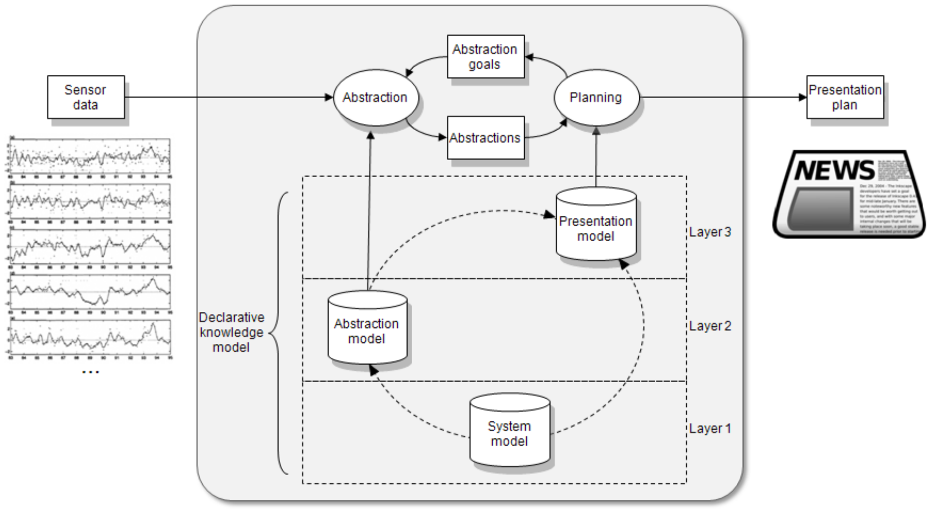


Fig. 5. Main VSAIH architecture components

VSAIH also uses a presentation model to generate the final descriptions. This model is represented as the knowledge base of a hierarchical planner following an HTN approach [8]. The model represents discourse strategies (called discourse patterns) to present the information. Figure 4 shows example discourse patterns. The planner generates the text description using partial text templates that are combined according to the hierarchical planning process. The planner also decides the most appropriate graphic to combine with texts (e.g., map or animation).

Figure 5 summarizes the main components of the VSAIH architecture. It includes the three models (the system, abstraction and presentation models) organized in three layers. The figure shows the main processes at the top (data analysis and presentation planning) and input/output relationships.

VSAIH generates news every hour using approximately 45,000 numerical measures (each page is generated in under 30 seconds). The first operational version of VSAIH was installed online in June 2008 using models corresponding to 3 river basins (Ebro, Segura and South of Spain). A second version was installed in May 2010 with improvements to the user interface and models corresponding to 9 river basins.

5 Discussion

Simulating data journalism can be an adequate solution for better communicating relevant information from sensor networks. VSAIH follows this approach to communicate information about water to a wide audience, using data measured by a national sensor network and presenting general descriptions that are useful for water management (e.g., for agriculture and energy production) and flood risk.

VSAIH can be considered a pioneer system that has successfully implemented the automated data journalism approach using artificial intelligence techniques. Some characteristics of VSAIH include the following:

- *News generators*. VSAIH includes 30 news generators to capture different types of news according to different goals (e.g., water management, risk floods or sensor validation in different geographic areas).
- *Interactive graphic newspaper*. VSAIH presents sets of news in an automatically generated digital newspaper with interactive graphics, a type of presentation that is familiar to most users. This user interface design contributes to better communication with general users.
- *Efficient computational architecture*. VSAIH processes thousands of sensors, combining different artificial intelligence techniques (e.g., abstraction, discourse planning, text generation, graphic generation) in an efficient computational architecture.
- *Practical utility*. The VSAIH evaluation showed that this system can save to operators up to 5 hours, especially in emergency situations (see evaluation details in [12]).

We are currently working on more advanced approaches to simulate data journalism. In general, our future work aims to design more general and flexible solutions. One challenge in designing this type of application is finding a balance between its efficiency (combining different AI techniques) and its generality and flexibility.

VSAIH efficiently generates text descriptions using a hierarchical planner with partial text templates. Instead of using templates, natural language generation techniques could provide more flexibility when generating descriptions [16, 17]. We are applying natural language generation techniques to automatically generate more complex and reusable narratives and geographic references [13]. We are also interested in identifying and formalizing discourses applicable to different domains.

VSAIH represents the hydrologic network using a system model following a component-approach. This is useful for certain sensor networks similar to the SAIH network. However, other networks (e.g., domains with moving sensors) can require different representations (e.g., event-based approaches) [14].

We are also interested in using general components for data analysis. VSAIH uses a particular data analyzer that was designed for this purpose. However, it is possible to use software libraries (for spatial and temporal reasoning) and tools (e.g., Weka or *R*) to implement data analyzers. Online data sources can help facilitate the development of data analyzers, thus reusing domain knowledge (e.g., geographic knowledge such as Open Street Maps [19]).

Finally, VSAIH presents the information using text in combination with graphics. Hyperlinks relate part of the text to the graphics. The relation between text and graphics could be improved with more flexible coordination, using solutions proposed in multimedia presentation systems [1].

6 Related Work

In general, researchers have explored the idea of combining computer systems and journalism from different perspectives. Computational journalism is a general field that involves applying information technologies to journalism activities [4, 7].

A more specific approach is summarizing news from documents by applying text summarization techniques and web mining [15]. This approach contrasts to the goal of VSAIH, as VSAIH automatically creates news stories by analyzing non-linguistic data instead of summarizing news from different documents.

The goal of a recent project called Stats Monkey is to develop a news generator for baseball games. This generator is similar to the news generators in VSAIH. Stats Monkey was initiated in 2009 and seems to apply a statistics-based approach with text templates[1] [10]. In contrast, VSAIH is an operational system, first installed online in 2008, that follows a complex knowledge-based approach with a collection of 30 news generators. Each news generator uses a model-based data analyzer and a hierarchical planner. VSAIH also includes a dynamically generated multimedia user interface as a multi-page digital newspaper to present the collection of generated news.

The VSAIH architecture includes components that are present in intelligent multimedia presentation systems [1] (e.g., presentation planning, text generation). Bateman et al. [2] explore how to automatically construct the page layout (using rhetorical relationships and design heuristics) with the Dart$_{\text{bio}}$ prototype. VSAIH uses a presentation planner to select the best page layout (from a prefixed set of layouts), which can efficiently and flexibly combine different types of text descriptions and graphics.

[1] We have not found scientific publications related to this project or an operational web application. It is thus difficult to know the degree of success of this project.

The text generation in VSAIH is similar to the goal of data-to-text systems, which generate natural language text from non-linguistic data [9, 18]. Compared to other data-to-text systems, VSAIH can generate complex narratives (with rhetorical relationships such as contrast, exemplify, cause and elaboration) because it uses a model-based data analyzer and a discourse planner with discourse patterns. As mentioned above, VSAIH uses text templates, i.e., it does not apply advanced natural language generation techniques that could provide more flexibility to the generated sentences.

VSAIH uses a system model to capture qualitative descriptions, as in physical reasoning [5]. As a main difference from other qualitative reasoning approaches, the VSAIH representation is designed to simulate abstraction instead of other tasks that require more precision (e.g., diagnosis or prediction). Compared to qualitative reasoning representations, the VSAIH representation is simpler and thus more efficient for both inference and knowledge acquisition.

7 Conclusions

Sensor networks for monitoring natural resources usually generate large amounts of data that are useful for many users. Monitoring water in natural environments can help different types of users, e.g., municipalities, civil protection, consultants, scientific researchers and educators, make decisions related to agriculture, energy production and flood risk. An adequate communication of this type of information can help increase awareness about the limited availability of natural resources and promote their better use with sustainable practices.

In this paper, I argue that simulating data journalism with artificial intelligence techniques can help communicate information from sensor networks to the general public. I illustrate this concept with the VSAIH system, a successful pioneer web-based application that follows this approach in hydrology.

VSAIH generates news from thousands of hydrological sensors using 30 specialized news generators and presents the news on a dynamically generated user interface designed as a digital newspaper. VSAIH combines an intelligent data analyzer using a model-based approach and a presentation planner to generate text-graphic presentations.

This type of solution has great applicability, especially for sensor networks that monitor natural environments. Our future research in this area aims to build general and flexible solutions that can help developers implement this type of application in different domains.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement PIOF-GA-2009-253331 (Project INTERACTIVEX). This work was also partially supported by the Ministry of Science and Innovation of Spain within the VIOMATICA project (TIN2008-05837). This work was also possible thanks to the support of the Ministry of Environment of Spain (*Dirección General del Agua, Ministerio de Medio Ambiente, Medio Rural y Marino*).

References

1. André, E.: The Generation of Multimedia Presentations. In: Dale, R., Moisl, H., Somers, H. (eds.) A Handbook of Natural Language Processing Techniques and Applications for the Processing of Language as Text, pp. 305–327. Marcel Dekker Inc., New York (2000)
2. Bateman, J., Kleinz, J., Kamps, T., Reichenberger, K.: Towards Constructive Text, Diagram, and Layout Generation for Information Presentation. Computational Linguistics 27(3), 409–449 (2001)
3. Bradshaw, P.: How to Be a Data Journalist. The Guardian (October 1, 2010)
4. Cohen, S., Li, C., Yang, J., Yu, C.: Computational Journalism A Call to Arms to Database Researchers. In: 5th Biennial Conference on Innovative Data Systems Research (CIDR 2011), pp. 148–151 (2011)
5. Davis, E.: Physical Reasoning. In: van Harmelen, F., Lifschitz, V., Porter, B. (eds.) Handbook of Knowledge Representation, pp. 597–620. Elsevier, Oxford (2007)
6. Dirección General del Agua: El Programa S.A.I.H. – Descripción y Funcionalidad, el Presente y el Futuro del Sistema. Ministerio de Medio Ambiente y Medio Rural y Marino (Spain) (2009), `http://www.mma.es`
7. Flew, T., Daniel, A., Spurgeon, C., Swift, A.: The Promise of Computational Journalism. Journalistic Practice 6(2), 157–161 (2012)
8. Ghallab, M., Nau, D., Traverso, P.: Automated Planning: Theory and Practice. Morgan Kaufmann (2004)
9. Hunter, J., Gatt, A., Portet, F., Reiter, E., Sripada, S.: Using Natural Language Generation Technology to Improve Information Flows in Intensive Care Units. In: 18th European Conference on Artificial Intelligence (ECAI 2008), pp. 678–682 (2008)
10. Infolab. Intelligent Information Laboratory at Northwestern University. Stats Monkey project (2012), `http://infolab.northwestern.edu/projects/stats-monkey/`
11. Lorenz, M.: Data Driven Journalism: What is There to Learn? In: Innovation Journalism Conference (IJ-7), Stanford, CA, June 7-9 (2010), `http://datadrivenjournalism.net`
12. Molina, M., Flores, V.: Generating Multimedia Presentations that Summarize the Behavior of Dynamic Systems Using a Model-Based Approach. Expert Systems with Applications 39, 2759–2770 (2012)
13. Molina, M., Stent, A.: A Knowledge-Based Method for Generating Summaries of Spatial Movement in Geographic Areas. International Journal on Artificial Intelligence Tools 19(4), 393–415 (2010)
14. Molina, M., Stent, A., Parodi, E.: Generating Automated News to Explain the Meaning of Sensor Data. In: Gama, J., Bradley, E., Hollmén, J. (eds.) IDA 2011. LNCS, vol. 7014, pp. 282–293. Springer, Heidelberg (2011)
15. Radev, D.R., Otterbacher, J., Winkel, A., Blair-Goldensohn, S.: NewsInEssence Summarizing Online News Topics. Communications of the ACM 48(10), 95–98 (2005)
16. Reiter, E.: NLG vs. Templates. In: 5th European Workshop on Natural Language Generation (1995)
17. Reiter, E., Dale, R.: Building Natural Language Generation Systems. Cambridge University Press, Cambridge (2000)
18. Reiter, E., Sripada, S., Hunter, J., Yu, J., Davy, I.: Choosing Words in Computer-Generated Weather Forecasts. Artificial Intelligence 67(1-2), 137–169 (2005)
19. Roth, M., Frank, A.: A NLG-based application for walking direction. In: ACL-IJCNLP 2009 Software Demonstrations (ACLDemos 2009), pp. 37–40 (2009)