"Vasile Alecsandri" University of Bacău
Faculty of Sciences
Scientific Studies and Research
Series Mathematics and Informatics
Vol. 23 (2013), No. 1, 85 - 96

MEASURES FOR UNCERTAIN DATA. CASE STUDY ON DATA EXTRACTED FROM MASS MEDIA

ELENA NECHITA

Abstract. In most of the applications related to real world situations, especially in those dealing with large data sets, uncertainty is unavoidable. Depending on the sources where data come from, uncertainty also comes in different facets. For the computer scientists, the challenge is to process these types of data in such a way that the end user receives the needed information with as much accuracy as possible. Therefore, specific methods had to be developed to deal with the uncertainty characteristic of the data.

This paper investigates and compares the information given by two measures of uncertainty, namely uncertainty density and answer decisiveness, when applied to a set of data extracted from mass media. As well, several issues related to uncertainty are going to be discussed for digital media, together with their implications.

1. INTRODUCTION

In most of the applications related to real world situations, especially in those dealing with large data sets, uncertainty is unavoidable. During the last years, the field of data management addressed special attention to uncertain data because of the new technologies for collecting data in an imprecise way.

Although uncertainty has been primarily expressed by means of probability ([19] presents an extended review on this approach) there are a lot of other ways to model it, due to different approaches on uncertain data [8, 10, 18]. However, there is a great need for tools to handle databases containing uncertain data and to mine them [1].

Keywords and phrases: uncertainty, data quality, digital media.
(2010) Mathematics Subject Classification: 68P20.

In many situations, such as those requiring information from different sources, data integration is a complex task. As an example, let us consider a database whose aim is to support the healthcare system in assisting people who have complex needs, usually arising from a combination of two or more issues such as mental illness, physical disability, behavioural difficulties, drug and/or alcohol use, etc. The sources where the data come from may include: police (population records), health services, corrective services, community services, housing facilities, justice [3].

Another example comes from social sciences. Social phenomena are complex and opened to a large number of influences, requiring attention on multiple dimensions. When specific parts of a big picture are studied with different methods and put together in order to get a better understanding of the phenomena, inconsistent data can easily be produced. In such cases, specialists talk about "linking data" or "meshing methods" rather than of integration [15].

Scientific measurements or sensors may lead to imprecise data [20], the monitoring of the environment implies values which are inherently imprecise [2], the information gathered in contexts related to law and justice can easily rise contradictions [21], and the list of the examples could be extended significantly. As applications need to reduce uncertainty by correcting for systematic error and minimizing random errors, appropriate algorithms and heuristics [6] have to be provided by computer scientists.

2. UNCERTAINTY VERSUS QUALITY DATA IN THE NEW MEDIA

We generally regard information as a good thing, but too much information – as often happens in media - may be conflicting [7]. Of course, it is better to find contradictory information on a search topic rather than finding no information at all, because this situation shows interest in that topic. Sometimes, the differences in the content of certain information are provided by different sources. A relevant example is described in the paper of Rubin, "How the News Media Reported on Three Mile Island and Chernobyl" [17], related to crucial information after the nuclear power plant accident at Three Mile Island: amount of radiation, rate, time, duration and location of the release, type of radioactive materials, impact. Both the American Agencies as well as the western European governments provided contradictory information, sometimes in unclear formats, or no information at all.

Although there is no consensus on what "quality information" means when it is provided by media, some of the most important characteristics that we think of are: accessibility, accuracy, availability, completeness, integrity, redundancy, reliability, timeliness, trustworthiness, usability. For some fields, uncertain information acquired through media channels may have important

consequences. In marketing, conflicting information about products influence decision makers in deciding weather to persist or abandon product usage, sometimes with vital effects on human heath, such as in drug use [12]. Related to this topic, there is also an increasing concern on health information presented in media. For example, in [16] the authors analyze the effects of such (potential) media exposure to nutritional information by means of four measures, considering two dimensions: obtrusiveness (the presence of conflicting or contradictory information) as opposing to content specificity. This is the reason why, in the United States, there has been introduced partial guidance regarding to the information that the federal agencies disseminate. The Data Quality Act (also referred to as the Information Quality Act [4]), enacted in 2000, was such an attempt to define the key concepts that might ensure quality, objectivity, utility, and integrity of information (including the statistical one).

Of course, a general conceptual framework on this issue is not a simple task to accomplish, due to the different characteristics of data sources, attributes, interests, and needs in processing the information in different fields of human expertise. In what follows we shall refer the large category of information provided by mass media.

It is well known that mass media influences public perception and (re)actions. Broadcast media, print media, outdoor media and - lately - more and more digital media reach a large audience. Social, mobile and other digital technologies allow access at any time and any place, on any digital device, and also facilitate interactivity with the users [11]. Therefore, information provided through these channels is obliged to be at least relevant, reliable, and accurate.

New media is now a part of the main press offices. Web pages, Twitter, Facebook, the blogosphere are expected to be updated and give the users trustable information. Segregating traditional and “new” media can result in conflicting messages, which is why both should be integrated for efficient communication tasks. Newspapers (traditional and online), wire reports and other journalistic accounts offer an immense volume of information, where uncertainty in its different facets is certainly present. The problem of inconsistencies in media is not new and has been extensively studied, as inappropriate handling of information may lead to media scepticism. As it was defined in [5], this quality indicator may be regarded as the degree to which individuals are sceptical toward the reality presented in the mass media.

3. A CASE STUDY ON CONTRADICTORY INFORMATION ABOUT DESTINY OF HOSTAGES IN ALGERIA, IN JANUARY 2013

Through the page http://english.ruvr.ru/2013_01_17/Contradictory-information-about-destiny-of-hostages-in-Algeria/, the (online) radio "The Voice of Russia" signalled contradictory information reporting on the Algerian army's operation of freeing hostages who have been captured by Islamists in a Sahara Desert gas complex, in January 2013. The sources, the core information as well as the type of items that this information addresses are given in the following table:

Variant	Source	Information (in Italics, missing information)	Describes
a	Official Algerian	The army is storming the house where the hostages are held.	Action
b	Algerian news agency APS	600 Algerian hostages have already been released. (*Nothing has yet been announced about the destiny of foreign hostages*).	Action, Nationality, Number of hostages
c	Militants (Islamists)	35 hostages have been killed in an Algerian military raid.	Action, Number of hostages
d	Algerian news service ANP	The Algerian military conducted air strikes and a ground operation to free the hostages, who were picked up by military helicopters. (*It remained unclear if any of the hostages were injured*).	Action, Consequence
e	Hostage takers (Islamists), witness and media reports	The helicopter strikes allowed some 200 Algerian workers and several foreigners to flee and killed several of the militants.	Action, Nationality, Number of hostages, Consequence
f	Mauritania's ANI news agency	Militants claimed that 35 hostages and 15 captors were killed. (*Not confirmed*).	Nationality, Number of hostages, Consequence
g	Mauritania's ANI news agency	Militants were still holding two Americans, three Belgians, a Japanese and a British. The	Nationality, Number of hostages,

		original group also included several Norwegians, a Romanian and an Austrian. The militants have demanded a end to France's military operatic against Islamist rebels neighbouring Mali and said the wanted to also punish Algeria f allowing French warplanes overfly the country	Consequence, Cause, Request
h	Ireland representative	An Irish had escaped and four hostages from Britain, France and Kenya were freed. (*It was unclear how many peopl were injured or killed*).	Nationality, Number of hostages
i	The "Blood Signatories" brigade	The brigade claimed to be holding 41 hostages.	Number of hostages
j	Algerian radio	The military had launched Thursday's attack after the hostage-takers attempted to flee with a number of captives.	Action, Cause

In order to measure the quality of this data collection, we shall use the approach presented in [13], taking into account the semantics of the data and the impact that it could have on the reader.

At first, we shall integrate the information gathered in the previous table into a relational system, choosing the following attributes: *Nationality* (representing the nationality of the hostages or the country of origin; the value “foreigners” has the meaning “non-Algerians”), *State* (representing the state of the hostages as a consequence of an action: released, held, etc.), and *Number* (representing the number of hostages that are subject to the state described in the previous column). An *Id* (first column) has also been introduced for further reference of the tuples, while the last column indicates the variant each tuple comes from and is just a comment, not an attribute of the relational system.

When transferring the information variants into the relational system, some data (such as context, causes, types of actions, fight/rescue means) have been deliberately avoided, as the information is complex and highly conflicting and had to be simplified for the aims of the paper. Empty cells correspond to null (in this case, unknown) values. We also allowed lists of values for the attribute *Nationality*. Therefore, the table is not in the First Normal Form; as we shall

further eliminate some tuples, this is not an issue and has been done for an accurate mapping of the information from the first table to the second one, denoted *News*.

News

Id	***Nationality***	***State***	***Number***	*Comment*: Coming from Variant
1		held		a
2	Algerian	released	600	b
3	foreigners			
4		killed	35	c
5		freed		d
6	Algerian	fled	200	e
7	foreigners	fled		
8	Islamists	killed		
9	Algerian, foreigners	killed	35	f
10	Islamists	killed	15	
11	American	held	2	g
12	Belgian	held	3	
13	Japanese	held	1	
14	British	held	1	
15	Norwegian	held		
16	Romanian	held	1	
17	Austrian	held	1	
18	Irish	escaped	1	h
19	British, French, Kenyan	freed	4	
20		injured		
21		killed		
22		held	41	i
23	Islamists, Algerian, foreigners			j

The next step is to perform some processing of the data, in order to avoid information loss: the list of values “Algerian, foreigners” of the attribute *Nationality* (tuple 9) will be assimilated with a new value, “hostages”; tuples 11, 12, 13, 14, 16, 17 with the *State* value “held” are summed up in line 11 and given “hostages” for *Nationality*; the list of values “British, French, Kenyan” of the attribute *Nationality* (tuple 19) will be also assimilated with the value “foreigners”; the null value of *Nationality* (tuple 22) will be also assimilated with the value “hostages”. All these data transformations take into account the

significance of the data, in a context that a human analyst is familiar with (it is well known that such a task is not at all simple for an automated text analysis [9]). After selecting the tuples having only atomic, non-null values, the result comes in the table below, denoted *Selected_News*.

In the table *Selected_News*, *Confidence* is a probability value representing the perceived amount of uncertainty for the set of the attributes of each tuple. Logically, the confidence values must correlate with the trust upon the source of the information, but in this case study the values have been arbitrarily assigned, as we do not hold any specific information related to the sources mentioned in the previous table. However, the values have been chosen so as to cover correctly the possible representations or information about the same real object. When there is no conflicting information about an object, the confidence value is 1 (for example, there is only one piece of information about the Islamists, about the Irish hostage, and about the freed foreigners). With this remark, the tuples to be analysed are: 2, 6, 9, 11, and 22. The analyses will be performed on two sets, given the semantics of the data.

Selected_News

Id	***Nationality***	***State***	***Number***	**Confidence**
2	Algerian	released	600	0.8 (for *State* and *Number*)
6	Algerian	fled	200	0.2 (for *State* and *Number*)
9	hostages	killed	35	0.3 (for *State*)
10	Islamists	killed	15	1 (*for all attributes*)
11	hostages	held	9	0.6 for *State*, 0.8 for *Number*
18	Irish	escaped	1	1 (*for all attributes*)
19	foreigners	freed	4	1 (*for all attributes*)
22	hostages	held	41	0.1 for *State*, 0.2 for *Number*

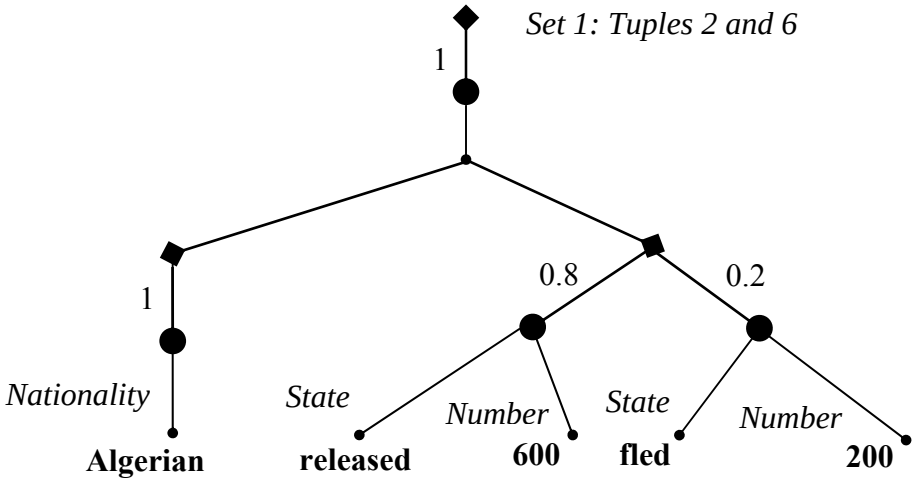
The data contained in the first two tuples (*Set 1*) can be represented into the tree below (the attributes *State* and *Number* are considered dependent, as they come from the same source and no other information related to Algerians is available). In this tree, the nodes marked with squares are *probability* nodes which induce *choice points*, while those marked with big disks are *possibility* nodes.

The following notations, as introduced in [13], describe the structure of the tree and are needed for the computation of the uncertainty measures:

Ncp – the number of choice points in the data

Nposs,cp – the number of possibilities or alternatives of choice point cp and

Pmax,cp – the probability of the most likely possibility of choice point cp.



For the tree above, these values are:

Ncp = 3

Nposs,1 = Nposs,2 = 1, Nposs,3 = 2 and

Pmax,1 = Pmax,2 = 1, Pmax,3 = 0.8

As introduced in [13], two values that give a measure of the uncertainty amount for the information represented in the tree are computed as follows:

i. Uncertainty density

$$\text{Dens}(\textit{Set 1}) = 1 - \frac{1}{\text{Ncp}} \sum_{j=1}^{\text{Ncp}} \frac{1}{\text{Nposs, j}} = 1 - \frac{1}{3}\left(\frac{1}{1} + \frac{1}{1} + \frac{1}{2}\right) = 0.16$$

ii. Answer decisiveness

$$\text{Dec}(\textit{Set 1}) = \frac{1}{\text{Ncp}} \sum_{j=1}^{\text{Ncp}} \frac{\text{P max, j}}{(2 - \text{P max, j}) \cdot \log_2(\max(2, \text{Nposs, j}))} =$$

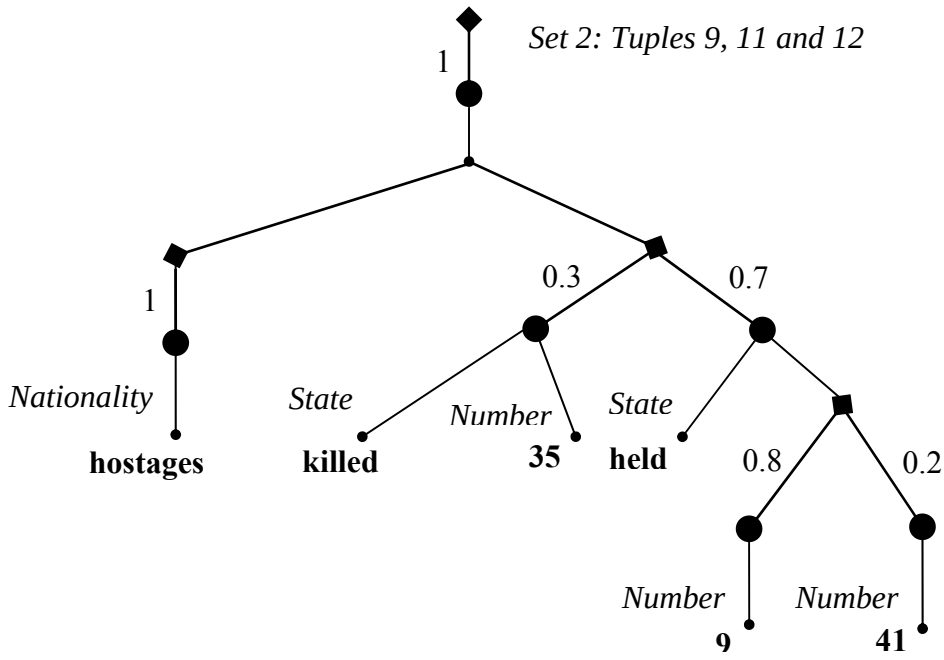
$$= \frac{1}{3}\left(1 + 1 + \frac{0.8}{2 - 0.8}\right) = 0.88$$

The second set of information (*Set 2*) that exhibits uncertainty is made of tuples 9, 11, and 22. We mapped the data contained in these tuples into the following tree, whose structure is reflected in:

Ncp = 4

Nposs,1 = Nposs,2 = 1, Nposs,3 = Nposs,4 = 2 and

Pmax,1 = Pmax,2 = 1, Pmax,3 = 0.7, Pmax,4 = 0.8



The uncertainty values for the corresponding data are:

i. Uncertainty density

$$\text{Dens}(\textit{Set 2}) = 1 - \frac{1}{4}\left(\frac{1}{1} + \frac{1}{1} + \frac{1}{2} + \frac{1}{2}\right) = 0.25$$

ii. Answer decisiveness

$$\text{Dec}(\textit{Set 2}) = \frac{1}{4}\left(1 + 1 + \frac{0.7}{2 - 0.7} + \frac{0.8}{2 - 0.8}\right) = 0.80$$

The relation Dens(*Set 1*) < Dens(*Set 2*) is easy to interpret, as the *uncertainty density* is based on the number of choice points and on the number of alternatives in each choice point, and does not depend on the confidence values. *Set 2* has more uncertain data than *Set 1*. If no uncertainty would exist in the data, the uncertainty density would be 0.

The value $1 - \frac{1}{n} = 1 - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{n}$ for this parameter corresponds to a tree with n choice points and n alternatives in each point, and approaches 1 as n goes to infinite.

On the other hand, Dec(*Set 1*) > Dec(*Set 2*), therefore the answer decisiveness and the uncertainty density are conversely correlated.

The parameter *answer decisiveness* considers the highest probability for each choice point, hence giving weight to the "most probable" possible world. The data describing this world is most likely to be returned in a query on the system. This is the case in *Set 2*, where we would intuitively deduce that the statement "9 hostages were held" is the most credible one.

Suppose we would have $N_{poss,4} = 5$ and all the five probabilities in this choice point would be 0.2. In this case, Dec(*Set 2*) would be

$$\text{Dec}(\textit{Set 2}) = \frac{1}{4}\left(1+1+\frac{0.7}{2-0.7}+\frac{0.2}{2-\log_2 5}\right) = 0.47$$

therefore the answer decisiveness of *Set 2* decreases. It is easy to prove that a high decisiveness value results from small numbers of alternatives in the choice points and, for the same number of possible worlds, the decisiveness increases if the probability of one of these possible worlds is close to 1. The answer decisiveness is 1 if there is no uncertainty in data.

4. CONCLUSIONS AND FUTURE WORK

Large volumes of heterogeneous data are nowadays gathered due to the proliferation of digital services, such as call centres, point-of-sale systems, websites, and social media. Many of these applications deal with data that is uncertain, but either ignore this characteristic or handle it themselves [13]. There is a need for solutions to assess the quality of an answer given by a system dealing with uncertain data.

In this paper we have extracted some information exhibiting uncertainty from online media, and computed two parameters to measure this uncertainty: *uncertainty density* and *answer decisiveness*. Our conclusion is that the values of these complimentary parameters give the analyst some information on the structure of the data set and, consequently, the possibility to give a proper interpretation of the answers of the system. *Uncertainty density* takes into account the amount of certain data, while *answer decisiveness* is an indicator of how well a most likely answer can be distinguished among a set of possible answers.

The case study approached in this paper also raised an educational issue. Given so much uncertainty of the information in mass media, and especially in the new media, we should pay more attention on how the young people extract the information they need. As some authors suggest [14], we have to reconsider the assumption according to which young people, as "digital natives", are able to use online information effectively. Moreover, research is required in every field of human expertise, in order to propose new, specific methods to deal with uncertainty [22].

In a further study we shall compute and compare the values of *uncertainty density* and *answer decisiveness* for larger data sets and explore the impact of the answers returned by regular queries on those data.

REFERENCES

[1] C.C. Aggarwal (Ed.), **Managing and Mining Uncertain Data**, Springer, 2008.

[2] K. Beven, **Towards integrated environmental models of everywhere: uncertainty, data and modeling as a learning process**, Hydrol. Earth Syst. Science 11(1) (2007), 460-467.

[3] L. Burnside, **Special Report for the Office of the Children's Advocate**, 2012, http://www.childrensadvocate.mb.ca/wp-content/uploads/Youth-with-Complex-Needs-Report-final.pdf, accessed online at May 10, 2013.

[4] Center for Effective Government, **Data Quality Act**, http://www.foreffectivegov.org/node/3479, accessed online at May 10, 2013.

[5] M. D. Cozzens, N. S. Contractor, **The Effect of Conflicting Information on Media Skepticism**, Communication Research 14 (4) (1987), 437-451.

[6] G.C. Crişan, C.M. Pintea, C. Chira, **Risk assessment for incoherent data,** Environmental Engineering and Management Journal 11(12) (2012), 2169-2174.

[7] J.M. Dunn, **Contradictory Information – Too Much of a Good Thing**, Journal of Philosophical Logic 39(4) (2010), 425-452.

[8] H. Garcia-Molina, and D. Porter, **The Management of Probabilistic Data**, in EEE Transactions on Knowledge and Data Engineering 4 (1992), 487-501.

[9] J. Grimmer and B. M. Stewart, **Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts**, Political Analysis 21 (3) (Summer 2013), 267-297, doi:10.1093/pan/mps028.

[10] E. Ioannou, **Entity-Aware Query Processing for Heterogeneous data with Uncertainty and Correlations**, Proceedings of the 2009 EDBT/ICDT Workshops, 170-176, ACM New York, 2009.

[11] H. Jenkins, S. Ford, J. Green**, Spreadable Media: Creating Value and Meaning in a Networked Culture**, New York University Press, 2012.

[12] A. Kalra, S. Li, W. Zhang, **Understanding Responses to Contradictory Information about Products**, Marketing Science 30(6) (2011), 1098-1114.

[13] A. de Keijzer, M. von Keulen, **Quality Measures in Uncertain Data Management**, H. Prade and V.S. Subrahmanian (Eds.): SUM 2007, Lecture Notes in Artificial Intelligence 4772, 104-115, 2007.

[14] T. Lumley, J. Mendelovits, **How well do young people deal with contradictory and unreliable information on line? What the PISA digital reading assessment tells us**, The Annual Conference of the American Educational Research Association (AERA), Vancouver, 2012. Available at: http://works.bepress.com/juliette_mendelovits/4.

[15] J. Mason, **Mixing methods in a qualitatively driven way**, Qualitative Research 6(1) (2006), 9-25.

[16] R.H. Nagler, R. C. Hornik, **Measuring Media Exposure to Contradictory Health Information: A Comparative Analysis of Four Potential Measures**, Communication Methods and Measures 6(1) (2012), 56-75.
[17] D.M. Rubin, **How the News Media Reported on three Mile Island and Chernobyl**, Communicating Risk: The Media and the Public, 1987, online at http://www.penelopeironstone.com/Rubin.pdf.
[18] A. das Sarma, O. Benjelloun, A. Halevy, and J. Widom, **Working Models for Uncertain Data**, in ICDE Conference Proceedings, 2006.
[19] M. Thimm, **Probabilistic Reasoning with Incomplete and Inconsistent Beliefs**, Dissertations in Artificial Intelligence, IOS Press, 2012.
[20] C.H. Wagner, **Uncertainty in Science and Statistics**, The Two-Year College Mathematics Journal 14(4) (1983), 360-363.
[21] D.A.Waterman, M.A. Peterson, **Expert Systems for Legal Decision Making**, Expert Systems 3(4) (1986), 212-225.
[22] C. Weiss, **Expressing scientific uncertainty**, Law, Probability and Risk 2 (2003), 25-46.

"Vasile Alecsandri" University of Bacău
Faculty of Sciences
Department of Mathematics, Informatics and Education Sciences
157 Calea Mărăşeşti, Bacău, 600115, ROMANIA
e-mail: enechita@ub.ro