A model of architecture for peer-to-peer systems based on information quality

Horacio Paggi
Universidad Politécnica de Madrid
Campus de Montegancedo
28660 Boadilla del Monte (Madrid)
+598 99165376
horacio.paggi@gmail.com

Javier Soriano
Universidad Politécnica de Madrid
Campus de Montegancedo
28660 Boadilla del Monte (Madrid)
+34 91 336 74 00
javier.soriano@upm.es

Juan A. Lara
Madrid Open University, UDIMA
Carretera de La Coruña km 38,5,
Vía de Servicio, nº 15
28400 Collado Villalba
+34 91 856 16 99 (ext. 3591)
juanalfonso.lara@udima.es

ABSTRACT
In this paper, we present a model of architecture based on flat peer-to-peer (P2P) networks. This architecture aims to reduce the impact of information vagueness and uncertainty on decision making using an information quality metric to manage its components. The model makes no assumptions about the homogeneity of the components or of the data they handle; however, all components must all be able to process the queries that they receive taking into account different qualities and response times, as well as limitations on the number of interactions that they can each undertake (limited number of messages, limited energy for communications, limited bandwidth or number of transmissible data). Part of the response process is to fuse information on the incoming data. Systems with these features have a wide variety of applications, ranging from decision support in critical environments (disaster areas, war, etc.) to mobile recommendation systems.

CCS Concepts
• Networks→Network types→Overlay and other logical network structures→Peer-to-peer networks • Mathematics of computing→Information theory.

Keywords
Adaptive P2P systems; multi-agent information fusion system; uncertain information handling; information quality metric.

1. INTRODUCTION
The proliferation of interconnectable devices, as well as the increasing interaction of people with intelligent machines has given rise to a host of challenges, for example, mobile communications equipment battery life, communications security, available data bandwidths, etc. As such interconnections can occur in any order (think of a mobile phone connecting to a random Bluetooth device in its environment) and seeing how the quality of information received can be very diverse, especially considering communications among human beings, the aim was to design a model of architecture capable of achieving improved performance, irrespective of its components, based on a flat network where all connections are possible (peer-to-peer) that had no specialized element (hubs, servers, etc.) [6]. To do this, we took into account the quality of the information circulating in the network. We tried to maximize information quality for each network element at any time, which, considering that resources are random and limited, will not necessarily be the overall maximum at that time.

This paper is organized as follows. Section 2 describes the state of the art with respect to information quality (IQ) metrics and forms of representing this IQ. Section 3 describes some characteristics of the proposed model such as the defined IQ metric and the dynamics of a system based on this architecture. Section 4 shows some preliminary results after instantiating the model as a specific system and applying it to a case study. Finally, Section 5 discusses some conclusions and outlines possible future work.

DATA'19, December 2–5, 2019, Dubai, United Arab Emirates 2019 Association for Computing Machinery. ACM ISBN 978-1-4503-7284-8/19/12…$15.00

https://doi.org/10.1145/3368691.3368692

2. STATE OF THE ART
2.1 Information quality
In this paper, we, like other authors, use information quality (IQ) and data quality as equivalent terms [5].

Data quality can be viewed as a set of dimensions describing the quality of the in-formation produced by the information system [2].

IQ usually accounts for accuracy, timeliness, completeness, consistency, relevance and fitness for use [7, 15]. Researchers have considered other alternative dimensions according to the different frameworks used for evaluation. An outline of 12 such dimensions is given in [5]. Hence, IQ influences the quality of the decisions made ([4, 10]).

We consider two of all the above dimensions of IQ regarded in the literature as critical: vagueness and uncertainty. They can be viewed as two facets of indeterminacy. Indeed, according to Novák, "uncertainty and vagueness form two complementary facets of a more general phenomenon which we may call indeterminacy" [9]. Such indeterminacy refers to how much we know about the consequences of receiving a message. Therefore, we are interested in vagueness and uncertainty insofar as they can affect agent beliefs and responses, as in [11]. For a thorough analysis of the different aspects of indeterminacy, see [1].

2.2 Characteristics of IQ metrics and tags
To manage quality effectively, it is crucial to provide decision makers with IQ metadata, data that describe the quality of the data used in the decision-making process. IQ is a metadata item that can

be represented in different ways (visual, discrete numerical, continuous numerical and linguistic variables, some of which are more usable for human beings. For a detailed study of this issue, see [13]. They suggest that information quality should be plotted as a chart, obviously for human use in decision making.

IQ metadata, taking the form of a [0, 1] ratio or as a percentage representing the quality, help decision makers to gauge data quality in the context of the decision-making task [3, 16].

We did not specifically consider human usability for the proposed architecture; we looked for the simplest human- and machine-readable representation that was at the same time easy to calculate and did not generate a major computational overload.

Decision makers may suffer from information overload as a result of the use of such metadata within the decision-making process. On this ground, the system uses a set of preferred agents.

The tags are usually designed to be created and used by people and not by ma-chines. This means that the right tags are not available for fast-growing fields like the IoT.

2.3 Use of information fusion to reduce uncertainty and vagueness

Information fusion (IF) has been used for reducing indeterminacy and improving results [8, 14].

Different approaches can be used to fuse data from the queried agents. Their description is omitted for reasons of space. This paper does not examine the properties of any particular fusion technique; algorithms refer only to "fuse" information, that is, fusion is taken as a generic function. This is one of the reasons why we say that we are proposing a system model or architecture rather than an actual system. As we are not interested in the technique as such, we opted for the one that was simplest and easiest to implement.

3. MODEL DESCRIPTION

In this paper, when we refer to "agents" or "peers" we are assuming that they are neither all of the same type nor all humans or machines.

A detailed model description can be read in [12].

As suggested, the model assumes that agents do not have a special-purpose organization but can communicate unrestrictedly with each other. On this ground, they are regarded as being part of a flat P2P, whose individuals have limited resources. Typical examples of this are:

a) a community whose members cooperate to find the best response to a query received from another system (such as hotel or tourist attraction recommendation system) where users (network peers) can use their smartphones, which have a limited battery life and possibly number of messages), to enter their real-time opinions in response to a query by someone about the best option;
b) a scenario where a group of people, which new members may join at any time, are helping to manage a disaster and communicate with each other using devices with limited resources (battery life, transmission length, etc.);
c) a scenario where a group is taking part in military manoeuvres in hostile territory, where the number of messages is, for example, limited on security grounds, and none of the members can be considered indispensable for the mission.

3.1 IQ metric used

In this model, the IQ metric can be applied for both numerical and symbolic data.

The IQ is calculated as a scalar using ad hoc formulas that are easily computed by each peer and include other dimensions apart from vagueness and uncertainty, thereby preventing a major computational overload for the peer. The above formulas are not specified here for reasons of space, but can be found in [12].

The dimensions considered for calculating the IQ (for one agent α querying another agent) are, among others: a) the relative importance of each of the constituent parts of an information item for a querying agent, b) the importance attached by the agent to the information being first, second or third hand, etc., that is, whether or not it is supplied directly by the queried agent c) the variation in the quality of the responses given by the queried agents c) the number of queried agents, which is associated with the credibility and reliability of the resulting value d) the number of agents that responded to the query, together with the number of queried agents e) the vagueness and uncertainty with which an agent α responds and the uncertainty of its response. These two values can be determined by the queried agent (through introspection), negotiated with α or assigned directly by α. We implemented introspection.

By using a single number for quality, we avoid multicriteria decision making.

3.2 Dynamics of a system with this architecture

In general terms, the system operates by receiving messages from an external agent (Ω). The external agent asks the system to estimate certain message fields (or even the full message), selected at random by the querying agent. The system returns this estimate or estimates within a preset time. If this time limit is not met, there is a time-out. Generally, neither the system nor any of its agents respond if they are of lower quality than the querying system or agent (although the probability of this happening is very low). This behavior is inspired by simulated annealing.

The general dynamics can be read in [12].

In the early stages, all the agents receive the queries. With the passage of time, that is, once it has been determined which agents provide Ω with better quality information for the specified fields (or full message), only a few (at most N — a system parameter — for each field or for the full message) receive the queries.

As their resources are depleted, the agents stop being able to respond. On this ground, new agents must be selected to respond to a query. Recursively, every agent in the system can query or be queried about the estimate of a field.

At the start, none of the agents have any preferences for any agents with respect to a field. All agents receive a query, as Ω does not yet have any agent preferences. This query is processed and generates others among the system agents. When these queries are processed, some agents may not be able to respond (it is as if they left the system, whereas others may join). The response is returned to Ω. With the subsequent queries, Ω and the other agents will determine which their favourite agents are. Likewise, as the queries are passed on, the resources of some of the agents will be depleted. As a result, Ω will determine the agents that are its new favourites. The new favourites of Ω and of the other agents will now respond to the queries. This process is repeated indefinitely as long as the system is operational; it is depicted by the state diagram of figure 1, for a given field (or part of the query). Note that agents can enter or leave the system at any time.

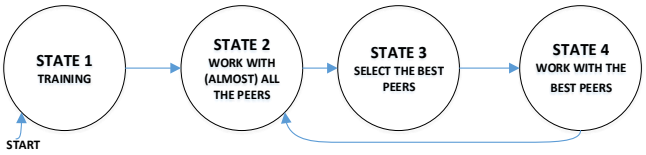

Figure 1. State diagram of the elements of the system.

The difficulty or cost of communications is not taken into account: they are all equally valuable, and there is no notion of topological neighborhood, that is, there is no such thing as a distant or close agent.

The proposed model can help to handle volume, variety and reliability by enabling distributed data processing, where each agent can specialize in a subset of data, especially as regards volume, variety and reliability.

4. PRELIMINARY QUANTITATIVE RESULTS

We instantiated and applied the architecture to three case study. The data were based on the data repository at the University of California in Irvine [16]. The selected case study were A) the classification of a day as of high ozone level B) the classification of a site as of phising C) the classification of whether or not to buy a car depending on characteristics such as number of doors, safety, price, etc. The network was simulated as an agent system, programmed in Java. A statistically significant number of runs were conducted. The behavior of the system was tested for five to 1000 agents.

The hypothesis "H_0: the system using the model described here performs equal or worse than a system which does not consider information quality to optimize communications" was rejected for different case studies as shown in Table 1.

It should be noted that it is not always possible to get a 100% of rejections of H_0, as in the case of obtention frequency of a high quality: when one model outperforms (gets a higher frequency) the other for a given quality, it will be outperformed for another values, because the sum of the frequencies is 1. Note also that the three cases used the same values of the model parameters, so it could be that with an adequate set of values for every case better results were obtained.

Table 1. Rejection of H_0 for different case studies

Case	Rejection	Number of set of tests
A - Ozone	80%	3
B - Websites	87.5%	12
C - Car	87.5%	12

5. CONCLUSIONS AND FUTURE WORK

We sketched a model of architecture for a system based on flat P2P networks whose elements are organized taking into account the circulating information quality. It is clear that the objectives of the proposed architecture are aligned with sense making, as the quality metric used penalizes both vagueness and uncertainty.

The described model appears to yield positive results with respect to several issues. Preliminary experimental studies have been conducted to assess system performance, and it was found to outperform a flat P2P architecture on several points, as described in Section 4. In the future, we propose to run tests on other case studies.

Future research can be divided as follows:

- Modify the IQ evaluation formula so it reflects the learning of the agents (this is, the formula becomes dynamic through time).

- In the model, an agent never knows whether its response was the best or whether it timed out. Aggregating feedback from the sender (the querying agent) on this point would increase the messages to be sent. However, it would trigger intelligent behavior on the part of the receiver (queried agent) earlier (an agent seeing that low quality results are being generated would decide to switch to intelligent mode), thus reducing the number of messages (as the receiver is prompted to query other agents to get a better quality response). Thus, a design issue worth analysing is whether, in view of the resultant additional learning by agents, such feedback should be enabled.

- We assumed that the agent has no doubt about how to decompose (or, like-wise, identify the sources of) a given field. One scenario worth investigating is when the agent has to pick one of several feasible decompositions. More generally, we assumed that all the agents share the message and field decomposition criterion (lazy or otherwise) (for example, by using one and the same ontology). We could study which changes should be made to the model if this does not occur (for example, fields that are not known by the queried agent).

6. ACKNOWLEDGMENTS

This work has been partially funded by the grant S-C-BE 55/18, Préstamo BID OCUR /1296-PDT, Ministerio de Educación y Cultura, R.O del Uruguay.

7. REFERENCES

[1] Bossé, É. and B. Solaiman, *Information Fusion and Analytics for Big Data and IoT*. 2016: Artech House Publishers.

[2] DeLone, W.H. and E.R. McLean, *Information systems success: The quest for the dependent variable.* Information Systems Research, 1992(3): p. 60-95.

[3] G. Shankaranarayanan, M.Z. and R.Y. Wang, *Managing Date Quality in Dynamic Decision Environment: An Information Product Approach.* Journal of Database Management, 2003(14): p. 14-32.

[4] Janssen, M., H. van der Voort, and A. Wahyudi, *Factors influencing big data decision-making quality.* Journal of Business Research, 2017. **70**(C): p. 338-345.

[5] Knight, S.-a. and J. Burn, *Developing a Framework for Assessing Information Quality on the World Wide Web.* Informing Science, 2005.

[6] Meer, H.D. and C. Koppen, *Self-organization in Peer-to-Peer Systems*, R. Steinmetz and K. Wehrle, Editors. 2005, Springer.

[7] Miller, H., *The multiple dimensions of information quality.* Information Systems Management, 1996. **13**(2): p. 79-82.

[8] Nakamura, E.F., et al., *Localized algorithms for information fusion in resource constrained networks.* Inf. Fusion, 2014. **15**: p. 2-4.

[9] Novák, V., *Are Fuzzy Sets a Reasonable Tool for Modeling Vague Phenomena?* Fuzzy Sets and Systems., 2005. **156 .** p. 341-348.

[10] O'Reilly, C.A., *Variations in decision makers' use of information sources: The impact of quality and accessibility of information.* Academy of Management Journal, 1982(25): p. 756-771.

[11] Paggi, H. and M. Cochez, *Indeterminacy Reduction in Agent Communication Using a Semantic Language*. WSEAS Transactions on Systems, 2015. **14**: p. 77-89.

[12] Paggi, H., J. Soriano, and J.A. Lara, *A multi-agent system for minimizing information indeterminacy within information fusion scenarios in peer-to-peer networks with limited resources.* Information Sciences, 2018. **451-452**: p. 271-294.

[13] Price, R.S., Graeme, *The Impact of Data Quality Tags on Decision-Making Outcomes and Process.* Journal of the Association for Information System, 2011. **12**(4): p. 323-346.

[14] Raol, J.R., *Data Fusion Mathematics. Theory and Practice*. 2015: CRC Press.

[15] Wand, Y. and R.Y. Wang, *Anchoring data quality dimensions in ontological foundations.* Commun. ACM, 1996. **39**(11): p. 86-95.

[16] Zhu, X. and S. Gauch. *Incorporating Quality Metrics in Centralized/Distributed Information Retrieval on the World Wide Web*. in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.