PANDA: a framework for reasoning with scientific Uncertainty

Larisa N. Soldatova[1], Oghenejokpeme I. Orhobor[2], Joseph French[2], and Ross D. King[2]

[1] Goldsmiths, University of London, UK
[2] University of Manchester, UK

Keywords: probabilistic reasoning · ontology · uncertainty

1 Introduction

Science is generating vast amounts of heterogeneous globally distributed data: millions of scientific publications, numerous databases, knowledge bases, and models. This data is fuelling technological progress and underpinning the modern economy. Unfortunately, data is often fragmented, incomplete, uncertain or contradictory, but nevertheless useful. A means by which to integrate these vast arrays of data to determine our confidence in an individual fact would be of immense benefit. For this purpose, we propose PANDA (ProbAbilistic kNowleDge Assembly), a framework for reasoning over heterogeneous globally distributed sources of information in order to assess the probability of a factual statement being true given the available relevant information. PANDA enables the coherent integration of multiple sources of uncertainty to computationally estimate the probability of scientific facts. PANDA employs a formalisation of the sources of uncertainty, first-order predicate logic, and Bayesian probability theory to determine the probability of truth using ProbLog, PANDA's measure for its confidence in the specified statement.

We consider the concept of uncertainty from a practical data integration perspective. We demonstrate the utility of PANDA on the area of cancer research and describe a worked example of using PANDA. PANDA has been designed as a generic framework, and it is suitable for various applications in different areas.

2 PANDA Framework

The PANDA framework is a knowledge representation and reasoning architecture designed for assembling fragmented pieces of information extracted from text into large, probabilistic knowledge models. PANDA can be broadly split into three main sequential processes: parsing index cards from the National Center for Text Mining (NaCTeM), calculating the uncertainties of the statements in the index cards, and generating a unified knowledge model.

The index cards are json files which attempt to describe the interactions between elements in a statement extracted from the literature. PANDA can extract

from them the following information: the event type (e.g. gene expression), confidence (a numeric value indicating the probability a statement has been correctly extracted by text mining, or extraction probability $P(E)$) and uncertainty (the probability a statement is true given how it is linguistically expressed, or textual probability $P(T)$). Each index card works from a single extracted statement from the source paper. This statement is referred to as the "evidence" in the index card. Each interaction involves two participants: a chemical and either a protein or a gene.

We evaluated a total of 2,176 index cards. However, we constrained our uncertainty evaluation to the 530 of them which described interactions between a chemical compound and protein/gene. For each index card evaluated we retrieved additional information from the NCBI using their PMC identifiers. We were particularly interested in the journal or conference from which the corresponding paper originated as this enables one to modify their confidence in an extracted statement. We estimated the values for every journal using their average journal impact factor from the science citation index which is downloadable from the Web of Science, as estimates for provenance probability $P(J)$.

We calculate the probability of a statement being true, $P(X)$ as follows:

$$P(X) = P(J) \times P(E) \times P(T) \tag{1}$$

In the PANDA framework, we update our level of uncertainty that a statement in an index card is true given new information using a Bayesian approach. This new information takes the form of sources external to the index card, in this case laboratory-based experiments or a third-party database. The overall aim is to calculate $P(X)$ given the full set of available additional information pertaining to the truthfulness of a statement in an index card: $P(X|Ex)$, where Ex is a set containing the evidence being considered. Conceptually, $P(X)$ should increase as additional information supports the statement, while results refuting the statement should cause it to decrease.

The PANDA system was written in Python [2], with the probabilistic reasoning modelled and integrated using Problog [1]. We used the Chemical Entities of Biological Interest (ChEBI), Universal Protein Resource (UniProt) and HUGO Gene Nomenclature Committee [3] (HGNC) databases to map entity identifiers assigned by text mining to ontological names for both grounding and verification.

3 Example

An index card was created from the paper PMC2249593. In this example, we consider single event: the upregulation of gene P53 (UniProt:P04637) by Curcumin (CHEBI:3962) with the following estimates of uncertainty:

- The extraction probability is 0.66.
- The textual probability $P(T)$ is 0.8.
- The provenance probability $P(J)$ is 0.71, using the scaled impact factor of the Journal of Molecular Cancer.

Therefore, the probability before experimental validation, $P(X)$, is 0.356 using equation 1.

We tested this statement in our lab at the University of Manchester. This task was complicated by the fact that text mining tools are not sophisticated enough to extract contextual information from the literature, i.e. cell line and other experimental conditions required for validation, assuming such information even reported in the source paper to begin with. In our lab we constrained our experimental validation to two commonly used cell lines: MCF7 and MDA-MB-231. We found the opposite to be true for 5 out of 7 tests in reference to MCF7 and 8 out of 9 in reference to MDA-MB-231. Considering this, PANDA updates the new probability of the statement being true from 0.356 to 0.195 in MCF7 and 0.085 in MDA-MB-231, giving new uncertainty values of 0.805 and 0.915 respectively as PANDA decreases the level of confidence given our experimental evaluation.

4 Conclusions

PANDA is a prototype with a clear path for expansion for collation a variety of heterogeneous data, including a defined means of incorporating the varying level of confidence in each source of data. We envision that future expansion of this framework would combine technological tools such as machine learning to build on the currently demonstration calculation of initial probabilities. The work on PANDA highlights the necessity to push the boundaries of existing artificial intelligence technology and develop efficient solutions for consuming heterogeneous pieces of information and assembling them into reach, actionable knowledge models, based on robust formal foundations facilitating scalable data integration and use of automated reasoning techniques for explanation, prediction, and discovery.

5 Acknowledgments

This work was supported by the Big Mechanism project funded by DARPA.

References

1. De Raedt, L., Kimmig, A., Toivonen, H.: Problog: A probabilistic prolog and its application in link discovery. In: IJCAI. vol. 7, pp. 2462–2467. Hyderabad (2007)
2. Sanner, M.F., et al.: Python: a programming language for software integration and development. J Mol Graph Model **17**(1), 57–61 (1999)
3. Yates, B., Braschi, B., Gray, K.A., Seal, R.L., Tweedie, S., Bruford, E.A.: Genenames. org: the hgnc and vgnc resources in 2017. Nucleic acids research p. gkw1033 (2016)