Qoan Manifesto

Qoan Wissenschaft & Software

1 What is Qoan?

Qoan in its most basic is an open-source information management system which is designed to deal with large data exploiting distributed resource storage and processing strategies. It is our belief, distributed computing is the only feasible way for dealing with large data.

In order to expand on this, the relationships between individual bits of data will also be taken into account, in order to represent it in a mathematical graph model. This will allow for organizing data, as well as ranking them in their degrees of importance. In Qoan graph algorithms will find other usages as well.

Furthermore, vast amounts of data make very good source as training data for artifical neural networks. Neural networks are widely used in different fields of artificial intelligence and have also common applications. With the help of ranking, very specialized neural networks can be trained to improve their performance in problem specific applications.

It must also be stressed, that usage of specialized networks and graphs only, makes the problem of computational feasibility solvable. The whole network is out of reach, parts of it is doable.

Qoan is therefore a framework, where you can make use of data available in Wikipedia for training neural networks, or exploiting in the same manner biochemical data in online gene databases to create biochemical reaction networks, or making use of online stock quotes to create predictive models of the current market.

In order to demonstrate how this can be useful, we have designed four *homework assignments* where these possibilities will be explored. These are picked out of publications and applications from the fields of finance, network and graph theories, artificial neural networks, machine translation, computational linguistics and biochemistry. With each homework assignment, we will expand on

those works publishing our results, while implementing these as new features in Qoan. With each completed homework assignment, Qoan will have additional features which we hope will be useful for routine tasks as well.

In realization of the software Java language was chosen, as it is very commonly used in all of the worlds we want to bring together, making it also a language of choice for the existing excellent open-source projects, which we make abundant use of. The web-interface which is the Qoan project, has been implemented with Vaadin, which allows for a modern web-application and all the visual goodies we are used to and fond of. This is a simple web-application which commands and visualizes the computing grid and the data stored there.

On the backend, Qai project, is designed to make available all different types of networks, their mathematical representations and data processing, in a distributed computing environment. The distributed data storage and computing capabilities are thanks to Hazelcast platform. Graph algorithms are carried out by GRPH and as an implementation of artificial neural network algorithms ENCOG was picked. These are excellent libraries which offer many more uses of both graph theory and neural-networks, with which we want to encourage developers who want to develop their own applications easily using those.

In order to index Wikipedia data and other resources Apache Lucene is employed, as well as Tika, Apache OpenNLP, BlikiWiki and, last but not least, Apache-Jena for usage of semantic-nets and SPARQL queries.

Qoan stores data distributed on different nodes as files, as well as in relational databases, like MySQL and/or in SPARQL capable datastores, like OpenLink Virtuoso, whichever is convenient. The current configuration makes use of all of those, but this is merely for testing purposes. Depending on needs of the system the whole data can be stored with just one or in any combination of those.

The demo-system, Qoan.org, currently is only meant to give you an idea how the user-interface looks, without any of the homework assignments actually functioning. With the support we raise, we will be able to work out our homework assignments and present our results to you, as well as a functional open-source framework for distributed artificial intelligence.

Qoan.org is currently running on a collection of computers consisting only of Raspberry PI boxes, in order to demonstrate that this can also be done, a proof of concept grid. With this, we want to make grid computing available to a wider spectrum of users and to make usage of artificial intelligence more widespread.

We are painfully aware, many wonderful applications of these technologies are not being adressed with the given set of features we have chosen. But we hope to hear from you about your ideas and wishes, where the future development should head to and which features be added.

2 Homework Assignments

2.1 How fair are the markets?

With this homework assignment we wish to discuss a model of stock markets and analyse the resulting relationships which give insight about the market. In their work Borysov et.al[7] have

shown, with a certain neural-network architecture, stock data can be used for training an interaction network of the stock markets which can be interpreted as a quantifier of the relationships between stocks.

At this point, we turn to the work of Venkatasubramanian et.al[1] and take their interpretation of entropy from game theory and micro-market simulations point of view.

Using these tools together, we are expecting to show that market data can be divided into different episodes where these differ significantly in both their entropy values and how the relationships between companies are. As well as the toolkit in Qoan for doing it yourself too, obviously.

2.2 Machine-translation as shortest-path problem

Google's automatic translator is an excellent example of a neural-network which has been trained with data out of two different languages to convert one sentence into the other language. When a sentence is represented in its semantic-network form, we end up with a path in a graph. Wiktionary entries have word translations into most other languages, therefore switching the nodes in the path of the origin sentence to represent words in target language, and finding the shortest-path in the target semantic-network which contain these words, will sum up to translating the text.

Since Google's neural networks doing the translations are designed to be as general as possible to address unspecified needs, the results can sometimes get disappointing. We want to demonstrate that this can of course be compensated if specialized networks on specific topics could be trained and applied in a particular field.

Upon completion of this homework assignment, you will be able to translate text using expert networks which are trained with the specific needs of your field. Our work will be based on works of Bahdanau et.al.[3] and Sutskver et.al.[4].

2.3 You and me and everyone we know

It has already been pointed that Wikipedia data can be used for ranking purposes as well[5, 6]. Therefore if you were suddenly taken with the question *"Just how important is Kilgore Trout really?"*, this example is for you. Wikipedia, not only has information about him, but also many references in different articles, which makes it possible to make a ranking of all persons, real or fictionary, with help of a simple query. This will be possible, thanks to ontology developed by DBpedia project.

But this is not the only data this homework assignment will be arranging in form of graphs, also biochemical and genetical databases will be included so that molecules and genes can be ranked as well. These graphs will be particularly useful later in fourth homework assignment, which is also the reason why we chose this assignment.

And who knows we might indeed be surprised to find out just how important Kilgore Trout actually is.

2.4 Towards the virtual-cell

When it comes to large networks, with twenty-thousand or so genes and biochemicals in humans, chemical reaction networks, as they are found in living organisms, are certainly among the most complex. Researchers of many different fields are constantly looking just for one reaction, which could be the cure for cancer, or be the next large step in malaria treatment. Predictive searches using computer simulations have been developed constantly in order to aid this search undertaking.

With this homework assignment, we want to address this problem. Employing machine learning approach we will be developing on the work of Brockherde et.al[2].

Upon completion of this homework assignment, with the help of Qoan, neural networks will be trained with online data in order to run molecular dynamics simulations of biochemical reactions, which can predict other reactions. These neural networks can then be used for searching reactions and running their simulations.

The wonderful visualizations which you will find on the demo system, are thanks to NGLViewer project.

References

[1] *"How much inequality in income is fair? A microeconomic game theoretic perspective"*, Venkatasubramanian V., Luo Y., Sethuraman J., Physica A-435 (2015) 120-138.

[2] *"Bypassing the Kohn-Sham equations with machine learning"*, Brockherde F., Vogt L., Li L., Tuckerman M.E., Burke K., Müller K., Nature Communications DOI: 10.1038/s41467-017-0089-3

[3] *"Neural Machine Translation by jointly learning to align and translate"*, Bahdanau D., Cho KH., Bengio Y., arXiv:1409.0473v7 [cs.CL] 19 May 2016

[4] *"Generating Text with Recurrent Neural Networks"*, Sutskever I., Martens J., Hinton G., Proceedings of the 28 th International Conference on Machine Learning, Bellevue, WA, USA, 2011.

[5] *"Wikipedia Ranking of World Universities"*, Lages J., Patt A., Shepelyansky D.L., arXiv:1511.09021v2 [cs-SI] 4 Feb 2016

[6] *"Time evolution of Wikipedia network ranking"*, Eom YH., Frahm K.M., Benczúr A., Shepelyansky D.L., arXiv:1304.6601v2 [physics.soc-ph] 31 Oct 2013

[7] *"US stock market interaction network as learned by the Boltzmann Machine"*, Borysov S.S., Roudi Y., Balatsky A.V., arXiv:1504.02280v1 [q-fin.ST] 9 Apr 2015