Fortschritt-Berichte VDI

VDI

Reihe 10

Informatik/
Kommunikation

Nr. 862

Mario M. Kubek,
Zhong Li (Eds.)

Autonomous Systems 2018

Proceedings of the 11th Conference

FernUniversität in Hagen
Schriften zur Informations- und Kommunikationstechnik

Fortschritt-Berichte VDI

Reihe 10

Informatik/
Kommunikation

Mario M. Kubek,
Zhong Li (Eds.)

Nr. 862

Autonomous Systems 2018

Proceedings
of the 11th Conference

FernUniversität in Hagen
Schriften zur Informations-
und Kommunikationstechnik

Kubek/Li (Eds.)
Autonomous Systems 2018 – Proceedings of the 11th Conference
Fortschr.-Ber. VDI Reihe 10 Nr. 862. Düsseldorf: VDI Verlag 2018.
176 Seiten, 60 Bilder, 12 Tabellen.
ISBN 978-3-18-386210-8, ISSN 0178-9627,
€ 62,00/VDI-Mitgliederpreis € 55,80.

Keywords: Autonomous Systems – Data Interferences – Data Assessments – Decentralised Search – Machine Learning – Deep Learning – Simulation – Energy Systems – Safety – Security

To meet the expectations raised by the terms Industry 4.0, Industrial Internet and Internet of Things, real innovations are necessary, which can be brought about by information processing systems working autonomously. Owing to their growing complexity and their embedding in ever-changing environments, their design becomes increasingly critical. Thus, the many topics addressed in this book range from data integration on hardware level to methods for security and safety of data and to stochastic methods, data interferences as well as machine learning and search in decentralised systems. Their validity is proven by extensive simulation results. Also, applications for methods from deep learning and neurocomputing are presented. The sustainable management of energy systems using intelligent methods of self-organisation and learning is dealt with in the second major part of this book. As in these particular settings, the assessment of network vulnerabilities plays a crucial role, respective methods are discussed as well. Finally, the establishment of trustbased machine-to-machine communication is elaborated on in detail.

Bibliographische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet unter www.dnb.de abrufbar.

Bibliographic information published by the Deutsche Bibliothek
(German National Library)
The Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliographie (German National Bibliography); detailed bibliographic data is available via Internet at www.dnb.de.

Schriften zur Informations- und Kommunikationstechnik
Herausgeber:
Wolfgang A. Halang, ehemaliger Lehrstuhl für Informationstechnik
Herwig Unger, Lehrstuhl für Kommunikationstechnik
FernUniversität in Hagen

Als Manuskript gedruckt. Printed in Germany.

ISSN 0178-9627

ISBN 978-3-18-386210-8

Preface

Choose a job you love, and you will never have to work a day in your life.

Unnamed source

'Publish or Perish' is probably the most famous slogan of deans, university presidents and administration staff forcing their researchers to increase their numbers of written publications in highly ranked international conferences and journals. Rather sick brains created over the years more and more complex evaluation and control measures, neglecting that the own reputation of a researcher is the best quality assurance. Last but not least, an overwhelming number of grants and other third-party research money applications drastically reduce the time for state-of-the-art science once more. In fact, all those measures have a negative impact on the scientists' inspiration and creativity, do not contribute to a more successful scientific work and just hide the permanent underfinancing of today's universities.

The editors of the recent proceedings of the 11th Conference on Autonomous Systems, held again in the scenic environment of the small village of Cala Millor, wonder in this year for the first time, why the number of (optional) written contributions has reduced while the numbers of participants is almost constant. We found the answer to that question in an increasingly powerful wish of most scientists to slow down, go away from bureaucratic, unneeded and incomprehensible paperwork and senseless meetings. We think that it is the strong need to return to the roots of science, have time for a dispute or an interesting, but maybe not immediately goal-directed chat, a competition of ideas and last but not least to find an inspiring place away from more and more irrational world politics.

A consideration of the contents of the papers over the years reflects the changing interests of our participants triggered by a rapidly changing world around us. For the first time, huge amounts of data have been made available to various system levels by different sensors and applications all over the world in the last decade. Its exploitation became a huge business field and its influence is reflected within science in general as well as in the contributions of the first big part of our proceedings. Topics range from data integration on hardware level to methods for security and safety of data and to stochastic methods, data interferences as well as machine learning. The second part of contributions is closer connected with the hardware and system level: data assessments, simulation results and new algorithms including intelligent methods of self-organisation and learning increasingly influence the design of systems and are presented by the authors.

We are thankful to our two invited speakers, Mrs. C. Yuan and Mr. P. Meesad, for accepting the difficult task to give us an overview of and some perspectives on those complex, often interdisciplinary developments. Furthermore, we wish to thank the members of the steering committee Mr. W. A. Halang and Mr. H. Unger for their constant support and trust extended to us. Last but not least, we have to thank again Mrs. Düring and Mrs. Kleine for their enduring support in all organising tasks of our event. In addition, we appreciate the support of FernUniversität in Hagen given to publish this volume.

Finally, we hope that we meet with our event the needs of the participants, can all together spend some quiet, inspiring and peaceful days in the Sabina Playa Hotel on Majorca Island with a lot of deep and fruitful discussions. We hope that all participants feel well in the special atmosphere of our event and will join us again in the next years.

Hagen, August 2018

Mario M. Kubek
Zhong Li

Contents

Hardware, Energy and Systems

Keynote 1: Deep Learning and Applications

Phayung Meesad

Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand

Abstract:
Deep Learning has been extremely successful in many fields such as image processing and natural language processing. Convolutional Neural Network (CNN) and Long Short Term Memory Recurrent Neural Network (LSTM-RNN) are probably most search hit in Deep Learning fields. CNNs are popular in image processing while LSTMs seem to play big roles in Time series data and natural language processing. This talk gives brief reviews about Deep Learning focusing on CNNs and LSTMs as well as their applications.

Keynote 2: Mobile Autonomous Systems: Sensing, Reasoning and Acting

Chunrong Yuan

Department of Information, Media and Electrical Engineering
Technische Hochschule, Cologne, Germany

Abstract:
The three important capabilities of mobile autonomous systems are sensing, reasoning and acting. Since the development of the first mobile robot Shakey around 1970, research in mobile and autonomous systems has been making steady progress in all three aspects. However, it is still a challenging task to design machines which are capable of safe and autonomous operations in the dynamic real world, particularly in situations where humans and other biological or artificial systems are involved or interactions among them are necessary. In this talk, I am going to present a few research questions relevant to mobile autonomy and provide our solutions which have been developed for a set of different application scenarios. With an analysis of the strength and limitations of these solutions, I will finish with a summary of our future focus and research strategy.

fastAN(BD) – a Fast Method for Integrity Checking and Decoding of AN(BD)-coded Data

Stefan Widmann

Abstract: To improve error detection capabilities of commercial off-the-shelf microprocessors, arithmetic coding in form of AN(BD) coding can be used. The standard scheme for integrity checking and decoding of coded data utilizes divisions, which are very time consuming operations and smaller microprocessors don't even provide native division instructions. Instead of using divisions, a new method for integrity checking and decoding called fastAN(BD) is presented, using multiplications only. Its exploitation of residual class ring arithmetic results in a runtime benefit of up to 89 % and a lower residual error probability compared to the standard integrity checking scheme.

1 Introduction

Safety-related applications that have been realized using mechanical or electromechanical systems before are being more and more realized by programmable systems today. Good examples are automotive and avionic applications like braking, steering and flying: brake-by-wire, steer-by-wire and fly-by-wire replace the proven and reliable mechanical systems by electrical and electronic sensors and actors, the sensors' signals being processed by microprocessors to calculate control signals for the actors [1, 8, 12].

The complexity of the hardware used in such applications is rising: the number transistors per die has reached 1.3 billion in 2015 [11]. The integration of growing numbers of components in integrated circuits is only possible due to a continuous reduction of the minimum feature size, making them more sensitive to environmental influences like radiation, especially neutrons, even on ground level [3, 9].

The complexity and the sensitivity results in a rising probability of errors. Commercial off-the-shelf (COTS) microprocessors commonly used in the systems described above are not designed to provide effective means for error detection. To improve the probability of detecting HW errors, arithmetic coding in form of AN(BD) coding can be used. But AN(BD) coding uses runtime intensive divisions and the residue error probability depends on the choice of the magic constant A.

The new fastAN(BD) method replaces the divisions by multiplications, resulting in faster integrity checking and decoding compared to the standard schemes, while providing a lower residue error probability and a consistent sequence of decoding and integrity checking at the same time. This makes the fastAN(BD) method ideal for usage on smaller processors used in applications like the Internet-of-Things (IoT).

2 Errors in Data Processing

Schiffel [10] extended Forin's error model [5] to cover the following error types:

- erroneous operations, where the result of an operation isn't correct,
- corruption of data values,
- processing of wrong operands,
- lost updates resulting in outdated operands and
- application of wrong operations to the input operands.

These errors must be detected before they can lead to dangerous outputs that could endanger human lives, the environment or investment goods.

3 State of the art

The relevant state of the art shall be illustrated by AN coding [4] and its extension to ANBD coding [5]. The basic encoding, decoding and integrity checking functionality is explained for both.

3.1 AN Coding

The arithmetic coding scheme AN coding was introduced by Brown in [4] and is capable of detecting corruption of data in memory or registers, as well as erroneous operations, where the ALU of a processor does not produce a correct result. Encoding is done by multiplying the magic constant A to a data value x, resulting in the coded $x_{c_{AN}}$:

$$x_{c_{AN}} = A \cdot x$$

One big advantage of AN coding is the fact that normal arithmetic operations can be applied to coded data and the integrity of the data processing operations and the coded data words itself can be easily verified by using a modulo operation, where the equation

$$x_{c_{AN}} \equiv 0 \mod A \quad \text{or} \quad x_{c_{AN}} \bmod A = 0$$

must be satisfied.

The data value can be decoded by dividing the coded value x_c by the magic constant A:

$$x = \frac{x_{c_{AN}}}{A}$$

For a long time primes have been recommended as a good choice for the magic constant A, e. g. by Forin in [5]. But Ulbrich empirically demonstrated in [13], that there are several non-prime 16 bit wide As, that have a minimum hamming distance d_h of 6: 58659, 59665, 63157, 63859 and 63877. Using these "Super-As" – as Ulbrich called them – allows detection of 5-bit-errors. In this paper, the Super-A 58659 will be used in all examples.

3.2 ANBD Coding

In order to be able to detect additional types of errors, the AN coding scheme was extended to ANBD coding by Forin in [5]. Additionally to the multiplication by A, an identifier B and a timestamp D are added during the encoding of a data value. During decoding, both B and D have to be subtracted from the coded value prior to dividing by A.

$$x_{c_{ANBD}} = A \cdot x + B + D \qquad\qquad x = \frac{x_{c_{ANBD}} - (B + D)}{A}$$

The integrity of an ANBD-coded value $x_{c_{ANBD}}$ can be checked using given expected B' and D' by verifying, that the equation

$$x_{c_{ANBD}} = A \cdot x + B + D \equiv B' + D' \mod A$$

or

$$x_{c_{ANBD}} = A \cdot x + B + D - (B' + D') \equiv 0 \mod A$$

is satisfied.

In addition to erroneous operations and data corruption which can be detected using AN-coded data values, ANBD coding can unveil the usage of wrong or outdated operands and the application of wrong operations during data processing.

4 Drawbacks of the State of the art

AN(BD) coding is using divisions for integrity checking and decoding operations, which are very runtime consuming operations on most processor architectures, compared to multiplications. A number of low-cost processors that are used in embedded applications do not even have native instruction set support for divisions, some do not even support multiplications. In such cases, these instructions have to be emulated by SW, which is a very cycle-intensive task. A comparison of clock cycles needed to execute multiplication and division instructions is shown in table 1 for various processor architectures.

Table 1: Comparison of number of clock cycles on different architectures

Processor / -architecture	**Multiplication**	**Division**
Intel x86 (Sandybridge) [6]	4 cycles (mul r)	26 cycles (div r)
Intel x86 (Haswell) [6]	4 cycles (mul r)	28 cycles (div r)
Intel x86 (Broadwell) [6]	4 cycles (mul r)	31 cycles (div r)
ARM Cortex-M3 [2]	1 cycle	2 - 12 cycles[1]
dsPIC33 [7]	1 cycle	18 cycles
MSP430 (no HW-mult.)	≈ 25 - 150 cycles[1,2,3]	≈ 430 - 460 cycles[1,2,3]
MSP430 (HW-mult.)	14 cycles[2,4,5]	≈ 430 - 460 cycles[1,2,3]

[1]: number of clock cycles depends on operands; [2]: own evaluation based on IAR EW430 compiler; [3]: no HW support, SW emulation necessary; [4]: integrated HW-multiplication peripheral; [5]: HW multiplication takes 1 cycle, but register loading and storing prior and posterior consume additional cycles

The residual error probability of the AN(BD) coding depends on the choice of the magic constant A. If an equally distributed error probability can be assumed, the residual error probability $p_{AN(BD)}$ is:

$$p_{AN(BD)} = \frac{1}{A}$$

Additionally, integrity checking and decoding is done in two separate calculation steps, which can both be affected by errors, so one can never be sure that e. g. an transient error didn't corrupt the decoding operation, while the integrity checking operation verified the data's integrity.

5 The fastAN(BD) Method

The fastAN(BD) method exploits the residue class ring arithmetic of processor ALUs in order to replace the divisions needed for integrity checking and decoding by multiplications. An important parameter the fastAN(BD) method needs is the multiplicative inverse of the magic constant A called A^{-1}. There's no A^{-1} for every A in a residue class ring, so the choice of A for use with the fastAN(BD) method depends on the availability of a corresponding A^{-1} which can be found e. g. using the extended euclidian algorithm.

The parameters x, A and A^{-1} are elements of the residue class ring $\mathbb{Z}/2^n\mathbb{Z}$.

$$x, A, A^{-1} \in \mathbb{Z}/2^n\mathbb{Z}$$

The coded value $x_{c_{AN(BD)}}$ has double the bit width and is thus an element of the residue class ring $\mathbb{Z}/2^{2n}\mathbb{Z}$.

$$x_{c_{AN(BD)}} \in \mathbb{Z}/2^{2n}\mathbb{Z}$$

5.1 fastAN

The fastAN method is a fast algorithm for integrity checking and decoding of AN-coded data values. It is not suitable for ANBD-coded data values, for which the fastANBD method must be used that will be explained in the next chapter. The key of the fastAN method is the fact that the equation

$$x_{cAN} \cdot A^{-1} = A \cdot x \cdot A^{-1} \bmod 2^n = x \quad \forall x \in \mathbb{Z} \mid 0 \leq x < 2^n$$

applies in a residue class ring. That means that multiplying the multiplicative inverse A^{-1} to the coded data value $x_c = A \cdot x$ calculates the uncoded data value x again.

The fastAN method integrity checking flow chart is shown together with the encoding scheme of AN coding in figure 1. The function low() returns the lower half of the data contents it is applied to.

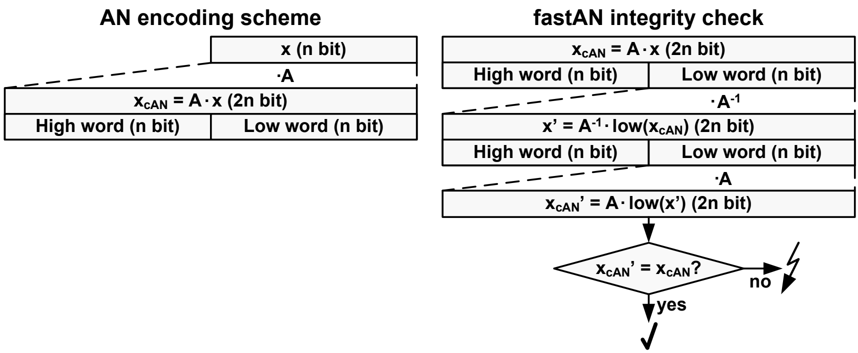

Fig. 1: Flow chart of AN encoding and new fastAN scheme

Expressed in a mathematic equation, the fastAN method looks like shown below. Solving the equation shows that it is satisfied for all $x \in \{0, ..., 2^n - 1\}$.

$$\begin{aligned}
& \left(\left((x_{c_{AN}} \bmod 2^n) \cdot A^{-1}\right) \bmod 2^n\right) \cdot A = x_{c_{AN}} \\
\Leftrightarrow \quad & \left(\left(((A \cdot x) \bmod 2^n) \cdot A^{-1}\right) \bmod 2^n\right) \cdot A = A \cdot x \\
\Leftrightarrow \quad & \left(\left(A \cdot x \cdot A^{-1}\right) \bmod 2^n\right) = x \\
\Leftrightarrow \quad & x \bmod 2^n = x \\
\Leftrightarrow \quad & x \in \{0, ..., 2^n - 1\}
\end{aligned}$$

The function low() that was shown in figure 1 has been replaced by the mathematical correct term mod2^n. This might look like the single modulo operation of the standard integrity checking scheme has now been replaced by two ones, resulting in additional runtime effort. But selecting specific bytes, words, double words etc. is a very simple task done by addressing the affected memory address or addresses in memory directly, not involving any arithmetic operation.

Although the equation above has already shown which values will satisfy the integrity checking equation, an error e can be explicitly added to a coded value x_c, transforming it to an erroneous coded value x'_c.

$$\begin{aligned}
& \left(\left((x'_c \bmod 2^n) \cdot A^{-1}\right) \bmod 2^n\right) \cdot A = x'_c \\
\Leftrightarrow\quad & \left(\left(((A \cdot x + e) \bmod 2^n) \cdot A^{-1}\right) \bmod 2^n\right) \cdot A = A \cdot x + e \\
\Leftrightarrow\quad & \left(\left((A \cdot x \bmod 2^n + e \bmod 2^n) \cdot A^{-1}\right) \bmod 2^n\right) \cdot A = A \cdot x + e \\
\Leftrightarrow\quad & \left(\left(A \cdot x \cdot A^{-1} \bmod 2^n + e \cdot A^{-1} \bmod 2^n\right) \bmod 2^n\right) \cdot A = A \cdot x + e \\
\Leftrightarrow\quad & \left(\left(x + e \cdot A^{-1} \bmod 2^n\right) \bmod 2^n\right) \cdot A = A \cdot x + e \\
\Leftrightarrow\quad & (e \cdot A^{-1}) \bmod 2^n \cdot A = e \\
\Leftrightarrow\quad & (e \cdot A^{-1}) \bmod 2^n = \frac{e}{A}
\end{aligned}$$

The substitution $e = e' \cdot A$ gives the possibility to solve the equation:

$$\begin{aligned}
\Leftrightarrow\quad & (e' \cdot A \cdot A^{-1}) \bmod 2^n = \frac{e' \cdot A}{A} \\
\Leftrightarrow\quad & e' \bmod 2^n = e' \\
\Leftrightarrow\quad & e' \in \{0, ..., 2^n - 1\}
\end{aligned}$$

If $e = 0$, no error has corrupted the coded data value, in all other cases, where e is a multiple of A ($e = e' \cdot A$), the error cannot be detected by the fastAN integrity checking scheme, but cannot be detected by the standard scheme neither.

The first example shown in listing 2.1 illustrates the new fastAN integrity checking scheme.

Listing 2.1: fastAN pseudo code example without error

```
Magic constant: A = 58659 = 0xE523
Inverse of A:   A^-1 = 29323 = 0x728B
Data value:     x = 0x0333

x_c = A · x = 0xE523 · 0x0333 = 0x02DD0EF9

### Integrity checking ###
low(x_c) · A^-1 = 0x0EF9 · 0x728B = 0x06B30333
x* = low(low(x_c) · A^-1) = 0x0333
x*_c = x* · A = 0x0333 · 0xE523 = 0x02DD0EF9
x*_c = x_c ⇒ x_c is error-free ⇒ x* = x.
```

The second example in listing 2.2 shows the detection of the multi-bit error f_1, transforming the error-free x_c into the erroneous value x_c'.

Listing 2.2: fastAN pseudo code example with error

```
Magic constant: A = 58659 = 0xE523
Inverse of A:   A^-1 = 29323 = 0x728B
Data value:     x = 0x0333

x_c = A · x = 0xE523 · 0x0333 = 0x02DD0EF9

Error syndrome: f_1 = 0x01007100

x_c' = x_c XOR f_1 = x_c XOR 0x01007100 = 0x03DD7FF9

### Integrity checking ###
low(x_c') · A^-1 = 0x7FF9 · 0x728B = 0x39425E33
x*' = low(low(x_c) · A^-1) = 0x5E33
x*_c' = x*' · A = 0x5E33 · 0xE523 = 0x54507FF9
x*_c' ≠ x_c' ⇒ x_c' is erroneous ⇒ x*' ≠ x!
```

All errors transforming a valid x_c to another multiple of A cannot be detected, as already shown above. The example in listing 2.3 illustrates how the multi-bit error f_2 transforms the error-free x_c into the erroneous value x_c', which is erroneously considered being correct.

Listing 2.3: fastAN pseudo code example with undetected error

```
Magic constant: A = 58659 = 0xE523
Inverse of A:   A^-1 = 29323 = 0x728B
Data value:     x = 0x0333

x_c = A · x = 0xE523 · 0x0333 = 0x02DD0EF9

Error syndrome: f_2 = 0x0003D7C6

x_c' = x_c XOR f_2 = x_c XOR 0x0003D7C6 = 0x02DED93F (≡ 0 mod A)

### Integrity checking ###
low(x_c') · A^-1 = 0xD93F · 0x728B = 0x61340335
x*' = low(low(x_c) · A^-1) = 0x0335
x_c*' = x*' · A = 0x0335 · 0xE523 = 0x02DED93F
x_c*' = x_c' ⇒ x_c' is erroneously assumed to be correct!
```

5.2 fastANBD

The fastANBD method shown in figure 2 uses the same integrity checking equation as the fastAN method does, but needs two additional calculation steps. Prior to multiplying the lower part of $x_{c_{ANBD}}$ by A^{-1}, the expected sum $B + D$ must be subtracted. After the last multiplication by A, $B + D$ is added to the result again and it is compared to the original coded value $x_{c_{ANBD}}$.

$$
\begin{aligned}
& \left(\left(\left((x_{c_{ANBD}} - (B + D) \bmod 2^n) \cdot A^{-1}\right) \bmod 2^n\right) \cdot A\right) + (B + D) = x_{c_{ANBD}} \\
\Leftrightarrow \quad & \left(\left(A \cdot x \cdot A^{-1}\right) \bmod 2^n\right) = x \\
\Leftrightarrow \quad & x \bmod 2^n = x \\
& \Leftrightarrow x \in \{0, ..., 2^n - 1\}
\end{aligned}
$$

The initial subtraction and the final addition of $(B + D)$ results in a lower runtime reduction compared to the one achieved with the fastAN scheme.

5.3 fastAN(BD) Usage with Different Data Type Widths

Higher data bit widths, e. g. casting to 2^{2n} introduce the need to use a A^{-1} determined for this bit width. In the examples shown up to now, multiplicative

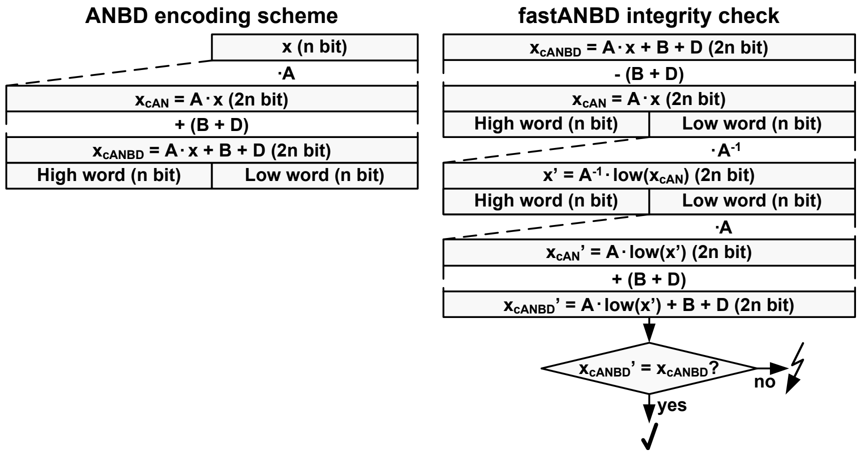

Fig. 2: Flow chart of ANBD encoding and new fastANBD scheme

inverse A^{-1} with the same bit width 2^n as the magic constant A have been used. But for uncoded data values z with

$$\log_2 z > n\,,$$

that can lead to encoded data values with

$$\log_2 z_c > 2n$$

respectively, a multiplicative inverse suitable for the extended bit width of the data value must be used. The example in listing 2.4 shows the addition of two coded operands, where the result does not fit in the bit width 2^{16} of the source operands. The 16 bit wide A_{16}^{-1} cannot be used for integrity checking, its 32 bit counterpart A_{32}^{-1} is applied instead. Please note that A_{32}^{-1} can be determined like A_{16}^{-1} using the extended euclidian algorithm.

If the overflow shown in the example in listing 2.4 is not handled correctly by increasing the bit width of the encoded result, the result is invalid. This is detected by the fastAN(BD) scheme as shown in listing 2.5.

Listing 2.4: fastAN pseudo code with change of bit width

```
Magic constant:          A = 58659 = 0xE523
Inverse of A (16 bit): A_16^-1 = 29323 = 0x728B
Inverse of A (32 bit): A_32^-1 = 2839442059 = 0xA93E728B
Data values:             x = 0xF000, y = 0xE000

Addition of x and y results in 32 bit wide z:
z = x + y = 0xF000 + 0xE000 = 0x0001D000

x_c = A · x = 0xE523 · 0xF000 = 0xD6D0D000
y_c = A · y = 0xE523 · 0xE000 = 0xC87EA000

Addition of x_c and y_c results in 64 bit wide z_c:
z_c = x_c + y_c = 0xD6D0D000 + 0xC87EA000 = 0x000000019F4F7000

### Integrity checking ###
low(z_c) · A_32^-1 = 0x9F4F7000 · 0xA93E728B = 0x69524D750001D000
z* = low(low(z_c) · A_32^-1) = 0x0001D000
z_c* = z* · A = 0x0001D000 · 0xE523 = 0x000000019F4F7000
z_c* = z_c ⇒ z_c is error-free ⇒ z* = z.
```

Listing 2.5: fastAN pseudo code of unhandled overflow

```
Magic constant:          A = 58659 = 0xE523
Inverse of A (16 bit): A_16^-1 = 29323 = 0x728B
Data values:             x = 0xF000, y = 0xE000

Addition of x and y results in 32 bit wide z:
z = x + y = 0xF000 + 0xE000 = 0x0001D000

x_c = A · x = 0xE523 · 0xF000 = 0xD6D0D000
y_c = A · y = 0xE523 · 0xE000 = 0xC87EA000

Unhandled overflow occurs during addition:
z_c = x_c + y_c = 0xD6D0D000 + 0xC87EA000 = 0x9F4F7000

### Integrity checking ###
low(z_c) · A^-1 = 0x7000 · 0x728B = 0x321CD000
z* = low(low(z_c) · A^-1) = 0xD000
z_c* = z* · A = 0xD000 · 0xE523 = 0xBA2C7000
z_c* ≠ z_c ⇒ z_c is erroneous!
```

6 Evaluation of the fastAN(BD) Method

The new fastAN(BD) method introduced in this paper shall now be evaluated regarding reduction of runtime needed for integrity checking, the residual error probability and the consistency of the decoding process.

6.1 Runtime Reduction

The runtime reduction of fastAN compared to the standard integrity checking scheme is shown in table 2. The magic constant $A = 58659$ was used to measure the runtime reduction, together with its multiplicative inverse $A^{-1} = 29232$.

For all x86- and ARM-based tests the code shown in listings 2.6, 2.7 and 2.8 was used and compiled with gcc. On x86 machines, the code was compiled using the option "-m32" to force the creation of a 32-bit binary, without any non-default optimization settings. The value x to be encoded was $x = 4772$. In listing 2.8, the subtraction and addition of `u16_BD` was put in square brackets, used in the fastANBD scheme only. For $x = 0$ only, the ARMv6 processor is showing an increased runtime for the fastANBD scheme.

Listing 2.6: Standard AN integrity check

```
if(0U != u32_AN % u16_A)
{
  u32_num_errors++;
}
```

Listing 2.7: Standard ANBD integrity check

```
if(u16_BD != u32_AN % u16_A)
{
  u32_num_errors++;
}
```

Listing 2.8: fastAN[BD] integrity checking scheme

```
u32_temp = (((unsigned short) u32_AN) [- u16_BD]) * u16_inverse_A;
u32_temp = ((unsigned short) u32_temp) * u16_A;
if(u32_AN != u32_temp [+ u16_BD])
{
  u32_num_errors++;
}
```

The runtime reduction on MSP430 devicews has been empirically derived from the minimum and maximum number of cycles needed to calculate the standard and the new fastAN integrity checking schemes $\forall x \in \mathbb{Z}/2^{16}\mathbb{Z}$. The significant runtime reduction shows the suitability of the new fastAN(BD) schemes for low-cost microprocessors, e. g. in Internet of Things (IoT) applications.

Table 2: Runtime reduction results by using fastAN(BD) integrity checking scheme

Processor	**fastAN**	**fastANBD**
Intel i7-8700K	$\approx 45\,\%$	$\approx 20\,\%$
Intel i5-3320M	$\approx 61\,\%$	$\approx 35\,\%$
Intel i5-2520M	$\approx 67\,\%$	$\approx 40\,\%$
ARMv6 (Raspberry Pi)	$\approx$ 9 - 66 %[1]	$\approx$ -12 - +59 %[1,2]
MSP430G2553 (no HW-mult.)	$\approx$ 29 - 84 %[1]	$\approx$ 31 - 84 %[1]
MSP430F247 (HW-mult.)	$\approx 89\,\%$	$\approx 88\,\%$

[1]: Runtime reduction is input data dependent; [2] runtime increases for $x = 0$

6.2 Residual Error Probability

The residual error probability $p_{AN(BD)}$ of AN(BD)-coded data is often cited as being

$$p_{AN(BD)} = \frac{1}{A}$$

for equally distributed errors. This is due to the fact that every multiple of A will satisfy the integrity checking equation. This applies even to those multiples of A, that would lead to an overflow in case of decoding of a 2^{2n}-bit-wide encoded data value to a 2^n-bit-wide decoded one.

$$x_c = x \cdot A = 65537 \cdot A = 0 \mod A$$

$$x' = \frac{x_c}{A} \bmod 2^{16} = \frac{x \cdot A}{A} \bmod 2^{16} = \frac{65537 \cdot A}{A} \bmod 2^{16} = 1 \neq x = 65537$$

The new fastAN(BD) scheme provides a lower residual error probability $p_{fastAN(BD)}$ compared to the standard scheme for all $A < 2^n - 1$. It detects overflows of the coded data values which would result in an erroneous decoded value as an error – a benefit, the standard integrity checking scheme does not provide. This results in the residual error probability $p_{fastAN(BD)}$ being

$$p_{fastAN(BD)} = \frac{1}{2^n - 1} \leq p_{AN(BD)} = \frac{1}{A}$$

for equally distributed errors, where n is the bit width of data words to be encoded.

Another benefit is the fact, that $p_{fastAN(BD)}$ is not dependent on the choice of A. This gives the possibility to use lower As without any unwanted rise of the residual error probability.

6.3 Consistency of the Decoding Process

The new fastAN(BD) scheme has a significant advantage compared to the standard scheme: decoding and integrity checking is being done in one sequence. The data value is first decoded and then re-encoded to do the integrity check, allowing to verify the integrity of the decoded value.

7 Recommended Values of A and their Multiplicative Inverses

For the sake of completeness, table 3 shows the 16-bit Super-As identified by Ulbrich [13] together with their specific multiplicative inverses A_{16}^{-1} and A_{32}^{-1} for usage in the introduced fastAN(BD) schemes for data values with $n = 16$ or $n = 32$ bit and encoded data values of $2n = 32$ or $2n = 64$ bit width, respectively.

Table 3: Super-As [13] and their multiplicative inverses A_{16}^{-1} and A_{32}^{-1}

Super-A $\in \mathbb{Z}/2^{16}\mathbb{Z}$	$\mathbf{A_{16}^{-1}} \in \mathbb{Z}/2^{16}\mathbb{Z}$	$\mathbf{A_{32}^{-1}} \in \mathbb{Z}/2^{32}\mathbb{Z}$
58659	29323	2839442059
59665	10225	2574460913
63157	51101	4099590045
63859	54203	1536218043
63877	5965	3510769485

As mentioned earlier, there can be multiplicative inverses A^{-1} for the chosen As in a residue class ring, but their existence is not guaranteed. As table 3 shows, all the Super-As $\in \mathbb{Z}/2^{16}\mathbb{Z}$ do have valid $A_{16}^{-1} \in \mathbb{Z}/2^{16}\mathbb{Z}$ and $A_{32}^{-1} \in \mathbb{Z}/2^{32}\mathbb{Z}$ and can thus be used in the fastAN(BD) integrity checking and decoding schemes.

References

[1] AIRBUS: Fly-by-wire; `http://www.airbus.com/innovation/proven-concepts/in-design/fly-by-wire/`

[2] ARM: Cortex-M3 Technical Reference Manual, 3.3.1. Cortex-M3 instructions; Revision r2p0; `http://infocenter.arm.com/help/index.jsp?topic=/com.arm.doc.ddi0337h/CHDDIGAC.html`

[3] R. C. Baumann, E. B. Smith: Neutron-Induced Boron Fission as a Major Source of Soft Errors in Deep Submicron SRAM Devices; Reliability Physics Symposium, 2000.

[4] D. T. Brown: Error Detecting and Correcting Binary Codes for Arithmetic Operations; IRE Transactions on Electronic Computers; Vol. EC-9, Issue 3; 1960

[5] P. Forin: Vital Coded Microprocessor Principles and Application for Various Transit Systems; IFAC Control, Computers, Communications; S. 79–84; 1989

[6] T. Granlund: Instruction latencies and throughput for AMD and Intel x86 processors; 24.04.2017; `https://gmplib.org/~tege/x86-timing.pdf`

[7] Microchip: 16-bit MCU and DSC Programmer's Reference Manual; DS70157F; `http://ww1.microchip.com/downloads/en/DeviceDoc/70157F.pdf`

[8] NISSAN: Nissan Pivo Concept Press Kit: Overview; `http://nissannews.com/en-US/nissan/usa/releases/435dd488-658e-433a-a57a-cd0184e4b51c`

[9] E. Normand: Single Event Upset at Ground Level; IEEE Transactions on Nuclear Science, Vol. 43, Issue 6; pp. 2742–2750; 1996

[10] U. Schiffel: Hardware Error Detection Using AN-Codes; PhD thesis; Technische Universität Dresden; 2011

[11] SEMI: Why Moore Matters; `http://semi.org/en/node/55026`; 2015

[12] N. Shimizu: Nissan Puts Steer-by-Wire on the Road: An In-Depth Look at the Technology; Nikkei BP Japan Technology Report / A1403-058-005

[13] P. M. Ulbrich: Ganzheitliche Fehlertoleranz in eingebetteten Softwaresystemen; PhD thesis; Friedrich-Alexander-Universität Erlangen-Nürnberg; 2014

Notes on the Design of a Statically Safe Microprocessor

Marcel Schaible

FernUniversität in Hagen, Germany
Faculty of Mathematics and Computer Science
email: marcel.schaible@fernuni-hagen.de

Abstract: Off-the-shelf microprocessors with their steadily increasing complexity prevent the construction of verifiable systems. The deployment of these processors in safety critical environments, where faulty systems can harm major infrastructures and human lifes, must be considered negligent. Formal, mechanised theorem proving software cannot solve this problem, because verification is, just like mathematical proofs, a social process which relies on the consensus of their community members. This paper advocates abandoning superfluous complexity and suggests a microprocessor architecture with the major design goal: simplicity. The whole design will be accessible online for consensual analysis and verification.

1 Introduction

The microprocessors used in embedded systems are becoming more and more complex, preventing the construction of testable and safe systems. The remedy is the described microprocessor architecture with the major design objective to remove all superfluous complexity. This design cuts out consequently all "artificial" complexity like e.g. caches and out-of-order execution. The complete design will be publicly available for consensual analysis and verification.

Safety-critical systems are the main target of the proposed architecture. Especially these systems are demanding for correctness, reliability, availability and deterministic time behaviour [9, 20, 21].

The VLIW (Very Long Instruction Word) architecture is a variant of a instruction set architecture (ISA) with the goal to increase the execution speed by issuing multiple operations in parallel at the instruction level. A compiler analyses sequential programs and determines operations, which can be performed in parallel. These operations are grouped together and encoded in the instruction word. The size of the various groups is determined by the available functional units.

The proposed architecture adopts this idea by composing an instruction word with several operations not to increase the execution performance but rather to simplify the instruction decoder.

The control unit (CU) is the central part of a microprocessor. It initiates sequences of controlling signals in a predefined order to coordinate the signal flow of the various parts of the processor. Control units are designed with hardwired or microprogrammed control logic [28]. Hardwired control units, especially with a significant amount of instructions, are tedious to verify, because of their hardly comprehensible wired connections between different parts. On the other hand microprogrammed control units model most of their logic in a read-only control store (CS). The sequence of signal transitions is stored in a control store as a two dimensional table and each operation is represented by a set of control signals.

The major requirements of the control unit presented in chapter 2.1 are *consensus-oriented* and *crowd-verifiable* [5, 12]. Both terms are defined in chapter 3.

2 Architecture

The proposed microprocessor architecture (see fig. 1) consists (like the Harvard approach) of strictly separated program (ROM: read only memory) and transient memories (RAM: random access memory), the control unit (CU) and several basic functional units (FU). In contrast to the Harvard architecture the execution of all instructions is strictly sequential to simplify the understanding of the internal state changes of the processor. Because ROM and RAM are housed completely on-chip the memory access is relieved from complicated external memory control logic. The functional units are implementing the basic operations e.g. adding or comparing two data values.

The processor can address 2^m memory cells. The program memory is n-bit wide where n depends on the number of operations (see fig. 4). The transient memory consists of m-bit wide cells and reserves a part of the memory address space to support memory mapped input/output.

The presented microprocessor architecture includes the following properties:

- Instructions are executed strictly sequential.

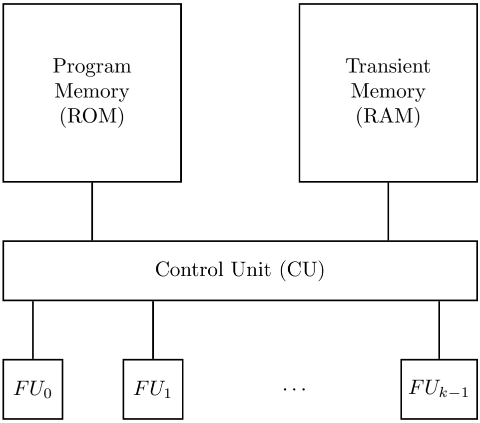

Fig. 1: Architecture Overview

- Code and data memory are housed on-chip.
- Code and data memory are completely separated for safety reasons (Harvard Architecture [11]) .
- Code memory is read-only during execution time.
- The table-driven control unit is designed to guarantee deterministic execution time per instruction.
- The instruction decoding is simplified by defining all instructions equally long.
- Each instruction points directly to a distinct row of the control store.

2.1 Datapaths and Control Units

The data path (DP) in fig. 3 consists of a set of functional units (FU).

The transient memory is connected through the memory address register (MAR) and the memory content register (MDR) to the data bus. The registers A, B, C, program counter (PC) and the program status word (PSW) are also connected to the data bus. The code memory interfaces via the code address register (CAR) and the instruction register (IR) to the various functional units (FU).

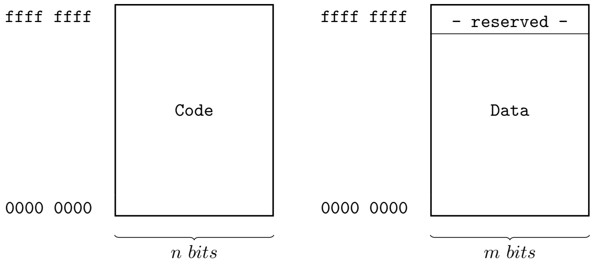

Fig. 2: Memory Layout

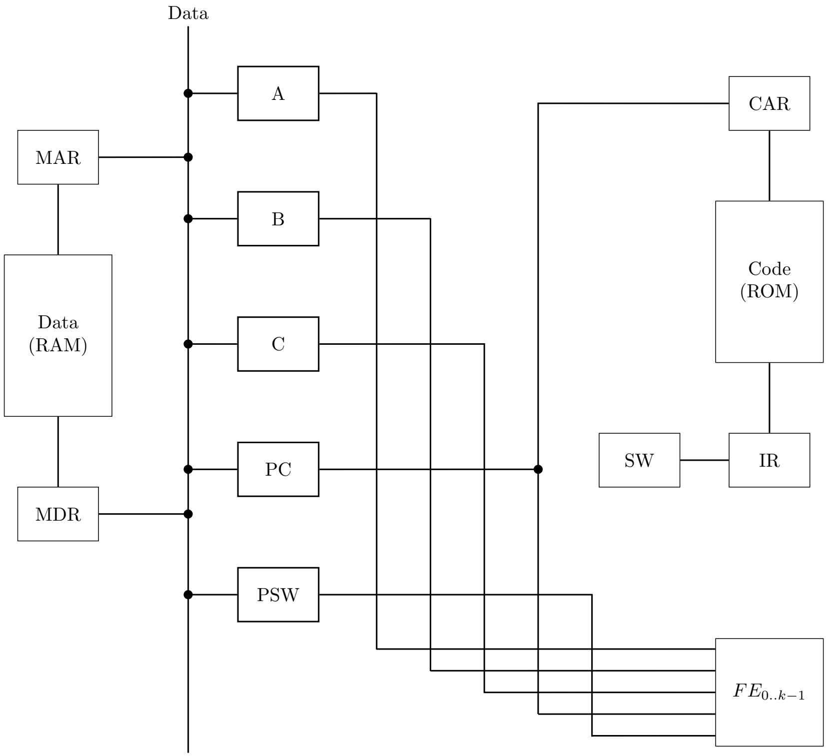

Fig. 3: Datapaths

The major part of the control unit (CU) consists of the control store (CS). CS is a read-only table of control signals where the columns define a set of control signals for enabling certain registers and setting read- or write-flags.

The general execution of the CU is described by algorithm 1.

The control unit is designed with the following properties:

1. The first three operations are stored in ROM cells and each instruction is implicitly prefixed with them.
2. Fetch cycle: The program counter (PC) points to the next instruction and is transferred into the program address register (CAR). This initiates the reading of the memory content addressed by CAR. After completion the memory content is available in the instruction register (IR). After loading the IR register the PC is incremented.
3. Execute cycle: Loop over the operations and enable the various functional units (see fig. 5).

A decode phase is unnecessary, because each operation points directly at exactly one row of the control store. All unused operation codes will immediately lead to a processor halt when loaded into the IR register of the control unit.

Algorithm 1: Execution Model

Reset;
while *forever* **do**
  foreach $i \in \{-3, -2, -1, 0, ..., l-1\}$ **do**
    $CU_{addr} \leftarrow \text{Operation}_i$
    Apply selected control signals which e.g. enable FU
  end
end

2.2 Instructions and Operands

All instructions share the uniform layout in fig. 4 and consists of $0 \ldots l-1$ operations followed by three operands.

The m-bit wide operands are embedded in the instruction word and can contain constants, direct or indirect memory addresses.

Table 1 list all supported addressing modes.

$$\underbrace{\boxed{\mathit{Elementary\ Operation}_0}\ \cdots\ \boxed{\mathit{Elementary\ Operation}_{l-1}}}_{\text{Elementary Operations}}\ \underbrace{\boxed{\mathit{Operand\ A}}\ \boxed{\mathit{Operand\ B}}\ \boxed{\mathit{Operand\ C}}}_{\text{Operands}}$$

Fig. 4: Instruction Format

Table 1: Addressing Modes

Addressing mode	Description
constant	$register \leftarrow operand$
direct	$register \leftarrow M[operand]$
indirect	$register \leftarrow M[M[operand]]$

Table 2: *Opcodes*

Opcode binary	sedecimal	Mnemonic	Description
0000 0000	00	`SKIP`	Skip instruction
0000 0001	01	`LDPC`	$CAR \leftarrow PC$
0000 0010	02	`LDIR`	$IR \leftarrow CDR$
0000 0011	03	`INCPC`	$PC \leftarrow PC + 1$
0000 0100	04	`MARD`	$MAR \leftarrow MDR$
0000 0101	05	`LDAC`	$A \leftarrow constant$
0000 0110	06	`MARA`	$MAR \leftarrow A$
0000 0111	07	`AMDR`	$A \leftarrow MDR$
0000 1000	08	`LDBC`	$B \leftarrow constant$
0000 1001	09	`MARB`	$MAR \leftarrow B$
0000 1010	$0a$	`BMDR`	$B \leftarrow MDR$
0000 1011	$0b$	`LDCD`	$C \leftarrow addr$
0000 1100	$0c$	`MARC`	$MAR \leftarrow C$
0000 1101	$0d$	`CMDR`	$C \leftarrow MDR$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
0010 0000	20	`ADD`	$C \leftarrow A + B$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
0100 0000	40	`COPR`	$M[C] \leftarrow R$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1111 1111	ff	Halt	Halts the execution

The encoding of the various addressing modes is done by issuing the appropriate operation like $LDAC$ for loading the register A with a constant.

2.3 Instruction Register

The instruction register (IR) stores the active set of operations and operands. Because one requirement of this architecture is to relieve the CU from complex logic, the fetch phase is integrated into the IR. The opcodes with indices -3 through -1 implement the fetch of the next instruction and are hardcoded. With this approach the CU does not implement a distinct fetch phase.

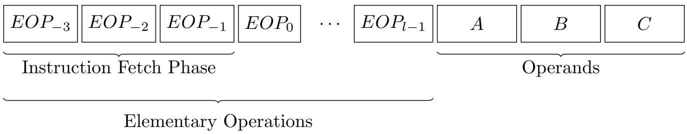

Fig. 5: Instruction Register

2.4 Functional Units

All functional units (see fig. 6) share the following structure:

- Inputs: Datapath to the operands A and B, program counter (PC) and program status word (PSW)
- Processor clock
- Enable signal activates the processing of the FU.
- Outputs: Datapath to the result register, PC and PSW

Because some functional units do not use all of the connections (e.g. the adder does not need to access the PC) it is feasible not to implement these paths.

2.5 Example

Consider the following instruction for adding the constant `#42` to the value at memory address `$10` and storing the result pointed to by the address `$20`.

```
ADD #42, $10, [$20]
```

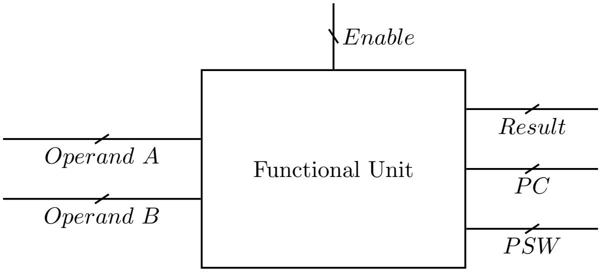

Fig. 6: Functional Unit

LDPC	LDIR	INCPC	LDAC	MARB	BMDR	MARC	MARD	CMDR	...	42	10_{16}	20_{16}

Fig. 7: ADD Instruction Format

The instruction encoding is described in fig. 7.

The execution of this instruction is divided into the following steps:

ADD	#42, $10, [$20]	
LDPC		Load `PC` into `CAR` register
LDIR		Load `CDR` into `IR` register
INCPC		Increment the `PC`
LDAC	42_{10}	Load the `A` register with the constant `42`
MARB	10_{16}	Load 10_{16} into the `MAR` register
BMDR		Store M[10_{16}] into the `B` register
MARC	20_{16}	Load 20_{16} into the `MAR` register
MARD		Load M[20_{16}] into `MAR`
CMDR		Store M[M[20_{16}]] into the `C` register
ADD		Enable the adder
STR		Store the value in `R` into `M[C]`

3 Verification

The standard model of chip verification is described in fig. 8. Starting with the designer's intention a functional specification is produced, which leads to a microarchitecture, a Register-Transfer-Level (RTL) description, a netlist and finally from a physical layout to a silicon dye. In each step the correctness of the transformations must be ensured.

Chip manufactures utilise formal verification methodologies in some areas like floating-point unit for risk minimisation [14]. Because of the complex designs of modern processors, automated theorem proving systems (ATPS) [2, 4, 13, 29] tend to generate large proofs with thousands of deductions which cannot be checked even by human experts. It is the state of the art that sufficiently complex software programs like ATPS are unlikely to be error-free and therefore the confidence into the correctness of the generated proofs is questionable. This implies that errors or misconceptions regarding the specification will remain undetected.

The main drawbacks of automated and mechanised theorem prover are:

- Sophisticated designs produce long and hard-to-follow reasoning.
- Theorem provers like all other software programs are not error free. Therefore the generated proofs can contain errors which are hard to reveal.
- Intentions of designers cannot be formalised.

Understanding the nature of a proof cannot be achieved by reading it line-by-line and checking each line for syntactically correct transformations. A proof is more a guide for reasoning [22, p. 317].

Confidence and trust in technical artefacts can only be gained by transparent designs and hence are the result of a social process. The key is *intelligibility* and as a consequence *simplicity* of the architecture. Comprehensible and verifiable designs by humans are inevitable.

3.1 Consensus-oriented Verification

Because of the clean and straightforward design of the above described architecture the verification will be performed consensus-oriented and crowd-based [5, 12, 24, 25]. Consensus-oriented verification is the process of examination of a community (crowd-based), which comes to a common understanding and agreement that a design can be considered correct. This approach is well established in science. Mathematicians have been working consensus-oriented and providing their work to the mathematic community for examination with great success for thousands of years.

The proposed architecture will be published online for crowd-based evaluation and verification.

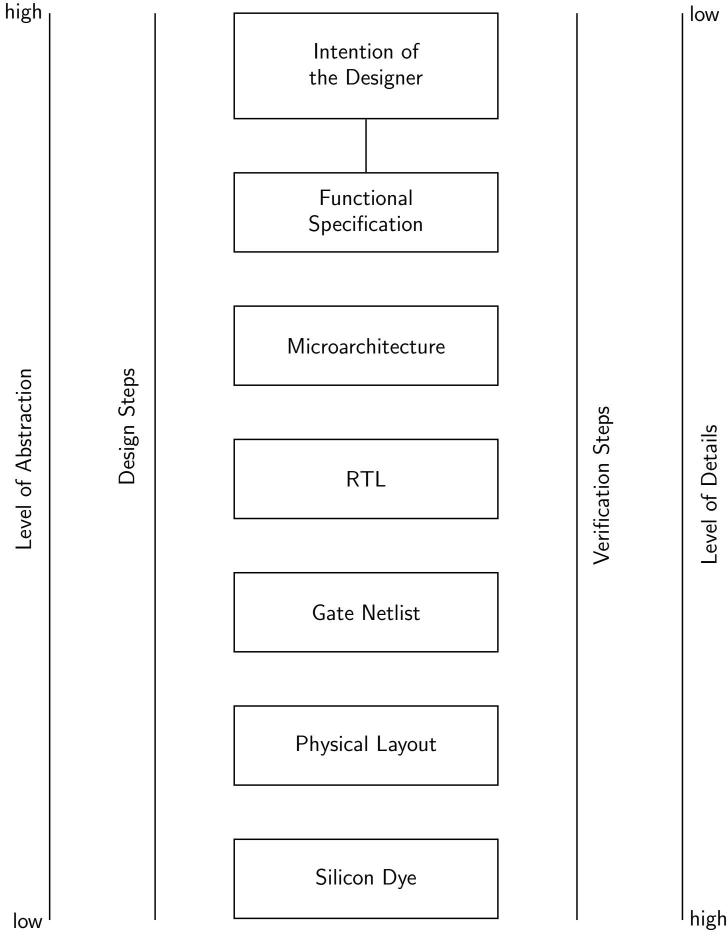

Fig. 8: Verification Model

But because of the inherent design properties of the presented architecture it is feasible to apply the consensus-oriented and crowd-based methodology to a reversed verification process: Starting from the physical layout the verification can be done backwards to the netlist all the way up to the specification. This approach was first applied as a software verification methodology by TÜV Rheinland [19] to acquire permission for commissioning of the nuclear power plant in Halden (Norway). Although this method was used with software it can be adapted to hardware verification. An important ancillary effect of the diverse backward analysis methodology [9] is the intrinsic verification of the correctness of the applied transformation tools.

The core characteristics (see [9, pp. 152-153]) of the diverse backward analysis methodology are

1. The process of verification is diverse in relation to the implementation.
2. The process of verification has the character of a proof.
3. Each step of the verification is strictly documented and checkable.
4. The verifier is not defencelessly delivered to potential systematic errors of the verification process.

4 Summary

The objective of this work is to develop a verifiable microprocessor architecture for usage in safety-critical systems. The architecture has consequently cut out all artificial "complicatedness" [10].

The described control unit architecture with its table-driven composition and strictly sequential execution logic provides a coherent architecture which can be verified by field and non-field experts in a consensus-oriented and crowd-based approach. Due to the characteristics of the table-driven design the verification can be literally performed by checking each set of generated control signals.

The crowd-based verification can be performed both forward and backward in a consensus-oriented discourse with the objective to gain confidence and trust in the correctness of the examined design. Diverse backward analysis as a powerful verification methodology, which establishes trust not only in the correctness of the design itself but also in the tools used to generate it, can be applied.

References

[1] Andersen, B. Scott and Romanski, George: Verification of Safety-critical Software, In: *Queue, ACM*, 9, 8, pp. 50–59, 2011

[2] Berg, C. and Beyer, S. and Jacobi, C. and Kröning, D. and Leinenbach, D.: Formal Verification of the VAMP Microprocessor (Project Status), Symposium on the Effectiveness of Logic in Computer Science (ELICS02), pp. 31–36, 2002

[3] Berg, C. and Jacobi, C. and Kröning, D.: Formal Verification of a Basic Circuits Library, In: *Proc. of the IASTED International Conference on Applied Informatics, Innsbruck (AI 2001)*, pp. 31–36, 2001

[4] Beyer, S. and Jacobi, C. and Kroening, D. and Leinenbach, D. and Paul, W. J. : Putting it all together - Formal Verification of the VAMP, In: *Software Tools for Technology Transfer (STTT), Special Section on Recent Advances in Hardware Verification* , 8, pp. 411–430, 2006

[5] Brabham, Daren C.: Crowdsourcing, *The MIT Press*, 2013

[6] Cohn, A: A proof of correctness of the viper microprocessor: the first level, In: *University of Cambridge, Computer Laboratory*, 1987

[7] Dijkstra, E. W.: The next fifty years, 1996

[8] Gilreath, W. F. and Laplante, P. A.:Computer Architecture: A Minimalist Perspective, *Kluwer Academic Publishers*, 2003

[9] Halang, W. A. and Konakovsky, R.: Sicherheitsgerichtete Echtzeitsysteme, *Oldenbourg*, pp. 150–152, 1999

[10] Halang, W. A.: Simplicity Considered Fundamental to Design for Predictability, In: *Perspectives Workshop: Design of Systems with Predictable Behaviour, 16.-19. November 2003*, 2004

[11] Hennessy, J. and Patterson, D.: Computer Architecture - A Quantitative Approach, Morgan Kaufmann, 2011

[12] Howe, Jeff: Crowdsourcing: Why the Power of the Crowd is Driving the Future of Business, *Crown Business*, 2008

[13] Hunt, W. A. J.:FM8501: A Verified Microprocessor, In: *Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence*, Springer, 1994

[14] Intel Corp.: *Intel Xeon Processor E3-1200, v3 Product Family, Specification Update*, 2015

[15] Joyce, J.J.: Formal specification and verification of microprocessor systems, In: *Technical report (University of Cambridge. Computer Laboratory)*, 1998

[16] von Kaenel, P. A.: Designing and testing a control unit, In: *J. Comput. Sci. Coll.*, 19, 5, pp. 228–237, 2004

[17] Kaivola, Roope et al.: Replacing Testing with Formal Verification in Intel®

Core^{TM} i7 Processor Execution Engine Validation, In: *Proceedings of the 21st International Conference on Computer Aided Verification*, Springer Verlag, pp. 414–429, 2009

[18] Knuth, D. E.: An Empirical Study of FORTRAN Programs, In: *Softw., Pract. Exper.*, 1, 2, pp. 105–133, 1971

[19] Krebs, H. and Haspel, U.: Ein Verfahren zur Software-Verifikation. In: *Regelungstechnische Praxis*, rtp 26, pp. 73–78, 1984

[20] Laprie, J. C., Avizienis, A. and Kopetz, H.: Dependability: Basic Concepts and Terminology, Springer Verlag, 1992

[21] Laprie, J.-C. and Béounes, C. and Kanoun, K.: Definition and Analysis of Hardware- and Software-Fault-Tolerant Architectures, In: *Computer, IEEE Computer Society Press*, 23, 7, pp. 39–51, 1990

[22] MacKenzie, D.: Mechanizing Proof: Computing, Risk, and Trust, MIT Press, 2001

[23] Schaible M.: Towards the Verification of a Table-driven Microprocessor Architecture for Safety-critical Systems, In: *Fortschritts-Berichte VDI, Proceedings of the 6th GI Workshop Autonomous Systems 2013*, Vol. 835, pp. 24–30, 2013

[24] Schaible M.: A Consensus-oriented Crowd-verifiable Microprocessor Architecture, In: *Fortschritts-Berichte VDI, Proceedings of the 7th GI Workshop Autonomous Systems 2014*, Vol. 835, pp. 209–220, 2014

[25] Schaible M.: Design of a Consensus-oriented and Crowd-verifiable Control Unit, In: *Fortschritts-Berichte VDI, Proceedings of the 8th GI Workshop Autonomous Systems 2015*, Vol. 842, pp. 52–60, 2015

[26] Schaible M.: On the construction of a crowd-verifiable Microprocessor, In: *Fortschritts-Berichte VDI, Proceedings of the 9th GI Workshop Autonomous Systems 2016*, Vol. 848, pp. 46–54, 2016

[27] Stieger, H. and Halang W. A.: Eine hochsprachenorientierte Rechnerarchitektur ohne arithmetische Register, *IFB Verlag*, 2003

[28] Wilkes, M. V. and Stringer, J. B.: Micro-programming and the design of the control circuits in an electronic digital computer, In: *Mathematical Proceedings of the Cambridge Philosophical Society*, 2, pp. 230–238, 1953

[29] Phillip J. Windley: Formal Modeling and Verification of Microprocessors, In: *IEEE Transactions on Computers*, 44, 1, pp. 54–72, 1995

An Associative Ring Memory to Support Decentralised Search

Herwig Unger and Mario Kubek

Chair of Communication Networks, FernUniversität in Hagen, Germany

Abstract: Centroid terms are single words that semantically and topically characterise text documents and thus can act as their very compact representation in automatic text processing and mining. In order to make them a useful tool in textual search tasks as well, a concept for a novel associative memory with a ring-like structure storing and managing these terms and their associated contents is proposed. As this memory is designed to be applied in decentralised search systems, centroid terms will inherently support the efficient routing and forwarding of queries to matching nodes this way. Besides providing remarks on its implementation, necessary load balancing activities as part of the memory's management functions are discussed as well.

1 Motivation

Text-representing centroids [12] have been introduced to foster the categorisation of text documents and multi-word queries using a single, descriptive term. These terms can be used to compute of the similarity of documents [15] and to build document clusters as well as hierarchic structures to support the search of documents in the Internet [14, 17].

The practical application of centroid-based methods was made possible by a new graph-based method [16] for the fast calculation of centroid terms for texts of arbitrary length while relying on preferably large co-occurrence graphs as a knowledge base. Last but not least, the derived measures *speciality* and *diversity* of a given document or query indicate how general or detailed and topic-oriented the analysed content is [15].

In particular, multiple keywords of a search query and whole (longer) texts are represented by single centroid classifiers. Those single terms can be easily matched or compared [13, 14]. Thus, complex programs to combine a set of partial results for multi-keyword queries (like MapReduce of Google [5]) are no longer needed.

Additionally, the calculation of centroid terms can make use of personal co-occurrence graphs. Therefore, centroids may be obtained reflecting the recent knowledge and experience of the user. Last but not least, centroids can be lexically (alphabetically) ordered. The obtained order and relation ($<, =, >$) between any two centroid terms allows for the creation of well-ordered structures (like lists, rings, or trees), which can be efficiently searched by standardised algorithms. Only the expected, extremely high number of entries will require a distributed storage of such structures, preferably on peers of a flexible peer-to-peer (P2P) system.

After a short discussion of related works, a new associative ring structure shall be introduced, which enables a fast, fault-tolerant management of an almost unlimited number of entries and supports a fast textual search.

2 Review of Related Works

After the rise of classical, client/server-based content delivery systems, several approaches have been proposed to manage a large amount of contents within decentralised, P2P systems [2]. In order to avoid flooding, a slow access or replication of files, several approaches based on dynamic hash tables (DHT) have been introduced (for an overview see [3]). Chord [7][1] seems to be the most practicable and widely used approach since it reduces additional overhead and problems, especially in case of an unexpected leave of peers, to a minimum. A set of projects is available, using Chord and/or offering libraries to setup a Chord system, e.g. Open Chord `http://open-chord.sourceforge.net/`.

Two major problems could be identified using a standard Chord implementation: first, no load balancing is available. This may result in a concentration of entries on a single peer, if the hash value of many keys/entries is the same, but also if a concentrated access to a few, dedicated content items takes place. Second, due to the used hash function, similar content is usually not placed on neighbouring peers on the Chord ring. The latter said requires that a user must know the exact key (category) in order to locate and access the wanted content

[1] In order to avoid misinterpretations, 'Chord' is written with a capitalised 'C', if we mean the DHT-system, while 'chord' with a non-capitalised, first letter denotes the connection of two vertexes in a ring graph/circle.

on the Chord ring. Since similar document keys would not be placed in close proximity, a document search by human users would become more difficult.

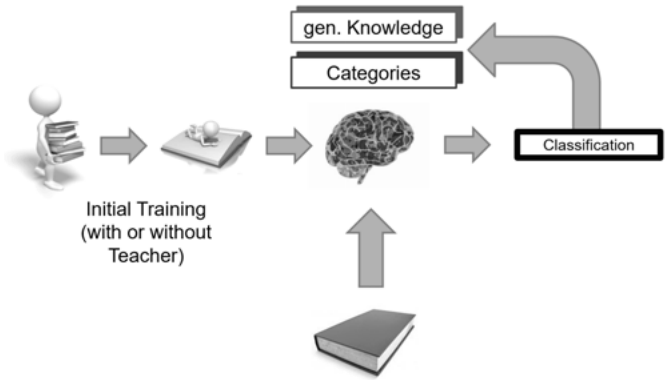

Fig. 1: Human learning and application of categories and classification

Humans usually learn about categories and their descriptors in a longer, supervised or unsupervised learning process Fig. 1. Starting with the research on WordNet [9] there is a consent that the relation between those categories can be well modelled and approximated by lexical graphs such as the so-called co-occurrence graphs. Different co-occurrence graphs for each individual may therefore reflect the state of knowledge of that person and are subject to a permanent change depending on the person's private experiences. As discussed in [10], these personal experiences are poorly considered in the search procedures of the big Internet search providers.

A data structure to be developed must take the above discussed facts and processes into account and support the following requirements:

1. Search must be understood as process, rather than a single event.
   Each document read and each search carried out qualifies the user.

2. The user's knowledge and experience must be represented in a data structure, which is stored –due to its personal character and for privacy reasons– on the user's machine, only.

3. Similar information shall be stored in a close proximity in the data structure to support a fast location.
4. The number of results returned shall be adjustable.
5. The used structure must be adaptable to a growing or huge number of data. Absolute addressing shall be avoided in order to support a hardware-independent, flexible work.

In the next sections, a respective data and communication structure shall be described and evaluated.

3 Design of Data- and Access Structures

3.1 Concept

Since the major concern is the support of search in the World Wide Web (WWW), it is intended to build up the data and access structures for the peer-to-peer (P2P) system, which has already been provided in form of a web server extension, described in [11] (Fig. 2).
Here, every web server also generates an own peer running in parallel to the web server system but is able to exchange data with it. Each peer obtains an initial neighbourhood derived from the links of pages hosted by the web server. This neighbourhood is continuously updated by the known, standard P2P PING-PONG protocol (what also allows for the inclusion of non-webserver peers in a more advanced stage of the system). In addition, every peer may index the locally offered webpages and may therefore immediately answer any incoming queries matching those documents. With the standard functions of a P2P-(file-)sharing system, a simple, decentralised search engine is already made available.

Fig. 3 shows the design of the proposed system. Its main component is a ring structure of peers, which are running on different machines.
It initially contains a single peer, only. A respective management functionality ensures that new, participating peers may be added and –if not needed anymore– be removed.
The ring of peers hosts a ring list of entries, whereby every peer can store a larger number of entries (partial list). Every entry represents one HTML document by a pair consisting of a key and a link (i.e. an URL) to the document. As searchable key, the corresponding centroid of the HTML document is used.
In such a manner, the $i-$th entry in the ring list has the form $[centroid_i, URL_i]$.

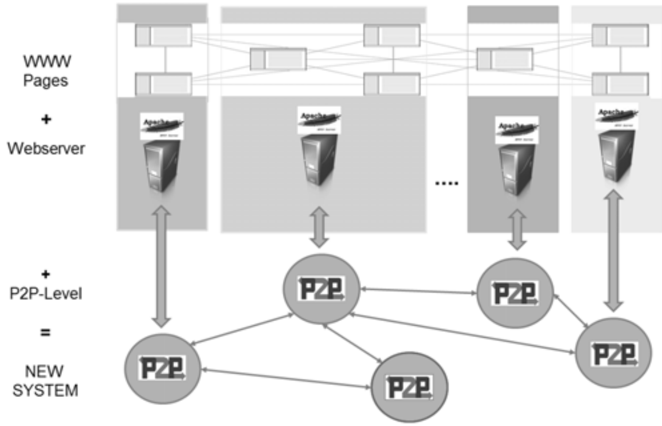

Fig. 2: Generating a P2P-system with the Apache Tomcat web server

Seen from the first entry $[centroid_1, URL_1]$, all items until the last one are lexically ordered[2], i.e. $centroid_i \leq centroid_j, \forall i, j$ with $i < j$.

Differing from Chord, as no fixed position can be given for any entry and depending on the number of entries and peers (as well as the load), this position is flexible. Queries as well as update operations of items (search, add and remove) can be directed to any peer participating in the ring structure and will be forwarded along the ring to the respective place, where the operation can be executed.

Furthermore, two types of chords are added to each node of the ring:

- F random chords with chosen destinations from all peers of the ring, shortening the access to any item, similar to the fingers in Chord.
- S chords connecting the peer with its K nearest neighbours for fault tolerance reasons. Therefore, each peer must mirror the contents of those K peers as a backup copy in case one of these peers suddenly leave the ring without carrying out a proper exit procedure.

Last but not least, every user may access the ring structure by a respective peer service. This peer service includes

[2]Note that for a position (x, y, λ) representing an evolving centroid $x \leq y$ is required ensuring that for every position a unique representation is given and a lexical ordering by x, y and last but not least followed by the value of λ will be possible.

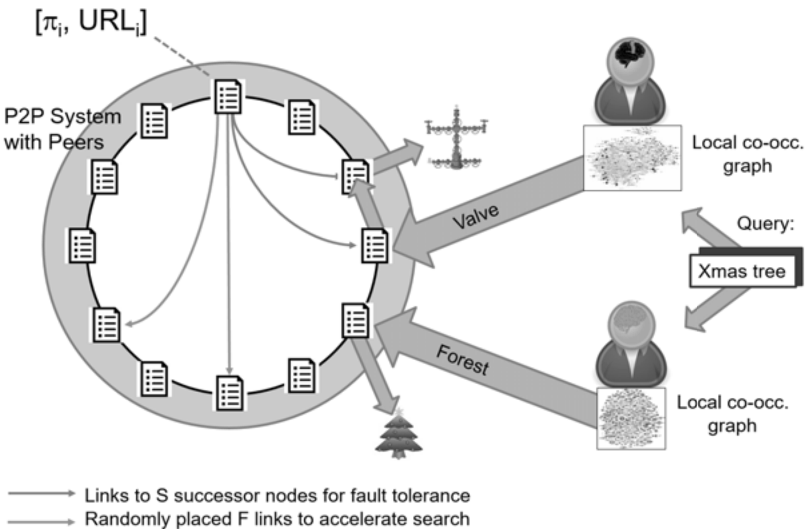

Fig. 3: The general concept

- an individual co-occurrence graph, which allows for the calculation of a centroid to handle a query or document, is built depending on the experience and history of the user and possibly his or her files on the local system.
  Following some considerations on the stability of co-occurrence graphs, in [15], no highly experienced user or librarian is needed to add a new document to the ring.
- a (bounded) broadcast mechanism (TTL 3..4) to locate at least one peer of the ring.
- functions with the respective communication protocol to execute the respective operations on the ring.

With these preliminaries, the functionalities and operations for the ring list memory can now be described.

3.2 Operations

The functionalities may be divided into the following three groups: user operations, ring management function and load balancing activities.

1. **User Operations**
   include the data operations *add*, *remove* and *search* for items on the ring. Therefore, a *lookup* function allows to locate any peer of the ring structure in the entire P2P-system. With the known address, any data operation may be send to this peer, while the ring itself provides the forwarding functionality to transfer and execute the requested operation on the right peer. Therefore, the given key $(centroid_r)$ must be routed to the peer p with $centroid_i$ such that $centroid_i < centroid_r \leq centroid_j$ for $j = i + 1$.
   (Depending on security and privacy needs, every entry may contain an owner and access right information in addition to the above defined version.)

2. **Ring Management Operations**
   adapt the size of the ring to the needed one depending on the relation between available and used memory and keep the system of chords.

   - Since the ring starts with a single peer only, new peers must be *added* and correctly linked in the ring structure, if the number of entries exceeds a given maximum depending on the size of the entire ring (due to the subsequently described load balancing, it is not usually necessary to consider a single peer only).

   - In the same manner, a peer can be *removed* if the number of entries on the ring is getting so low that (more than) one peer can be detached without exceeding the remaining capacity.
     After finding a respective candidate for removal, any remaining items must be moved to the remaining peers before the removal can be executed.
     It is self-evident, that the links of the successor and predecessor node must be adapted, if peers are added or removed from the ring.

   - Last but not least, chords to other nodes in the ring shall be periodically updated; the S links for fault tolerance in shorter periods, the F finger links supporting long-distance jumps after longer periods.
     For doing so, a *peer-lookup*-protocol is used, which consists of *peer-lookup*-operations and -messages. Of course, all peers in the ring must support the execution of all *lookup*- operations as well as the generation and forward of needed communications in the ring.

In order to stabilise the system, all activities shall be executed only, if the respective conditions are fulfilled for a (randomly chosen) period.

In case the management does not prevent the existence of several ring structures after system initialisation, a mechanism must allow the survival of the older structure and the entry-by-entry inclusion of the newer one; a ring identifier with a respective time stamp given to all participating peers is useful for doing so.

3. **Load Balancing Activities**
represent the most significant difference of this new approach to the classic Chord rings. It is therefore intended to use the physical analogue of a drop of water, wetting a surface (see Fig. 4).

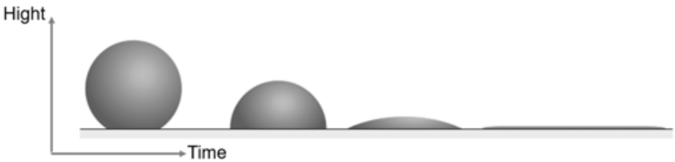

Fig. 4: A drop of water dissolving on a surface

The effects of the different acting powers and surface tension are modelled by the communication of each peer in the ring with its direct neighbours. If a peer contains β percent more entries than the average of its neighbours, it sends entries to both of its neighbours to equalise the powers and to level the height (Note that this realises a sender-based load balancing).

By doing so, it must be ensured that the right lexical order of the entries is kept. To simplify communication procedures, the levelling only involves the one neighbouring peer with the highest difference of items. For an example, see Fig. 5. Load balancing activities can also be activated, if a peer experiences a significant access load, i.e. an extremely high volume of communication must be handled.

As a result of this load balancing, no item is assigned to any fixed peer (and respectively no fixed IP address). Moreover, this position usually will change over time. Therefore, any item can be addressed only by its key, i.e. the realised memory is an associative one.

Time	Peer1	Peer2	Peer3	Peer4	Peer5
0	20	0	100	20	20
1	20	50	50	20	20
2	35	35	35	35	20
3	35	35	35	28	27
4	31	35	32	31	31
5	33	33	32	31	31
6	32	33	32	31	32

Fig. 5: Example: balancing the number of entries on a ring with 5 peers

In the following section, the working principles of the described associative ring memory shall be considered.

4 Remarks on Implementation

4.1 Ring Management

The algorithm for the ring management shall ensure that enough memory is available to keep and insert all entries on request but also avoid that to many peers are involved and resources remain unused. Therefore ring management operations can be executed resulting in an addition to or removal of peers from the ring structure. Of course, frequent changes of the ring size (i.e. oscillations) shall be avoided, i.e. without those changes, the size of the ring shall converge to a constant one.

To avoid central instances and allow a fully decentralised ring management, a token-/(random walker-) based procedure carried out by all participating peers is suggested. The set of all tokens represent the state of the system. The system strives to combine tokens and derive ring management activities from the obtained information.

Therefore, the content of each token stands for the balance of a part of the system: a value of 0 represents a fully balanced system having enough resources to add new data items to the list, a positive value denotes (especially with increasing numbers) that the system is getting filled up more and more. Negative values describe an underused system having too much unused resources (peers) employed.

Now, the system can work and adapt its size in a suitable manner by following the subsequent rules on each peer:

1. Every token i stands for an integer value $T(i)$.
2. Inserting into and removing an entry from the ring list result in the creation of a token with the value $T(i) = 1$ and $T(i) = -1$, respectively,
   - after the operation was successful and
   - by the peer hosting the respective item.
3. Tokens are forwarded in a randomly chosen direction clockwise or counter-clockwise along the doubly connected peer ring and remain for a randomly chosen time on each peer. This time will be extended if more than one token are on a peer or if an entry is added or removed from the ring list at the respective peer.
4. Two tokens i and j meeting on one and the same peer are merged into a single token by adding their values, i.e. $T(i) := T(i) + T(j)$ and token j will be removed.
5. Let $INS(p)$ be a constant corresponding to each peer p describing its memory capacity (usually 80% of the peer's memory capacity, which is made available to the ring).
   If a token with a value $T(i) \geq INS(p)$ is recognised on a peer p at any time, this peer inserts a new peer p' to the ring list.
   After doing so, the token value is reduced by $T(i) := T(i) - INS(p')$, where $INS(p')$ is representing the space on the new peer p', which can be different from those on p (Note, that by doing so, the token value may become negative, if peers with different resources (memory sizes) are used).
6. Let $DEL(p)$ be a (negative) constant, which is equal to $-(Capacity\_of\_Peer)\ p$. If a token value is $T(i) \leq DEL(p)$, the peer p on which the token is located organises the its removal from the ring list. The token value is increased by $T(i) := T(i) + DEL(p)$.
   Not that
   - the last peer (i.e. a peer pointing on itself as predecessor and successor) can not be removed and

- the token will be forwarded without any operation, if the data items of the peer cannot be successfully moved to its neighbouring peers within a given time limit.

7. Any token with $T(j) = 0$ will be removed immediately.

8. *Emergency rule:* If the capacity of any peer is used and an entry insertion is requested, a new peer p'' will be immediately added. In addition, a token k with the value $T(k) = -INS(p'')$ is created.

The described token game is susceptible to token losses, however, the S fault tolerance chords can be used to avoid problems from lost tokens. Therefore, the following handling and protocol is suggested (as a simplification of the procedures published in [1] for the use on ring structures).

1. Every token gets an unique Token ID in addition to its value.

2. The S predecessor peers $P_{i-1}..P_{i-|S|}$ of peer P_i keep a log of the passed token in a special registry.

3. After moving a token from peer P_i to peer P_{i+1}, P_i sends a message to $P_{i-|S|}$ to cancel the entry about this token in its registry and adds the token to its own.

4. If token in any registry gets older than a deadline T_{out}, the last node in the chain (i.e. the peer whose predecessor does not contain an entry on that token) re-creates that token.
   If $|S| \geq 2$, the following nodes confirm the re-creation of the token and adjust their registries in an appropriate manner.

It is easy to be seen that $|S|$ nodes must fail at the same time, in order to cause faults in the token management. i.e. with the above described methods, a stable, fault-tolerant execution shall be achievable.

As a matter of last resort, selected administrator nodes shall be allowed to

1. collect and stop all incoming tokens as well as process them as described in the basic token procedure,

2. initialise a special status token, which counts the available (free) memory space while completing one round in the ring and

3. generating an adjusted token after processing the status token and all stopped token on the administrator peer.

On need, more administrative tasks may be subsequently added to this status token protocol, which anyhow realises a decentrally working yet central control of the ring memory structure.

4.2 Finger Management

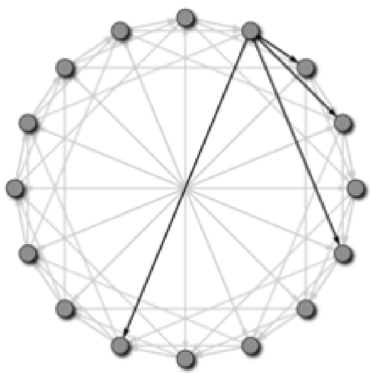

Fig. 6: The Chord Ring with its finger assignment

To ensure a fast, scalable search, Chord implements a method based on finger tables on each node containing up to m entries, where m denotes the number of bits in the hash key. Hereby, the i^{th} entry of node n will contain a pointer to the successor $((n + 2^{i-1}) mod(2^m))$ of n (see Fig. 6). A stabilisation protocol is used to update all finger links and generates additional overhead. [8] showed that a random distribution of the fingers will not significantly influence efficiency.

The overhead may also be generally avoided, if two not yet considered circumstances are used:

1. An unstructured but well-connected P2P-system is running underneath the described ring structure. It makes use of frequent *PING-PONG* messages to keep the network connected.

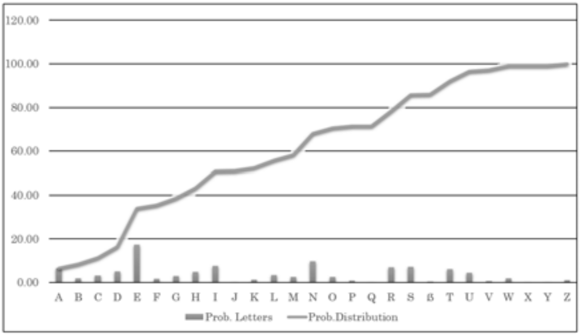

Fig. 7: Probability of letters and the respective distribution

2. The items on the peers along the ring are lexically ordered and the distribution of letters is known (for a large number of entries), see [6] and Fig. 7.

To increase the speed of search in the ring, especially the long-distance fingers are important, i.e. those connecting a peer to other peers having distances of $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$,.. of the ring length (number of peers in the ring) to the origin.

It shall be avoided, that the corresponding end peers of the chords shall be determined by (frequently repeated) 'counting'-procedures along the ring.

Therefore, let us consider the probability of letters as well as the corresponding probability distribution in Fig. 7. Lets assume that l_i, l_j are two letters with $l_j > l_i$. In this case

$$p = F(l_j) - F(l_i)$$

determines the percentage p of words starting with initial letters between l_i and l_j. Thus, finding a peer with an approximate distance of $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$,.. of the total ring length starting from the origin l_i means to divide the set of words in fractions with the size of $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$,.. and then, finally, find the peer having a key starting with the letter l_j such that $p = \frac{1}{2}, \frac{1}{4}, \frac{1}{8}, ..$ and so on. Let us consider for instance the first peer in the ring storing the very first key with the initial letter **A**. From the probability distribution in Fig. 7, it can be derived that 50% of all words will have initial letters between **A** and **I**. Consequently the chord from the first peer to the opposite side of the ring must end (approximately) on a peer having the key **I** on it. If finally the message content (payload) of the *PONG*-reply messages in the underlying P2P-system is extended to $[IP, First\_key\_on\_peer, Last\_key\_on\_peer]$, the (same) peers participating in the ring may obtain the IP-addresses of the needed fingers without any additional overhead by simply listening to the already existing communication.

5 Conclusion and Outlook

A concept of an associative memory with a ring-like structure for P2P-systems has been introduced. The entire memory structure is managed in a completely decentralised manner and can adapt itself to changing needs of the user as well as to a changing system environment. Anyhow, only the basic concept has been developed so far. Thus, many refinements may make the functionality of the proposed structures more efficient and will be elaborated on in future contributions.

References

[1] Unger, H., Böhme, T.: A probabilistic money system for the use in P2P network communities, In: *Proceedings of the Virtual Goods Workshop*, Ilmenau, pp. 60–69, 2003

[2] Lv, Q., Cao, P., Cohen, E., Li, K., Shenker, S.: Search and replication in unstructured P2P networks, In: *Proceedings of the 16th International Conference on Super Computing*, ACM, pp. 84–95, 2002

[3] Castro, M., Costa, M., Rowstron, A.: Peer-to-peer overlays:structured, unstructured, or both, *Technical Report MSR-TR-2004-73*, Microsoft Research, System and Networking Group, Cambridge (UK), 2004

[4] Castro, M.; Druchel, P.; Hu, Y.C.; Rowstron, A.: Topology-aware routing in structured peer-to-peer overlay networks, *Technical Report MSR-TR-2002*, Microsoft Research, 2002

[5] Lämmel, R.: Google's Map Reduce programming model – Revisited, In: *Science of Computer Programming*, Elsevier, Vol.70(1), pp. 1–30, 2008

[6] Meier, H.: Deutsche Sprachstatistik, In: *Olms Paperbacks 31*, 2nd edition, Olms, Hildesheim, 1967

[7] Stoica, I., Morris, R., Karger, D., Kaashoek, M. F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications, In: *ACM SIGCOMM Computer Communication Review*, 31(4):149, 2001

[8] Trompeter, M.: Evaluierung der Auswirkung von gleichverteilten Fingertabellen bei Chord, Masterthesis, University of Hagen, Chair of Communication Networks, Hagen, 2014

[9] Fellbaum, C.: WordNet and wordnets, In: *Brown, Keith et al. (eds.): Encyclopedia of Language and Linguistics*, 2nd edition, Oxford: Elsevier, pp. 665–670, 2005

[10] Kubek, M., Unger, H., Loauschasai, T.: A Quality- and Security-improved Web Search using Local Agents, In: *Intl. Journal of Research in Engineering and Technology (IJRET)*, Vol. 1, No. 6, 2012

[11] Eberhardt, R., Kubek, M., Unger, H.: Why Google Isn't the Future. Really Not., In: *H. Unger and W. Halang: Proceedings der Autonomous Systems 2015*, Fortschritt-Berichte VDI, Series 10: Informatik/Kommunikation, VDI, Düsseldorf, 2015

[12] Kubek, M., Unger, H.: Centroid Terms as Text Representatives, In: *Proceedings of the 2016 ACM Symposium on Document Engineering*, DocEng '16, New York, NY, USA, ACM, pp. 99–102, 2016

[13] Kubek, M., Unger, H.: Centroid Terms and their Use in Natural Language Processing, In: *Autonomous Systems 2016*, Fortschritt-Berichte VDI, Reihe 10 Nr. 848, VDI-Verlag Düsseldorf, pp. 167–185, 2016

[14] Kubek, M., Unger, H.: Towards a Librarian of the Web, In: *Proceedings of the 2nd International Conference on Communication and Information Processing (ICCIP 2016)*, New York, NY, USA, ACM, pp. 70–78, 2016

[15] Kubek, M., Böhme, T., Unger, H.: Empiric Experiments with Text Representing Centroids, In: *Lecture Notes on Information Theory*, Vol. 5, No. 1, pp. 23–28, 2017

[16] Kubek, M., Böhme, T., Unger, H.: Spreading Activation: A Fast Calculation Method for Text Centroids, In: *Proceedings of the 3rd International Conference on Communication and Information Processing (ICCIP 2017)*, New York, NY, USA, ACM, 2017

[17] Kubek, M., Unger, H.: A Concept Supporting Resilient, Fault-tolerant and Decentralised Search, In: *Autonomous Systems 2017*, Fortschritt-Berichte VDI, Reihe 10 Nr. 857, VDI-Verlag Düsseldorf, pp. 20–31, 2017

Dynamic Data Management for an Associative P2P Memory

Supaporn Simcharoen and Herwig Unger

Chair of Communication Networks
FernUniversität in Hagen, Germany

Abstract:
Dynamic hash tables (DHT, [1]) are efficient tools to improve the performance and reduce the overhead of search in peer-to-peer (P2P) systems [2].

However, Chord [3] – the most famous and most used DHT – has a set of significant disadvantages. First, entries are stored on a fixed member peer of the ring determined by the obtained hash value of their keys and also the position of a peer in the ring is fixed by the respective hash value of its IP address. Load imbalances may appear not only if many entries are related to one and the same key or a small set of keys only, but also if a high number of users to access one and the same data items (what often may happen following Zipf's law, indicating that 10% of all content are addressed by 90% of all search queries).

To overcome this situation, a new ring-like associative memory is introduced. Differing to Chord, the size of the ring is solely depending on the number of stored items and will be strictly adapt to it. In addition, the items, forming a linear, lexical ordered ring-list, are not fixed to a determined peer, but maybe shifted along the ring, whereby only their lexical order must be kept.

In the contribution, a fully decentralized token-game is presented, which allows the adaption of the number of peers in the ring without any global instances. Furthermore, a suitable load-balancing basing on the physical analogue of communicating tubes is derived. Finally, first simulation results are used to show that the introduced mechanisms are working well.

References

[1] Castro, M.; Costa, M.; Rowstron, A.: Peer-to-peer overlays: structured, unstructured, or both. Technical Report MSR-TR-2004-73, Microsoft Research, System and Networking Group, Cambridge (UK), (2004)

[2] Lv, Q.; Cao, P.; Cohen, E.; Li, K.; Shenker, S.: Search and replication in unstructured P2P networks. In: Proceedings of the 16th International Conference on Super Computing, pp, 84-95, ACM, (2002)

[3] Stoica, I.; Morris, R.; Karger, D.; Kaashoek, M. F.; Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for internet applications. In: ACM SIGCOMM Computer Communication Review. 31 (4): 149. doi:10.1145/964723.383071, (2001)

Time Series Imputation and Prediction Based on Machine Learning

Phayung Meesad[1] and Kornsirinut Rojanawan[2]

[1]Department of Information Technology Management
[2]Department of Information Technology
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok, Thailand

Abstract: Missing data in time series have the consequent analysis such as time series prediction cannot be done efficiently. This paper proposes a framework for machine learning based model imputation and prediction for time series data. In the proposed framework, first missing data are removed from the time series data and remain only the available data for training machine learning models. The proposed framework takes advantages of machine learning to perform both imputation of missing values and prediction for future values. Firstly, the model is trained and validated using available data in the time series. The trained model is applied to perform missing values imputation as well as prediction of the future values. In the experiment, machine learning models based on K-Nearest Neighbor (KNN), Multilayer Perceptron (MLP), Support Vector Regression (SVR), and Adaptive Neuro Fuzzy Inference System (ANFIS). It is found that SVR is superior on time series imputation and prediction.

1 Introduction

Time series data related to sequence of data with time stamped play very important roles in large domain of applications. The applications of time series data appear in many fields such as science, engineering, business, finance, etc. Scientific applications are scientific workflow systems [14], Scientific Exploitation of Operational Missions [4], and fluid-structure simulation [20]. Examples of engineering applications are road traffic forecasting [12], wind power prediction [15], quality control [3]. Some examples of business and finance applications include economic forecasting, sales forecasting, budgetary analysis, and stock price prediction [5].

One of the challenge in multivariate time series data is the large amount of missing values due to human errors, data transmission errors, machine errors, or IoT sensor faults. To deal with time series data, researchers have proposed methodologies how to impute missing values in time series data. Previous researches focus on solving multivariate missing data problems in time series data using K-Nearest Neighbor [1], ARIMA [17], as well as Dynamic Bayesian Network [18].

Predictability of the future values of time series data is possible. Time series data have seasonal recurrence values. In addition, a time series may have trend, seasonal effect, and remaining variability assumptions: stationarity, uncorrelated random error, and no outliers. Since the time series repeats itself, the future values can be predicted by using the historical values of the same series. Furthermore, the cross-related multivariate time series can be influenced in the prediction of other time series.

In this paper, an time series impuation and prediction framework based on machine learning is proposed herein. The proposed work takes advantages of learning machnisisms of machine Learning methods for both missing values imputation and prediction of the future values. The time series data are rearranged such that they are properly fed into the multivariate imputation models, which is later used for prediction. The machine learning methods used in this work are Multilayer Perceptron (MLP), Adaptive Neuro Fuzzy Inference System (ANFIS), and Support Vector Machine (SVR).

2 Literature Reviews

2.1 Time Series Data

Time series data are composed of a sequence of data points and an equally time stamp. Lets $x(t)$ is a data point at time t. A time series data X may be given by Eqs. (1) and (2). An example of time series data is shown in Fig. 1.

$$S(t) = \ldots + x(t-2) + x(t-1) + x(t) + x(t+1) + x(t+2) + \ldots \tag{1}$$

$$S(t) = \sum_{i=-\infty}^{\infty} x(t+i) \tag{2}$$

where $S(t)$ is an infinite time series, $x(t)$ is a datum at time t, and i is in range $-\infty \ldots \infty$.

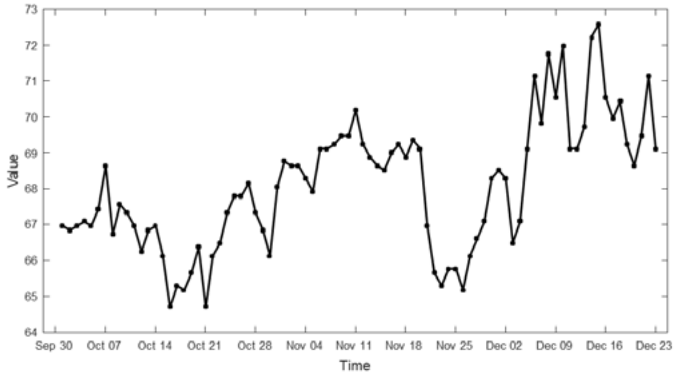

Fig. 1: A Sample of Time Series Data

2.2 Time Series Missing Values

Missing value problem is a burden in machine learning model construction. Missing data may come from many reasons such as human, machine, equipment, or sensor errors. To deal with the problems, researchers may choose to ignore the missing records and use only available data to train the machine learning models. By doing such a way, the number data records available should be enough for training and testing to make sure that the resulting trained models are not under fit and generalized with unseen data. In addition, instead of throw the missing records away, ones can solve the problems the filling in a new value to replace the missing value. The simplest methods are to use basic statistical methods such as min, max, mean, or mode. More advance techniques can be machine learning such as K Nearest Neighbor (KNN) [11], Artificial Neural Network (ANN) [13], Adaptive Neuro Fuzzy Inference System (ANFIS) [5]. Figure 2 displays an example of missing data in time series.

2.3 K-Nearest Neighbor (KNN)

KNN [9] is a machine learning based on supervised learning mechanism. It is a well known lazy learner. It seems to be fast in learning phase because it does not need to learn the training data. Instead, it just keeps all the training in the table. To predict the output KNN takes an input record and compares it with the kept data records in the table. The output can be received an average from those K closest to the input records. Distance metrics such as Euclidean

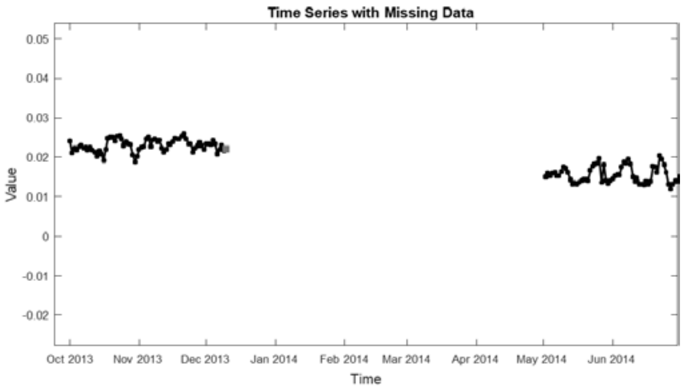

Fig. 2: Missing Data in Time Series

distance, Manhattan distance or any other can be applied for dissimilarity or similarity measurement. KNN algorithm can be shown in Fig. 3.

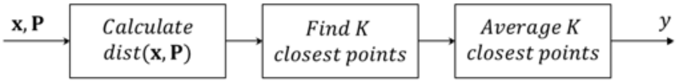

Fig. 3: K-Nearest Neighbor in Systematic Format

2.4 Multilayer Perceptron (MLP)

An artificial neural network is computational units mimicking the human brains, which are able to adapt and learn new information. Each neural unit, imitating human brains, comprises of input signals (dendrites) flowing via weights (synapses) and passing to activation function (axon). MLP is one of many kinds of artificial neural networks consisting of an input layer, one or more hidden layers, and an output layer. It receives input signals at the input layer and feeds forward information through hidden layers until reaching the output layer. The number of hidden layers can be more than one; however, one or two hidden layers are most common seen. MLP is sometimes called Feed-forward Network. In addition, the Backpropagation technique is used to train MLP. It is called backpropagation neural network (BPNN) [8]. Figure 4 illustrates an MLP neural network with two hidden layers.

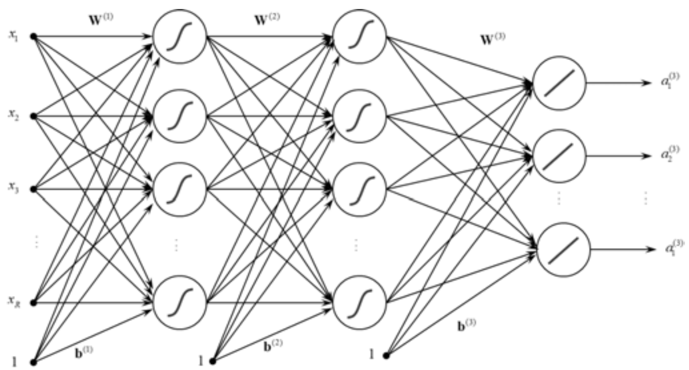

Fig. 4: Multilayer Perceptron Neural Network with two Hidden Layers

2.5 Adaptive Neuro Fuzzy Inference System (ANFIS)

ANFIS [10] is a hybrid combination of artificial neural network and fuzzy inference system. ANFIS has an ability to learn as in artificial neural networks while maintain the concept of fuzzy logic that can handle vague information in fuzzy systems. ANFIS can be read as IF-THEN rule as in fuzzy logic system thus it can be useful to draw reasoning for decision-making about the training data. ANFIS has five layers: input layer, membership function layer, T-norm layer, Normalized layer, and output layer. ANFIS can be trained by either standard backpropagation or hybrid technique. Figure 5 illustrates an example of ANFIS architecture.

2.6 Support Vector Regression (SVR)

Support Vector Machine (SVM) [6] is a famous machine learning for classification with the robustness feature. It can be used for regression. SVM learns to perform classification task by finding the hyperplane maximizing margins and

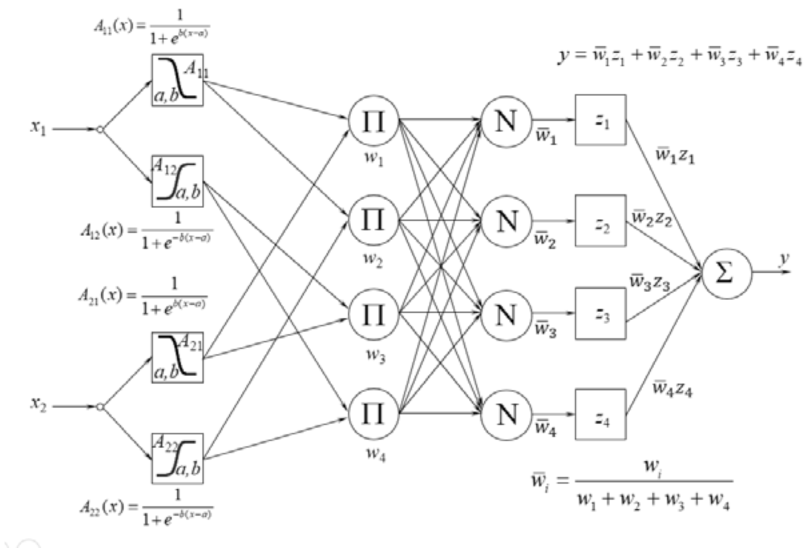

Fig. 5: Adaptive Neuro Fuzzy Inference System

minimizing errors, as well as mapping data to high dimension through kernel functions so the nonlinear decision surface became linear separable. The SVR maintains the same features of SVM; however, instead of generate output as classes, it outputs are real values. SVR is a top candidate of machine learning for time series prediction [7]. SVR have been widely used in time series prediction [16, 19, 21]. Figure 6 shows the architecture of support vector regression.

3 Research Design Framework

Time series imputation and prediction are challenging works due to the dynamics behavior of time series. However, with the assumption that the time series data have trend and occur in seasonal patterns thus it is possible to predict. This paper proposes time series imputation and prediction framework based on machine learning. The machine learning methods in this work include KNN, MLP, ANFIS, and SVR. The proposed time series data imputation and prediction framework is shown in Figs. 7 and 8. Figure 7 is the flowchart of training phase and applying phase of the imputation and prediction models. Figure 8 shows systematic view of the proposed time series, imputation and prediction framework.

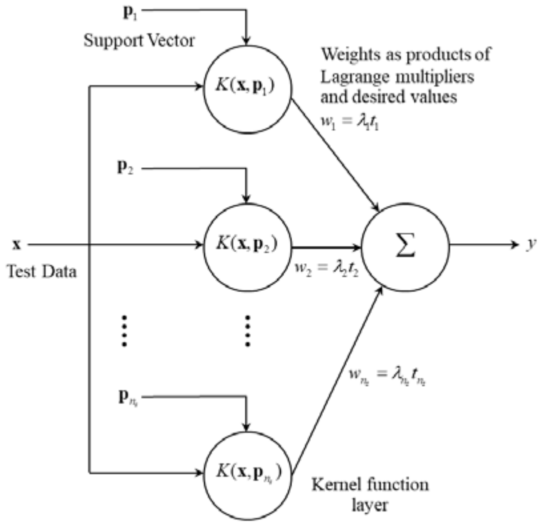

Fig. 6: Support Vector Machine

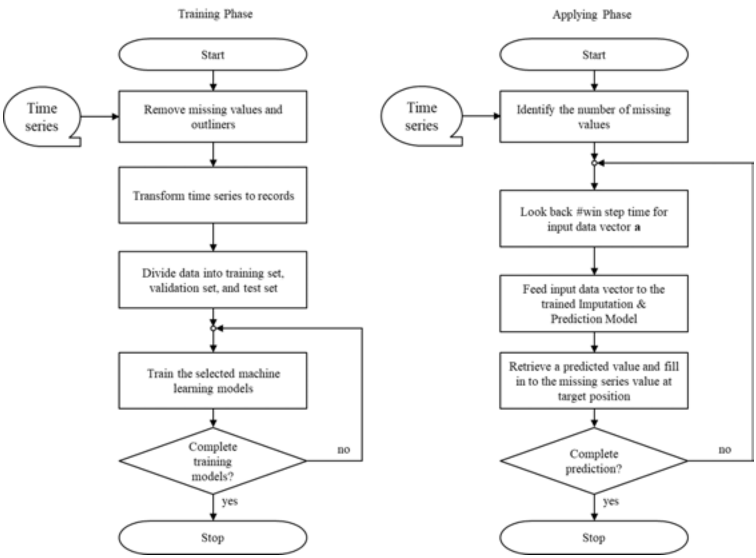

Fig. 7: Flowchart for Imputation and Prediction of Time Series

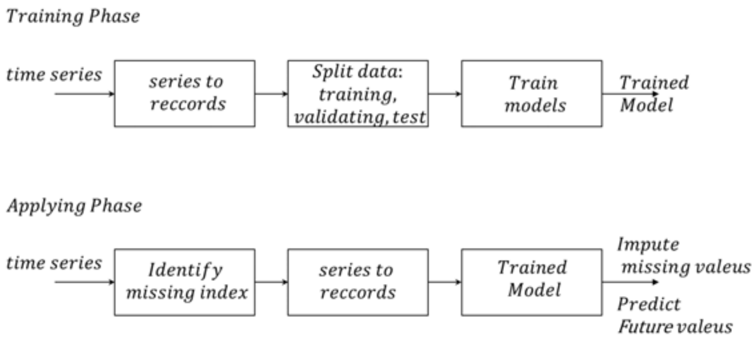

Fig. 8: Research Framework for Training Phase and Applying Phase

3.1 Training Phase

In order to impute the missing values and predict the future values in time series, the imputation and prediction model must be trained from the available data in the time series. The algorithm for training is as follows.

Step 1: Read in time series data for training an imputation and prediction model.
Step 2: Remove the missing values and remain only the available data as training data set. Any outliners must be identified and treated as missing values.
Step 3: Rearrange time series data by transforming from time series to records using a sliding window technique.
Step 4: Divide data into training set, validation set, and test set. The training set is used for model tuning, the validation set is used for validating model while training, and the test set represent the future data values or missing values.
Step 5: Train the selected model based on KNN, MLP, SVR and ANFIS.
Step 6: Retrieve the trained models.

It is worth notice that, usually, MLP, ANFIS, and SVR may have different performance in each run because they randomly choose the initial values and optimized to reach a minimum error. The training process may repeat many times to find the best model of each technique. For MLP the designer must design the number of hidden layers, the number of neurons in each layer, as well as the type of activation functions in each layer. In addition, ANFIS can be trained by specify number of rules needed as well as type of membership functions. In SVR, the designer must specify the number neurons in the hidden layer while the number of time delay step is set to be the same as window length as it uses to transform time series. KNN may get the same result as it keeps all training data as a model without optimizing.

3.2 Applying Phase

After the models are well trained. It is time to apply to impute the missing data or to predict future values in the time series. The procedure for the applying phase can be described as follows.

Step 1: Read in unseen missing time series data or time series to be predicted its future values.
Step 2: Identify the index of missing values.
Step 3: Look back in time with the number win step time from the current position of target that needs to fill in the missing value. Rearrange time series data by transforming from time series to records using a sliding window technique.
Step 4: Feed the unseen data to trained model based on KNN, MLP, ANFIS, and SVR.
Step 6: Retrieve the imputed values or predicted future values.
Step 7: If there are any data remain missing, go to Step 3; otherwise stop.

3.3 Time Series to Records Transformation

Machine learning models such as KNN, MLP, SVR, and ANFIS need to receive data in a record format. Time series data must be transformed to record data sliding window technique. Figure 9 illustrates the sliding window techniques that transform time series to vector matrix form suitable for machine learning. In each step, a sliding window will result in one record of data.

The missing data in time series must be filled with new imputed values. In the training phase, the missing data are ignored or removed out from the training set. Only available data in the series are maintained in the training set validation set, and test set. Training set is used for learning process in which the parameters of the training machine are adjusted to minimize the prediction error, which is the difference between the targets and the real output values. Missing values will be determined after the imputation and prediction model has been completely trained. In addition, future values also can be predicted using the same trained model, if the data points in the series are fed to the model.

Fig. 9: Sliding Window Technique through Time Series Data

Mathematic notations of the proposed framework are shown in Eqs. (3) thru (4).

$$X(t) = \{x(t-R), \ldots, x(t-2), x(t-1), x(t)\} \tag{3}$$

$$y(t) = x(t+1) \tag{4}$$

where $X(t)$ represents a sequence of time series data at time t; $X(t)$ is input vector; $y(t)$ represents an output target of input vector $x(t)$.

Thus, we will have a pair of input and target as $\{x(t), y(t)\}$. If the inputs have additional series, each series will be transformed in the same fashion by adding in the same vector for each sliding window. For example, in case of two time series inputs the transformed vector will be resulted in Eq. (5).

$$X(t) = \{x_1(t-R), \ldots, x_1(t-1), x_1(t), x_2(t-R), \ldots, x_2(t-1), x_2(t), y(t)\} \tag{5}$$

4 Experimental Results

4.1 Experimental Setting Up

To evaluate the performance on imputation efficiency of the proposed framework, we conducted 30 runs for all selected machine learning models. Input sliding window length was set to three days, current date and two previous two consecutive days for the input sequence of each input variable, and target was set to one day predicting in the future. Time series data set used in the experiments of this study was obtained from Pak Panang Royal Irrigation Department. The data used has 12 variables, where some variables were missing due to the persons in charge did not collect the data regularly. Some example plots of missing time series data are displayed in Fig. 10.

For the machine, learning models were set as follows. KNN set the K parameters equal to the window length. MLP has one hidden layer with 30 sigmoidal neurons and one output linear neuron. SVR parameters were $C = 2$, $Epsilon = 0.1$ and kernel function $= Gaussian$. Lastly, ANFIS used Sugeno type with three Gaussian membership functions for fuzzy sets. Fuzzy C-means (FCM) [2] wes employed to perform clustering task and transformed to fuzzy rules. The parameters of the ANFIS were then adapted using hybrid learning method [10].

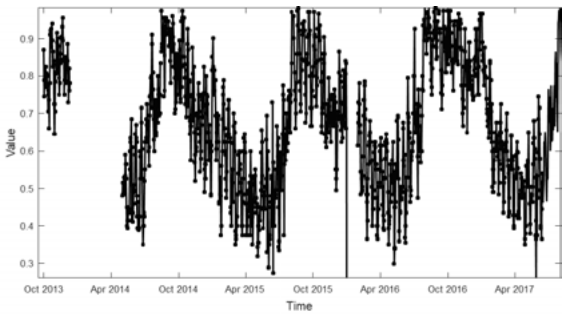

Fig. 10: Samples of Missing Time Series Data used in the Study

4.2 Imputation and Prediction Results

The missing data in the time series must be replaced with new values. In order to accomplish that, first the imputation model must be generated and trained using the available data in the time series. Multiple series can be used as input a select target variable. The process for building imputation models is shown in Fig. 7. Runing 30 experiments for each model, the statistical results based on Root Mean Squared (RMSE) are shown in Table 1. SVR performs best in imputation with smallest RMSE equal 0.061 on average. KNN ranked second on average. MLP and ANFIS are the third and fourth, respectively.

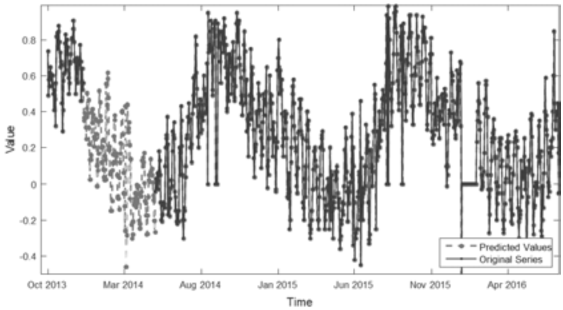

Fig. 11: An Example of Imputation for Missing Data in Time Series

Table 1: Comparisons on Root Mean Squared Errors

RMSE Stat	KNN	MLP	SVR	ANFIS
Minimum	0.072	0.066	0.000	0.100
Maximum	0.299	0.419	0.132	1.033
Average	0.159	0.162	0.061	0.483
Standard Dev.	0.059	0.069	0.039	0.334

After an imputation model is built, it can be applied to impute the missing values. Based on the test data, Figures 11 shows an example of the imputation results for the time series missing data in Fig. 10. The results show that machine learning can fill the new data to the missing points quite reasonable considering that the filled data are distributed nicely along with the available data in the series.

5 Conclusion

Existing techniques of replacing large amount of multivariate missing time series data are not effective enough. This research proposed an imputation for missing values in time series data using machine learning methods. The proposed model can function both imputation and prediction time series data. To build imputation and prediction model, firstly leaving out the missing, the available time series data are used for training, valuating, and testing. The trained model is later used to produce the new replacing values in each time series. In addition, the same model can be used for prediction of future values in the same series in which they will be used to predict unseen and new coming data.

The experiments have shown that the proposed model is well performed both the imputation part and the prediction part measured by the mean squared errors.

References

[1] Bashir, F., Wei, H. L.: Handling Missing Data in Multivariate Time Series Using a Vector Autoregressive Model Based Imputation (VAR-IM) Algorithm.*IEEE Int. Conf. on Computational Science and Engineering (CSE) and IEEE Int. Conf. on Embedded and Ubiquitous Computing (EUC) and 15th Int. Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, 459–463, 2016

[2] Bezdek, J. C., Ehrlich, R., Full, W.: FCM: The fuzzy C-Means Clustering Algorithm. *Science Direct*, 10, 2–3, 191–203, 1984

[3] Brooks, E. B., Wynne, R. H., Thomas, V. A., Blinn, C. E., Coulston, J. W.: On-the-Fly Massively Multitemporal Change Detection Using Statistical Quality Control Charts and Landsat Data. *IEEE Transactions on Geoscience and Remote Sensing*, 52, 6, 3316–3332, 2014

[4] Bruzzone, L., Bovolo, F., Paris, C., Solano-Correa, Y. T., Zanetti, M., FernÃ¡ndez-Prieto, D.: Analysis of multitemporal Sentinel-2 images in the framework of the ESA Scientific Exploitation of Operational Missions. *9th International Workshop on the Analysis of Multitemporal Remote Sensing Images (MultiTemp)*, 1–4, 2017.

[5] Chen, M.-Y.: A hybrid ANFIS model for business failure prediction utilizing particle swarm optimization and subtractive clustering. *Information Sciences*, 220, 180–195, 2013

[6] Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning*, 20, 3, 273–297, 1995

[7] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., Vapnik, V.: Support Vector Regression Machines, 7, 1996

[8] Hagan, M. T., Demuth, H. B., Beale, M. H.: *Neural network design*, 2014

[9] Hand, D. J., Mannila, H., Smyth, P.: *Principles of data mining*, Cambridge, Mass: MIT Press, 2001

[10] Jang, J. S. R.: ANFIS: adaptive-network-based fuzzy inference system, *IEEE Transactions on Systems, Man, and Cybernetics*, 23, 3, 665–685, 1993

[11] Keerin, P., Kurutach, W., Boongoen, T.: Cluster-based KNN missing value imputation for DNA microarray data. *IEEE International Conference on Systems, Man, and Cybernetics (SMC2012)*, 445–450, 2012

[12] Lana, I., Ser, J. D., Velez, M., Vlahogianni, E. I.: Road Traffic Forecasting: Recent Advances and New Challenges, *IEEE Intelligent Transportation Systems Magazine*, 10, 2, 93–109, 2018

[13] Layanun, V., Suksamosorn, S., Songsiri, J.: Missing-data imputation for solar irradiance forecasting in Thailand, *56th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, 1234–1239, 2017

[14] Liu, X., Chen, J., Liu, K., Yang, Y.: Forecasting Duration Intervals of Scientific Workflow Activities Based on Time-Series Patterns, *2008 IEEE Fourth International Conference on eScience*, 23–30, 2008

[15] Mahmoud, T., Dong, Z. Y., Ma, J.: Advanced method for short-term wind power prediction with multiple observation points using extreme learning machines, *The Journal of Engineering*, 1, 29–38, 2018

[16] Niu, L., Wu, J.: Nonlinear L-1 Support Vector Machines for Learning Using Privileged Information, *IEEE 12th International Conference on Data Mining Workshops*, 495–499, 2012

[17] Rizwan, M. O. D., Raj, R. J. R., Vasudev, M.: A novel approach for time series data forecasting based on ARIMA model for marine fishes, *2017 International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies (ICAMMAET)*, 1–4, 2017

[18] Susanti, S. P., Azizah, F. N.: Imputation of missing value using dynamic Bayesian network for multivariate time series data, *2017 International Conference on Data and Software Engineering (ICoDSE)*, 1–5, 2017

[19] Wu, J., Wei, J.: Combining ICA with SVR for prediction of finance time series, *2007 IEEE International Conference on Automation and Logistics*, 95–100, 2007

[20] Zhang, J., Yuasa, S., Fukuma, S., Mori, S. I.: A real-time GPU-based coupled fluid-structure simulation with haptic interaction, *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 1–6, 2016

[21] Zhang, W., Du, Y., Yoshida, T., Wang, Q., Li, X.: SamEn-SVR: using sample entropy and support vector regression for bug number prediction, *IET Software*, 12, 3, 2018

Blind Censoring for Instant Messaging

Günter Fahrnberger

FernUniversität in Hagen, 58084 Hagen, Germany

Abstract: Nowadays, cyberbullying and shady data mining represent two serious concerns for IM (Instant Messaging) users. Blind censoring as an amalgamation of blind computing and censoring appears as the most auspicious solution to get rid of perilous instant messages and eavesdroppers in one fell swoop. The planning of a framework for blind censoring of instant messages needs a detailed threat model at the outset to completely counter them with rigorous approaches. This paper establishes such a threat model based on well-known fictional characters in the style of former scientific and engineering literature about cryptology and IT (Information Technology) security. Thereupon, it merges the NIST framework for improving critical infrastructure cybersecurity and Boyd's OODA (Observe-Orient-Decide-Act) loop into an abstract framework for blind censoring of instant messages. With it, future work can instantiate concrete frameworks for black- and whitelisting of instant messages.

1 Introduction

Separate mitigation strategies against cyberbullying and extra ones against shady data mining characterize most of the approaches implemented in practice.

With regard to cyberbullying, home office task forces for child protection over the Internet were designed to train parents and children on sharing no personal information or passwords, arranging no face to face meetings, and ignoring messages that make them feel threatened or uncomfortable [9, 15, 17]. For instance, in June 2017 Europol launched its *Say NO* campaign that aims to help potential victims [7].

Blocking the causing IM accounts seems to be the most obvious technical approach to inhibit further affronts [9]. Sadly, nothing can prevent the miscreants to launch Sybil attacks, viz. to spawn new accounts, (re)inspire the mobbed sufferers with trust, and continue to afflict them [6].

A text filter that at least inspects all instant messages addressed to cyberbullying-endangered IM receivers shapes up as a more promising resort [9]. Such a sifter either cleans objectionable messages or drops them before their consignees get hold of them. Of course, adult addressees respectively legal guardians of minor recipients must consciously agree to such a screening rather than being forced to it.

If they have decided to employ a text sieve, then it always shall be on duty irrespective of their location and used IM client application. Aside from location-independence, they anticipate a sieve that also includes colloquialisms, vogue terms, and cutting-edge swearwords in its decisions. These two requirements have motivated centralized screening solutions (best directly integrated in IM core platforms) with an updatable vocabulary and proper resilience rather than a trivial client-side approach.

In regard to shady data mining, about 90% of computer security research delved into privacy and integrity problems, about 9% dealt with authenticity issues, and approximately only 1% addressed shortcomings in resilience [1]. Actual strikes and companies' defense expenses tend to be the other way round: more funds flow into the retention of resilience than into the other three security aims together. Both statistics miss proportions of conceptions that embrace all four IT security aims.

For the safe bidirectional transport of instant messages, many applicable cryptosystems have been proposed so far [9]. As earlier mentioned, IM core platforms hosted by public clouds cannot be regarded as trustworthy environments, which makes transport ciphering insufficient and requires further efforts. For this reason, a policy maker for IM must ponder to involve a public cloud in such a way that an interloper cannot figure out which instant messages an IM core platform stores or computes.

Instant messages intended just to be stored and relayed without any reading or writing computations on them easily can be protected by employing one of the many available sorely approved-as-secure end-to-end cryptosystems.

Furthermore, there have been constant inquiries to access and modify encrypted data directly with the so-called blind computing paradigm. Blind computing indicates an application that computes ciphertext data without becoming aware of the meaning of input, output, and intermediate results.

There already exist useful SESs (Searchable Encryption Schemes) in masses for (keyword) searches on encrypted texts. Most of them tackle issues in enciphered databases and in encrypted documents, mainly by converting (key)words into characterizing hash values with the aid of parameterized hash functions. The unavailability of granular ciphertext character string modifications (such as replacing or cutting out substrings) disqualifies them as a prudential filtering technique for IM.

Disguising techniques and homomorphic cryptosystems succeed in blind operations on numerical values, but both fail at calculations on encrypted character strings, viz. as well on ciphertext instant messages. Although character strings can be represented by binary, octal, decimal, or hexadecimal numbers, these representative codings can only be used for identification, because they do not designate any rank or quantity.

The use of trusted third parties, multiple parties, or cryptographic hardware has led to appropriate solutions, but obviously all of them depend on substantial hardware efforts. Reasons, like the reduced or lost flexibility of clouds, higher costs, or the unwillingness of integration by the cloud providers, make the use of additional hardware unattractive, in particular for operators of secure IM services.

On this account, IM urgently requires a novel methodology that attracts both IM providers and IM users to commit themselves to censoring instant messages based on blind computing.

For that reason, section 2 commences the fundamentals to make the further sections graspable. Section 3 specifies the challenge of escaping cyberbullying and shady data mining at once. Section 4 devotes itself to the construction of an abstract framework for blind censoring of instant messages. Section 5 summarizes this scholarly piece and proposes worthwhile future work.

2 Fundamentals

In the style of former scientific and engineering literature about cryptology and IT security, the fictional characters Alice and Bob denominate communication endpoints. Bob designates the endpoint that receives messages, which Alice has originated. Rivest, Shamir, and Adleman have coined these archetypes in 1978 [14]. Commonly, two parties bidirectionally communicate to each other, viz. they switch the roles of Alice and Bob dependent on the direction of the recent message. Imaginative authors augmented Alice and Bob with backstories and personalities. Others added new characters with names and roles, for instance, Bennett, Brassard, and Robert with the passive eavesdropper role Eve [3].

A design model with Alice and Bob merely regards an unidirectional message flow from Alice to Bob. The real world comprises billions of communication-willing people who can alternately slip into the roles of Alice and Bob. While BBSs (Bulletin Board Systems) accessible through dial-up connections were allaying the hunger for personal interaction between a couple of technology freaks in the old days, large globally omnipresent CSPs (Communication Service Providers) have let spring up IM platforms like mushrooms with millions of international ordinary subscribers [8]. Such centralized IM platforms have prevailed over P2P (Peer-to-Peer) communication because of two reasons. Firstly, a central IM platform can more easily retain pending instant messages for offline receivers, viz. it actualizes the so-called store-and-forward functionality. Secondly, the devices of IM users just need to act as clients and initiate outgoing sessions towards their IM platforms rather than keep imperiled source ports open and wait for inbound session requests. Formerly pure IM projects, e.g. ICQ (I seek you), have abated and made room for multimedia-enriched social networks (like Facebook, Twitter, or Google+) where members satisfy their friends by sharing nearly every minute of their lives. Nevertheless, the mere instant messenger WhatsApp has proved the opposite with its prosperity up to now.

Since Alice, Bob, and their CSP(s) reside far away from each other, they depend on technical transmission facilities for message conveyance. Transmission operators, for instance, ISPs (Internet Service Providers) and backbone operators, run facilities and networks that abet the traction of interchanged messages between Alice and Bob.

At this point, Eve and the active eavesdropper Mallory come into play (see figure 1). Mallory tops Eve's passive grabbing behavior by varying or denying valuable data in transmissions between Alice and Bob. The unimpeded propagation of freely available packet sniffers combined with careless or antagonistic transmission providers (which ignore or negate any prophylactic measures or responsive countermeasures, conspire with intelligences, or exploit customer data) permits even a low-skilled Eve equipped with access to transmission elements direct glimpses of transferred plaintext information. A well-versed Mallory conducts more brute intrusions with man-in-the-middle attacks into data streams to rip off authentication flaws in link protocols. In other words, she initially cracks poorly secured IM sessions, studies the spelling style of Alice, and replaces desired messages towards Bob with feigned ones in order to elicit confidential knowledge (such as Bob's watchwords or credit card details) or personal information (like nude pictures of Bob). It goes without saying that the latter in the wrong hands let prosper blackmails with eventual ransom demands.

Fig. 1: Transmission Attack

Latter-day instant messengers thwart such transmission attack vectors by enforcing transport encryption with a cryptographic suite for every session between an Alice and a CSP or a Bob and a CSP, e.g. with the famous TLS (Transport Layer Security) [5]. For instance, in the case of a web-based instant messenger, the waiving of HTTP (HyperText Transfer Protocol) in favor of the TLS-based HTTPS (HyperText Transfer Protocol Secure) drives into transport encryption. Transport encryption means the establishment of a secure connection with a state-of-the-art hybrid cryptosystem. A hybrid cryptosystem combines the pros of an asymmetrical cryptosystem with those of a symmetrical cryptosystem by letting them interact. The asymmetric cryptosystem with different keys for encryption (public key) and decryption (private key) comes into operation for secure session key stipulation between two communicating parties. The symmetric cryptosystem applies the negotiated session key for the safe conveyance of instant messages between them. Transport encryption almost safeguards an instant message on its complete pathway between Alice and Bob, solely interrupted during its detention in a CSP. As long as a CSP respects the privacy of instant messages and takes all feasible IT security precautions for them, transport encryption would suffice.

Regrettably, the reality looks different with CSPs that spurn data protection laws and commit data mining with instant messages to illegally commercialize private data. These CSPs still stick to their proffered IM protocols, but spy out traversed data. Such a conduct lets an ordinary CSP morph into the semi-honest(-but-curious) [4, 13] Wendy, an insider with privileged access capable of divulging information (see figure 2).

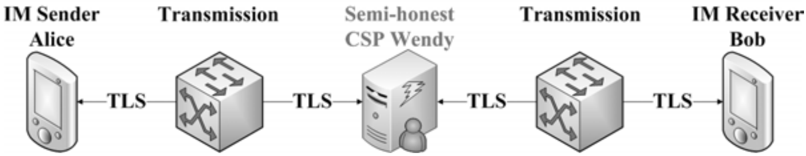

Fig. 2: Shady Data Mining

End-to-end encryption as the most facile remedial measure bridges this gap (like EndSec [11] for mobile telecommunication networks or WhatsApp). It can be contemplated as an extension of transport encryption that takes place between an Alice and a Bob rather than between an Alice and a CSP or between a Bob and a CSP.

Albeit such an end-to-end encryption shields an instant message during its whole course of life, it cannot deter Alice from becoming Chuck [16] and creating maleficent instant messages to bully Bob (see figure 3).

Fig. 3: Cyberbullying

The easiest countermeasure against end-to-end-scrambled instant messages with injurious payload seems to be an application in Bob's receiver device. This application must somehow inspect inbound instant messages after their decipherment and decide about their censoring.

Two spoilsports can upset the plan of Bob's local censoring software, the first of them is even Bob himself (see figure 4). This thinkable case bechances once Bob's custodian has installed and configured a local IM analyzer in Bob's terminal to his chagrin. Bob might fiercely attempt to undermine this app or to elude to IM clients or IM services without censorship due to his tiredness of being patronized [8]. Chuck strengthens the case against local censoring if he mutates into Sybil and launches so-called Sybil attacks, viz. Sybil subverts Bob's analytic software by adopting several identities and insistently harassing Bob [6].

Fig. 4: Sybil Attack and Local Censoring Evasion

The embedment of a centralized IM scanner in a CSP might incapacitate both spoilsports, providing that such a scanner can interpret the meaning of analyzed instant messages. Since the previously imposed end-to-end encryption thwarts that, the reversion to transport encryption seems to be the sole way out. Conformable to figure 2, that would again benefit Wendy with her racketeering, and the troubleshooting goes round in circles.

The successional section precises the challenge of escaping this seeming impasse.

3 Challenge

Section 2 motivates amalgamating end-to-end encryption and CSP-based IM analysis to avert the mischiefs of Bob's, Chuck's, Eve's, Mallory's, Sybil's, and Wendy's wrongdoings.

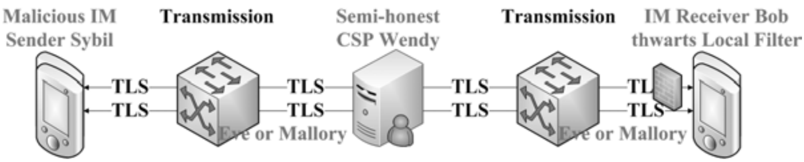

Fig. 5: Threat Model

Alice's and Bob's requirements below can be inferred from the threat model in figure 5.

- Alice and Bob are connected via a transmission network that they deem insecure.

- Bob and/or his legal guardians want him to merely receive instant messages with inoffensive content from Alice.
- Objectionable instant message parts must be exchanged for placeholders (e.g. asterisks) or eliminated before Bob can receive them.
- Alice and Bob cannot always assure to clean these instant messages locally, because they may use terminals (such as workstations in public Internet cafés) that disallow downloading and running a fitting sifter applet with an up-to-date repository of expressions supposed to be erased or replaced.
- Bob does not want Alice to be responsible for screening her sent instant messages.
- Alice and Bob by default mistrust an external entity (for example, located in a public cloud) that would sieve their interchanged plaintext instant messages.

Computations must not work on ciphertext that has been obfuscated with original end-to-end encryption schemes. Else, their existence would be reduced to absurdity, because they solely exist to assure at least the integrity and privacy of transported data. Canny researchers adapted existent cryptosystems or devised novel ones that permit a limited set of meaningful computations on ciphertext. The literature embraces such extraordinary cryptosystems by the term *blind computing*. *Blind censoring* as an enhancement of a blind computing scheme with an IM sieve or topic recognition technique relying on black- or whitelisting acts as an appropriate remedy against Wendy's semilegal profiteering.

4 Methodological Basis

The design science in information systems research of Hevner and Chatterjee abstractly advises to educe and evaluate a design artifact for a relevant problem with rigorous methods [10]. This abstraction necessitates to adopt well-proven IT security concepts, or adapt them to ultimately finalize an adequate framework design.

The final framework design in subsection 4.3 fuses two such IT security concepts. The first aids to address all subtleties of the threat model in section 3. The second specifies the functioning of a system model that resists all these identified threats.

4.1 Framework for Improving Critical Infrastructure Cybersecurity

The NIST (National Institute of Standards and Technology) published an excellent framework as guidance for organizations that must manage IT security risks for critical infrastructure [12]. It owes its existence to the former US (United States) president Barack Obama's EO (Executive Order) 13636. This EO calls for the development of a voluntary framework to manage IT security risks for those processes, information, and systems directly involved in the delivery of critical infrastructure services. The EO defines critical infrastructure as those physical and virtual systems and assets whose incapacity or destruction would have a debilitating impact on security, economic security, public health, safety, or any combination of those matters. The affiliation of an instant messenger to critical infrastructure appears to be debatable, but the guidance of the NIST can also be airily exerted as a methodological template for the framework design of non- or semi-critical infrastructure. That renders any discussion about the criticality of an instant messenger superfluous in this context.

The NIST shaped their framework technologically neutral, not industry-specific, and not country-specific to supply a maximal number of public and private organizations with a common taxonomy and mechanism to

- describe their current IT security posture,
- describe their target state for IT security,
- identify and prioritize opportunities for improvement within the context of a continuous and repeatable process,
- assess progress toward the target state,
- and communicate among internal and external stakeholders about IT security risk.

The framework either complements the existent IT security program of an organization or serves as a reference to establish a new one. It enfolds the three parts *framework core*, *framework profile*, and *framework implementation tiers*.

- Figure 6 presents the **framework core** as a four-stage hierarchy of cybersecurity activities (functions and categories), desired outcomes (subcategories), and applicable references (informative references). Each element on each hierarchic level can be selectively incorporated or ignored for further contemplations, i.e. the core does not correspond with a checklist of mandatory actions.

- **Functions** spearhead the hierarchy and align with existent methodologies for incident management. They ought to operate concurrently and continuously rather than form a serial path or lead to a static desired end state. The NIST distinguishes the undermentioned five functions.

  * **Identify** works the threats for organizational systems, assets, data, and capabilities out.

  * **Protect** develops and deploys adequate safeguards against the identified threats.

  * **Detect** includes the development and deployment of proper activities for the detection of cybersecurity events and incidents.

  * **Respond** revolves around pondering and implementation of suitable actions against detected cybersecurity events and incidents.

  * **Recover** encompasses the elaboration and commissioning of maintenance plans for resilience and of restoration processes for services, which were impaired due to cybersecurity events or incidents.

- **Categories** as the subdivisions of functions concern programmatic needs and particular activities. The NIST suggested five categories for *Identify*, six for *Protect*, three for *Detect*, five for *Respond*, and three for *Recover*, i.e. 22 categories in all.

- **Subcategories** further subdivide categories into specific outcomes of technical and/or management activities. In total, the NIST exemplarily recommended 98 subcategories, 24 of them assigned to the categories of *Identify*, 35 of them assigned to the categories of *Protect*, 18 of them assigned to the categories of *Detect*, 15 of them assigned to the categories of *Respond*, and six of them assigned to the categories of *Recover*.

- **Informative References** of a subcategory gather standards, guidelines, and practices that suit as solvers for its outcomes.

• The **framework profile** refers to the selection and treatment of fitting functions, categories, and subcategories. The current profile indicates the presently achieved condition, while the target profile does the same for the wanted situation. Gaps between the current profile and the target profile necessitate an action plan to address them. For an up-to-date action plan, the NIST suggests to carry out the seven steps iteratively as follows.

1. Prioritization and scope determination
2. Orientation
3. Creation of current profile
4. Risk assessment
5. Creation of target profile
6. Determination, analysis, and prioritization of gaps
7. Implementation of action plan

- **Framework implementation tiers** are different classes for the characterization of the degree of rigor and sophistication in cybersecurity risk management practices of an organization. *Partial (Tier 1)* reaches the lowest degree, surpassed by *Risk Informed (Tier 2)*, *Repeatable (Tier 3)*, and, ultimately, *Adaptive (Tier 4)* with the highest degree. Tiers do not amount to maturity levels. The progression to a higher tier merely makes sense if it cost-effectively weakens IT security threats. The successful realization of the framework depends on the action plan rather than on the tier determination.

Fig. 6: Framework Core Hierarchy

4.2 OODA Loop

US Air Force Colonel John Boyd has highly influenced the disciplines science, military, sports, business, and litigation with his theories. He authored an essay and six successive presentations between 1976 and 1996, which revolve around the so-called OODA loop, a scalable system and process model [2]. Figure 7 delineates a simplified OODA loop with the four activities *observation*, *orientation*, *decision*, and *action*.

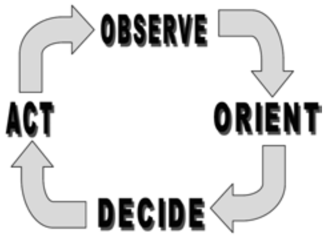

Fig. 7: Simplified OODA Loop

- **Observation** gathers sensory inputs from an observer's vicinity.
- **Orientation** makes sense of the observational input in light of existent knowledge.
- **Decision** picks the optimum of all available options.
- **Action** puts the decision into practice.

Figure 8 gives particulars to this rudimentary model to respect complex interdependent interactions that ordinarily occur within an OODA loop. The penultimate subsubsection in subsection 4.3 dwells on the OODA loop details with a view to a system model for blind censoring of instant messages.

Boyd constructed 16 different literary domains that rely on his OODA loop. US Air Force Colonel Scott Angerman appended eleven new items to this list during his postgraduate studies as Captain at the Air Force Institute of Technology.

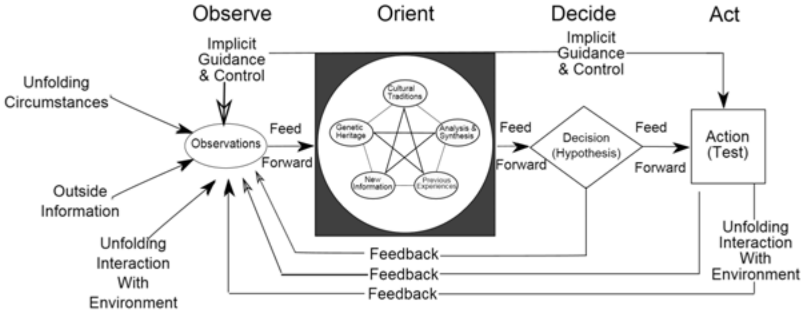

Fig. 8: OODA Loop

Three of Boyd's 16 themes (*competition*, *decision-making*, and *information processing*) concern duties that a system model for blind censoring of instant messages must render. *Competition* sounds far-fetched and devious, but Boyd's intention of gaining competitive advantages against adversaries with the aid of the OODA loop renders this topic plausible for a system in the focus of offenders.

Angerman identified, analyzed, coded, and categorized 224 OODA-related literature documents authored between 1992 and 2003. *Competition* emerged as front-runner in 165 of them, *decision-making* ranked third in 125 writs, and *information processing* lay in fifth place in 114 writings. Consequentially, Angerman ascertained an increasing use of the OODA loop in technical implementations. They manifested themselves in three ilks as hereafter given.

- **Physical computer system architectures** to either be defended or attacked from cyberspace
- Information cycle times or **information flows**
- Intelligent behavior in **computer systems**

All three can be seen as an intersection between *information*, *system*, and *process*, i.e. an OODA loop emerges where a system processes information.

- **Information** as input for orientation allegorizes the fuel for the OODA loop.
- Boyd looked upon a computer as closed **system**. The OODA loop palpably meets all of the descriptive characteristics of a generic **system**, for example,

inputs as observations into the OODA loop or outputs as actions stemming from the OODA loop.

- Every OODA loop constituent (observation, orientation, decision, and action) and also the OODA loop itself can be viewed as a separate **process**. The OODA loop processing permits a system to make use of information to collaborate with its proximity.

4.3 Framework Design

This treatise plainly pursues the recommended seven-step tutorial of the NIST for establishing or improving a cybersecurity program to have an abstract framework available for substantiating black- and whitelisting in future.

Prioritization and Scope Determination

The initiatory step requires determinating the priority and scope of mission-critical systems and assets. A holistic framework must ascribe worthiness of protection to all objects that encounter instant messages. Derived from figure 1, this implicates Alice, Bob, the CSP, and all interjacent transmission facilities. The more IM users a unit affects, the higher its precedency ranks. In the present case, the CSP heads the order of precedence in front of transmission, Bob, and Alice.

Orientation

The second step requisitions an overall risk estimation by identifying threats to and vulnerabilities of the previously selected systems and assets. Fahrnberger reports on the two chief vulnerabilities [9], section 2 uncovers further threats, and figure 5 sketches a top-level threat model.

Creation of Current Profile

The third step aims at assessing the currently fulfilled categories and subcategories of the framework core. Evidently, pedantic scrutinizing for appropriateness and fruition of all 98 subcategories would be the preferred approach for an authentic CSP in operation. In the hypothetical case of this disquisition, such pedantry appears tedious beyond doubt. On that account, the current profile ought to reflect an adaptive implementation tier with an instant messenger yet endangered by Bob, Chuck, Eve, Mallory, Sybil, and Wendy (see section 2). This results in the achievement of all subcategories with the exceptions as noted in the second right column in table 1. The seventh unmet subcategory has appositely been renamed *The network is monitored to detect potential cybersecurity events* into *The network is monitored to detect cyberbullying events*.

Table 1: Deficiencies of Current Profile

Function	Category	Subcategory	Risk
PROTECT	Access Control	Identities and credentials are managed for authorized devices and users	Authenticity
PROTECT	Data Security	Data-at-rest is protected	Privacy
PROTECT	Data Security	Data-at-transit is protected	Privacy
PROTECT	Data Security	Adequate capacity to ensure availability is maintained	Resilience
PROTECT	Data Security	Protection against data leaks are implemented	Privacy
PROTECT	Data Security	Integrity checking mechanisms are used to verify software, firmware, and information integrity	Integrity
DETECT	Security Continuous Monitoring	The network is monitored to detect cyberbullying events	Cyberbullying
DETECT	Detection Processes	Event detection information is communicated to appropriate parties	Cyberbullying
RESPOND	Mitigation	Incidents are mitigated	Cyberbullying
RESPOND	Mitigation	Newly identified vulnerabilities are mitigated or documented as accepted risks	Cyberbullying

Risk Assessment

The fourth step rounds off table 1 by discerning the impact of violated subcategories. For this purpose, the right column juxtaposes the IT security goal at risk if the subcategory in the second right column remains unfulfilled. The terms *cybersecurity events*, *incidents*, and *vulnerabilities* in the last four table rows allude to cyberbullying situations in this context.

Creation of Target Profile

The fifth step pays attention to determining the targeted subcategories that ought to be realized in the future. Since the current profile as the upshot of the third step already fulfills 88 subcategories by definition, it stands to reason that the target profile thoroughly consists of all 98.

Determination, Analysis, and Prioritization of Gaps

Understandably, the gap between the current and the target profile amounts to the ten unrealized subcategories in table 1. The major task of this sixth step attends to the compilation of a (prioritized) action plan against this unveiled gap. For that purpose, Boyd's OODA loop comes into play. The boxes and arrows of the flowchart in figure 8 must be furnished with wise actions and information to overcome all gap minutiae. The itemization below masters this *furnishing* from *observation* on the left to *action* on the right.

- **Observation:** A CSP has to process credentials for access control and instant messages as the two cardinal sorts of observed outside information. Both must be guarded at all costs from data leaks during transit between Alice and Bob and during rest to maintain their privacy. The conveyance of passwords as the sensitive part of credentials can airily be handled with any approved-as-secure hybrid cryptosystem. Just as well, saved representative (parametrized) hash values instead of plaintext passwords acceptably provide ample integrity and privacy for passwords at rest. The handling of instant messages proves to be more difficult. Mere hybrid cryptography would suffice for the integrity and privacy of harmless instant messages during transit between Alice and Bob and at rest. Collaterally, closing the gap for the monitoring, discovery, and mitigation of cyberbullying during the orientation and decision phase sanctifies the higher efforts of blind censoring as a merger of blind computing and censoring.

- **Orientation:** While the box *New Information* bears on the sensed instant messages during the observation phase, the box *Previous Experiences* references a blacklist with forbidden character strings or a whitelist with explicitly allowed words. Both boxes concur as input factors for *Analysis & Synthesis*. The latter (blindly) tests an instant message for the inclusion of black- or whitelist patterns.

- **Decision:** One or more included instant message words in the blacklist indicate the sleaze of the message. The same applies to an instant message that is composed of at least one non-whitelist word. On the contrary, instant messages can only be presumed to be offenseless if they do not embody blacklist items, or if they solely include whitelist elements.

- **Action:** Blind censoring unambiguously relays disagreeable and agreeable instant messages. While the latter merit their delivery to Bob without ifs and buts, the CSP must purge junk messages. It either simply obliterates such an instant message or it cuts off all its seedy ingredients prior to its relaying to Bob. Obligatory, lawful data preservation can compel the CSP to retain instant messages and logs of taken actions for a predetermined quarantine period. Optionally, the CSP informs Alice about the annihilated instant message respectively submits the trimmed instant message to Alice to obtain her affirmation. Further on, to boost the black- and whitelist accurateness, Bob can voluntarily report false positives (received undesired contents) and false negatives (missing hazard-free instant messages) to the CSP in an optional training mode.

Figure 9 expands the threat model in figure 5 by assigning a sphere to each OODA loop activity. While orientation and decision-making just take place in the CSP, observation stretches from Alice to the CSP, and action spans from the CSP to Bob. The graphical nesting of orientation in decision and decision in action does not implicate any hierarchy among them.

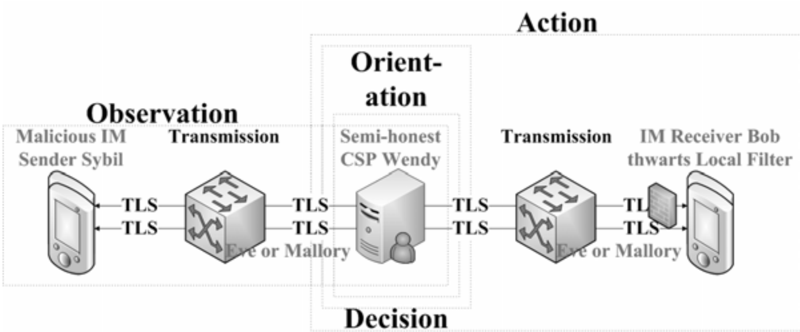

Fig. 9: Action Plan

Implementation of Action Plan

Future work has to implement the elaborated action plan, i.e. concretize the abstract framework of the sixth step to serviceable expedients for blind censoring of instant messages.

5 Conclusion

The approach presented in this paper introduces an abstract framework for blind censoring of instant messages that incorporates all demands of Alice and Bob in section 3. Upon the recommendation of the design science in information systems research of Hevner and Chatterjee [10], section 4 distills such a demanded framework by uniting two existing, well-tried IT security concepts. The NIST penned one of them for ameliorating critical infrastructure cybersecurity [12]. It consummately lends itself to the seeking of sustainable counteractions against the manifest perils in figure 5. An action plan has to organize all descried counteractive measures. Boyd's OODA loop as the second contributive concept leaves its mark by offering a crafty template for the activities of this action plan [2]. Future work can easily take advantage of this template with the construction of concrete action plans for black- or whitelisting of instant messages.

Acknowledgments

Many thanks to Bettina Baumgartner from the University of Vienna for proofreading this paper.

References

[1] R. J. Anderson. *Security engineering - a guide to building dependable distributed systems (2. ed.)*. Wiley, jan 2008.

[2] W. Angerman. Coming full circle with boyd's ooda loop ideas: An analysis of innovation diffusion and evolution. Master's thesis, Defense Technical Information Center, mar 2004.

[3] C. H. Bennett, G. Brassard, and J.-M. Robert. Privacy amplification by public discussion. *SIAM Journal on Computing*, 17(2):210–229, apr 1988.

[4] Q. Chai and G. Gong. Verifiable symmetric searchable encryption for semi-honest-but-curious cloud servers. In *Communications (ICC), 2012 IEEE International Conference on*, pages 917–922, jun 2012.

[5] T. Dierks and E. Rescorla. The transport layer security (TLS) protocol version 1.2. RFC 5246 (Proposed Standard), aug 2008.

[6] J. R. Douceur. The sybil attack. In *Proceedings of the IPTPS Workshop*, Cambridge, MA, USA, 2002.

[7] Europol. Internet organised crime threat assessment (IOCTA) 2017.

[8] G. Fahrnberger. SIMS: A comprehensive approach for a secure instant messaging sifter. In *Trust, Security and Privacy in Computing and Communications (TrustCom), IEEE 13th International Conference on*, pages 164–173, sep 2014.

[9] G. Fahrnberger. Secure filtering techniques for instant messaging. In H. Unger and W. A. Halang, editors, *Autonomous Systems 2017*, volume 857 of *Fortschritt-Berichte Reihe 10*, pages 226–240. VDI Düsseldorf, oct 2017.

[10] A. Hevner and S. Chatterjee. *Design Science Research in Information Systems*, volume 22, pages 9–22. Springer US, Boston, MA, mar 2010.

[11] K. Kotapati, P. Liu, and T. F. LaPorta. EndSec: An end-to-end message security protocol for mobile telecommunication networks. In *World of Wireless, Mobile and Multimedia Networks, 2008. WoWMoM 2008. 2008 International Symposium on a*, pages 1–7, jun 2008.

[12] National Institue of Standards and Technology (NIST). Framework for improving critical infrastructure cybersecurity. National Institue of Standards and Technology (NIST), feb 2014.

[13] C. Örencik and E. Savas. An efficient privacy-preserving multi-keyword search over encrypted cloud data with ranking. *Distributed and Parallel Databases*, 32(1):119–160, mar 2014.

[14] R. L. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Commun. ACM*, 21(2):120–126, feb 1978.

[15] U.S. Department of Education. Parents guide to the internet, nov 1997.

[16] M. van Steen and A. S. Tanenbaum. *Distributed Systems*. CreateSpace Independent Publishing Platform, feb 2017.

[17] J. Williams. Promoting internet safety through public awareness campaigns - guidance for using real life examples involving children or young people, nov 2005.

Distributions of Points

Hanno Lefmann

Fakultät für Informatik, TU Chemnitz, Germany

Abstract: Distributions of n points in the unit-interval $[0, 1]$ or unit-square $[0, 1]^2$ are considered, where the minimum mutual distance between two points in $[0, 1]$ or minimum area of a triangle formed by three points in $[0, 1]^2$ is as large as possible, respectively.

1 Introduction and Distances in the Unit-Interval $[0, 1]$

Given a positive integer n, n points should be positioned in the unit-interval $[0, 1]$, such that the minimum Euclidean distance between two points is as large as possible. This is the *offline-situation*, as n is known in advance. Let $d(n)$ be the maximum possible value for the distance that one can achieve. This is easy. If we distribute the n points equidistant in the unit-interval $[0, 1]$, i.e., at positions $0/(n-1); 1/(n-1); \dots (n-1)/(n-1)$, we get $d(n) \geq 1/(n-1)$.

Given n arbitrary points in $[0, 1]$, this gives $(n-1)$ pairwise disjoint intervals, whose sum of the lengths is at most 1. Therefore, there exists at least one interval of length at most $1/(n-1)$, and we get $d(n) \leq 1/(n-1)$. Thus the maximum value is

$$d(n) = \frac{1}{n-1}.$$

In this example, the number n of points was given in advance. In the *online-situation*, the number of points to be placed in the unit-interval $[0, 1]$ is not known in advance. Points have to be placed one by one, and these placements cannot be changed later anymore. The process of placing points can terminate any time. For an online-algorithm A for this problem, its *c*ompetivity $c_A(n)$ (or quality), when n points in the unit-interval have been distributed, is defined as the quotient of the minimum distance between two points in the off-line case and the by algorithm A achieved minimum distance.

The following algorithm PLA in principle places an arbitrary number of points $P_1, P_2, \dots$ in the unit-interval.

Starting with the first point P_1, the points $P_1, P_2, P_3, \ldots$ will be placed at positions $0; 1; 1/2, 1/4; 3/4; \ldots$. In general, for $n = 2^\ell + i$ with $2 \leq i \leq 2^\ell + 1$, $\ell \geq 0$, the point P_n will be placed at position $(1 + 2 \cdot (i-2))/2^{\ell+1}$. The obtained minimum distance $d_{PLA}(n)$ between two points among the first n points is equal to $1/2^{\ell+1}$. The optimal solution in the offline-case has value equal to $1/(n-1)$, as seen above, and we get for the competivity $c_{PLA}(n)$ of algorithm PLA for each $n = 2^\ell + i, 2 \leq i \leq 2^\ell + 1$:

$$c_{PLA}(n) = \frac{\frac{1}{n-1}}{\frac{1}{2^{\ell+1}}} = \frac{\frac{1}{2^\ell+i-1}}{\frac{1}{2^{\ell+1}}} = \frac{2^{\ell+1}}{2^\ell + i - 1} \leq \frac{2^{\ell+1}}{2^\ell + 1} = 2 - \frac{2}{2^\ell + 1} < 2 - \frac{2}{n} < 2.$$

Thus, the algorithm PLA achieves in the online-case a solution, which differs less than a factor of 2 from the optimal solution in the off-line case.

We mention here another approach for the offline-situation using the independence number of graphs. Divide $[0, 1]$ into $(K-1)$ intervals of equal size, thus discretizing, where $K \geq n$ is sufficiently large. The endpoints of these small intervals are possible positions for the n points. The distance of two points is 1 plus the number of endpoints of the small intervals strictly between them. We identify the K equidistant points $P_1, \ldots, P_K$ in the unit-interval $[0, 1]$ with the vertices of a graph $G = (V, E)$. Two vertices P_i and P_j are joined by an edge in G if and only if their distance is less than B. Then the maximum degree of a vertex in G is at most $(B-1)$. (Here we assume for simplicity that B is an integer.) We recall that a subset of the vertex set of a graph is called *independent* if it does not contain any edges. An independent set in the graph G yields a set points in the unit-interval $[0, 1]$ of the same size with minimum distance at least B between two of the selected points. It is known (Greedy algorithm) that the independence number $\alpha(G)$, which is the largest size of an independent set in G, satisfies

$$\alpha(G) \geq \frac{1}{2}\frac{K}{B}.$$

Now

$$\frac{K}{2B} \geq n$$

for $B \leq K/(2n)$. After rescaling, this approach yields $d(n) \geq 1/(2n)$, which is a factor of approximately 2 worse than in the optimal solution but still we have the correct order.

2 Average Minimum Distance

Next we consider the average value of the minimum distance among all distributions of n points in the unit-interval $[0, 1]$.

Theorem 1 *If n points are distributed at random in the unit-interval $[0, 1]$, then on average the minimum distance between two of the n points is $\Theta(1/n^2)$.*

Proof. Divide $[0, 1]$ into $(K - 1)$ intervals of equal size, thus discretizing, where K is sufficiently large. The endpoints of these small intervals are possible positions for the n points. The distance of two points is 1 plus the number of endpoints of the small intervals strictly between them.

First we show a lower bound. There are $\binom{K}{n}$ distributions of n points at all. Pick $n - 1$ points in $\binom{K}{n-1}$ ways. Choose one of these points, say P, in less than n ways. Another point at distance at most B from P can be chosen in at most $2B$ ways, thus there are less than

$$\binom{K}{n-1} \cdot n \cdot 2B$$

such configurations. Then, it is

$$\binom{K}{n-1} \cdot n \cdot 2B \leq \frac{1}{2} \cdot \binom{K}{n}$$

for $B \leq (K - n + 1)/(4n^2)$.

For K sufficiently large, we have

$$\frac{K - n + 1}{4n^2} \geq \frac{K}{5n^2}.$$

Thus for at least $(1/2) \cdot \binom{K}{n}$ configurations the minimum distance is at least $K/(5n^2)$, hence on average the minimum distance is at least $K/(10n^2)$.

Concerning an upper bound, we first determine an upper bound on the probability $Prob(B)$ that among a distribution of n points the minimum distance between two points among the n points is exactly B. The first point P_1 can be chosen in K ways. A second point P_2 with distance equal to B from P_1 can be chosen in at most 2 ways. Having chosen such points $P_1, \ldots, P_j$, $j \geq 2$, a point P_{j+1} with distance at least B to any of these points can be chosen in at most

$(K - jB/2)$ ways, as in such configurations any chosen point forbids at least $B/2$ other points. (We assume here for simplicity that B is not an integer, otherwise take the term $(K - j(B/2 - \varepsilon))$ for a small $\varepsilon > 0$.)

Taking care of ordering, i.e., of the choice that points P_1 and P_2 have distance exactly B by a factor of n^2, with $1 + x \leq e^x$, the probability $Prob(B)$ is at most

$$Prob(B) \leq \frac{2Kn^2 \cdot \prod_{j=2}^{n-1}(K - jB/2)}{K^n} \leq \frac{2n^2}{K} \cdot e^{-\frac{\binom{n-2}{2}B}{2K}}$$

Therefore, the expected value E_n of the minimum distance between two of n random points is at most

$$E_n = \sum_{B=1}^{K/n} Prob(B) \cdot B \leq \sum_{B=0}^{K/n} B \cdot \frac{2n^2}{K} \cdot e^{-\frac{\binom{n-2}{2}B}{2K}} = \frac{2Bn^2}{K} \cdot e^{-\frac{\binom{n-2}{2}B}{2K}}.$$

To estimate this sum, we upper bound it by

$$\frac{2n^2}{K} \int_0^{K/n} B \cdot e^{-\frac{\binom{n-2}{2}B}{2K}} dB.$$

Set $\alpha = \binom{n-2}{2}/(2K)$, $a = 0$ and $b = K/n$. We determine an integral

$$\int_a^b x e^{-\alpha x} dx,$$

using

$$\int_a^b u'v dx = [uv]_a^b - \int_a^b uv' dx.$$

Then,

$$\int_a^b x e^{-\alpha x} dx = -\frac{b}{\alpha} e^{-\alpha b} + \frac{a}{\alpha} e^{-\alpha a} - \frac{b}{\alpha^2} e^{-\alpha b} + \frac{1}{\alpha^2} e^{-\alpha a}.$$

Inserting $\alpha = \binom{n-2}{2}/(2K)$, $a = 0$ and $b = K/n$, this gives

$$\frac{2K^2}{n\binom{n-2}{2}} e^{-\frac{\binom{n-2}{2}}{n}} - \frac{4K^2}{n\binom{n-2}{2}^2} e^{-\frac{\binom{n-2}{2}}{n}} + \frac{4K^2}{\binom{n-2}{2}^2}.$$

Thus, for K sufficiently large, we infer

$$E_n \leq \frac{2n^2}{K} \cdot \left(\frac{2K^2}{n\binom{n-2}{2}} e^{-\frac{\binom{n-2}{2}}{n}} - \frac{4K^2}{n\binom{n-2}{2}^2} e^{-\frac{\binom{n-2}{2}}{n}} + \frac{4K^2}{\binom{n-2}{2}^2} \right) \leq \frac{33K}{n^2}.$$

Rescaling by the factor K yields the desired result. •

3 Triangles in the Unit-Square $[0,1]^2$

Now we sketch the situation, where we want that the minimum area of a triangle formed by three points of a configuration of n points in the unit-square $[0,1]^2$ is as large as possible: Heilbronn's triangle problem. Similar techniques, but more sophisticated ones, as above for considering distances in the unit-interval, were successfully applied here.

In the offline-case, where the number n of points to be distributed is known in advance, put for a prime $p \geq n$, a $p \times p$-grid on $[0,1]^2$ and consider the points $(1/p)(j \bmod p, j^2 \bmod p)$, $j = 0, \ldots, p-1$. No three of these points can be collinear (which would give a triangle of area equal to 0), so the minimum area formed by three of these points is at least $1/(2p^2)$. As the set of primes is sufficiently dense in the set of positive integers, for every n we have a configuration of n points in $[0,1]^2$ with minimum area formed by three of these points of order $\Omega(1/n^2)$. This construction is due to Erdős.

The best known lower bound so far available for the offline-situation, which can be achieved also deterministically in polynomial time [2], is of the order $\Omega(\log n/n^2)$ [6]. An upper bound of $O(1/n^{8/7-\varepsilon})$, where $\varepsilon > 0$ is small, is only known here [5].

For distributions of n points in $[0,1]^2$ on the average the smallest area of a triangle is of the order $\Theta(1/n^3)$ and was proved using Kolmogorov complexity [4], compare [3].

Finally, in the online-situation one can find configurations of points, where for every positive integer n every triangle among the first n points has area at least $\Omega(1/n^2)$ [1]. Comparing this with the best known upper bound $O(1/n^{8/7-\varepsilon})$ for the offline-case we can currently upper bound the competivity only by $O(n^{6/7+\varepsilon})$.

More details will be given in the talk.

References

[1] Barequet, G. and Shaikhet, A.: The on-line Heilbronn's triangle problem in d dimensions, *Discrete & Computational Geometry* 38, pp. 51–60, 2007

[2] Bertram-Kretzberg, C., Hofmeister, T. and Lefmann, H.: An algorithm for Heilbronn's problem, *SIAM Journal on Computing* 30, pp. 383–390, 2000

[3] Blum, W., Geometrisch Eingekreist – Forscher sind dem Zufall auf die Schliche gekommen, *Die Zeit*, 13.04.2000, `www.zeit.de/2000/16/200016.zufall_.xml`

[4] Jiang, T., Li, M. and Vitany, P.: The average case area of Heilbronn-type triangles, *Random Structures & Algorithms* 20, pp. 206–219, 2002

[5] Komlós, J., Pintz, J. and Szemerédi, E.: On Heilbronn's triangle problem, *Journal of the London Mathematical Society* 24, pp. 385–396, 1981

[6] Komlós, J., Pintz, J. and Szemerédi, E.: A lower bound for Heilbronn's problem, *Journal of the London Mathematical Society* 25, pp. 13–24, 1982

On Library Services in Decentralised Web Search Systems

Mario Kubek

Chair of Communication Networks
FernUniversität in Hagen, Germany

Abstract: If the World Wide Web is considered to be a huge library, it would need a librarian or at least services that can carry out this person's tasks in a comparable manner and quality, too. Google and other web search engines are more or less just keyword databases and cannot even remotely fulfil the manifold tasks of this person such as the cataloguing of publications as well as the mediation between library resources and users. Therefore, librarians can provide much better support during long-term and in-depth research tasks. This article discusses the many benefits of the services libraries and librarians provide and explains, how they can be realised in librarian-inspired and therefore decentrally organised web search systems that aim at sustainably supporting users in those research tasks.

1 Motivation

Public libraries are often lonesome places these days, because most of the information, knowledge and literature is made available in the omnipresent Internet, especially in the World Wide Web (WWW, web) as a major part of it. It seems that the times are forgotten, when librarians collected giant amounts of books, made them refindable by huge catalogue boxes containing thousands of small cards and archived them using their (own) special scheme in the right place in many floors consisting of a maze of shelves. In addition to these tasks and loaning books, they had time to support library users by giving them advises on where to find the wanted information quickly and maybe to even tell them the latest news and trends as well.

Currently, there is no information system available that can even remotely fulfil all of these tasks in an acceptable manner. However, as the amount of available textual data (especially in the WWW) is steadily growing and the data traffic volume for private web usage and sending e-mails reached 9170 PetaByte per month in 2016 [1], there is an urgent need for such a system which is –similar to the human librarian in her/his role as a caregiver– to be regarded as an active technical intermediary between users and resources such as textual documents. A respectively designed information system has to be able to autonomously

- provision, archive and manage information in many formats,
- provide an efficient information access (e.g. offer topic suggestions and topical 'signposts' as well as generate bibliographies),
- proactively procure information on the basis of identified information needs (demonstrate information literacy [2] i.e. 'the ability to know when there is a need for information as well as to identify, locate, evaluate, and effectively use that information for the issue or problem at hand' and become a pathfinder for possibly relevant information) and
- carry out search tasks while filtering out unimportant or even unsolicited information (one could simply speak of data in this case, too) by applying classification methods.

This implicitly means that the system must be able to perform search tasks on its own and on behalf of the user when requested. A system offering these features would especially facilitate in-depth research in a sustainable manner which strongly differs from short-term or adhoc search tasks. Particularly, in-depth research

- is an iterative and interactive process,
- has a context and history,
- consists of different search paths and directions,
- means to learn from positive and negative feedback and
- influences the objects being searched for as well (as an example: the most requested objects by experts in a field would likely be of relevance given respectively categorised queries, would therefore be returned first and would more easily become subjects to further scientific investigations).

These points imply that the mentioned information system is able to cope with and learn from dynamically changing contexts such as (even short-term) shifts in information needs and topical changes in the local document base as well as to identify possibly upcoming new trends or new concepts from various information streams. When the system takes into account the history of past and ongoing search processes in form of search paths consisting of queries and result sets, a navigation in previous search steps is possible. By this means and based on the learned concepts and their relationships, the system can also interactively provide alternative search directions as well as topic suggestions to follow.

While these functionalities are particularly of benefit for users conducting research, they address the problem of refinding information which is aggravated by the so-called 'Google effect', sometimes referred to as 'digital amnesia' [3], too. The main findings related to this effect are that people tend to forget information when they assume that it can be found again using digital technology and that they are more likely to remember how they previously found a certain information using search engines (the search path) instead of the information itself. This indicates that people are generally satisfied with the obtained search results (the relevant ones are presented first); otherwise the feeling would arise that it could be hard to find them again at a later time and it could be better to memorise the respective information. As this development –at least to a certain extent– affects the way research is carried out today [4], the mentioned system's functionalities of storing, retrieving and suggesting search paths should also facilitate the recovery of previous search processes and their results.

The following section reviews the many services provided by libraries with a focus on the activities of librarians. Afterwards, it is discussed, how these services are supported and extended by technical means such as cataloguing and information systems. Then, it is further analysed which of those activities and services can and must be technically realised in modern web search systems that are inspired by the working principles of librarians and thus should be organised in a decentralised manner. Finally, the first technical realisation of such a decentralised and fully integrated web search system, called 'WebEngine', is introduced.

2 Library Services

2.1 The Tasks of a Librarian

A library is popularly considered a collection of books or a building in which they are stored and cared for. While this understanding is generally correct, it is –at the same time– somewhat limited. There are a large number of common services that libraries provide. Here, the foremost function of libraries is to supply the public and institutions with information [5]. In order to be able to do so, they collect, catalogue and make published literature available in form of various types of media or resources such as books, magazines, newspapers and digital storage media like CDs and DVDs. The access to literature libraries make available is open, unrestricted and usually provided free of charge or for a reasonable price. Furthermore, the archiving of these resources is another important business of libraries which ensures the continuity of literary works. This step usually comprises additional tasks, especially when preserving book collections. For example, they have to be properly restored (when needed) and specially cared for (e.g. select a dry storage location with steady temperature and humidity levels). Also, their digitisation might be part of an archiving process in order to make contents searchable and easily transferable in electronic form. These tasks are usually carried out by librarians and archivists, depending on their specialisation.

The activities of librarians can be roughly classified into collection-centered and user-centered ones. While the collection-centered activities are related to the management of collected media and comprise the

- selection,
- acquisition,
- processing,
- cataloguing,
- care and
- archiving

of media, the user-centered activities include

- providing information,
- giving advice,
- organising and carrying out training courses,
- lending media and
- stocktaking (includes the recording and management of local as well as inter-library book loans).

A librarian's focus can therefore be either on collection-centered or user-centered activities. In the following subsections, the two most important activities of librarians are discussed in detail.

2.2 Librarians as Intermediaries

As information needs are constantly rising, the most important task of librarians is to mediate between requesting patrons and proper literature as well as information that satisfy their information needs. In that regard, they become active intermediaries in a search or research process. Therefore, they must be able to instruct library users properly on how they can find relevant information in the library (location of literature) with respect to the field of the subject at hand.

This role as a 'knowledge mediator' becomes even more important in the digital era [6] since it does not only require a solid educational background and good communication skills but encompasses the ability to deal with information technology and to make use of respective tools for data management and manipulating data, too. This changing role makes it possible for librarians to play an active and greater part in research processes and is thus especially of importance for the profession of the librarian as such in the future. Specialised librarians such as the so-called teaching librarians even give lectures on information literacy or competency [2].

Summarising, the provision of information and information sources is the most important service offered by librarians from the library user's point of view. For that reason, the quality of this service should be measured [7] by the following five indicators:

- Is the information desk visible and easily approachable?
- Does the librarian show interest in the user's request?
- Does the librarian listen carefully to the user and inquire openly if needed?
- Does the librarian make use of proper information resources and the correct research strategy (covers respective explanations to the user as well)?
- Are follow-up questions asked in order to determine if the user unterstood the information provided?

Thus, the librarian has to be friendly, ready to help, supportive and patient in order to provide good service. At the same time, the librarian must be able to adapt to a user's needs while keeping professional distance and providing equals service to all users (avoiding to outsmart particular users). Also, besides a good general education, language expertise and communication skills, a librarian has to have the ability to think in a structured manner and an interest in modern information technology.

2.3 Cataloguing Collections

Besides handling user requests, librarians usually are actively involved in the process of cataloguing media which is needed to keep track of the library's holdings and ultimately to make them refindable and is therefore at the core of the mentioned collection-centered activities. From a historical viewpoint, catalogues in book form, card catalogues and modern Online Public Access Catalogues (OPAC), which have widely replaced the two former kinds, can be distinguished.

In literature [5], two types of cataloguing approaches are distinguished: formal cataloguing (usually simply referred to as cataloguing) and subject indexing. Formal cataloguing means to apply formal rules to describe books and other media using formal elements such as their author and title. These elements are inherently drawn from the media themselves. Therefore, formal cataloguing means to transform [8] data into a form compliant to rules. Older rule sets for doing so are

- the RAK (Regeln für die alphabetische Katalogisierung),
- the AACR (Anglo-American Cataloguing Rules) and
- the AACR2.

The new standard RDA (Resource Description and Access) for cataloguing introduced in 2010 has a broader scope and is aimed to be applied by museums and archives besides libraries as well. Furthermore, this rule set (`https://access.rdatoolkit.org/`) provides extensive guidelines to extract attributes of entities such as a particular edition of a book as well as to determine their relationships to other entities in order to support applications that rely on linked data.

On the other hand, subject indexing means to describe resources based on their contents and content-related criteria and without relying on bibliographic or other formal data. Thus, subject indexing means to interpret contents and therefore implicitly requires methods that can transform data into information. The two most common methods for doing so are keyword assignment and content classification. Keywords for a resource can be directly drawn from it or by relying on external contents such as reviews and annotations assigned by users. Categories make it possible to distinguish e.g. between person-, time- and location-related keywords. Content classification relies on a given, usually hierarchic classification scheme and aims to assign resources to categories and subcategories and therefore to ultimately group them based on their topical orientation. Both approaches can be applied together.

Aside from this rather formal and theoretic distinction of cataloguing approaches, the practical establishment of a library, which at its core means to turn textual data into information and ultimately knowledge, requires librarians to make considerable efforts and is definitely a time-consuming learning process in which the interaction with their users plays an important role, i.e. it is a process with a determined history. This implicitly means that two librarians ordering documents such as books may end up with completely different arrangements depending on their own knowledge gathered and the experienced process of knowledge acquisition.

It usually requires a deep study of the texts (if not even special knowledge on particular subjects) in order to find out important terms as well as to determine their context-dependent meanings which are subsequently to be used in the assignment of categories to previously unseen contents and in the determination of their relations. This process also involves an estimation of the semantic similarity and distance to other terms and texts locally available. Thus, only after a larger amount of knowledge is gathered, a first classification of documents may be carried out with the necessary maturity and a first, later expansible catalogue and archiving system may be established. The resulting catalogue is a small and compact abstraction of details in each book and in a condensed form even a representation of human intelligence that was used in connecting related books with each other and, in case of a card catalogue, in deciding on the card placements accordingly.

Technically speaking, this construction process follows –in contrast to Google's top-down approach– a bottom-up approach as the arrangement as well as the classification and sorting of books is carried out in a successive manner starting with an initially small set of books and is mainly determined by the specialised (local) and the common knowledge of the librarian. As the library is growing this way, its now existing classification scheme makes it easier to catalogue and order incoming books. Furthermore, it is –besides the librarian's own knowledge– the knowledge base for giving advices on where to find particular books or information of interest to library users. This approach is likely more expedient and successful than the mentioned top-down approach of Google and Co., especially when domain knowledge is needed to handle inquiries with particular terminology, literally speaking when it comes to finding the 'needle in the haystack' of information. As already pointed out in [9] (for the field of marketing), the usage of 'small data' (in the case at hand represented by the specialised knowledge of the librarian used to properly guide users to their requested information) is often more beneficial than to rely on improper big data analyses. Furthermore, based on this local knowledge, topically similar and related documents can be identified fast and are therefore typically assigned the same category in the library.

3 Information Technology in Libraries

3.1 The Electronic Information Desk

The currently most important form of library catalogues is the so-called 'Online Public Access Catalog' (OPAC), an electronic bibliographic database, which has largely made the former physical card catalogues obsolete.

While OPACs make it possible for users to access and search for a library's resources using its respective online presence at any place and at any time, the integration and usage of Integrated Library Systems (ILS) [10] (which OPACs are a part of) has made the maintenance of these catalogues (management of metadata and information) along with the acquisition of media and management of loans more convenient for librarians, too. Especially, these systems make it possible to loan digital publications such as e-books, e-journals, e-papers (electronic newspapers and magazines) as well as digitised books and electronic course materials online.

Also, the cooperation between libraries has become easier by the introduction of data formats which foster the usage, exchange and interpretation of bibliographic information in records. For this purpose, the MARC (MAchine-Readable Cataloging) standard has been adopted widely. By this means, libraries cannot only offer their users local holdings, but can provide them records of associated libraries as well as additional services like inter-library book loans, too. This means that library users can access the preferred library's catalog locally (online and by physically going there) and are provided nationwide or even global information. The term 'hybrid library' [11] has been coined to indicate that a particular library offers classic and online services alike.

By the usage of ILS, spatial and temporal restrictions of classic libraries can be overcome as the provided electronic information desk is usually available around the clock. Thus, an adaptation to the users' communication behaviour is given. Furthermore, these systems support librarians by automatically analysing and forwarding user requests to the appropriate assistants. This assistance is even more extended by answering standard requests e.g. for opening times autonomously without the involvement of assistants. The integration of further electronic communication services such as chat and instant messenger services, microblogging sites, online social networks and Internet telephony has greatly facilitated the communication with library users. Online training courses can be easily provided and carried out by these means, too. Even so, it is always necessary to respect the protection of personal and private data, especially when a potentially large audience is addressed.

3.2 Searching the Web and OPACs

As mentioned in the introduction, web search engines can be helpful for finding relevant documents on short notice, especially when known items (e.g. the location of a shop) are searched for. However, when it comes to conducting comprehensive research on a topic of interest, users are –for the most part– not properly supported by them or even left for themselves. In such a case, usually referred to as subject search, users have to inspect the returned links to web documents by themselves, evaluate their relevance and possibly reformulate the initial or even subsequent queries in order to hopefully and finally satisfy their information need. This process is tedious and time-consuming alike, particularly when the user is unfamiliar with a subject and the proper terminology is yet unknown. Furthermore, most of the web documents are not catalogued or aimed for publication –in opposite to the literature– in libraries and as their trustworthiness cannot be taken for granted, it needs to be actively questioned at all times.

In these situations, libraries along with their services provided by both librarians and OPACs are definitely of more help. The reasons for this are obvious:

1. A library's literature has been intentionally selected and acquired.
2. A library provides literature in a well-ordered and structured form such that relevant contents in a field of interest can be found fast. A search process can be carried out in a more focussed way than it would be possible in a web search session.
3. Besides fields to formally classify a publication, OPACs provide dedicated fields for publications that are filled by applying methods of subject indexing. Faceted search is therefore a common feature in OPACs.

4. OPACs return a rich set of bibliographic information which simplifies continued searches of related materials. As an example, the author's name and the title of a publication are assigned to dedicated fields or elements with meaningful designations in the catalogue which makes it therefore easily possible for users to correctly interpret a publication's bibliographic information.

5. The information users are provided by libraries is usually trustworthy which includes both the literature found or suggested as well as other references to the subject of interest.

Subject searches are therefore more likely to succeed when relying on dependable library services. Furthermore, ILS are able to automatically extract citations from publications and link them to the literature referred to. The generated graph of related materials can then be the basis for content- or feature-based recommendation functions that users are accustomed to from e.g. web shops.

4 The Librarian of the Web

4.1 Decentralised Web Search with Library Services

Libraries have been early adopters of information technology. They applied information systems from the 1950s, a time during which the term 'information retrieval' (IR) [12] has been coined, too. At that time, professional searchers have been employed to act as 'search intermediaries' in order to translate users' requests into the respective system's language [13]. Nowadays, this function has been mostly replaced by search engines in various forms.

However, in order to realise a modern, librarian-inspired and decentralised web search engine as motivated in the introduction, the translation of users' information needs into their proper and promising technical representations as well as their matching with textual resources become again challenging core tasks of such a system. These tasks need to be carried out autonomously and automatically if necessary. Furthermore, in a decentralised setting, these representations must be routed and forwarded to peers that likely will be able to positively fulfil the mentioned information needs. Therefore, the next chapter gives an introduction to the general working principles of web search engines which covers the results of important research from the field of Peer-to-peer information retrieval (P2PIR), too.

The routing decision has to be made based on semantic considerations that also librarians would (unconsciously) take into account when guiding library users to relevant information and their sources. This is an especially crucial task as information in the web is largely unorganised as well as sparsely and often inconsistently annotated (if at all) by humans and machines alike and thus differs from information in catalogued library publications. In order to be able to do so, each peer of the proposed decentralised web search engine has to rely on a local knowledge base whose organisation closely matches the one of human (in this case the librarian's) lexical knowledge and therefore must be able to extract and index valuable information from textual sources and to put them into relation. This learning, ordering and cataloguing process can be implemented by applying specific algorithms and technical solutions known from the fields of natural language processing and text mining. They are particularly helpful for tasks such as automatic, high-quality key- and search word extraction as well as term and document clustering. The previously described cataloguing approaches can thus be applied in automatic form. However, the resulting catalogue or index is hardly comparable with rather monolithic, manually created OPACs as it is decentralised as well as automatically created and maintained.

At the same time, it is needed to account for implicit language-related dynamics in the web. Especially in online social networks and weblogs, language change is recognisable. This does not only mean that topics gain or loose public interest during a specific time period and at specific locations but that the wording to describe them changes as well. Also, youth language and slang has to be dealt with accordingly. While librarians in particular are easily able to adapt to these changes due to their constant interaction with users of all ages and their growing knowledge of subject-related developments in the library's publications, these changes of meaning are up-to-now not properly accounted for by current semantic approaches for web search. Particularly, (usually) specialised (i.e. domain-related) ontologies or taxonomies cannot properly reflect language dynamics as they are normally manually created by humans, such as librarians, using a fixed terminology. Here, a new take on automatically handling these dynamics is needed.

When it comes to conducting in-depth research in the web, the interaction with librarians is of great help due to their professional experience. In these cases, search becomes an information seeking process consisting of possibly numerous intermediate steps such as analysing presented information and reformulating queries. The proposed decentralised web search engine should be of similar usefulness in these situations. This means to interactively support the user in her or his current search task by giving instant feedback e.g. on the quality of a query or by suggesting (groups of) topically related query terms as well as by grouping similar and related web search results. In doing so, the system is able to learn from the interaction with its user and therefore to even make context-based predictions or to give recommendations on suitable next search steps. In this sense, an important step towards real 'information literacy' in information systems is taken[1].

4.2 Realising the Librarian of the Web

Based on these considerations and identified shortcomings of current web search engines, a new concept for decentralised web search, subsumed under the name 'Librarian of the Web' [14, 15], has been derived which comprises novel, librarian-inspired approaches, methods and technical solutions to decentrally search for text documents in the WWW. Its first implementation in form of an interactive, peer-to-peer (P2P) web search system, called 'WebEngine' [16], has already been made available.

The client software of this system consists of several components responsible for the storage, retrieval and semantic analysis of text documents, for the P2P-network construction and maintenance as well as for the execution of local and network-wide search tasks. Thus, a decentralised web search system is created and formed that –for the first time– combines modern text analysis techniques with novel and efficient search functions and a semantically induced P2P-network construction and management. In a more abstract and general view, the system makes use of analysis (text mining and query interpretation) and synthesis (library and network construction) methods, whereby the latter ones depend on the former ones.

The WebEngine has been realised as a Java-based P2P-plug-in for the popular Apache Tomcat (`http://tomcat.apache.org/`) servlet container and web server with a graphical user interface (GUI) for any standard web browser. Due to its integration with the web server, it uses the same runtime environment and may access the offered web pages and databases of the server with all related meta-information. Therefore, the system follows an alternative, integrated approach to web search under the motto 'The Web is its own search engine.' and was –as motivated before– conceived to inherently and actively support users with particular search and research tasks. Moreover, the structure of the generated

[1] The author is aware that real 'information literacy or competence' presented by humans is likely not to be reached by machines anytime soon.

P2P-network is directly induced by exploiting the web's explicit topology (links in web documents). The P2P-network is further able to restructure itself by means of self-organisation such that it becomes maintainable and searchable without any central authority.

5 Summary

This article reviewed and classified the main services provided at libraries and the activities of librarians working there. Especially, their two main tasks of mediating between information and library users as well as of cataloguing the (incoming) library's resources have been described in detail. As these tasks are usually backed by electronic information and cataloguing systems, they have been elaborated on as well. Furthermore, it has been analysed which activities carried out by librarians can be technically realised in decentralised web search systems to sustainably support research tasks. Finally, the new concept of the 'Librarian of the Web' as well as its first technical, P2P-based realisation, called WebEngine, have been outlined.

References

[1] Website of Statista, Monatliches Datenvolumen des privaten Internet-Traffics in den Jahren 2014 und 2015 sowie eine Prognose bis 2020 nach Segmenten (in Petabyte), 2016 `https://de.statista.com/statistik/daten/studie/152551/umfrage/prognose-zum-internet-traffic-nach-segment/`

[2] A. L. A. P. Committee (1989), Presidential committee on information literacy: Final report, American Library Association, 1989

[3] Sparrow B., Liu J., Wegner D. M.: Google Effects on Memory: Cognitive Consequences of Having Information at Our Fingertips, In: *Science*, 333:776–778, 2011

[4] Varshney, L. R.: The Google effect in doctoral theses, In: *Scientometrics*, 92(3):785–793, 2012

[5] Gantert, K.: Bibliothekarisches Grundwissen, 9th Edition, De Gruyter, 2016

[6] Bell, J., The developing role of librarians in a digital age, 2016 `http://www.infotoday.eu/Articles/Editorial/Featured-Articles/The-developing-role-of-librarians-in-a-digital-age-110185.aspx`

[7] American Library Association, Guidelines for Behavioral Performance of Reference and Information Service Providers, 2013 `http://www.ala.org/rusa/resources/guidelines/guidelinesbehavioral`

[8] Eberhardt, J.: Was ist (bibliothekarische) Sacherschliessung?, In: *Bibliotheksdienst 46.5*, pp. 386–401, 2012
[9] Lindstrom, M.: Small Data: The Tiny Clues That Uncover Huge Trends, Hachette UK, 2016
[10] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval, 2nd Edition, Addison-Wesley Publishing Company, 2011
[11] Oppenheim, C., Smithson, D.: What is the hybrid library? In: *Journal of Information Science*, 25(2):97–112, 1999
[12] Mooers, C. N.: Zatocoding applied to mechanical organization of knowledge, In: *American documentation*, 2(1):20–32, 1951
[13] Witschel, H. F.: Global and Local Resources for Peer-to-Peer Text Retrieval, PhD thesis, Leipzig University, 2008
[14] Kubek, M., Unger, H.: Towards a Librarian of the Web, In: *Proceedings of the 2nd International Conference on Communication and Information Processing (ICCIP 2016)*, New York, NY, USA, ACM, pp. 70–78, 2016
[15] Kubek, M., Unger, H.: A Concept Supporting Resilient, Fault-tolerant and Decentralised Search, In: *Autonomous Systems 2017*, Fortschritt-Berichte VDI, Reihe 10 Nr. 857, VDI-Verlag Düsseldorf, pp. 20–31, 2017
[16] Kubek, M., Unger, H.: The WebEngine – A Fully Integrated, Decentralised Web Search Engine, In: *Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval (NLPIR 2018)*, New York, NY, USA, ACM, 2018

How tall can be a Swiss Guardian, before he looses control?

About Wave Interference Properties in Nerve System

Gerd K. Heinz

Gesellschaft zur Förderung Angewandter Informatik (GFaI)
Berlin, Germany

Abstract: Behind Boolean[1] algebra and bus-protocols that carry the informatics of our micro-electronic devices like smart phones or personal computers there is an unknown, different type of information processing used by animals and human. Although the research about nerves has brought lots of advances in detail, neuro-computing lacks the great break-through. The author successfully investigates since 1992 wave interferences in nerve-like networks. Some examples of measurable properties in nerve-system shell give an impression, how valuable the theory of interference networks can be for an understanding of information processing in nerve systems. Behind a hyperbolic projection [H1, Kap.3], producing the Homunculus, we will discuss holistic properties of wave interference systems as shown by Lashley's rat experiments [11]. MacDougall's nerve circuit shows the *impossibility to learn with weights*. At hand of Hebbian weights learning we discuss the general problem of current neural network theory. The paper assumes a substantial understanding of interference systems and interference networks. Find introductions in [H0], [H2] or [H9].

1 Introduction

Information can be processed only, if the inputs are at the same time at the same place. Nerves carry the information very slowly and pulse-like. Compared to microelectronic information propagation it is one million times slower. But the cortical number of nerve cells is comparable very high. Men have around 100 billion cortical nerve cells. Any synchronisation of this giant supercomputer

[1] Gottfried Wilhelm Leibniz investigated binary logic and binary algebra in the way, modern microelectronics uses them, 170 years before Georg Boole's remarks. We should better call it Leibniz-Algebra (omnia ad unum).

using clocks fails because of the slow transfer velocities of signals (pulses) ranging from $\mu m/s$ and $120\,m/s$. The supercomputer has neither D-latches nor bus-protocols. But nerve systems work, but how? Modern brain research seems to be farther away from answers then ever before. Modern informatics gives no answer. Is there an unknown second informatics? Any nerve impulse tries to creep into each branching of nerve, exciting each destination. Communication can not work this way. Where is the solution?

Ionic signals in nerve systems can be electrically measured as pulse-like time-functions, flowing slowly through different nerve cells, through many stages of information processing. Compared to electron- velocity in organic materials the flow-velocity of ionic pulses in nerve is slow. Why nerves use slow moving pulses? Isn't it necessary for survival to be fast?

Because nerve pulses creep into any branch, any communication needs interference of lots of pulse-waves that reaches the destination location per coincidence exact at the same time. Researching in this field, we find information transfer that reminds to an optical style; we find mirrored interference projections, holograms, frequency and code matching circuits, all with nerve-like properties [H0].

In opposite to optical or acoustical wave-theories, we use single Gaussian waves (not sinoidal waves) for simulations. Only for demonstration purposes we plot neuro-projections on homogeneous 2-dimensional fields (nerve nets are supposed to be inhomogeneous).

Data addressing needs the *self-interference* condition [H1, Kap. 2, p. 42], frequency or code detection needs an understanding of *cross-interferences* [H1, Kap. 2, p. 53]. Like projections with optical lens systems, neural projections can occur under certain circumstances at defined places, the locations of self- and cross- interference [H7].

The term *self-interference* is used, if any pulse wave meets itself on a certain place in the net again. Like an optical lens system, self interferences can only produce *mirrored* projections, see the cover of [H1]. We talk about *cross-interference*, if subsequent pulse waves meet a following or a preceding pulse wave, necessary to detect sounds, codes or frequencies.

At hand of some examples we will demonstrate, that interference systems and wave interference networks (IN) can give a better understanding of nerve nets. Starting 1992 with the thumb experiment [H1, Kap. 6, H6] the author wrote different papers about nerve-like wave interference systems. The book "Neuronale

Interferenzen" [H1] has 2018 its 25th Birthday. Next year I'm retired. I hoped all the years, to get grants in this field. I tried different times, but failed. So I investigated them like a hobby behind the job. A first application of a simplest, technical interference network, called "Acoustic Camera" got lots of reports, radio-talks and TV-shows [H8]. The acoustic photo- and cinematography was born, but did not push the IN research.

2 Origins of Interference Networks

Different researchers found lots of views on interfering networks between holography and experimental sciences. With his rat experiments Karl Spencer Lashley found a direct visible holographic property of the brain. Independently, which part of the brain he removed, rats could remember a way through a labyrinth. Holograms are reasoned by signal interferences. So Lashley [11] used the terminus "interference" mutually for the first time, Karl Pribram [7] sent me this excerpt of one of his books:

> Lashley (1942) had proposed that interference patterns among wave fronts in brain electrical activity could serve as the substrate of perception and memory as well. This suited my earlier intuitions, but Lashley and I had discussed this alternative repeatedly, without coming up with any idea what wave fronts would look like in the brain. Nor could we figure out how, if they were there, how they could account for anything at the behavioral level. These discussions taking place between 1946 and 1948 became somewhat uncomfortable in regard to Don Hebb's book (1948) that he was writing at the time we were all together in the Yerkes Laboratory for Primate Biology in Florida. Lashley didn't like Hebb's formulation but could not express his reasons for this opinion: "Hebb is correct in all his details but he's just oh so wrong." (Karl Pribram in 'Brain and Mathematics', 1991, [7])

Lloyd A. Jeffress [5] was the first, who showed an interference circuit of the inner ear and Mark Konishi [6] was 1993 the one, who brought the Jeffress model of sound localization to a wide audience. Penfield [10] investigated body projections into the cortex – the so called 'Homunculus' was found as a coupling port between brain and body. We will discuss this finding later. Hodgkin and Huxley [12] investigated the ionic and electric behavior of nerve cells. Karl Pribram [7] and Walter Freeman [9] characterized nerve nets to be holomorphic and Andrew Packard found color waves on animals (squids) [8], showing the wave-like nature of pulse-propagation in nerve nets. The author found projections in IN to be "image-like" mirrored and holomorphic. He analyses interference circuits on nerve-like networks since 1992 [H0]. Basic properties of IN can be investigated with simple circuit configurations.

3 An Idea behind Penfield's Homunculus

Wilder Penfield [10] found the so called motoric and sensoric body projections in the gyrus precentralis of cortex, see Fig. 1, left. As neuro-surgeon he used electric stimulations to excite specific parts of the body. Nerve cells in gyrus precentralis map the whole body surface in all details, the drawing was mutually called 'Penfields Homunculus' by Love & Webb 1992 [16], see Wikipedia.

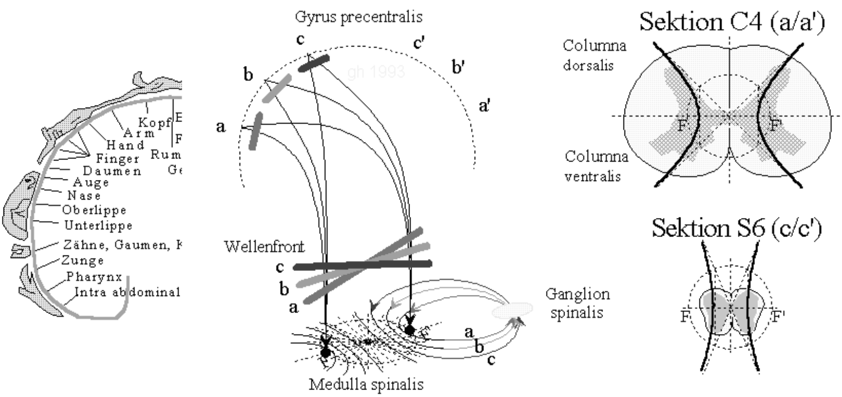

Fig. 1: Penfields Homunculus left. Middle and right: Hyperbolic interference projections create mutually the Homunculus

To analyze the wave theoretical nature of the Homunculus, let's have a look on the middle figure. If we mark three different wave fronts of the model with red, green and blue, we find a relation between the entrance points into the medulla spinalis and the output to the gyrus precentralis. The thickness of the spinal cord varies. Section C4 (image right) is bigger then section S6. If we suppose, that the projection type 'a' corresponds to S6 and type 'c' corresponds to type C4 of the spinal cord, this simple interference model produces the Homunculus.

In evolution of nature nothing is without of sense. What could be the sense of cortical body projections? If we read the thumb-experiment [H6], we get an idea. Is it possible, that the flexibility of the spinal cord makes problems to carry (straight) projections?

The spinal cord is very flexible and interference projections running through the spinal cord became shifted, if we turn the head. Reasoned by delay shift (movement of projections, see Bionet 1996, Fig. 8, [H4]) it is not simple to ”hold the screen”. It is the same, if we try to project an image through a long Berliner U-Bahn train, it is impossible if the train is in a curve. The single solution is, to use semi-transparent projection screens between all wagons and to transfer the image wagon by wagon through the whole train. This way the Homunculus appears as the last station - the projection screen in front of the train. On the other hand the correction of the projection field is simpler as with lens systems. We only need to control potentials at the embedding glia cells to make a neuron faster or slower – so we can shift the projection dependent of control potentials [H1]. Warning: Because the model is not verified by experts, these findings can be pure coincidence!

4 Understanding Lashley's Rat Experiments and Pribram's Holonomy

Subsequent pulses flow with specific velocity v, the width of the pulse and pause interval ($T = 1/f$) corresponds to a geometric distance, the *geometric wave length* $\lambda = v\,T$. Inspecting the nerve system, we do not find any nerve connection with negligible delay, each signal needs time to reach any destination. Thinking about projective interference systems we find answers [H1]: the cross-interference distance must be greater than the field size, thus the pulse velocity has to be slow. Because pulses expand in each direction, we will call them discrete waves on wires that flow in inhomogeneous nets of wires, so called ”Interference Networks” (IN). Excitement locations – interference integrals (I^2) are coupled to places, were lots of waves interfere at the same microsecond. Part of the solution is that all the different pathways (dendrites, axons) have different length and velocities – thus they have different delays.

The average distance between self- and cross-interference pattern is reasoned by the average delay between the waves. We call the corresponding distance *cross-interference radius* R of a plain field

$$R = v\,T/2 \tag{1}$$

Here the average nerve velocity is v, the pulse interval is $T = 1/f$ and the average fire frequency of generating neurons is f [H2]. The results of Lashley's rat experiments demanded a holomorphic memorization of brain content. (If a holographic glass plate is broken, we find the whole information on every glass fragment). Beginning 1994, the author made first simulations of this hologram-like behavior [H4] using pulse waves.

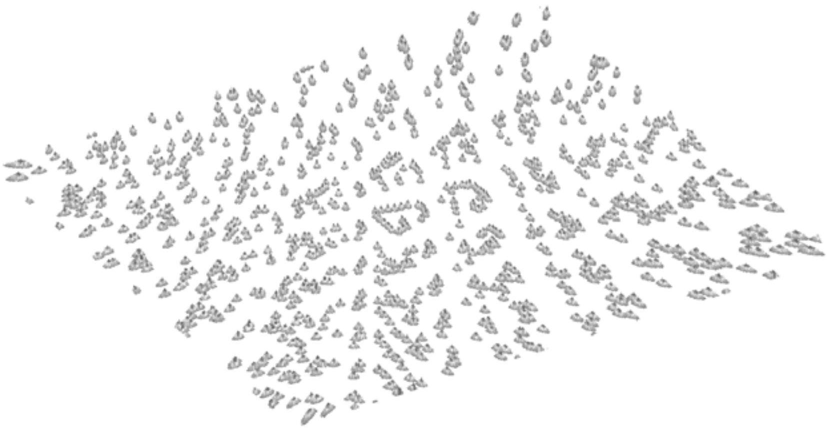

Fig. 2: Cross-interference pattern around the central self interference figure 'G' in a 3-channel, mirrored interference projection [H3], [H2] with holographic properties – all figures around have partial properties of the 'G'. (Simulation with PSI-Tools [H0]. The pulse generator field had permanent firing neurons in form of an mirrored 'G')

If many pulses flow through different nerves and re-combine, a specific pattern shows the so called cross-interference distance: Around a self-interference figure (the 'G' in the middle) subsequent following pulses form a cross-interference pattern, which has a distance (the cross-interference distance) to the self-interference figure. Fig. 2 implies, that each learning task produces a comparable holographic pattern, the labyrinth of Lashley's rat training could be found in each region of the brain. So independent of which part of the rats brain he removed, the rats could remember the way through the labyrinth. Again, because the model is not verified by experts, these findings can be pure coincidence.

5 How tall can be a Swiss Guardian – Nerve Velocity calculated by Cross-interference Distance

Changing the view, we can ask for the velocity for a cross interference distance of a two meter long Swiss Guardian [19].

The cross-interference distance R can be used to calculate the velocity v [H9]:

$$v = 2\,R/T = 2\,f\,R \quad \text{with} \tag{2}$$

$$R = 2\,m\,; \quad f = 30\,Hz \quad \text{we get} \tag{3}$$

$$v = 2\,f\,R = 2 * 30\,Hz * 2\,m = 120\,m/s\,. \tag{4}$$

Because of pure coincidence this is the value of the maximum velocity measurable in myelin-isolated nerves of human body! So, if the Swiss Guardian likes to be a fast boy ($f = 30\,Hz$), he should never be taller then two meters. If he will become faster, he has to reduce his lengths! Otherwise he can not address his feet's very well, cross interferences would produce an effect that remembers on Parkinson disease, the cross interference figures overlay the self interference figure, so every excitation into the self-interference area produces wrong excitements at unwanted locations.

A look to Leonardo da Vinci's "Vitruvianian Men" shows an arm length of approximately the half body length. That means: for $R = 1\,m$ and $v = 120\,m$ we get $f = v/(2\,R) = 120\,m/s/2\,m = 60\,Hz$, that means, the maximum fire frequency of sensory neurons of the hand can be two times higher. It seems, the peripheral nerve system can be calculated as an interference network! Is this pure coincidence again?

6 MacDougall's impossible Reflex Pathway

The inventor of the term "synapse" Charles Scott Sherington wrote 1906 about a discovery of MacDougall [2]. He published the drawing Fig. 3, where a reflex pathway was investigated. Lots of discussions followed about the possibility or non-possibility of this circuit, ending with: "No axon makes Type1 synapses (exciting) at some sites while making Type2 (inhibiting) at others." [H5]. The circuit has to work with only one kind of synapses; they have to be inhibiting or exciting. This implies, if the neurons have only one type of synapses the circuit can not work.

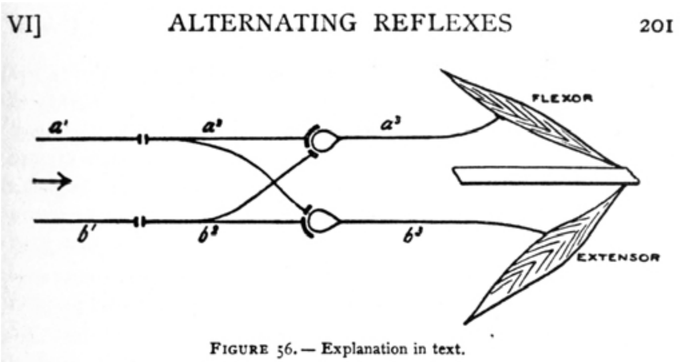

Fig. 3: Mac Dougalls reflex pathway, published by Sherington 1906, source [3].

If we observe the circuit as a *delaying, pulse-interference circuit*, Fig. 4, we need only exciting synapses of AND-type with a threshold to make it work. If the threshold of neuron N1 and neuron N2 is 3/2, and we suggest pulses of unity high = 1, then each neuron needs two pulses, arriving exactly at the same time to open the pathway. If we suggest delays on nerves proportional to the length, the network delays play the rule of the decoder. The patterns of Fig. 4 address the flexor or the extensor of Fig. 3 by delay shifts. Like lots of others, pure coincidence let us solve this problem again?

7 Weight or Delay Learning – why the Hebbian Rule is "oh so wrong"

If we have a look into the giant field of neuro-science carried of *synaptic weights*, we find learning weights from Hebb's rule over McCulloch-Pitts neurons over different Perzeptrons to Kohonens SOM, for example in [14]. McCulloch/Pitts neurons [18] and Hebb's rule [13] dominated fundamentally the new field of Artificial Neural Nets (ANN) [15] over 60 years with millions of papers and thousands of books. Everywhere we find the same introduction: "It is generally believed, that Hebbian learning ... ".

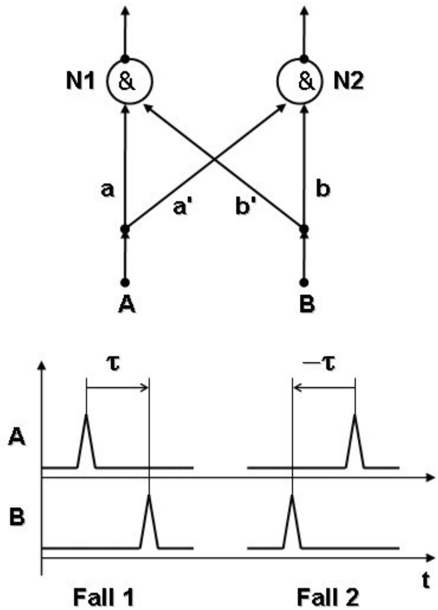

Fig. 4: Addressing the extensor (N2, Fall 1) or the flexor (N1, Fall 2) [H5] with pulse interferences. Attention: Nerve wires have length proportional delays (not drawn). Wires a' and b' are longer then a and b, they show by τ higher delays.

It was Donald Hebb, who introduced the most popular static learning rule in neuro-science, called Hebb's rule [13]. Learning was for Hebb the learning of synaptic weights with threshold gates. In general speaking we agree.

But in the case of MacDougalls reflex pathway we can learn and learn and learn the weights, and nothing happens! The pathway needs pulses and pulse-interference. If the pulse timing is different to the delays of the receiver circuit, it is absolutely impossible to learn anything with weights! That means, *learning is never only weight learning*! It is delay learning first; only the fine-tuning can be done with weights. Please listen again what Karl Pribram told us over his teacher and the teacher of Donald Hebb:

> "Lashley didn't like Hebb's formulation but could not express his reasons for this opinion: *"Hebb is correct in all his details but he's just oh so wrong.""*

Our simple flexor/extensor example shows, that Lashley had the right feeling:

Delays dominate over weights!

Only, if the delay structure of a network is well established for the solution, it will be possible to learn the details with weights: *Form codes behavior*.

So billions of dollars have been burned for millions of scientific works on weight learning. This was mutually on of the greatest disasters in modern science. It is the disaster of an international science policy, that is dominated by political correctness and majority believe.

Today we know, that dendrites grow and find the path through a soma in a way, that biologists directly could call "delay learning" [1]. On the other hand, by changing the thickness any axon or dendrite can change its velocity [2]. These questions will get the greatest relevance for future research in the age of wave interference network theory.

8 Summary

Karl Spencer Lashley observed directly holographic properties in rat's brain. He was mutually the first, how asked for interferences. Reported by Mark Konishi, Lloyd A. Jeffress had drawn mutually the first interference circuit. Andrew Packard was mutually the first, who filmed pulse waves on animals (squids).

A look to Penfield's Homunculus shows, that a simple interference network models the Homunculus in the gyrus precentralis over hyperbolic interference projections coming from the spinal cord.

The calculation of the cross interference radius of a tall Swiss Guardian with help of interference networks theory shows measurable pulse properties. It shows the peripheral nerve system can be calculated as an interference network.

MacDougall's reflex pathway cannot work as a threshold circuit. It works only with pulses and correct delays. For wrong delays, we can modify the weights without of the possibility, to make the circuit working.

So we found: Delays dominate over weights. If the delay structure of a network is not established for the solution, it will not be possible to learn anything with weights. McCulloch/Pitts "neurons" and Hebbian weights learning fails mutually for all delaying circuits (for example for nerve nets). It is not possible, to model interference systems (nerve nets) with weights learning only.

Details about biological delay learning could become the great advantage for interference network research in the future.

Asking for the problems in the middle of the 1990th the name of the scientific field was changed: The name "Neural Networks" (with weights learning) today are called "Artificial Neural Networks" (ANN). Now, we find this name also confusing. Weights' learning has nothing to do with delay-learning, nerve-like systems.

Warning: All findings can be pure coincidence!

Acknowledgments

Thanks for the invitation to the organizers of the 2018 conference "Autonomous Systems", especially Jutta Düring, Dr.-Ing. Mario Kubek, Prof. Dr.-Ing. Herwig Unger and Prof. Dr. Dr. Wolfgang A. Halang, FernUniversität in Hagen.

"Es ist schwierig, jemanden dazu zu bringen, etwas zu verstehen,
wenn er sein Gehalt dafür bekommt, daß er es nicht versteht."
(Upton Sinclair)

"Phantasie ist wichtiger als Wissen."
(Albert Einstein)

References

[1] Hubel, D.N., Wiesel, T.N.: Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195 (1968), 215–243

[2] Crick, F., Asanuma, C.: Certain aspects of the anatomy and physiology of the cerebral cortex. In: McClelland, J.L., Rumelhart, D.E.: *Parallel Distributed Processing*, Vol. 2, pp. 333–371, MIT-Press, 1986

[3] Sherington, Charles: The Integrative Action of Nervous System, 1906, Fig. 56, p. 201, Ref. 262: MacDougall, W.: *Brain*, Part cii, p. 153

[4] Mainen Z.F., Sejnowski T.J.: Reliability of spike timing in neocortical neurons. *Science* 1995, 268: 1503–1505.

[5] Jeffress, L.A.: A place theory of sound localization. In: *Journ. Comparative Physiol. Psychol.*, 41, (1948), pp. 35–39

[6] Konishi, M.: Listening with two ears. *Scientific American*, April 1993, pp. 66–73. German: Die Schallortung der Schleiereule. *Spektrum der Wissenschaft*, Juni 1993, S. 58–71, discussion at `http://www.gheinz.de/historic/intro/intro.htm`

[7] Pribram, K.H.: Brain and Mathematics. 1991, Ch.12 in Brain and Being: At the boundary between Brain, Physics, Language and Art. 2004. Eds. *Globus*. (personal communication)

[8] Packard, A.: Organization of cephalopod chromatophore systems: a neuromuscular image-generator. In: Abbott, N.J., Williamson, R., Maddock, L., Cephalopod Neurobiology, *Oxford University Press*, 1995, pp. 331–367, see historic `http://www.gheinz.de/biomodel/squids/index.htm`

[9] Walter J. Freeman III: Mass Action in the Nervous System, *Academic Press*, New York, 1975

[10] Penfield, W., Rasmussen, T.: The Cerebral Cortex of Man. A Clinical Study of Localization of Function. New York, *Macmillan Comp.*, 1950

[11] Lashley, K.S.: In search of the engram. Society of Exp. Biology Symp., No. 4 (1950), *Cambridge University Press*, pp. 454–480

[12] Hodgkin, A.L., Huxley, A.F.: A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve. *Journ. Physiology*, London, 117 (1952) pp. 500–544

[13] Hebb, D. O.: The Organization of Behavior. *Wiley*, New York, 1949

[14] Andersen, J.A., Rosenfeld, E. (Edts.): Neurocomputing : Foundations of research. Cambridge, *MIT-Press*, 1988, 729 pp.

[15] Sejnowski, T. J. and Tesauro, G. (1989). The Hebb rule for synaptic plasticity: algorithms and implementations. In: J. H. Byrne, W. O. Berry (eds.), Neural Models of Plasticity, *Academic Press*, 1989, Ch. 6, 94–103

[16] Love, R.J., Webb, W.G.: Neurology for the speech-language pathologist. *Butterworth-Heinemann Ltd.*, 1992, 309 p.

[17] Singer W, Gray CM: Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci.* 1995, 18 : 555–586

[18] McCulloch, W.S.; Pitts, W.: A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5 : 115–133

[19] See `https://en.wikipedia.org/wiki/Swiss_Guards`

Related Publications of the Author

Find the homepage www.gheinz.de and a list of publications under [H0]:

[H0] www.gheinz.de/publications

[H1] Heinz, G.: Neuronale Interferenzen. Eigenverlag, 300 S., 1993, www.gheinz.de/publications/NI/index.htm

[H2] Heinz, G.: Introduction to Wave Interference Networks. Workshop 2010 'Autonomous Systems' 24.-29.10.2010 Camp de Mar, Mallorca, Proceedings: Shaker-Verlag 2010, ISBN 978-3-8322-9514-1, Fig. 10

[H3] Heinz, G.: Cross interference distance: See Eqn. (1) at www.gheinz.de/historic/pressinf/bilder_d.htm#radius

[H4] Heinz, G., Höfs, S., Busch, C., Zöllner, M.: Time Pattern, Data Addressing, Coding, Projections and Topographic Maps between Multiple Connected Neural Fields – a Physical Approach to Neural Superimposition and Interference. Proceedings BioNet'96, GFaI-Berlin, 1997, pp. 45–57, www.gheinz.de/publications/1996_Bionet.pdf

[H5] Heinz über Sherington: siehe www.gheinz/historic/pressinf/bilder_d.htm

[H6] Heinz, G.: Relativität elektrischer Impulsausbreitung als Schlüssel zur Informatik biologischer Systeme. 39. Internationales Wissenschaftliches Kolloqium an der TU Ilmenau 27.-30.9.1994, printed in vol. 2, pp. 238–245, www.gheinz.de/publications/papers/1994_IWK.pdf

[H7] Heinz, G.: An investigation of 'Pictures of Thought' – properties of pulsating, short circuit networks in Theory and simulation. Int. School of Biophysics "Neuronal Coding of Perceptual Systems", Cassamicciola, Isle of Ischia, Naples, Italy, Oct. 12–17, 1998. Published in Backhaus, W.: Neuronal Coding of Perceptual Systems. Series on biophysics and biocybernetics, vol. 9 – Biophysics, World Scientific, New Yersey, London, Singapore, Hong Kong, 2001, ISBN 981-02-4164-X, pp. 377–391, http://www.gheinz.de/publications/papers/1998_Ischia.pdf

[H8] See www.gheinz.de/publications/presse

Architecture for Trust-based Machine to Machine Communication

Christoph Maget

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract: In existing major communication networks, authentication and authorization processes are primarily managed by hierarchically structured authorities, providing credentials of various types. Each user is obliged to provide the managing authority with user data and thus – hypothetically – can be held responsible for any kind of action taken within the network.

In the upcoming field of ubiquitous computing and the internet of things, with its huge number of connected elements and dynamic network structures, centralized management no longer seems appropriate. Instead of central authorities, the need for decentralized and horizontally organized methods for connecting items, based on the trustworthiness of the item, is stated.

In this paper we use the automotive and road transport sector as an example to propose a communication architecture, based on assigning a trust-based token to each user as a key element for participation. The metric to provide a particular node with trust follows an innovative combination of assignments by other users who act as guarantors. Maintaining these trust-levels and providing them among the network requires a distributed database (DDBMS). It is suggested that this DDBMS be implemented by a blockchain.

A proof of concept is given by an implementation based on vastly accessible consumer electronics devices. Such devices could be used to upgrade existing sensors or actuators, that are not yet connected. Proper configured, those devices can also meet rigorous security demands. Possible further applications of these concepts are all kinds of interacting artificial entities like software agents, devices in automation networks and varied sensor and actuator nodes, contributing to a decentralized automation process.

1 Introduction and Terminology

Information gathering about the physical environment and its sharing with other individuals are major human needs [1]. This interaction and the exchange of information takes place both in reality and in virtual worlds [3]. Although various types of technology have been developed that facilitate this information gathering and transmission, the endpoints of communication in the evolving networks usually remain human. For accessing digital communication networks, customized devices (also called "hosts") are necessary. Depending on the network structure, the source and destination of each type of information is thus connectable to a responsible person. In the case of the widespread "world wide web", an Internet Service Provider (ISP) is able to link necessary IP-addresses of subscribers with companies, households or persons. Since communication networks are usually represented as graphs, the acting entities are also considered nodes. Depending on the logical network structure, the nodes are distinguishable in "clients" and "servers" (central approach) or "peers" (decentralized approach).

Optimized software and increase in capacity of modern computers enhances the ability of computer systems to solve problems without additional human interaction [2]. This observation, combined with reduced hardware and communication equipment costs, can allow – and sometimes even justify – literally connecting every entity ("thing") of the physical world to each other. The entity would thus be upgraded to a host. Terms like the "Internet of Things" (IoT) or "Ubiquitous Computing" have been coined for that kind of technology [4]. Although some of those concepts are already implemented partly within modern industrial facilities and automation networks [5], the central demand for network management by central authorities has never disappeared.

A possible further development is remote and slightly autonomous acting and transceiving machines as the above-mentioned destination of the information. These communication processes may be a necessity to handle minor tasks, while the human supervisor is only recipient of the final result. The underlying processes for achieving the final results may be formed by self-organized, autonomously acting computer systems. To avoid misuse, the necessary interaction of any host with another among a potentially large number of connected devices needs to be based on an authentication and authorization process. For a particular host it is very important to know about the "trustworthiness" of other nodes and thus to have a basis for deciding whether to share information or not. Moreover, due to the lack of a managing authority and a central storage, this knowledge about hosts has to be evaluated, coordinated and distributed among the network participants.

2 Specific Limits of distributed network management

Computer systems that are designed and programmed to operate autonomously and cooperatively must obviously be able to communicate with one another. This ability includes both technical and organizational aspects. As stated in the following paragraphs, the existing network technology can not yet meet these requirements.

2.1 Technical Infrastructure

Technical aspects of communication in computer networks are a well-established field of research including a variety of protocols and standards. There are already numerous implementations that form local and global networks [6].

Existing major communication networks are highly dependent on a kind of centralized management by the respective infrastructure owner, aiming at restricted access to these networks in order to avoid misuse. For example, a Wi-Fi operator issues SSID and password to trusted persons, a GSM operator issues SIM modules to authenticated persons both in order to grant access to the network using symmetric-key algorithms. The use of asymmetric-key algorithms requires the implementation and maintenance of a public-key-infrastructure (PKI) or – with some restrictions on revocation and personal data – the establishment a web of trust (WOT). Although those processes can more or less be automated, each host's initial access to those networks must be granted individually. Authentication along with authorization and assignment of specific roles need administration and thus (human) interaction. With raising size and potential autonomy of the operating computer systems, three major challenges are identified.

First, design, standardization and implementation of network protocols lead to a network of highly heterogeneous connected entities. Small, low power devices use the same network stack as high-end computers. Concerning security, the former might lack resistance to the latter, as far as security bases on the difficulty of a mathematical problem. Second, there is no physical barrier to accessing the (wireless) communication infrastructure. On the one hand this eliminates the need for a physical infrastructure for communicating computer systems. On the other hand, because certain frequencies of wireless communication channels are available to literally anyone, the only way to prevent messages from being eavesdropped is to use encryption. Third, any kind of managing authorities can be placed away from any access by local or national law. In a connected world, it is quite easy to migrate digital services around the world to places with laws that are more appropriate to the specific service. A common example of misuse may be the provision of masses of emails containing advertising or malicious content ("Spam"), sent from servers beyond the reach of appropriate law enforcement agencies. Penalties or simple shutdown of malicious services is thereby prevented.

It is thus stated that with the evolution of the internet of things, at a certain point, a central authority cannot or will not efficiently manage authentication and authorization for to technical or regulatory reasons. Consequently, hosts or "things" cannot rely on central authorities to manage both the identity and reliability of other hosts. They have to independently and autonomously manage authentication and authorization during the communication process. These problems demonstrate the necessity to consider about new methods for a future communication architecture. In a connected world across administrative boundaries, computer systems should be equipped with a decision model that allows to distinguish whether to establish a connection with another node. The exchange of data with the other node may thus be based on a parameterized trustworthiness of the other node. Therefore, it is necessary to consider a new kind of communication architecture based on trust in the individual subscribers rather than relying on central network management authorities of any kind.

2.2 Decentralized Authentication and Authorization

In the upcoming field of ubiquitous computing, the nodes of distributed computer systems must be equipped with some kind of "skill" to decide whether and what to communicate with other nodes in that system. In order to transfer aspects of "trustworthiness" in a social understanding to technical systems, three principles of trust assessment can be distinguished.

Active assessment: A peer can gain trust actively through own actions undertaken in the network. The merit of trust lies in the very own responsibility of the specific peer. A possible metric would be the level of participation as the quotient of transmitted and received data.

Passive assessment: A peer can gain trust passively through actions undertaken by other peers in the network. The peer must therefore rely on the cooperation of other peers. One possible metric would be the evaluation by other peers, as it is used in all types of existing reputation systems. Examples are the "Advogato" algorithm which evaluates authors and software developers in a likewise peer review process [7], or the "PageRank" algorithm [8] used by Google Inc. to classify the significance of web pages for specific topics during search.

Transitive assessment: A peer can gain trust with respect to another peer if there is a transitive connection among them. A binary relation R on a set X is defined as a transitive relation, if $\forall a, b, c \in X : (aRb \land bRc) \implies aRc$. By equating R with "trust", it can be determined whether peer a trusts peer c. A flaw of transitive assessment is the lack of an objective determination of trust itself. In any case, this concept forms the basis of the Web of Trust and the EigenTrust value [9].

Some of these aspects are already being addressed by peer-to-peer (P2P) networks [10]. Even with P2P networks basically still implemented as overlay networks on managed (IP) networks, they include equally privileged communication nodes – the peers – and no mandatory managing authority. Although the upcoming internet of things should require specific parameters that allow peers to decide whether and how they should interact with each other, only very rudimentary metrics for this topic were introduced so far.

3 Trust-based Network Establishing

The evolution from automation networks towards increasingly autonomous systems requires a more appropriate metric that considers the complex interactions of peers. In the following section, we will design such a communication architecture based on assignment and distribution of trust as proof for authentication and authorization in wireless networks.

In the previous chapter 2, three possibilities of assigning trust were introduced. Since automation networks are associated with high safety and security requirements, we will use the third one: Transitive trust. As we will see, this offers an effective method to achieve trustworthiness in an appropriate manner, while the first (active) the and second (passive) possibility that were introduced would always result in a threshold of some kind.It would therefore be difficult to discern whether one should trust or not trust an entity, as that threshold is always arbitrary.

For the general development of the communication architecture abovementioned, three main factors must be considered. First, a decentralized assignment of identifiers to hosts is necessary. Secondly, the decentralized assignment of "trust", the merging of appropriate parameters, is to be considered. Third, the trust of each peer must be distributed among the network in a way such that an arbitrary assignment of trust to specific peers without consensus among the other peers is avoided.

3.1 Identity Management

First of all, the identity (ID) of a "thing" in the network has to be considered. In accordance with the basic principles of information security, it is crucial that IDs of entities that form hosts in the network are unique, non-transferable and non-repudiable. This allows authentication and interaction with other nodes in the network. In a technical point of view, the ID allows addressing the host in order to send information. An example in present IP-networks is the implementation of fixed IP-addresses. In the proposed architecture, the identity is stored in form of a bit string according to RFC 4122 on an identification module. It has to be implemented in such a way that removal or manipulation results in the destruction of the identity module, so that its use for authentication is no longer possible. One way to create those unique IDs in a decentralized manner may be the "Universally Unique Identifier (UUID)" and its implementation "Globally Unique Identifier (GUID)" with a length of 128 bit. While GUID uses the MAC address of the network device as a unique part of the ID, we here use a hash function of the encryption keys to construct a unique element as one part for ID generation.

3.2 Trust Abstraction and Assignment

The introduced unique IDs, which allow to identify and address hosts in the network, are the basis for further assignment of values that determine whether a node is trusted or not.

Instead of using a metric combining active, passive and transitive trust, using only a transitive trust assignment is proposed in the presented architecture. Starting from an initial node, the network is set up by issuing trust by two other nodes to an applying node that wants to join the network. These two other nodes are the first contact nodes for the applying node and must already be part of the network. Different possibilities for the necessary authentication in peer-to-peer networks are presented in [11].

The network generation follows dedicated steps:

- node A is ready to form a network
- node A and node B authenticate to each other forming the initial part of the network
- node C authenticates to node A and node B
- node A and node B check their databases to see whether node C is trustworthy
- if both node A and node B trust node C, the latter is accepted to the network
- node D connects accordingly whilst it can decide whether to authenticate to A and B, A and C or B and C
- additional nodes are added accordingly

The described protocol is also depicted in figure 1.

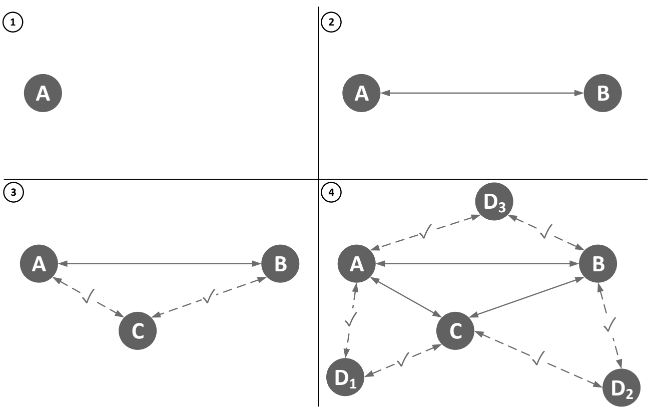

Fig. 1: Connection Protocol

If the applying node is not yet existing in the databases or is not marked as trustworthy, the issuing nodes must decide whether to add the applying node anyway by proving the trustworthiness with a separate method [12].

Using this protocol ensures that only those nodes, that have been trusted by at least two other nodes that are already part of the network, will enter the network. Any malicious acting node can be traced back to the initial granting nodes and possibly allow prosecution.

3.3 Provision and Storage of Trust

To provide knowledge about each peer's trust for every other node in the network, consistent records must be distributed among all peers. Technically spoken, a distributed database management system (DDBMS) needs to be established that contains consistent information that is accessible to each peer at all times.

The simplest form of any entry in such a database would be a tuple of ID and trust

$$(ID, trusted\{yes|no\})$$

which could be stored as dictionary data type:

$$nodes = \{"ID1" : "yes", "ID2" : "yes", "ID3" : "no"\}$$

Traditional distributed databases assume cooperation among the database fragments and suppose that the possibly insecure element lies within the communication channel. As a remedy the data may be encrypted before transmission, but this still doesn't handle with compromised participants manipulating the database content after decryption and storage: A personal replication or specific database for each node would be easy to alter and thus compromise the functionality of the overall concept. One example of such forgery is a P2P file-sharing software, in which the client software could be modified to allow an erroneous assignment of the maximum participation level to a particular node without the supposed contribution to the network. Furthermore, implementing a DDBMS according to the concepts of Master-Slave or Client-Server, would create the same central managing authority issues as discussed in the chapter 2.

This problem of generating consensus about on the trust of every peer in the network can eventually be solved by using distributed ledger technology. By implementing such a blockchain technology, it is extremely difficult for a single node, or even a significant bunch of the network participants, to mutate the database in an illegitimate manner. For a detailed description of Blockchain technology, including the essential concepts of proof-of-work and proof-of-stake, see [13].

Among the described disadvantages of a blockchain, two aspects have a special significance for the considered application:

Extensive energy consumption: Distributed computer systems may be equipped with a limited power supply. It is therefore not suitable to use these devices to perform the costly calculations that are necessary to create the blockchain. The vulnerability of these devices to powerful stationary hardware, which can carry out the aforementioned attack on the blockchain database, complicates this concept.

Limited transactions per time: Calculating and adding new blocks to the chain takes a specific amount of time which is not necessarily determined. The creation of blocks is then queued and is therefore not real-time capable.

As a workaround for the first issue, it is suggested to use trusted platform modules (TPM) to specify the hardware provided for trust assignment and distribution. This provides the possibility to exactly stipulate the specific hardware. As a result, none of the peers has a dedicated advantage at assigning, providing, or maintaining trust levels, even if poor computing modules are used. Instead, all peers meet identical requirements. At this point, it must be noted that the TPM is not used for securing the individual peer as may be the case in other implementations. The TPM is deployed to secure the network as a whole since it incorporates a disclosure of joining hosts. This prevents hosts from accessing the network with uncertified hardware being capable to mutate the DDBMS in an illegitimate way.

The second issue goes hand in hand with the problem of unpredictable time intervals for providing an updated trust status among the participating nodes. As a remedy, it is considered to set an arbitrary upper time limit for the generation of trust transactions and to accordingly reduce the difficulty of generating transaction blocks when appropriate. Another approach would be to use the proof-of-stake concept instead of the proof-of-work concept. In this case, established peers would have a significant advantage in adding newly admitted nodes to the DDBMS. New nodes would thus favor established nodes for their application, and prevent relatively new nodes from adding additional nodes to the communication architecture. This would result in a inefficient use of network capacity. A detailed investigation of these different approaches is planned with a simulation.

4 Experimental Considerations

As a proof of concept, the communication architecture is implemented as a prototype for the automotive sector. The resulting device is intended to be integrated with cars or roadside equipment to provide them with additional connection features. It is worth noticing that the presented device is vendor independent and based on open source components. This enables a wide provision without limitations by proprietary technologies or rights.

To implement the proposed concepts, standard consumer electronics is used. With the availability in vast quantities, those devices have the advantages of low cost, exchangeability and support by a huge community. An understandable deficit can be a lack of warranty and reliability. In the present work, the device serves as proof-of-concept and is not yet intended to be used to exchange safety- or security-critical information. Therefore, a failure has no serious consequences. Whether the device in its actual form can also be used for security crucial applications has to be evaluated in further research. Anyway, by using standardized parts, all components can easily be replaced by more robust types.

The logical structure of the device is shown in figure 2: Each device on board of a vehicle or roadside equipment is initially provided with keys for one-time encryption by a control module according to [14]. A hash value of the database is used to generate a unique element for the ID according to RFC 4122. A trust module, connected to a DDBMS, evaluates applying peers and stores known peers together with their respective trust values in this DDBMS. The trust module is also connected to a communication module to receive IDs from the applying peers as well as to establish the network connection. A microcontroller unit carrys out all further communication.

Figure 3 shows the implementation by assembling consumer electronic. This version is integrated into an electrical enclosure and thus intended for use in roadside equipment.

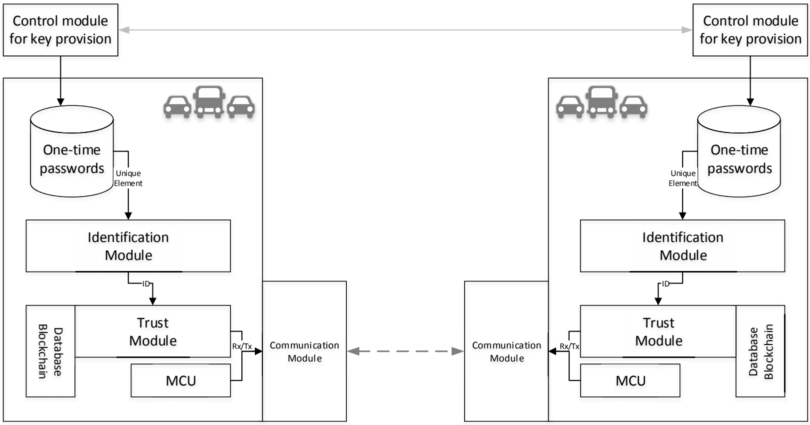

Fig. 2: Logical Structure

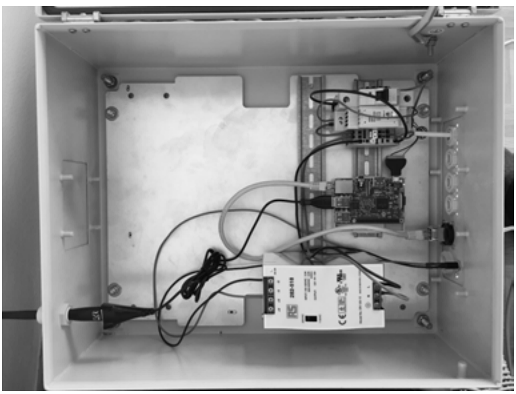

Fig. 3: Technical Implementation

5 Summary and Outlook

In this paper the technical and organizational limitations of present network management mechanisms were discussed. As a remedy, an approach to a self-organizing network architecture has been proposed. It is mainly set up by a peer-to-peer network, in which the connection is allowed based on a value representing the trustworthiness of the participating peers. Determining this trust of a peer applying for network access is provided by two other peers who act as guarantors. A protocol for the application process was introduced. It is said that this concept has the potential to be highly scalable and thus being applicable to large scale networks that are to be materialized in the upcoming "internet of things".

Key element of the distributed storage of peers and the trust assigned to them is a distributed database management system (DDBMS). The network-wide provision of one peer's trust level is protected by a distributed ledger implemented by a blockchain. Concerning hardware implementation, a first prototype was presented. The hardware used is still relatively powerful, but should be reduced and optimized to the specific application. Costs and energy consumption are expected to be further optimized.

As a first application, the automotive sector could be considered. The connection and autonomous interaction of cars is expected to significantly raise the capacity of existing road infrastructure through its more efficient use. Also called "car to x communication (C2X)", this sector seems to be predestined for such communication architectures.

Although one motivation for the development of the communication architecture has been the vast and ever-increasing use of wireless networks, it is not limited to those. Existing wired automation networks could be improved, particularly in terms of scalability, by equipping them with the proposed architecture.

References

[1] Tomasello, M.: *Origins of Human Communication*, MIT Press, 2008
[2] Russell, S. Norvig, P.: *Artificial Intelligence: A Modern Approach*, Prentice Hall, 2009
[3] Coltzau, H.: *Dezentrale Netzwelten als Interaktions- und Handelsplattformen*, FernUniversität in Hagen, 2012

[4] Weiser, M.: *The Computer for the 21st Century*, Scientific American, 1991
[5] Bauernhansl, T.: *Industrie 4.0 in Produktion, Automatisierung und Logistik*, Springer Vieweg, 2014
[6] Comer, D.: *Computer Networks and Internets*, Pearson Education Limited, 2014
[7] Ruderman, J.: *A comparison of two trust metrics*, University of California, San Diego, 2004
[8] Page, L.: *Method for node ranking in a linked database Abstract*, US-Patent US6285999B1, 1998
[9] Trung, S.: *On Trustworthiness Recommendation*, FernUniversität in Hagen, 2017
[10] Barkai, D.: *Peer-to-peer computing : technologies for sharing and collaborating on the net*, Intel Press, 2001
[11] Nassermostofi, F.: Toward Authentication between familiar Peers in P2P Networking Systems, In *Autonomous Systems 2016: Proceedings of the 9th GI Conference*, pp. 88–101, 2016
[12] Marti, S.: *Trust and Reputation in Peer-to-Peer Networks*, Stanford University, 2005
[13] Swan, M.: *Blockchain: Blueprint for a New Economy*, O'Reilly Media, 2015
[14] Schleupner, L.: *Perfekt sichere Kommunikation in der Automatisierungstechnik*, FernUniversität in Hagen, 2012

Research on Information Network Vulnerability of Intelligent Substation

Ruiwen He

Guangdong University of Technology, Guangzhou, P.R. China

Abstract: With the development of the smart grid, the power system protection, control and monitoring functions of substation automation system rely more and more on the communication networks. Because of the increasing dependence on information flow, network security has great influence on the reliability of power system. In this paper, by comparing the existing information network vulnerability assessment methods, the network evaluation model suitable for intelligent substation is found, and an improved vulnerability assessment based on state attack graph. And the degree of vulnerability associated with protection mal-operation and mis-operation in intelligent substation is also obtained.

On Hierarchical Clustering using Random Walks in Microgrid

Yudha Nurdin

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract:
This paper investigates the hierarchical clustering in Microgrid using Random Walker algorithm. In order to decentralize the electrical grid, the interaction between several entities such as Distributed Energy Resources (DER), Energy Storage Systems (ESS) and consumers will play its role in determining the best hierarchical cluster in the electrical grid in order to optimize energy distribution locally.

This paper proposed the use of random walker as a proxy for energy flow in the grid. This approach will help in viewing the real network structure as the realization of distance that creates links between groups of nodes. It helps to identify the most likely group cluster as a part of multilevel structures in the grid. This cluster which considered as a microgrid can further be represented by a higher level random walker in more large integrated grid systems.

The visualization of the concept will be discussed as part of the model hierarchical organization in the microgrid. Several scenarios where nodes overlap in the different cluster also discussed and solution alternative was presented. The paper concludes with remarks on future works.

A novel Microgrid coined

Zhong Li

Faculty of Mathematics and Computer Science
FernUniversität in Hagen, Germany

Abstract: The intermittence of renewable energy from e.g. photovoltaic cells and wind power results in problems like unstability, harmonic pollution, unpredictability, inefficiency, etc., which pose a great challenge for scientists and engineers for constructing a smart grid. To address these problems, this paper proposes a novel structure of a microgrid, where the loads are categorized into sensitive and non-sensitive ones, and the latter ones are used together with storage devices to balance the energy. To solve the harmonic and unpredictable problems, the microgrid should present pure-resistance; and for the main grid easily to predict and manage the microgrid, it should present a piecewise constant power feature, which means that the microgrid generates constant energy at a certain time period. For this purpose, a collective control method, that is the particle swarm optimization (PSO) control method, is adopted to take five control factors into account. Finally, simulations will be conducted to verify the effectiveness of the proposed microgrid structure and the control methodology.

Design, Analysis and Implementation of High-Step-Up Converters in Renewable Energy Systems

Guidong Zhang, Zhiyang Wang and Yun Zhang

Guangdong University of Technology, Guangzhou, China

Abstract:
This paper devises a novel single-switch n-cell high-step-up converter. Compared to the existing cascaded boost converters and traditional n-cell converters with same components and voltage gain, the proposed ones have lower components stresses. It is noteworthy that the novelty of the proposed converters lies in lower components stresses only by re-connecting the same components. For a high voltage with multi-cell or multi-stage, one should choose components with strict parameters with high currents or voltages, but the proposed one can realize the same functions with normal components, which are very valuable for renewable energy applications of boosting low input voltages to desired high ones. Then, the proposed converters unique features are analyzed and demonstrated, and it is followed with a comparison with other high-step-up converters. Finally, simulation results are conducted to validate their effectiveness.

A Fully Neurocomputing based Traffic Modelling-and-Simulation Concept

Nkiediel Alan Akwir[1], Muhindo Kule Mutengi[2], Witesyavwirwa Vianney Kambale[1,3], Jean Chamberlain Chedjou[1] and Kyandoghere Kyamakya[1]

[1]Institute of Smart Systems Technologies, Transportation Informatics Group
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
[2]Institut Supérieur de Techniques Appliquées (ISTA), Goma, DR Congo
[3]Tshwane University of Technology, Pretoria, South Africa

Abstract: This paper proposes a fully neuro-computing based novel concept for traffic modelling-and-simulation which does better fit to the particularly hard requirements (amongst others w.r.t. accuracy and computational speed in view of real-time constraints) of the so-called "online traffic simulation". Neurocomputing has been already shown and demonstrated its efficiency in several applications in science and engineering. Here, we do use this paradigm to build a robust traffic modeler-and-simulator. Different candidate neural networks paradigms/architectures are reviewed and briefly discussed in the paper such as MLP ANNs (Multi-Layer Perceptrons feed forward Artificial Neural Networks), RNNs (Recurrent Neural network), CNNs (Cellular Neural network) and Echo-State (Reservoir computing RC) Neural Networks. All of them may be exploited to design and realize the neurocomputing traffic model-and-simulator that we call "NeuroTrafficSim" (Neurocomputing based Traffic Simulator). The extensive training of NeuroTrafficSim is done using a supervised training mode, whereby either real-world or field data from different traffic scenarios are generated by one of the best microscopic traffic simulators, the VISSIM Simulator, and used for training and testing the NeuroTrafficSim. The concept implementation and some experimental results are briefly discussed in this paper.

1 Introduction

Road traffic management is necessary to solve inherent traffic problems such as congestion, traffic jams, pollution and reduce accidents just to name a few and thus increasing the quality of life of the society. "Online traffic simulation" is one possible solution used to overcome traffic problems since it is fast and consumes less memory.

In this work, we propose and develop a novel traffic flow modelling-and-simulation fully involving neurocomputing, especially using the cellular neural network paradigm. It is exemplary applied on complex junction for illustration of the concept: one lane + ramp metering. Online real-time traffic simulation is useful in various contexts such as traffic planning and optimized adaptive and proactive traffic management. Regarding the online-simulation capability, it is generally judged/assessed and validated while considering the following performance metrics: (a) accuracy, (b) high computation speed (for ensuring real-time computing), (c) low memory consumption, and (d) enabling and efficient control/optimization and/or the forecasting capability of traffic flows.

The main hypothesis of this work is that the scenario(s) envisaged correspond to real operational configurations of a single lane road segment with ramps. The resulting "calibrated macroscopic/microscopic & neuro-computing" traffic model can be further used to perform various online-simulations of the spatiotemporal dynamics of traffic flows under diverse configuration contexts of a signalized/unsignalised road segment. Here, we do develop and exemplary validate a fully neuro-computing based (involving Cellular Neural Networks – CNN) black-box calibrated spatiotemporal traffic flow modeling-and-simulation model.

For an appropriate accurate training, we do involve and use real or quasi-realistic data generated by one of the best microscopic the traffic flow simulators (VISSIM). VISSIM is capable of simulating and predicting the spatiotemporal states of traffic flows and their evolution on a road-segment; this is thus considered as real traffic data for referencing and calibration. The inputs of the neuro-computing model are the same as for a macroscopic traffic model of the same context, i. e.: initial conditions, boundary conditions and control parameters of the traffic flow model expressed in the form of second order coupled partial differential equations (PDEs). Regarding the positions and the time for spatiotemporal prediction capabilities, as boundaries conditions, we shall propose different conditions (as input values for traffic volumes for both the main segment and the ramp) and finally, for initial conditions the road segment is considered as empty (no vehicles).

The scenario under consideration in this paper is made of a 1-lane road segment of length 1 km with different max speeds. The ramp flow (flow generation rates) is set up for both low and high values (e. g. ″give low value″ and ″give high value″) to emphasize different traffic behaviors. The boundary and initial conditions may further be chosen to provide more scenarios such as the effect due to ″the pedestrian crossing″ and ″the traffic signal″; these last-named scenarios are more relevant for a city/urban road context.

2 A Critical Review of the Relevant State-of-the-art

To date, several alternatives or approaches or perspectives have been proposed to address specific issues related to traffic flow modelling, e. g.: specific phenomena (such as shockwaves, bottleneck, congestion, etc.), the level of details (see microscopic versus macroscopic), traffic states (see: under saturation, saturation and oversaturation), traffic disturbing conditions (e. g.: weather, accidents), and the efficiency (e. g.: accuracy, robustness, stability, and online simulation capability).

2.1 A Brief Survey of Traditional Traffic Modeling Models/Concepts

The basic/seminal model for traffic flow was proposed by Lighthill-Whitham (1955) and Richards (1956), the so-called LWR's model. This model, also known as first order model, is based on the continuity equation from compressible dynamics theory which expresses the conservation of a flowing quantity from one point to another. Despite the fact that the LWR model can reproduce the formation, propagation and the evolution of shockwaves, it faces a lot of drawbacks, amongst others the following ones: the ″constant speed″ and the ″infinite acceleration of vehicles″.

To face this above-named issue w.r.t. to the drawbacks, Payne proposed the first high-order model which, in addition to the LWR model, adds a second equation expressing the dynamics of the speed. The drawback due to the fact that the so-called anisotropic principle is not respected leads to negative speeds. New models that can solve this issue started to be constructed. The first model in this respect have been proposed by Zhang [1]. Later, Jiang et al. [2] developed a speed-gradient (SG) model from the full velocity difference model FDV. Furthers, Gupta and Katiyar [3] did develop and suggest a modified anisotropic model.

Overall, traditional models involving differential equations do not fulfill the hard requirements fixed for a truly online and realtime traffic-model-and-simulation. Neural networks do however offer more potential to come near to these hard requirements; see next section.

2.2 State-of-the-art Related to Neural Network-based Traffic Modelling

Over the past few years, besides models involving differential equations, artificial intelligence techniques have played an important role in the design of sophisticated traffic management systems [4] while exploiting data driven models such as the ones based on neural network paradigms and which are trying to capture a "function fitting" with data (traffic information). Neural network paradigms have already been exploited for example to derive traffic flow models for heavy traffic conditions [4], or to model a freeway traffic flow [5], or to derive a real-time short term traffic flow forecasting model [6]; all are core components of relevance for the core challenged addressed by this paper work. Overall, several models have been proposed and do involve different sources of information/data such as non-vehicle variables (weather, accident or day time just to name a few). Most of the models involving neural networks did however rather focus on short and long-term forecasting.

Different kinds of neural networks have been used for traffic forecasting such as the artificial neural network (ANN), which is the basic architecture, by Guan et al. [7]. Guozhen Tan et al. [8] exploited the first with the generalized neural network (GNN) on traffic flow prediction. Furthers, Gu and Yu [9] demonstrated that chaotic neural networks show better results when compared to the classical ANN that uses the traditional back propagation algorithm (BP) for training, this while predicting/assessing data related to road junction.

Chan et al. [10] proposed a generalizes ANN which improves the features of the classical ANN by using hybrid exponential smoothening. Fusco et al. [11] compared and found that ANN and Bayesian Networks (BN) are similar regarding the respective accuracy characteristics.

Lately, the use of deep learning becomes more and more popular due to its capability of hierarchical extraction of features, which helps for a good training process towards better classification and prediction. Several papers such as Lv et al. [12] and , Huang et al. [13] and Huang et al. [14], just to name few, have exploited different deep learning techniques for traffic flow forecasting. Generally, it outperforms the classical neural network learning techniques, ARIMA (auto regression integrated moving average) and SVR (support vector regression) models especially for the forecasting task of traffic flow values.

3 Traffic Modelling Scenarios Definition

3.1 Modelling Scenarios Definition

As mentioned before, the scenario under consideration in this work is made of one lane road segment coupled to a ramp. The ramp is one of the responsible sources/causes of congestion problems on a highway due to merging and diverging maneuvers in a limited space. A good ramp modelling and design helps to apply better control strategies towards avoiding congestion and improve the traffic flow on a highway.

Regarding a mathematical modelling, ramps are basically modeled by a nonzero source term from a system of conservative equations that govern the traffic flow phenomenon. As an example, the following is a mathematical model of a lane with a ramp:

$$\begin{cases} \dfrac{\partial k}{\partial t} + \dfrac{\partial (ku)}{\partial x} = r_{rmp}(x,t) \\ \dfrac{\partial u}{\partial t} + u\dfrac{\partial u}{\partial x} = \dfrac{u_e - u}{\tau} + c_0 \dfrac{\partial u}{\partial x} \end{cases} \tag{1}$$

where k and u stand respectively for traffic density and speed; τ and c_0 are the driver's reaction time respectively the propagation speed of disturbance.

u_e stands for the fundamental diagram of traffic flow which expresses the relationship between the density and speed. Several models have been proposed such as the Greeshields's model [15], Greenberg model [16], underwood model [17] just to name a few. Let us mention that the Greeshields's model is the most used and expressed as follows: $u_e = u_f - (u_f/k_j)\,k$ with u_f stands the free flow speed and k_j stands the jam density.

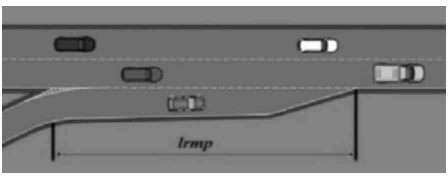

Fig. 1: A road segment with ramp – illustrative scheme

$r_{rmp}(x,t)$ is the term expressing the ramp as follows:

$$r_{rmp}(x,t) = \begin{cases} \dfrac{q_{rmp}(t)}{l_{rmp}} & \text{if } x \text{ is within merging or diverging zones} \\ 0 & \text{elsewhere} \end{cases} \tag{2}$$

with l_{rmp} the length of the ramp.

Several algorithms have been proposed for ramp metering (ramp control) to optimize the traffic flow at the ramp area such that ALINEA [18], METALINE7 [19] or SWARM [20] just to name a few. This paper does however not handle ramp metering, which is a traffic control task. This paper is just performing a modelling and simulation of the traffic flows in this scenario; the resulting model is then of relevance for the relevant traffic control concepts.

This work proposes a fully neurocomputing based concept for modelling (i. e. the equivalent of a macroscopic/microscopic traffic model) and simulation of traffic flows, which shall help to reduce traffic congestion on highways with ramp by exploiting its online and forecasting capabilities (the respective traffic controller will use the output of the traffic flow modelling-and-simulation function brick). Different scenarios with different traffic states (i. e.: undersaturation, saturation, congestion) will be considered.

3.2 Scenarios Settings

We consider a road of 1 km length with an on-ramp localized in the middle of the road. We exploit different traffic flow scenarios which allow us to experiment different traffic states to obtain the data necessary for training the model based on a neural network paradigm developed. Different traffic volumes and speeds have been considered on the main road to fit with the different traffic states (under saturation state, saturation state and over saturation state) and different traffic volumes on the ramp have been also chosen such that we can also experience different traffic states. Let us denotes $u_{(x_0\_max)}$ as the maximum speed of vehicles, $Q_{(x_0)}$ the generation volume rate of vehicles on the main road and $q_{(x_{L/2})}$ the generation volume rate on the ramp. The following are different combinations of values of speeds and generation rates for scenarios settings.

Table 1: Different combinations of settings values

$u_{(x_0\_max)}$ in km/h	$Q_{(x_0)}$ in veh/hour	$q_{(x_{L/2})}$ in veh/hour
30	500	50
40	700	100
50	1000	250
60	1500	350
80	2000	500

4 Generation of Reference Data by Using a Microscopic Traffic Simulator

4.1 VISSIM Brief Description

The well-known VISSIM[1] (microscopic) traffic simulator is used to generate the data used for (supervised) training and testing of the model proposed in this work. The tool VISSIM models the movement of individuals vehicles driving in a road network by exploiting different simulation models such cars following models, lane changing or gap acceptances rules. It is more accurate and realistic than the so-called empirical models.

4.2 Reference Data Generator/Production

The VISSIM simulator generates different kinds of data according to the user requirements/settings. We need to obtain the spatiotemporal information related to the fundamental traffic parameters which are the "density" and the "flow". Let us consider a road segment with the following features:

- A length of 1 km;
- Data collections points at every 10 m;
- A ramp in the middle of the road.

The data collection points (are equivalent to the "virtual" presence of sensors are those points) help to obtain directly the information related to the time, the position and the speed. The flow and the density are obtained indirectly by using the time headways obtained from consecutively following vehicles and

[1]See Link: `http://vision-traffic.ptvgroup.com/en-us/products/ptv-vissim/`

the speed of vehicles (from data collected at the "data collection points"). Different simulations have been performed to generate data according to various traffic conditions as mentioned in the subsection 3.2, see Table 1.

5 Neural Network Architectures

Neural networks are computational model implemented as computer programs to process information by taking inspiration from the biological nervous system, such as the human brain, to process information. Like human beings, they do learn by examples; see the so-called supervised learning. Inputs and their corresponding outputs data (example, training data) are trained to define/model an unspecified link/relationship between the input and the output especially when one cannot derive an explicit mathematical expression. The training (supervised learning) aim is to derive neural network parameters through an optimization process (training process) to minimize the discrepancy between the neural network (modeled) outputs and the measured/real system outputs.

5.1 Artificial Neural Network

A single neuron (perceptron) is made of inputs, weights, and an activation function in a relationship through a mathematical expression as follows (see also Fig. 2):

$$y_i = f\left(\sum_i w_i\, x_i + b\right) \tag{3}$$

with $X = [x_i]$ the set of inputs, $W =$ the set of weights. The output signal is $Y = [y_i]$ and b is the bias. The activation function is given by $f = f(W, X)$ [21].

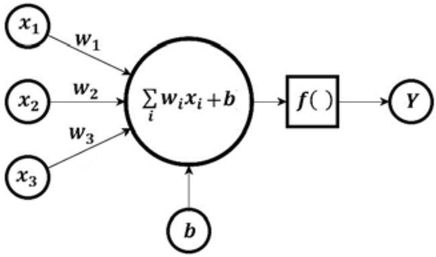

Fig. 2: A sample neuron model scheme

Several perceptrons are interconnected by following a certain architecture (with layers) to form a neural network architecture; see for example Fig. 3.

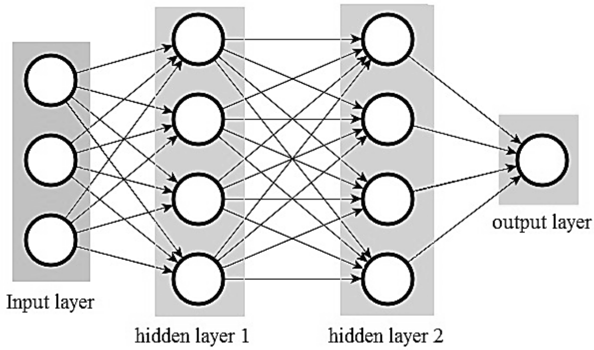

Fig. 3: A sample artificial neural network configuration (multi-layer perceptron, feed-forward)

Several optimization methods can be used for training, examples: backpropagation algorithm [22], Levenberg-Marquardt algorithm [23, 24] etc., just to name a few.

Considering that the target is given as $\overline{Y} = [\overline{y}_i]$, a training algorithm is performed (iteratively in most of the cases by randomly initializing the weights which are decision variables) to minimize the discrepancy E between the modeled output data and the target data.

$$\min_{w} E(Y, \overline{Y}, w) \tag{4}$$

A new set of weight is obtained $W^* = [w_i^*]$ which minimizes the discrepancy E. Thus, the neural network can be used to perform various task related to the problem modeled.

5.2 Recurrent Neural Networks

In the previous neural network models (MLP, ANN) mostly known as traditional or artificial neural network ANN they assume that the inputs and the outputs are independent of each other. For many tasks such as the ones with sequential information, we may need previous information to predict the next information. They do have a sort of "memory" which captures information about previous computations. The inputs and outputs are thus dependent. Figure 4 illustrates an example a RNN principle.

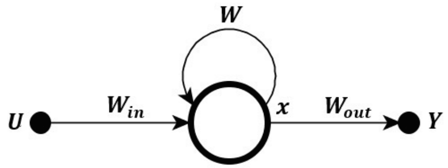

Fig. 4: Recurrent neural network with feedback connection

with $x(k) = f[W_{in}\, u(k) + W\, x(k-1)]$ and $y(k) = softmax[W_{out}\, x(k)]$;

Furthers, $U = [u(k)]$ is the input, x stands for the hidden state and $Y = [y(k)]$ is the output. W_{in}, W, W_{out} are weights within the RNN network architecture.

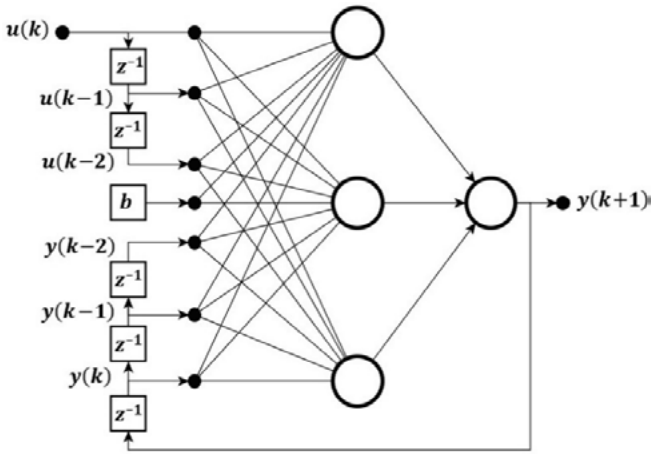

Fig. 5: NARX configuration NN architecture

Different types of recurrent networks exist and are often used such as the following ones: Elman RNN, NARX, LSTM, just to name few. In this paper, we do exploit the NARX neural network, mainly for illustration, but also due to its capability to deal with non-linear data. Figure 5 shows the NARX architecture.

5.3 Cellular Neural Networks

Cellular neural Networks (CNNs) are nonlinear, recurrent and locally interconnected arrays of dynamical cells located on a multidimensional grid [25]. CNNs operate in parallel and their structure allows to develop analog chips which can be used to process complex signals where low power consumption is needed. The following figure (see Fig. 6) shows the architecture of CNNs.

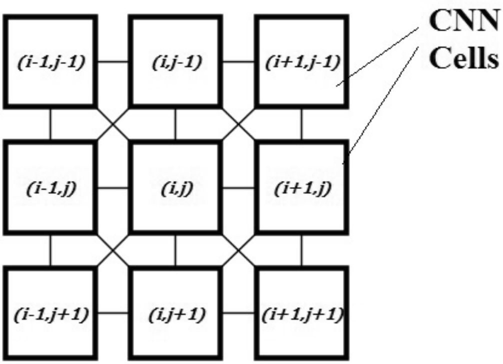

Fig. 6: A cellular neural network grid connection showing the neighborhood concept amongst CNN cells

CNN Mathematical Model

The basic mathematical model of a CNN-cell has been proposed by the related seminal paper in Ref. [4] (see also Ref. [5]). The model is expressed as follows:

$$\dot{x}_i = -x_i + \sum_j \left(A_{ij}\, x_j + B_{ij}\, y_j + C_{ij}\, u_j\right) + I_i \tag{5}$$

$$y_i = f(x_i) \tag{6}$$

The following is the compact-form (matrix form) of the CNN-model:

$$\dot{\vec{X}} = -\vec{X} + A\vec{X} + B\vec{Y} + C\vec{U} + I \tag{7}$$

$$\vec{Y} = f(\vec{X}) \tag{8}$$

where $\vec{X}$ is the vector of neural network states, $\vec{U}$ the input of the CNN (e.g. input data) and $\vec{Y}$ the output (output data) of the CNN. The parameters A, B, C, and I stand for the so-called CNN-templates. Indeed, the parameter A denotes the state-controlled template, B stands for the feedback template, C is the forward template, and I stands for the threshold vector of CNN. f is the activation function and it is the same as the one for ANNs.

Consider training. An error $\boldsymbol{E}$ (generally mean squared error) function is derived exploiting the CNN model. Further, the error is minimized by using different optimization methods such as gradient descent, genetic algorithm (GA), and particle swarm optimization (PSO), just to name a few. The optimization problem is defined as follows:

$$\min_{w=A\cup B\cup C\cup I} E(Y, \bar{Y}, w) \tag{9}$$

After applying the optimization procedure, new templates $w^* = A^* \cup B^* \cup C^* \cup I^*$ are obtained which minimize the RMSE (as objective function).

More CNN model have been proposed in the literature according the various applications (signal and/or image processing, classification, prediction etc.) and the complexity of the problem (continuous or discrete input data, non-linearity, etc.).

The following is the CNN model we are going to consider in this work.

$$\dot{x}_i = -x_i + \sum_j \left(A_{ij}\, x_j + B_{ij}\, x_j^2 + C_{ij}\, x_j^3 + D_{ij}\, y_j + E_{ij}\, u_j \right) + I_i \tag{10}$$

$$y_i = f(x_i) \tag{11}$$

The quadratic and cubic nonlinearity "$B_{ij}\, x_j^2 + C_{ij}\, x_j^3$" increase the robustness of the model and help to cope with the complexity of the data.

5.4 Echo State Network (Reservoir Computing)

The Echo State Network (ESN) is a recurrent neural network and one of the most popular reservoir computing (RC) approaches [26]. In Reservoir computing, the input is applied to a random dynamical system called reservoir, which maps the input to a higher dimension. The readout layer is trained to rebuild the target output from the reservoir state. The advantage of this approach is that the reservoir is fixed and the training is performed only at the readout stage. The following Figure 7 illustrates the RC architecture.

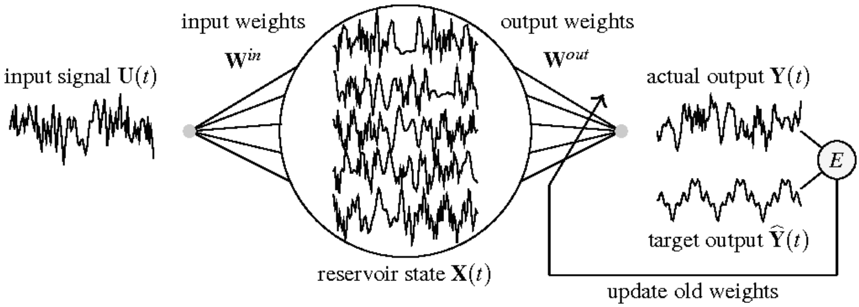

Fig. 7: Standard scheme of the reservoir computing architecture [26]

The reservoir state $X(t)$ is given as follows:

$$x_j(t+1) = \tanh\left[W_j^{res}.\,X(t) + W^{in}.\,U(t)\right] \tag{12}$$

where $U(t) = [u_i(t)]$ are the input signals, $X(t) = [x_j(t)]$ is the reservoir state, $W^{in} = \left[w_{i,j}^{in}\right]$ is the input weights matrix with size $I * N$ (where I is the number of input nodes and N the number of reservoir nodes), $w_{i,j}^{in}$ is the weight of the connection from input node i to reservoir node j. $W^{res} = \left[w_{j,k}^{res}\right]$ stands for the reservoir connection weights matrix with size $N * N$. $w_{j,k}^{res}$ is the weight connection from node j to the node k in the reservoir.

All the weights are samples of i.i.d (independent identically distributed) random variables from the normal distribution; with a normal distribution with mean μ_w and standard deviation σ_w.

The design matrix X is obtained by the concatenating the vectors $X = [1, u_i(t), x_j(t)]$; thus the reservoir output is given by:

$$Y = W^{out}.\,X \tag{13}$$

The output weights are trained also to minimize the MSE $E = \| Y - \bar{Y} \|^2$ with $\bar{Y}$ considered as target output data. This is done using the ordinary linear regression by performing the pseudo-inverse of the design matrix (X^+) as follows:

$$W^{out} = X^+ . \bar{Y} = (X^T X)^{-1} . X^T . \bar{Y} \tag{14}$$

In case the matrix product $(X^T X)^{-1}$ is not invertible, we can use the Moore-Penrose pseudo-inverse which always exits and given as follows:

$$X^+ = \lim_{\delta \to 0} X^T (X . X^T + \delta I)^{-1} = \lim_{\delta \to 0} (X^T X + \delta I)^{-1} . X^T \tag{15}$$

The performance of the Echo-state architecture mostly depends on the choice of its parameters such as the combination of the reservoir size N (number of reservoir nodes) along the standard deviation which define the randomness of all the weights. The optimal standard deviation σ_w^* is given as follows:

$$\sigma_w^* = \arg_{\sigma_w} \min RMNSE(\sigma_w, N) \tag{16}$$

6 The "NeuroTrafficSim" Concept

6.1 General Principle

The NeuroTrafficSim is a traffic flow modeler-and-simulator which is based on the neural network paradigms. The neural network paradigms have already shown their efficiency in many areas in science and engineering due to their learning and lately deep learning potentials which give more capabilities of accurate prediction and classification in complex problems. As mentioned before, traffic is a very complex system which involves a lot of phenomena. We exploit an Echo-State based CNN architecture which has already shown its effectiveness and robustness to capture the traffic flow behaviors. The data used for training have been generated using VISSIM Simulator (generate quasi realistic data) which mimics the reality.

6.2 Training Concept and Test Data Selection

VISSIM generates data which are processed to fit the Echo-State/CNN architecture.

As inputs, we have the "position" and the "time" (spatiotemporal capability), and the different scenarios parameters settings which are the generation volume rates of vehicles on both the main road and the ramp, and the maximum speeds of vehicles.

As targets, we have the flow, the density and the speed. The processed data have been divided into training and testing data as follows.

Table 2: Different combinations of settings values (**Training data**)

$u_{(x_0\_max)}$ in km/h	$Q_{(x_0)}$ in veh/hour	$q_{(x_{L/2})}$ in veh/hour
30	500	50
50	1000	250
80	2000	500

The combination of these parameters settings indicated in Table 2 gives 27 scenarios.

Table 3: Different combinations of settings values (**Testing data**)

$u_{(x_0\_max)}$ in km/h	$Q_{(x_0)}$ in veh/hour	$q_{(x_{L/2})}$ in veh/hour
40	700	100
60	1500	350

The combination of these parameters settings indicated in Table 3 gives 8 scenarios.

6.3 "NeuroTrafficSim" Implementation Scheme

The NeuroTrafficSim has 5 inputs and 3 outputs. The following figure 8 illustrates the different inputs and outputs:

In Figure 8, x and t are the spatiotemporal inputs. The total length of the road under consideration is 1000 m with 100 data collection points every 10 m ($M = 100$). The simulation time duration is of 600 s with $\Delta t = 1\,s$, which implies $N = 600$. For a given combination $u_{(x_0\_max)}$, $Q_{(x_0)}$ and $q_{(x_{L/2})}$ (see Tables 2 and 3) which are also inputs, for each position we have information about the density k, the speed u and the flow q during the whole simulation duration; this gives $M * N$ samples for each combination.

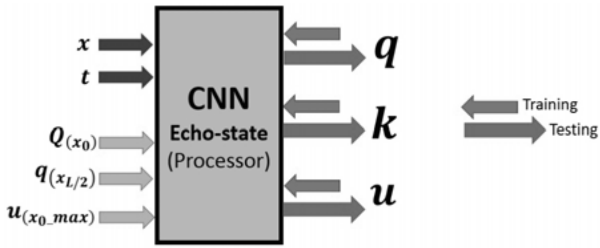

Fig. 8: CNN based Reservoir computing architecture – the "NeuroTrafficSim" related black-box Model

7 Simulation Results and their Interpretation

7.1 Two Selected Scenarios Descriptions

We have considered 2 different scenarios in this work. We generate data from a single segment with and without ramps. They are described as follows:

a) One single lane: $L = 1000\,m$, $\Delta x = 10\,m$ (from data collection point), different maximum speeds and different generations rates from the main segment to obtain different traffic states (under saturation, saturation and over saturation)

b) One single lane with ramp: Same settings as one single lane and different generation rates from the ramp to obtain different traffic situations

7.2 First Illustrative Result on First Scenario

The data obtains from the first scenario are illustrated in figure 9 above.

For evaluation, we use the normalized root mean square error which is calculated as follows:

$$NRMSE = \frac{\sqrt{1/n * (y - \bar{y})^2}}{\max(\bar{y}) - \min(\bar{y})} \tag{17}$$

The NRMSE normalizes the error between 0 and 1. It also compares the error to the maximum speed (free flow speed), maximum density (jam density) and maximum flow.

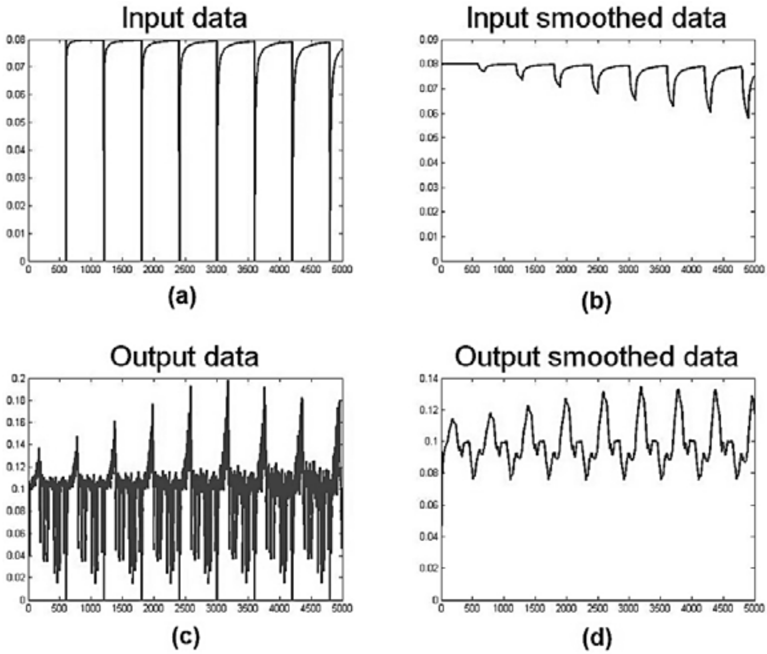

Fig. 9: Sample data involved in the illustrative experiments

Table 4: Preliminary results of simulations

Network Configuration	NRMSE in %	
	Normal Data	Smoothed Data
Data generated by d PDE solver	28.02	27.51
1 layer, 50 neurons (NARX)	6.37	5.61
10 hidden layers, 20 neurons per layer (NARX)	2.36	2.11

The preliminary results are given in Table 4. The results show that smoothed data gives a much better performance than non-smoothed data. This improvement is due to the fact that smoothing removes noises and thus decreases the complexity in the data.

8 Conclusion and Outlook

In this work, we have proposed a described a traffic flow modeler-and-simulator based on the neurocomputing paradigm called "NeuroTrafficSim". Different neural network paradigms have been discussed. Let us mention that the artificial neural networks (ANNs) does performs with good accuracy several tasks such as calibration of traffic model; however, it doesn't perform very well when it comes for forecasting which exploits previous information to predict future information. The recurrent neural network (RNNs) has a feedback loop which has this capability of better predicting future information which depends on previous information.

Also, the cellular neural networks (CNNs) are much more efficient since they are robust, recurrent and flexible (white-box model). Finally, the reservoir computing is also suitable for robust forecasting since it is recurrent and present a lot of parameters that can be tuned to increase its performance.

The implementation of NeuroTrafficSim concept has been discussed and we have shown some preliminaries experimental results. A selected set of some good results has been presented and discussed. Nevertheless, more research is conducted to improve the NeuroTrafficSim concept since a lot of neural network architectures are being developed nowadays in this field, especially w.r.t. recurrent neural networks such as LSTM and deep learning.

References

[1] Z. HM, "A theory of nonequilibrium traffic flow," *Transportation Research Part B*, vol. 32, no. 7, pp. 485–498, 1998.

[2] Rui Jiang, Qing-Song Wu , Zuo-Jin Zhu, "A new continuum model for traffic flow and numerical tests," *Transportation Research Part B: Methodological*, vol. 36, no. 5, pp. 405–419, 2002.

[3] A.K. gupta and V.K. Katiyar, "A new anisotropic continuum model for traffic flow," *Physica A*, vol. 368, pp. 551–559, 2006.

[4] C. Ledoux, "A neural network traffic flow model for heavy traffic conditions," *Transportation Reseach Board*, pp. 265–279, 1998.

[5] Jian-Mei Xiao and Xi-Huai Wang, "Freeway traffic flow modeling based on neural network," in *Intelligent Transportation Systems*, Proceedings, 2003.

[6] Shan He, Cheng Hu, Guo-Jie Song, kun-qing, and Yi-zhou Sun, "Real-Time Short-Term Traffic Flow Forecasting Based on Process Neural Network.".

[7] Guan, W., Cai, X., Wei Guan, Xiaolei Cai, Guan, W., and Cai, X. , " A practical model of dynamic forecasting of urban ring road traffic flow," in *IEEE Conference on Intelligent Transportation Systems*, Proceedings, 2005.

[8] Guozhen Tan, Wenjiang Yuan, and Hao Ding, "Traffic flow prediction based on generalized neural network.," in *IEEE Conference on Intelligent Transportation Systems*, Proceeding, 2004.

[9] Gu, Y. and Yu, L. (, " Study on Short-Time Traffic Flow Forecasting Methods," in *International Conference on Logistics Engineering and Intelligent Transportation Systems*, 2010.

[10] Chan, K. Y., Dillon, T. S., Singh, J., and Chang, E, "Neural-Network-Based Models for Short-Term Traffic Flow Forecasting Using a Hybrid Exponential Smoothing and Levenberg-Marquardt Algorithm," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 644–654, 2012.

[11] Fusco, G., Colombaroni, C., Comelli, L., and Isaenko, "Short-term traffic predictions on large urban traffic networks: Applications of network-based machine learning models and dynamic traffic assignment models," in *International Conference on Models and Technologies for Intelligent Transportation Systems*, 2015.

[12] Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F.-Y. Y., "Traffic Flow Prediction With Big Data: A Deep Learning Approach.," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 1–9, 2014.

[13] Huang, W., Hong, H., Song, G., and Xie, K., "Deep process neural network for temporal deep learning," In *International Joint Conference on Neural Networks (IJCNN)*, pp. 465–472, 2014.

[14] Huang, W., Song, G., Hong, H., and Xie, K. , "Deep Architecture for Traffic Flow Prediction: Deep Belief Networks With Multitask Learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2191–2201, 2014.

[15] B D Greenshield, "A study of Highway capacity," in *Highway Research Board Proceedings*, 1935.

[16] H. Greenberg, "An analysis of traffic flow," *Oper. Res.*, vol. 7, no. 1, pp. 79–85, 1959.

[17] R. T. Underwood, "Speed, Volume, and Density Relationships: Quality and Theory of Traffic Flow, Yale Bureau of Highway Traffic," *New Haven, CT, USA: Yale Univ. Press*, 1961.

[18] Papageorgiou, M. , Hadj-Salem, H. and Blosseville, J.M., "ALINEA: a Local Feedback control Law for On-ramp Metering," Transportation Research Record, No. 1320, *Transportation Research Board*, Washington, D.C, pp. 58–64, 1991.

[19] Papageorgiou, M., Blosseville, J. M., and Hadj Salem, h. , , "Modeling and Real-Time control of traffic Flow on the Southern Part of Boulevard Peripherique in paris: Part II: Coodinated On-ramp Metering," *Transportation Research*, vol. 24A, no. 5, pp. 361–370, 1990.

[20] "System Wide Adaptive Ramp Metering Algorithm - high Level Design," Final Report, Prepared by NET for Caltrans and FHWA, March 1996.

[21] "Activation function," Wikipedia, 11 7 2018. [Online]. Available: `https://en.wikipedia.org/wiki/Activation_function`. [Accessed 24 7 2018].

[22] Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J, "Learning representations by back-propagating errors," *Nature Publishing Group*, vol. 323, pp. 533–536, 1986.

[23] K. Levenberg, "A Method for the Solution of Certain Non-Linear Problems in Least Squares," *Quarterly of Applied Mathematics*, vol. 2, pp. 164–168, 1944.

[24] D. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *SIAM Journal on Applied Mathematics*, vol. 11, no. 2, pp. 431–441, 1963.

[25] Hänggi, Martin, Moschytz, George S., Cellular Neural Networks: Analysis, Design and Optimization, Zurich: Springer, 2000.

[26] Alireza Goudarzi, Peter Banda, Matthew R. Lakin, Christof Teuscher and Darko Stefanovic, "A Comparative Study of Reservoir Computing for Temporal Signal Processing," *arXiv:1401.2224v1*, 2014.

Graph Theoretical Problems in Traffic Management – A Brief Survey

Nkiediel Alan Akwir[1], Muhindo Kule Mutengi[2], Witesyavwirwa Vianney Kambale[1,3], Jean Chamberlain Chedjou[1] and Kyandoghere Kyamakya[1]

[1]Institute of Smart Systems Technologies, Transportation Informatics Group
Alpen-Adria-Universität Klagenfurt, Klagenfurt, Austria
[2]Institut Supérieur de Techniques Appliquées (ISTA), Goma, DR Congo
[3]Tshwane University of Technology, Pretoria, South Africa

Abstract: This paper proposes a brief survey of graph theoretical problems which are involved in diverse modeling and simulation endeavors in modern advanced Traffic Management. Mostly they are the following: shortest path problems (SPP), shortest past spanning tree problem (SPSTP), travelling salesman problems (TSP) and maximum clique's problems. We present different real-world traffic management applications and scenarios where the graph theoretical problems are heavily involved in the related system modelling, for example the following ones: phase grouping for traffic control, vehicle routing problems or vehicle navigation, just to name few. Each traffic management scenario and/or application is presented with its corresponding graph models (corresponding nodes, corresponding edges and weights), the specific graph theoretical problem involved, and finally, wherever necessary, information about complexity or computational requirements/constraints is discussed or provided.

1 Introduction

Graphs are one of the most used tools for networks representation. It reduces networks representation into a mathematical matrix. Thereby, the nodes and the edges of a graph help to model, visualize and understand specific interactions between entities in a given network. Graphs have several applications in real-world scenarios. In Transportation, traffic flows/processes modeling provide several interesting cases where graph-based modelling is applied: road network topology, area-wide traffic light network, etc.

Transportation and logistics systems engineering does encounter a series of graph theory related problems involved in the solving of specific related problems. Indeed, some of the most popular graph theoretical problems include, amongst others, the "Shortest Path Problem (SPP)" and the "Travelling Salesman Problem (TSP)", just to name few.

The research related to transportation network models took (historically) a very crucial advance when the so-called "Dantzig's simplex method" was introduced for graph optimization [1, 2]. The last-named method is used amongst others for solving linear programming problems. Overall, several problems involving graphs (e. g.: SPP, TSP, maximum cliques, etc.) are solved by using optimization techniques such as the following ones: simplex method, genetic algorithm, ant colony algorithm, and the memetic algorithm.

A road network can be modelled by a graph. Hereby, edges do represent lane segments between junctions or links between locations; and nodes (also called vertexes) do represent either road junctions or road locations where the number of lanes changes or particular locations or towns. Further, road length, travel time, travel cost, and traffic conditions can be (individually or combined) considered as edges/arc weights depending on the special transportation problem under consideration.

An appropriate algorithm or optimization technique is (should be) used to find the either shortest path between two nodes (SPP) or the Hamiltonian cycle (a cycle involving all the nodes of a graph) which has lowest cost (TSP). Both SPP and TSP are involved in modeling and solving a series of key transportation/logistics problems such as traffic control (local and/or wide-area control), vehicle routing problems, vehicle navigation problems, etc., just to name a few.

One of the biggest issues regarding graphs is their complexity. The more we have nodes (refers to the magnitude of the graph) and/or edges (refers to the size of the graphs) the bigger the complexity will be increasing. The complexity related to several graph problems such as the "TSP problem" and the "clique problem" is NP-hard [3]. The big complexity leads to big computation challenges such as too long computation time and high memory consumption, which might be critical issues especially for real-time traffic management related endeavors such as traffic jams detection involving the travelling salesman problem [4].

In this survey, we do present some of the well-known traffic management applications/scenarios involving specific graph theoretical problems. The rest of this paper has the following structure. In Section 2, we present how the traffic control at a junction can be modeled using graph theoretical instruments. In Section 3, we do address the traffic assignment issue, which is an important part of the so-called ”four steps model” and thereby described the involved graph-theoretical problems. Section 4 is devoted to graph theoretical problems exploited by microscopic traffic simulation software tools. Section 5 presents a traffic control scheme over a global (city/urban) area with related involved graph theoretical instruments. Then, Section 6, and respectively Section 7, are devoted to the vehicle routing problem, and respectively vehicle navigation, which do both involve appropriate graph theoretical elements. And finally, Section 8 does present, besides concluding remarks, a comprehensive summary of this review.

2 Road Junction Modelling Based on Graph Theory

The control of traffic flows at a road junction must be properly designed because it does influence traffic congestion management, incidents clearance management, and can lead to a lot of driver's frustrations if not well done. Traffic lights are used to regulate the traffic at a road junction. The green and the red time phases must be setup optimally in order to ensure both maximum throughput and minimum delays while ensuring a smallest possible number stops at the junction.

Graphs can be used/involved in the modelling of traffic flow control at a road junction [5–7]. The junction is first represented with the different directions (they are called ”approaches”: left-turns, straight-turns, and right-turns) in such a way that that one can derive both conflicting and non-conflicting directions and/or phase groups (see Figure 1).

We do model the junction in form of a graph $G = (V, E)$, where V and E are respectively the set of nodes and that of edges, which are connected using following rule: the approaches are represented by nodes and edges do connect two nodes if and only if their corresponding directions are non-conflicting (and therefore are compatible directions). The graph $G(V, E)$ is therefore called ”compatible” or ”compatibility graph”; the compatible graph corresponding to Figure 1 is shown in Figure 2.

From the compatibility graph, we can derive a series different cliques (that is, subsets that are complete graphs). Those cliques correspond to the so-called ”phase groups”. A phase group is a set of directions that are non-conflicting and can therefore be assigned the same traffic light phase (green, red) [8]. The

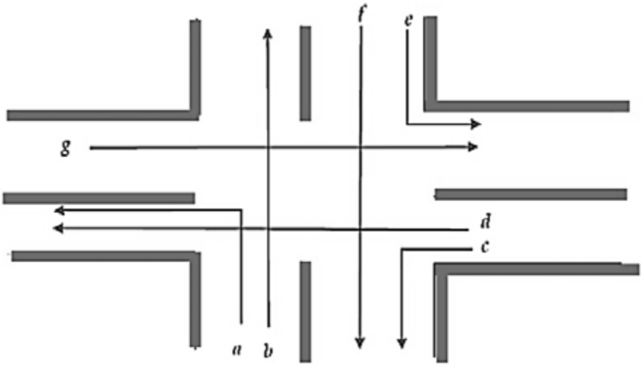

Fig. 1: A road traffic junction (an illustrative example) with its different approaches. The arrows indicated in this figure's caption show the different turns.

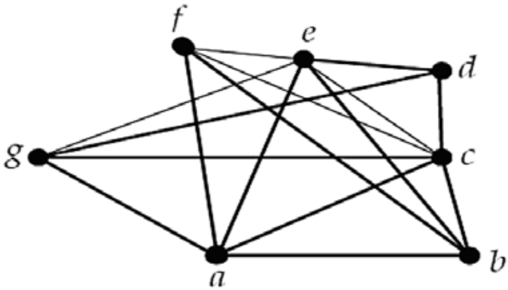

Fig. 2: The compatibility graph (G) corresponding to the road traffic junction in Figure 1.

following cliques are obtained from the graph G of Figure 2: $A = a, b, c, e, f$; and $B = c, d, e, g$.

In the case of a given compatibility graph, some specific algorithms are used to derive the different cliques, which do thereby provide the optimal phase grouping. Having defined the different phase groups, the phase group timings duration setting is formulated as an optimization problem and solved by using e. g. linear programming (or other appropriate algorithms) in order to find the optimal green phase durations, which do ensure an efficient and effective flow throughput of each the different traffic flow directions through the junction.

3 Traffic Assignment (Involving the so-called "Four-Steps Model") Involving Graph Theory

Graph theory is extensively exploited in traffic assignment. Thereby, it is part of the so-called of the "4-steps model" used in urban transportation planning for an efficient and effective use of the transportation infrastructures. Let us recall first the four-steps model [9, 10].

An adequate urban transportation planning is done by considering trips and/or travel forecasting models. The predicted trips are then used to predict traffic flows on different road segments, which are further involved in determining/-fixing a series of transportation infrastructure related parameters/issues: road capacity needs, transit service changes, definitions of strategies and policies to regulate the traffic, etc.

Trip/travel demand modelling involves a lot of variables in form of mathematical equations, which try to mimic the human behavior while traveling (steps taken to make decisions of travelling). Many assumptions are made concerning their choice of a transportation mode. The area under study is divided into so-called "zones" (i. e. locations; thus, the process is called "zoning") before using the four steps models.

The 79 wards of Dhaka city corporation shared into 10 zones known as TAZ (Traffic Analysis Zone).

The nodes in the graph of figure 5 are the traffic analysis zones (TAZ) of the study area (figure 3). The undirected edges in the graph of figure are the main roads which connect a zone with its neighboring zones. The arc weights are obtaining by computing the generalized travel cost (GTC) from travel time and cost from one zone to another.

The four-step process consists of the following 4 steps [10]:

Step 1 - **Trip generation:** Predict the number of trips originating or destinated to a specific zone by exploiting a series of socio-geographical information such as household characteristics.

Step 2 - **Trip distribution:** This step allows the derivation of the origin-destination matrix (O-D matrix), which is a trips table that shows the number of travels from each origin to each destination for a given hour amongst 24 hours of the day.

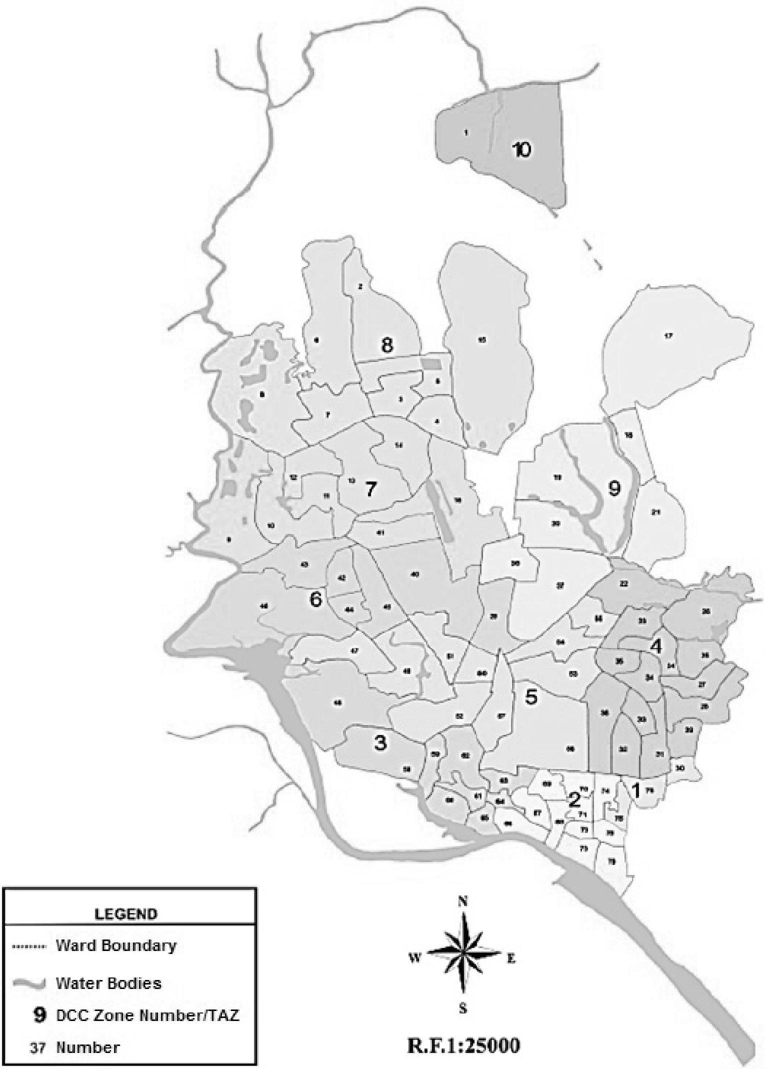

Fig. 3: Sample city zoning [10].

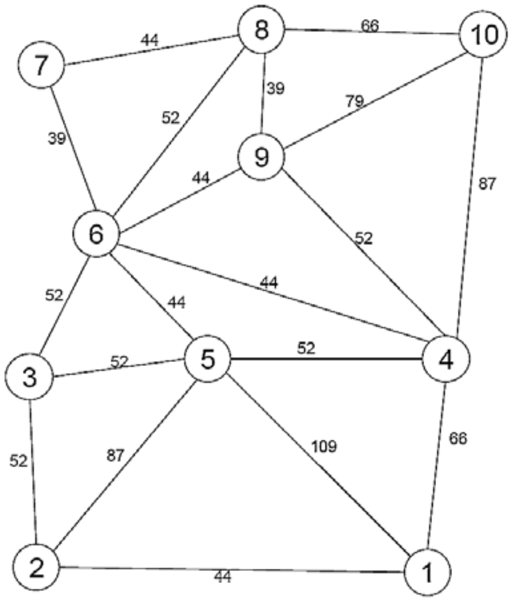

Fig. 4: The corresponding graph of the city zoning represented in Figure 3 [10].

Step 3 - Modal split: Also called "mode choice", it does tell which mode of transport (or which proportion of all travelers for a given O-D pair) will be used to travel from a given origin to a given destination. Travel time and travel costs are used here to derive a so-called "utility function" for each relevant available mode.

Step 4 - Traffic assignment: Also called "route choice", it refers to the selection of (shortest/best/practicable) routes (or paths) between given origin and destination pairs in the road network. Appropriate graph theoretical algorithms are used here to determine the paths that travelers do/can/will use between origins and destination pairs.

It is very important to predict the number of trips which are assigned to the different paths for the reason of optimizing the use of the transportation infrastructure. The shortest path is calculated from each origin to all destinations by exploiting usually the "travel times" as arcs weights; on this basis it is then possible to derive specifically the "minimum-time" path. The trips for each O-D pair are assigned to the links contained in the respective "minimum paths". For every path, the trip will be added-up for each involved link. As the links have an assigned volume (or capacity), if the number of trips exceed the capacity, the speed on that link will decrease and consequently also increase the travel time. Thus, the increase in travel time leads to a possible change of the shortest path. The process is repeated until one reaches the equilibrium between "travel demand" and "travel supply". In another words, trips on congested links will be transferred to uncongested links until one obtains/reaches the equilibrium.

Let us mention that the use of graph theory through the "shortest path problem" (SPP) is one of the most important tasks in this step (i. e. the traffic assignment step). Since the shortest path (SP) is calculated several times, the SP algorithm must be robust and fast to be efficient and effective for the traffic assignment step, especially if this step is performed in real or near real-time; this is the case in modern settings calling for online traffic management (over large city areas). Time-dependent shortest path problems (TDSPP) [11] are also considered for dynamic traffic assignment since the travel time on road links may change over time, for instance, according to daily-base traffic conditions [12].

4 Microscopic Traffic Simulation (e. g. VISSIM)

Traffic simulation is a very important instrument for transportation planning and decision making. Transportation planners, scientists and engineers use these tools to analyze, optimize and predict the traffic for an efficient management of the transportation infrastructures.

Several traffic simulation tools (especially the macroscopic ones) consider traffic flows at a network scale. For this reason, those tools use graph theoretical models and algorithms in the background. This enables them to deal with the networking related behavior of the simulator. Let us mention, for illustration purposes, the commercial tool VISSIM as an example. VISSIM is a microscopic, behavior-based multi-purpose traffic simulation instrument used to analyze and optimize traffic flows. It proposes a wide variety of urban and highway applications, and thereby integrating both public and private transportation [13]. In VISSIM, links and connectors which are mostly considered are roads and are represented in form of edges of the graphs. Also, the intersection between two or more links is considered as a node. A node does is also used when roads or lanes merge, cross and/or split.

For either vehicles or pedestrian paths and route choice, VISSIM uses the "traffic assignment concept" mentioned in the "four-steps model" exploiting the node-edges topology of the underlying road graph. The route choice (based on the shortest path) is based on a linear combination of ”travel time”, ”travel distance” and ”travel cost” (e. g. tolls).

5 Global or Area Traffic Control in a City (Urban Area)

5.1 Optimal Traffic Lights Timing

The first level of traffic management considers traffic control at a single junction. Here, that is for area traffic control, we consider a network of traffic controllers at several road junctions in a given part of a city. The idea proposed by [14] is to coordinate traffic lights at different junctions for optimizing the traffic in a given area/network. For this purpose, graph theory is exploited to find the optimal settings of traffic lights.

The traffic network is transformed into a graph in which nodes represent the road intersections (roads junction) and edges represent the road segments/links. Due to the complexity of the system (interactions of several junctions), in some simple cases, some assumptions are made such as the following ones [14]: all local traffic light controllers (at the different road junctions) have the same cycle time; vehicles are not allowed to turn left or turn right, and all junctions have only two phase groups.

The following information (amongst others) is used as input for the optimization problem for area traffic control:

- The cycle times of controllers at all junctions,
- A distance matrix, which is an adjacency matrix with weights (that is, time units required to move from one junction to another),
- A traffic flow matrix which is an adjacency matrix with weights (that is, the numbers of vehicles moving from one junction to another),
- A Boolean matrix which expresses the two-different lights states (green 1, red 0) at a given time for vehicles moving from one node to another.

This information is used to derive an objective function in which we must find the optimal traffic lights settings which minimize the number of stops considered as penalty of vehicles for the whole network and hence, increase the throughput through the network.

5.2 Optimal Routes for an Automated Traffic Management System

Consider a network of roads in which intersections are considered as nodes and the roads as edges [15]. For the edges, we consider the travel time as weight instead of the length due to the traffic jam. The travel time is dynamic and can be predicted depending on both the previous and the actual traffic situation.

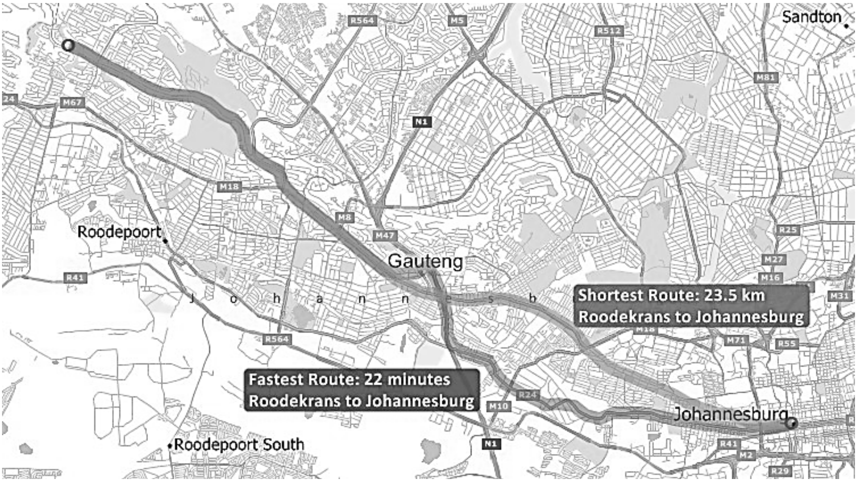

Fig. 5: Maptitude South Africa Route Mapping Software [16].

Assuming the road graph is in form of a grid of NxM nodes, the shortest path is computed between two extreme nodes (origin and destination). When a trip involves a lot of different stops, the best way to visit them is also computed [16].

The shortest path can be implemented in in such a system based on a distributed computation in which we have the following elements:

- A Client (vehicles) which sends queries to a server and thereby provide source and destination; it then receives an optimal path from the server;
- A host server for historical data and the city map in form of a road graph; this graph is used to compute the shortest paths; and finally

- Sensors devices to collect information in real-time, which is necessary to update the server database (in the meantime updating the edges weights according to the traffic situation in real-time, which leads to the path updates).

6 Vehicle Routing Problems

6.1 Congested/Accidental Network

The Vehicle Routing Problem (VRP) is a combinatorial optimization problem, which tries to answer the following question: "What is the optimal set of routes for a fleet of vehicles to traverse in order to deliver to a given set of customers?"

Road traffic faces a lot problem such as congestions, accidents and incidents, just to name a few. Despite the fact that these problems might be local, they may nevertheless affect the overall traffic network and thereby lead to socio-economic and environmental problems (e. g.: increase of travel time, higher fuel consumption, etc.) [17]. Efficient vehicles routing (VRP) and scheduling contribute to improve the overall traffic situation when, for instance, traffic congestion occurs due to accidents, construction, work zones or environmental events. Thereby, vehicle routing to safer roads segment and intersections can be used as a strategy to increase safety locally or globally (at network scale).

Graph theory is extensively used in the vehicle routing problems. Shortest Path (SP), Spanning Tree (ST) or Travelling Salesman Problem (TSP) algorithms are applied on road networks transformed into graphs to find safer/better routes whenever incidents occur in the road network [19].

As already mentioned, generally, road segments are considered as edges and intersections are taken as nodes. The edge cost is considered as distance and/or travel time. Edge costs used to be static for most cases of vehicle routing problems. But nowadays with very complex and uncertain road transportation contexts with changing travel times, dynamic edges (time-dependent) are the mostly used [20].

The modelling of vehicle routing problems exploits more constraints and functions such as the probability of crash based on traffic crashes records, maximum travelled distance, time windows, and customers prioritizing [21, 22].

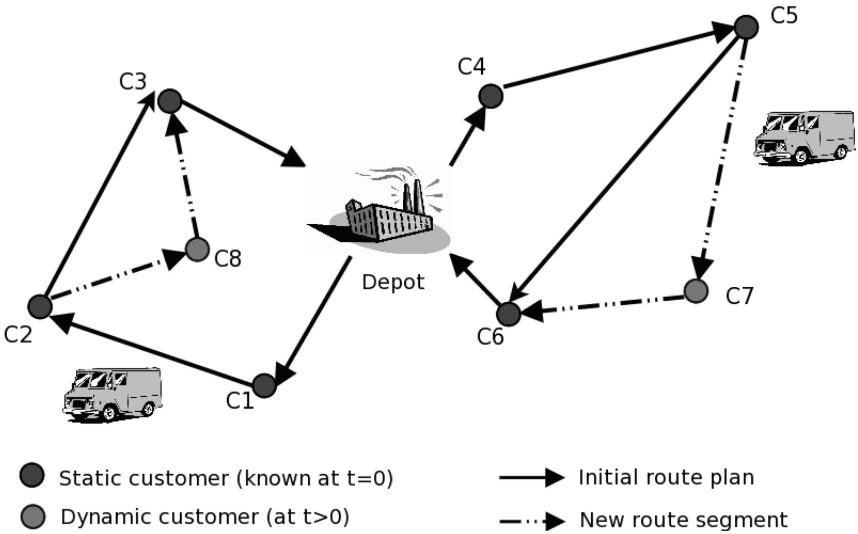

Fig. 6: A sample of routing problem: Dynamic VRP with dynamic customer [18].

Beyond the application example mentioned above, the vehicle routing problem has a lot of applications, especially in logistics and fleet management.

VRP has different variants such as the following ones [23]: Vehicle Routing Problem with Pickup and Delivery (VRPPD), Vehicle Routing Problem with Time Windows (VRPTW), Capacitated Vehicle Routing Problem (CVRP).

As the vehicle routing problem is an optimization problem; the objective function can be very different depending on some specific applications (e. g.: minimize the global distance travelled, minimize the number of vehicles necessary to serve consumers, minimize the penalties for low quality of service, etc.).

6.2 Smart City and Logistics

The deployment of smart cities and logistics will exploit more autonomous and semi-autonomous vehicles for both human and goods mobility [24]. The concept smart city does refer to sustainable cities, where the use of machine-to-machine communications, broadband connectivity anywhere and anytime, or big data analytics will be ordinary. Alongside the smart city technological needs, the autonomous vehicles will need efficient and effective (fast and accurate) route planning systems for better human and good mobility.

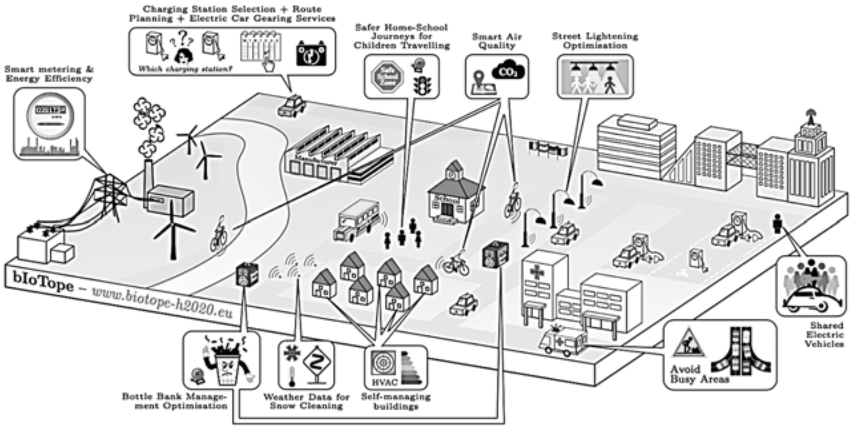

Fig. 7: Overview of the "biotope" project which involves efficient traffic management using smart object [25].

Graph theory is used for vehicle routing problems within a smart city through the shortest path, the spanning tree, and the travelling salesman problems. As an example, see [26], route planning may be done based on the processing of GIS (Geographic information System) road data combined with given graph theory algorithms.

7 Vehicle Navigation

7.1 Principle

Vehicle navigation system is a part of vehicle controls or a third-party add-on/service, which helps to find directions (towards destination) for a vehicle while driving [27]. It is very helpful to save drivers from getting lost when travelling in the dark, on a road that does not look familiar or just to give the shortest way from one point to another. It increases safety and decreases both travel time and fuel consumption. The navigation system has two important parts which are the following:

- The localization system: It uses localization technology like GPS (Global position system), which is most used nowadays, to locate a vehicle on a road network.
- The routing system: It proposes the shortest path from one point to another. In some cases, however, a route involving multiple destinations may be necessary (see VRP).

By combining the two parts, the driver always locates himself on the path from his origin on the way to his destination. The routing system of the navigation part is based on graph theory where either SP or TSP algorithms are involved to find the optimal path. The graph is made up, as mentioned before, by considering locations and road intersections as nodes and roads segments as edges.

Usually, the cost of edges (weights) considered are either travel time or travel distance. However, more advanced navigation systems involve more information (maybe from an integration of infrastructure and vehicle based intelligent transportation systems) of the real-time traffic situations such as the traffic density, traffic rules (restricted roads, blocked roads du to incidents/accidents, etc.) and proposes/calculate a best shortest path/tour [4] accordingly.

7.2 Navigation Vehicles for/in Smart Cities

As mentioned above, vehicle routing is and will play a crucial role for human and goods mobility in smart cities. Thus, vehicle navigation systems will be used even more. Autonomous vehicles will need efficient navigation systems to always follow the direction of its assigned routes. The aim is not only to increase the mobility, but in a global view, to optimize the overall traffic flows.

8 A Comprehensive Summary of Modeling and Computational REQUIREMENTS for Graph Theoretical Problems (SP, TSP, VRP, etc.) in Traffic Management

Table 1 summarizes the different traffic management scenarios. Their graph theoretical problems are explained according to corresponding applications. The nature of the graph (type and size) is also presented to give an idea about the complexity of the problem (computational requirements). We also propose some selected literatures for more details about each application.

Table 1: A brief overview of traffic management scenarios involving a modeling and computational requirements for graph theoretical problems. (Abbreviations: NP – non-polynomial; SP – shortest path problem; TSP – traveler salesman problem; TDSP – time dependent shortest path problem)

	Traffic Management Scenario	Nature (form, size, weight) of the Graph	Graph Theoretical Problem to be solved + Specific Requirements and Challenges	Selected Related Papers (References)
1	*Traffic Junction Modelling* **Nodes** are Approaches **Edges** connect non-conflicting or compatible approaches A **phase group** corresponds to a clique	**Nature:** Compatibility graph (undirected, connected graph) **Graph size:** Magnitude of 4 to 15 **Edges Weights:** Unity	Clique Problem. Challenge: an NP hard problem	[5] [8]
2	*Four Steps Model (Traffic Assignment)* The Graph is obtained from a city map GIS: **Nodes:** Locations **Edges:** Roads connecting locations **Weight** are road distances or time duration to travel between two connected locations	**Nature:** Undirected, connected graph **Size:** Magnitude of 10 up to more than 100 (for large urban areas) **Edges Weight:** Distance (constant weight), traffic conditions (time-dependent weights)	SP, TDTSP	[10] [12]
3	*Microscopic Traffic Simulation (VISSIM)* **Nodes:** Intersections (abstract nodes) **Edges:** links + connectors (traffic conditions on the road, travel time, travel distance)	**Nature:** Undirected, connected graph **Graph size:** Magnitude of 10 to hundreds of nodes **Weight:** Distance (constant weight), traffic conditions (time dependent weight), traffic history conditions and financial cost (Tolls)	SP, TDTSP	[28] [13]

	Traffic Management Scenario	Nature (form, size, weight) of the Graph	Graph Theoretical Problem to be solved + Specific Requirements and Challenges	Selected Related Papers (References)
4	*Global/Area Traffic Control in a city* Topology study	**Nature:** Undirected, connected graph **Graph size:** Magnitude of 10 to 20 nodes **Edges Weight:** Distance, travel time (constant weight), variable (traffic conditions)	SP, TSP	[29] [15]
5	*Vehicle Routing Problem*	**Nature:** Undirected, connected graph **Graph size:** Magnitude of 10 to 20 nodes **Graph Weight:** Distance (constant weight), traffic conditions (time dependent weight), traffic history conditions and financial cost (Tolls) and time window	TSP, TDTSP, SP, SPST	[17] [4]
6	*Vehicle Navigation (timed shortest path problem)* **Weights** $= f(t)$	**Nature:** Undirected, connected graph **Size:** Around 10 to 20 nodes **Weight:** Multiple edges costs (linear combination)	SP, TSP, TDTSP	[30] [4]

9 Concluding Remarks

Graphs are one of the best modelling approaches for real word problems. As mentioned before, it efficiently emphasizes the interaction or relationship between entities in a network. Traffic flow is one of several applications involving networks. This paper has comprehensively reviewed the use graph theoretical concepts for an effective and efficient management of a traffic flow networks and thus improve travel times by avoiding traffic jams, accidents, incidents and thereby increasing road safety.

Graphs which are represented by nodes and edges: we have shown that according to different applications we can propose different types of weights for edges such as fixed weight with constant value or time-dependent weight conforming to traffic situations provided by information from the traffic management centers. Referring to certain applications such as vehicle routing problem, the nodes also might be assigned values to express their capacity of their processing time-windows. These information details do increase the complexity of the network.

We have also mentioned the complexity involving graph problems. The worst well-known one is the NP-hard complexity which brings practical computational challenges when computing TSP or Clique problems, especially when the magnitude and the size of the graph are big and for real-time applications.

Further, we have reviewed a series of important well-known applications of graph theoretical problems in traffic management. We have presented each application by mentioning its scenario, its corresponding graph (size, nature of edges) and the specific graph theoretical problem exploited in the application. Several applications exploit travel-time instead of distance as weight for the shortest path calculation which considers more traffic conditions to find the shortest path. For some efficient modelling contexts arc weights are more complex, e. g. a linear combination of travel time, travel cost and distance.

References

[1] G. Dantzig, Maximization of a linear function of variables subject to linear inequalities (Chap. XXI), in *Activity Analysis of Production and Allocation*, New York: Wiley, 1951.

[2] D. G.B, Application of the simple method to a transportation problem, New York: Wiley, 1951.

[3] S. Cook, "The complexity of theorem proving procedures", in *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, Shaker Heights, Ohio, 1971.

[4] Taesu Cheong and Chelsea C. White, "Dynamic Traveling Salesman Problem: Value of Real-Time Traffic Information", *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, vol. 13, no. 2, pp. 619–630, 2012.

[5] Ekky Kurnia Setiawan, and I. Ketut Budayasa, "Application of graph theory concept for traffic light control at crossroad," 2017.

[6] Th.Riedel and U.Brunner, "Traffic control using graph theory", *Control Engineering Practice*, vol. 2, no. 3, pp. 397–404, 1994.

[7] Arun kumar Baruah, "trafic Conttrol Problems using Graph Connectivity", *International Journal of Computer Applications*, vol. 86, no. 11, pp. 1–3, 2014.

[8] Darshankumar Dave, Nityangini Jhala , "Application of Graph Theory in Traffic Management", *International Journal of Engineering and Innovative Technology (IJEIT)*, vol. 3, no. 12, 2014.

[9] "Handbook of Transport Modeling", in *The four step model*, Irvine, Institute of Transportation Studies, 2007.

[10] B. Ahmed, "The Traditional Four Steps Transportation Modeling Using Simplified Transport Network: A Case Study of Dhaka City, Bangladesh", *IJASETR* , vol. 1, no. 1, 2012.

[11] Abdelfattah Idri, Mariyem Oukarfi, Azedine Boulmakoul, Karine Zeitouni and Ali Masric, "A new time-dependent shortest path algorithm for multimodal transportation network," in *Procedia Computer Science*, 2017.

[12] Yi-Chang Chui, Jon Bottom, michael Mahut, Alex Paz, Ramachandran Balakrishna, travis Waller and jim Hicks, "Dynamic Traffic Assignment", Transportation Research Board, Washington, 2011.

[13] M. Fellendorf and P. Vortisch, "Microscopic Traffic Flow Simulator VISSIM", in *Fundamentals of Traffic Simulation*, Barcelona, Springer, 2010, pp. 63–93.

[14] Shiuan-Wen Chen, Chang-Biau Yang and Yung-Hsing Peng, "Algorithms for the Traffic Light Setting Problem on the Graph Model".

[15] Y. Tan, "Analyzing Traffic Problem Model with Graph Theory Algorithms," in *Science and Information Conference*, London, 2015.

[16] "Maptitude South Africa Route mapping Sotware", [Online]. Available: `https://www.caliper.com/maptitude/route_mapping_software/route-planning-software-south-africa.htm`. [Accessed 16.7.2018].

[17] A. Omidvar, E. E. Ozguven, O. A. Vanli, "A Two-Phase Safe Vehicle Routing and Scheduling Problem: Formulations and Solution Algorithms".

[18] ”roadME: Fundamentals for Real World Applications of Metaheuristics: The vehicular case,” [Online]. Available: `http://roadme.lcc.uma.es/problems.html`. [Accessed 16.7.2018].

[19] Gitae Kim, Yew Soon Ong, Taesu Cheong, and Puay Siew Tan, ”Solving the Dynamic Vehicle Routing Problem Under Traffic Congestion”, *EEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, vol. 17, no. 8, pp. 2367-2380, 2016.

[20] Li Yanfeng ; Gao Ziyou ; Li Jun, ”Vehicle routing problem in dynamic urban traffic network,” in *Service Systems and Service Management (ICSSSM), 8th International Conference on*, Tianjin, 2011.

[21] Gustavo Alfredo Bula, Fabio Augusto Gonzalez, Caroline Prodhon, H. Murat Afsar and Nubia Milena Velasco, ”Mixed Integer Linear Programming Model for Vehicle Routing Problem for Hazardous Materials Transportation”, in *IFAC Conference on Manufacturing Modelling, Management and Control*, Troyes, 2016.

[22] Ali Kourank Beheshti, Seyed Reza Hejazi and Mehdi Alinaghian, ”The vehicle routing problem with multiple prioritized time windows: A case study”, *Computers & Industrial Engineering*, vol. 90, pp. 402–413, 2015.

[23] ”Vehicle routing problem”, Wikipedia, [Online]. Available: `https://en.wikipedia.org/wiki/Vehicle_routing_problem`. [Accessed 25.2.2018].

[24] Jose M. Gutierrez, Tahir Riaz, ”Applied Graph Theory to Real Smart City Logistic Problems”, in *Complex Adaptive Systems*, Los Angeles, 2016.

[25] ”Biotope: Building An ioT Open innovation ecosystem for connected smart objects”, [Online]. Available: `https://biotope-project.eu/overview`. [Accessed 16.7.2018].

[26] I. Okhrin, ”Vehicle routing problem with real-time travel times”, *Int. J. Vehicle Information and Communication Systems*, vol. 2, no. 1/2, pp. 59–77, 2009.

[27] ”Automotive navigation system”, Wikipedia, [Online]. Available: `https://en.wikipedia.org/wiki/Automotive_navigation_system`. [Accessed 26.2.2018].

[28] K. Malecki, ”Graph Cellular Automata with Relation-Based Neighbourhoods of Cells for Complex Systems Modelling: A Case of Traffic Simulation”, *Symmwtry*, vol. 9, no. 322, 2017.

[29] Maciej Kurant and Patrick Thiran, ”Extraction and analysis of traffic and topologies of transportation networks”, *PHYSICAL REVIEW E*, 2006.

[30] Fuquan Pan, Lixia Zhang, and Fengyuan Wang, ”Algorithm for Optimal Path Accounted for Traffic Rules in Vehicle Navigation System”, in *International Conference on Industrial and Information Systems*, 2009.

Index of Authors

Die Reihen der Fortschritt-Berichte VDI:

1 Konstruktionstechnik/Maschinenelemente
2 Fertigungstechnik
3 Verfahrenstechnik
4 Bauingenieurwesen
5 Grund- und Werkstoffe/Kunststoffe
6 Energietechnik
7 Strömungstechnik
8 Mess-, Steuerungs- und Regelungstechnik
9 Elektronik/Mikro- und Nanotechnik
10 Informatik/Kommunikation
11 Schwingungstechnik
12 Verkehrstechnik/Fahrzeugtechnik
13 Fördertechnik/Logistik
14 Landtechnik/Lebensmitteltechnik
15 Umwelttechnik
16 Technik und Wirtschaft
17 Biotechnik/Medizintechnik
18 Mechanik/Bruchmechanik
19 Wärmetechnik/Kältetechnik
20 Rechnerunterstützte Verfahren (CAD, CAM, CAE CAQ, CIM ...)
21 Elektrotechnik
22 Mensch-Maschine-Systeme
23 Technische Gebäudeausrüstung

ISBN 978-3-18-3**86210**-8