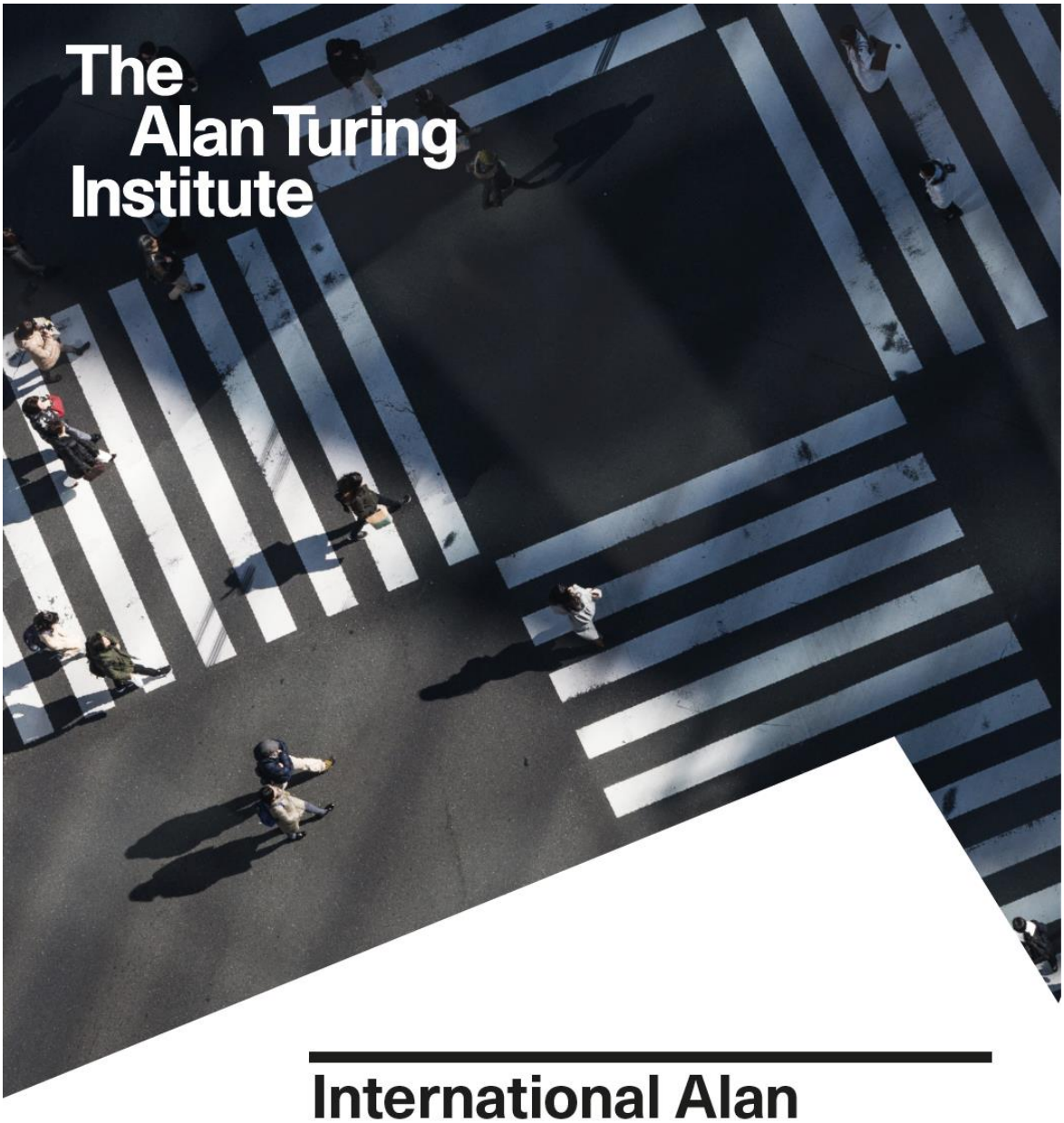


International Alan Turing Conference on Decision Support and Recommender Systems (DSRS-Turing'19)

Proceedings DSRS-Turing'19. London, 21-22nd Nov, 2019

Iván Palomares (Editor)

The Alan Turing Institute. London, United Kingdom.

ISBN (online): 978-1-5262-0820-0

Preface - Conference Chair's Message

The 1st Edition of the International 'Alan Turing' Conference on *Decision Support and Recommender Systems* (DSRS-Turing, 2019) successfully took place at the Alan Turing Institute, London, UK, on 21-22nd November 2019.

This conference was organised at the Alan Turing Institute in central London. The event provided a discussion forum both UK-wide and internationally. It attracted over 80 attendees from academic and research institutions, business and industry firms across 9 countries. The event focused on the latest research advances and challenges involving complex decision-making in real life scenarios, as well as cutting-edge decision support approaches based on AI and Data Science techniques to help alleviate the complexity in such (sometimes arduous and highly uncertain) decision making processes. Our goal was to bring together international researchers, industry professionals and domain experts on Data Science and AI, alongside leaders in the field at the Turing Institute. We discussed the latest trends and ongoing challenges in the aforesaid areas of research involving decision aid, reaching out to diverse disciplines and current societal challenges, e.g. Health and Wellbeing, Sustainability, Finance, Urban Analytics and Planning, e-government and, with a particular emphasis, Personalisation via Recommender Systems.

Six renowned speakers based across Europe accepted to give plenary talks not only on Decision Making predicated on AI and Machine Learning, but also on Human Decision Making in domains such as tourism, Decision Support Systems for surveillance and security, and Recommender Systems for music listeners, nutrition and wellbeing. We also introduced a side activity for young researchers, consisting of a call for extended abstracts led by students and early stage researchers. This initiative was successful and 14 papers have been accepted for oral presentations, posters and demos at the Alan Turing Institute, thereby providing a unique opportunity for these young researchers to discuss their ideas and network with leading experts in DSRS, AI and Data Science. These proceedings constitute a compilation of these 14 promising works. In addition, given the growing importance of protecting and making a fair and transparent use of citizens' data in a world where both human and machine-led decisions are increasingly data driven, as well as "explaining why the decisions made have been made", we had a remarkable panel with experts on *Ethics, Explainability and Interpretability in Decision Support and Recommender Systems* to give the perfect wrap-up to the conference.

I am very grateful to the Alan Turing Institute, with a special mention to the events and communications team, for supporting this project, and *no less grateful to the conference co-organisers and members of our research team at Bristol, for their generous help in organising this event – it has been the very first experience of this kind for all of us including myself, and without your valuable help, it would not have been possible to make this event a success, so thank you* ***James Neve, Hugo Alcaraz-Herrera*** *and* ***Benjamin Arana.*** I would also like to thank the six phenomenal speakers from various international institutions – **Prof. Peter Flach, Prof. Francisco Herrera, Prof. Mounia Lalmas, Dr Julia Neidhardt, Dr Christoph Trattner and Dr Matthijs Spaan** - who agreed to join us and share their expertise and latest research ideas at the conference. Of course, special thanks also to **Christina Hitrova**, Digital Ethics Researcher at the Turing, for proposing and coordinating an Expert Panel on Ethics, Explainability and Interpretability in DSRS: thanks to panellists **Dr David Leslie, Dr Ewa Luger, Prof Francesca Toni** and **Dr Florian Ostmann**. Last but not least, many thanks to all the authors and presenters of the accepted contributions at the conference which greatly enriched our technical programme, and thanks to you, conference attendee, for your genuine interest in this conference and for contributing to make it possible.

Yours sincerely,

Iván Palomares Carrascosa, General Chair of DSRS-Turing 2019.

University of Bristol. The Alan Turing Institute (UK)

Acknowledgments to the DSRS-Turing Programme Committee

We would like to give our most sincere thanks to those experts and colleagues in decision making, decision support systems and recommender systems, who kindly helped peer-reviewing the young researchers' contributions submitted to the conference, providing them with valuable and constructive comments to improve their research contributions.

- Hugo Alcaraz-Herrera (University of Bristol, UK)
- Benjamin Arana (University of Bristol, UK)
- Alejandro Bellogin (Universidad Autonoma de Madrid, Spain)
- Katharina Burger (University of Bristol, UK)
- Ivan Cantador (Universidad Autonoma de Madrid, Spain)
- Cristobal J. Carmona-del Jesus (Universidad de Jaen, Spain)
- Macarena Espinilla (Universidad de Jaen, Spain)
- Frank Hopfgartner (University of Sheffield, UK)
- Anna Jurek-Loughrey (Queen's University Belfast, UK)
- Mesut Kaya (Technological University (TU) Delft, Netherlands)
- Daniel Kershaw (Elsevier, UK)
- Sergey Kovalchuk (ITMO University, Russia)
- Zhiwei Lin (Ulster University, UK)
- Victoria Lopez (Universidad Complutense de Madrid, Spain)
- Eugenio Martinez-Camara (Universidad de Granada, Spain)
- Javier Medina (Universidad de Jaen, Spain)
- Rosana Montes (Universidad de Granada, Spain)
- Julia Neidhardt (Technological University (TU) of Wien, Austria)
- James Neve (University of Bristol, UK)
- Ivan Palomares (University of Bristol & The Alan Turing Institute, UK)
- Massimo Quadrana (Pandora Media, Italy)
- Oscar Gabriel Reyes-Pupo (Universidad de Córdoba, Spain)
- Raul Santos-Rodriguez (University of Bristol, UK)
- Hanna Schäfer (Technical University of Munich, Germany)
- Marko Tkalcic (Free University of Bolzano, Italy)
- Zhen Zhang (Dalian University of Technology, China)

On behalf of the organising committee, thank you!

Ivan Palomares Carrascosa, Conference Chair.

James Neve, Hugo Alcaraz-Herrera and Benjamin Arana. Co-organisers.

Table of Contents

A group decision-making procedure where agents with different expertise evaluate alternatives through qualitative assessments

José Luis García-Lapresta[1], Raquel González del Pozo[2]

[1] *PRESAD Research Group, BORDA Research Unit, IMUVA,*
Departamento de Economía Aplicada, Universidad de Valladolid,
Avenida Valle de Esgueva 6, Valladolid, SPAIN .
[2] *PRESAD Research Group, IMUVA,*
Departamento de Economía Aplicada, Universidad de Valladolid, SPAIN,
Avenida Valle de Esgueva 6, Valladolid, SPAIN .

[1] *lapresta@eco.uva.es*, [2] *raquel.gonzalez.pozo@uva.es*

ABSTRACT

Some decision-making problems use ordered qualitative scales formed by linguistic terms to evaluate alternatives or the knowledge of experts. Sometimes these scales can be considered as non-uniform, in the sense that agents may perceive different proximities between the terms of the scale. In this contribution, we propose a group decision-making procedure for ranking alternatives evaluated by a group of experts through a non-necessarily uniform ordered qualitative scale. To assign a weight to each expert according to their expertise, a decision-maker assesses the experts' expertise by means of another ordered qualitative scale. In the procedure, each of the two ordered qualitative scales is equipped with an ordinal proximity measure that collects the ordinal proximities between the linguistic terms of the scale. The procedure assigns scores to the linguistic terms of the scales taking into account the ordinal proximity measures of the scales. Afterwards, the scores are normalized and aggregated for generating the ranking of the alternatives.

1. INTRODUCTION

Some decision-making problems use ordered qualitative scales formed by linguistic terms to evaluate different aspects of a set of alternatives (e.g. quality control analysis, electoral polls or sensory analysis). For instance, the Pew Research Center conducted a survey on respondents' opinions about the candidates to U.S.A. presidential election in 2016. The respondents showed their opinions through an ordered qualitative scale formed by 5 linguistic terms: {"terrible", "poor", "average", "good", "great"}.

Balinski and Laraki [1] have proposed a voting system called Majority Judgment (MJ). In MJ voters assess candidates in political elections considering the following 6-term ordered qualitative scale {"to reject", "poor", "acceptable", "good", "very good", "excellent"}.

The American Wine Society uses a qualitative scale {"objectionable", "poor", "deficient", "acceptable", "good", "excellent", "extraordinary"} for evaluating some sensory aspects of wines. In the wine-tasting sheet of the Society, each linguistic term is identified with a numerical value: 0, 1, 2, 3, 4, 5 and 6, respectively.

Sometimes these ordered qualitative scales can be considered as non-uniform, in the sense that agents may perceive different proximities between the terms of the scale. For instance, taking into account the scales used by the American Wine Society, some agents may perceive that the term "Excellent" is closer to "extraordinary" than to "good" or that the term "poor" is closer to "deficient" than to "objectionable".

Generally, to manage ordered qualitative scales it is very usual translating qualitative into numerical information. However, these numerical codifications are not appropriate, since sometimes they do not reflect adequately how agents perceive the proximities between the terms, and also they may misrepresent the results coming from these scales (see Merbitz et al. [10], Franceschini et al. [2] and Gadrich et al. [4], among others).

To deal with non-uniform ordered qualitative scales we use the concept of ordinal proximity introduced by García-Lapresta and Pérez-Román [7]. Ordinal proximity measures deal with non-uniform qualitative scales in an ordinal way avoiding numerical codifications to linguistic terms of the scales. Other methods tackle non-uniform qualitative scales following cardinal approaches or fuzzy techniques (see Zadeh [11], Herrera-Viedma

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

and López-Herrera [9] and Herrera et al. [8], among others), that are practically equivalent to the use of numerical values (see García-Lapresta [5]).

On the other hand, in some decision-making problems it is necessary to aggregate the assessments given by a set of experts with different knowledge or expertise. In these cases, the importance of each expert in the collective decision can be expressed by means of weights, through a weak order over the experts or by qualitative assessments given by a decision-maker about the experts (see Franceschini and García-Lapresta [3]).

In this contribution, we propose a group decision-making procedure where a group of experts with different expertise evaluate a set of alternatives through an ordered qualitative scale with the purpose of generate a ranking of the alternatives. The procedure assigns a weight to each expert according to their expertise. To do that, a decision-maker assesses the experts' expertise by means of another ordered qualitative scale. Each of the two ordered qualitative scales is equipped with an ordinal proximity measure that collects the ordinal proximities between the linguistic terms of the scale (see García-Lapresta and Pérez-Román [7] and García-Lapresta et al. [6]). Depending on the considered ordinal proximity measures, the procedure assigns scores to the linguistic terms of the scales by means of a scoring function that represents, as faithfully as possible, the proximities between the linguistic terms. Subsequently, to generate the ranking of the alternatives the scores are normalized and aggregated taking into account the different weights of the experts.

2. PRELIMINARIES

Along this contribution we consider an ordered qualitative scale $\mathcal{L} = \{l_1, \dots, l_g\}$, where $l_1 \prec \cdots \prec l_g$ and $g \geq 3$.

We now recall the concept of ordinal proximity measure, introduced by García-Lapresta and Pérez-Román [4]. An ordinal proximity measure is a mapping that assigns an ordinal degree of proximity to each pair of linguistic terms of an ordered qualitative scale $\mathcal{L}$. These ordinal degrees of proximity belong to a linear order $\Delta = \{\delta_1, \dots, \delta_h\}$, with $\delta_1 \succ \cdots \succ \delta_h$, being δ_1 and δ_h the maximum and the minimum degrees of proximity, respectively. It is important to note that the elements of Δ are not numbers. They are only abstract objects that represent different degrees of proximity.

Definition 1 ([7])**.** An *ordinal proximity measure* (OPM) on $\mathcal{L}$ with values in Δ is a mapping $\pi\colon \mathcal{L} \times \mathcal{L} \to \Delta$, where $\pi(l_r, l_s) = \pi_{rs}$ represents the degree of proximity between l_r and l_s, satisfying the following conditions:

1. Exhaustiveness: For every $\delta \in \Delta$, there exist $l_r, l_s \in \mathcal{L}$ such that $\delta = \pi_{rs}$.
2. Symmetry: $\pi_{sr} = \pi_{rs}$, for all $r, s \in \{1, \dots, g\}$.
3. Maximum proximity: $\pi_{rs} = \delta_1 \Leftrightarrow r = s$, for all $r, s \in \{1, \dots, g\}$.
4. Monotonicity: $\pi_{rs} \succ \pi_{rt}$ and $\pi_{st} \succ \pi_{rt}$ for all $r, s, t \in \{1, \dots, g\}$ such that $r < s < t$.

A prominent class of OPMs, introduced by García-Lapresta et al. [3], is the one of metrizable OPMs which is based on linear metrics on ordered qualitative scales.

Definition 2 ([6])**.** A *linear metric* on $\mathcal{L}$ is a mapping $d\colon \mathcal{L} \times \mathcal{L} \to \mathbb{R}$ satisfying the following conditions for all $r, s, t \in \{1, \dots, g\}$:

1. Positiveness: $d(l_r, l_s) > 0$.
2. Identity of indiscernibles: $d(l_r, l_s) = 0 \Leftrightarrow l_r = l_s$.
3. Symmetry: $d(l_s, l_r) = d(l_r, l_s)$.
4. Linearity: $d(l_r, l_t) = d(l_r, l_s) + d(l_s, l_t)$ whenever $r < s < t$.

Definition 3 ([6])**.** An OPM $\pi\colon \mathcal{L} \times \mathcal{L} \to \Delta$ is *metrizable* if there exists a linear metric $d\colon \mathcal{L} \times \mathcal{L} \to \mathbb{R}$ such that $\pi_{rs} \succ \pi_{tu} \Leftrightarrow d(l_r, l_s) < d(l_t, l_u)$, for all $r, s, t, u \in \{1, \dots, g\}$. We say that π is generated by d.

Definition 4 ([6])**.** An OPM $\pi\colon \mathcal{L} \times \mathcal{L} \to \Delta$ is *uniform* if $\pi_{r(r+1)} = \pi_{s(s+1)}$ for all $r, s \in \{1, \dots, g-1\}$, and *totally uniform* if $\pi_{r(r+t)} = \pi_{s(s+t)}$ for all $r, s, t \in \{1, \dots, g-1\}$ such as $r + t \leq g$ and $s + t \leq g$.

 ISBN: 978-1-5262-0820-0

3. THE DECISION-MAKING PROCEDURE

Consider a set of experts with different expertise $E = \{e_1, \dots, e_m\}$, with $m \geq 2$, that evaluate a set of alternatives $X = \{x_1, \dots, x_n\}$, with $n \geq 2$, through an ordered qualitative scale $\mathcal{L}^a = \{l_1^a, \dots, l_{g_a}^a\}$ equipped with a metrizable OPM π^a: $\mathcal{L}^a \times \mathcal{L}^a \to \Delta^a = \{\delta_1^a, \dots, \delta_{h_a}^a\}$.

In the procedure, a decision-maker assesses the experts' expertise by means of another ordered qualitative scale $\mathcal{L}^e = \{l_1^e, \dots, l_{g_e}^e\}$ equipped with a metrizable OPM π^e: $\mathcal{L}^e \times \mathcal{L}^e \to \Delta^e = \{\delta_1^e, \dots, \delta_{h_e}^e\}$. In order to assign scores to the linguistic terms of ordered qualitative scales, the procedure introduces a scoring function. Subsequently, the scores are normalized following two different paradigms.

Definition 5. Given an ordered qualitative scale $\mathcal{L} = \{l_1, \dots, l_g\}$ equipped with a metrizable OPM π: $\mathcal{L} \times \mathcal{L} \to \Delta$, a *scoring function* on π is a function S: $\mathcal{L} \to \mathbb{R}$ satisfying the following conditions for all $r, s, t, u \in \{1, \dots, g\}$:

1. $S(l_r) < S(l_s) \Leftrightarrow r < s$.
2. $\pi_{rs} \succ \pi_{tu} \Leftrightarrow |S(l_r) - S(l_s)| < |S(l_t) - S(l_u)|$.

Proposition 1. Given an ordered qualitative scale $\mathcal{L} = \{l_1, \dots, l_g\}$ equipped with a metrizable OPM π: $\mathcal{L} \times \mathcal{L} \to \Delta$ and a scoring function S: $\mathcal{L} \to \mathbb{R}$ on π, the mapping d_S: $\mathcal{L} \times \mathcal{L} \to \mathbb{R}$ defined as

$$d_S(l_r, l_s) = |S(l_r) - S(l_s)| \tag{1}$$

is a linear metric on $\mathcal{L}$, and π is generated by d_S.

Proposition 2. Given an ordered qualitative scale $\mathcal{L} = \{l_1, \dots, l_g\}$ equipped with a metrizable OPM π: $\mathcal{L} \times \mathcal{L} \to \Delta$, the scoring function S_0: $\mathcal{L} \to \mathbb{R}$ is defined as

$$S_0(l_r) = C_g + \sum_{s<r} \rho(\pi_{sr}) - \sum_{s>r} \rho(\pi_{rs}), \tag{2}$$

where $\rho(\delta_k) = k$, $C_3 = 10$, and $C_{g+1} = C_g + g + 2$, is a scoring function on π.

Remark 1: If π: $\mathcal{L} \times \mathcal{L} \to \Delta$ is a totally uniform OPM on $\mathcal{L}$, then

1. $S_0(l_1) = C_g - (2 + \cdots + g) = C_g - \frac{g^2+g-2}{2}$.
2. $S_0(l_g) = C_g + (2 + \cdots + g) = C_g + \frac{g^2+g-2}{2}$.

Taking into account Eq. (2), we now introduce two scoring function that normalize S_0 following two different paradigms.

Proposition 3. Given an ordered qualitative scale $\mathcal{L} = \{l_1, \dots, l_g\}$ equipped with a metrizable OPM π: $\mathcal{L} \times \mathcal{L} \to \Delta$, the function S_1: $\mathcal{L} \to \mathbb{R}$ defined as

$$S_1(l_r) = \frac{S_0(l_r)}{S_0(l_g)} \tag{3}$$

is a scoring function on π.

Notice that $S_1(l_g) = 1$.

Remark 2: If π: $\mathcal{L} \times \mathcal{L} \to \Delta$ is a totally uniform OPM on $\mathcal{L}$, then $S_1(l_r) = \frac{r}{g}$, for every $r \in \{1, \dots, g\}$.

Proposition 4. Given an ordered qualitative scale $\mathcal{L} = \{l_1, \dots, l_g\}$ equipped with a metrizable OPM π: $\mathcal{L} \times \mathcal{L} \to \Delta$, the function S_2: $\mathcal{L} \to \mathbb{R}$ defined as

$$S_2(l_r) = \frac{S_0(l_r) - S_0(l_1)}{S_0(l_g) - S_0(l_1)} \tag{4}$$

is a scoring function on π.

Notice that $S_2(l_1) = 0$, $S_2(l_g) = 1$ and, consequently, $S_2(l_r) \in [0,1]$ for every $r \in \{1, \dots, g\}$.

Remark 3: If π: $\mathcal{L} \times \mathcal{L} \to \Delta$ is a totally uniform OPM on $\mathcal{L}$, then $S_2(l_r) = \frac{r-1}{g-1}$, for every $r \in \{1, \dots, g\}$.

The procedure is divided in the following steps:

1. Assign a weight $w_k \in [0,1]$ to each expert according to their expertise. To do that, a decision-maker evaluates the set of experts through a metrizable OPM π^e: $\mathcal{L}^e \times \mathcal{L}^e \to \Delta^e = \{\delta_1^e, \dots, \delta_{h_e}^e\}$. Taking into account Eq. (4) the weights are determined as follows:
$$w_k = \frac{S_2(v_k)}{S_2(v_1) + \cdots + S_2(v_m)}, k = 1, \dots, m, \tag{5}$$
where $v_k \in \mathcal{L}^e$ is the assessment obtained by the expert $e_k \in E$. Notice that $w_1 + \cdots + w_m = 1$.
2. The experts' assessments are collected in a *profile*:

 ISBN: 978-1-5262-0820-0

$$\begin{pmatrix} v_1^1 & \cdots & v_i^1 & \ldots & v_n^1 \\ \ldots & \cdots & \ldots & \ldots & \cdots \\ v_1^k & \cdots & v_i^k & \cdots & v_n^k \\ \ldots & \ldots & \cdots & \cdots & \cdots \\ v_1^m & \ldots & v_i^m & \cdots & v_n^m \end{pmatrix},$$

where $v_i^k \in \mathcal{L}^a$ is the linguistic assessment given by the expert $e_k \in E$ to the alternative $x_i \in X$.

3. Calculate for each alternative $x_i \in X$ a global score taking into account Eq. (3) and the experts' weights introduced in Eq. (5):

$$U(x_i) = \sum_{k=1}^{m} w_k \cdot S_1(v_i^k) \in [0,1]. \quad (6)$$

4. Rank order the alternatives through the following weak order on *X*:

$$x_i \succcurlyeq x_j \Leftrightarrow U(x_i) \geq U(x_j).$$

4. CONCLUSIONS

In this contribution, we have introduced a new group decision-making procedure for ranking a set of alternatives from the opinions of a group of experts with different expertise. To determine the importance of each expert in the decision-making procedure, the experts are evaluated by a decision-maker.

In the procedure, experts and alternatives are assessed by means of ordered qualitative scales equipped with metrizable ordinal proximity measures that collect the ordinal proximities between the linguistic terms of the scales. Taking into account the proximities between the linguistic terms of the corresponding ordered qualitative scales, the procedure uses a scoring function that assigns numerical scores to the linguistic terms.

Once the numerical scores have been obtained, the procedure calculates a global score for each alternative: a weighted average of the normalized scores implicitly given by the experts to the alternatives. The weights associated with the experts are obtained from the linguistic assessments given by a decision-maker after a normalization process. Finally, the alternatives are ranked according to the global scores.

Acknowledgements: The financial support of the Spanish *Ministerio de Economía y Competitividad* (project ECO2016-77900-P) and ERDF is acknowledged.

5. REFERENCES

[1] Balinski M., Laraki R.: *Majority Judgment: Measuring, Ranking, and Electing*. MIT Press, Cambridge, 2011.

[2] Franceschini, F., Galetto, M., Varetto. M.: Qualitative ordinal scales: the concept of ordinal range. *Quality Engineering* **16**, pp. 515-524, 2004.

[3] Franceschini, F., García-Lapresta, J.L.: Decision-making in semi-democratic contexts". *Information Fusion* **52**, pp. 281-289, 2019

[4] Gadrich, T., Bashkansky, E., Zitikis, R.: Assessing variation: a unifying approach for all scales of measurement. *Quality and Quantity* **49**, pp. 1145-1167, 2015.

[5] García-Lapresta, J.L.: Reciprocal and linguistic preferences, in: R. Seising, E. Trillas, C. Moraga, S. Termini (Eds.), On Fuzziness. A Homage to Lotfi A. Zadeh, vol. 1, Springer-Verlag, Berlin, pp. 193-197, 2013.

[6] García-Lapresta, J.L., González del Pozo, R., Pérez-Román, D.: Metrizable ordinal proximity measures and their aggregation. *Information Sciences* **448-449**, pp. 149-163, 2018.

[7] García-Lapresta, J.L., Pérez-Román, D.: Ordinal proximity measures in the context of unbalanced qualitative scales and some applications to consensus and clustering. *Applied Soft Computing* **35**, pp. 864-872, 2015.

[8] Herrera, F., Herrera-Viedma, E., Martínez, L.: A fuzzy linguistic methodology to deal with unbalanced linguistic term sets. *IEEE Transactions on Fuzzy Systems* **16**, pp. 354-370, 2008.

[9] Herrera-Viedma, E., López-Herrera, A.G.: A model of an information retrieval system wit unbalanced fuzzy linguistic information. *International Journal of Intelligent Systems* **22**, pp. 1197-1214, 2007.

[10] Merbitz, C., Morris, J., Grip, J.C.: Ordinal scales and foundations of misinference. *Archives of Physical Medicine and Rehabilitation* **70**, pp. 308-312, 1989.

[11] Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning, Part I. *Information Sciences* **8**, pp. 199-249, 1975; Part II *Information Sciences* **8**, pp. 301-357, 1975; Part III *Information Sciences* **9**, pp. 43-80, 1975.

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Bip4Cast: Some advances in mood disorders data analysis

P Llamocca[1], D Urgelés[2], M Cukic[3], V Lopez[4]

[1,4]Complutense University, Madrid, SPAIN
[2]Hospital Ntra. Sra. de la Paz, Madrid, SPAIN
[3]3EGA, Amsterdam, NEDERLAND

[1]pavellam@ucm.es, [2]diego.urgeles@sjd.es, [3]cukic@3ega.nl, [4]vlopezlo@ucm.es

[1,4]www.ucm.es, [2]www.sjd.es, [3]www.3ega.nl

ABSTRACT

Mood disorders have been a relevant topic for the last decade. According to the World Health Organization, the cost of mood disorders and anxiety in the EU is about €170 billion per year. Bip4Cast is a project aiming at crisis prediction for patients with bipolar disorder. Traditional indicators like Hamilton and Yang are insufficient to predict and avoid crises. This study adds up to 149 new variables from data gathered from different sources as wearable devices and smartphones. The analytics include correlation between all the accelerometer variables and linear regression between variables that come from different sources. The results show the existence of a relationship between biological, psychological, physical indicators with the appearance of a crisis of depression or mania. These relations are the base of the predictive analytics that clinicians need in order to make better decisions on the future treatment plans.

1. INTRODUCTION

Bipolar disorder and depression are chronic and severe mental disorders and a major mental health problem in Europe. The Regional Office for Europe of the World Health Organization, in 2019, reports that 25% of the population suffer from depression or anxiety and that neuropsychiatric disorders account for 26% of the burden of disease in European Union countries. They have a lifetime prevalence rate of 0.6% (0.4%-1.4%) and a high suicide index. These mental disorders can be monitored by mobile devices. They can be useful to improve the efficiency of treatments because they allow the gathering of physical, biological and psychological indicators from patients. The importance of these indicators is discussed in recent publication (Vanello, 2012). As a continuation of this, the project Bip4Cast (bip4cast.org) focus on discovering the relationship among the indicators provided by the data gathered and developing a CAD tool (Lopez, 2016) to support the clinical decisions. These studies seek to find common patterns that could trigger a crisis of mania or depression. Bip4Cast is a project developed together with clinicians and patients from the hospital Ntra. Sra. de la Paz in Spain. Data from patients are collected from wearable medical devices (Actigraph, www.activinsights.com) and integrated into the dataset that psychiatrists collect during a regular consultation. The patients wear a smart band and periodically send the data to the system. In addition, the patients fill in a daily form on smartphones. All these data are collected in a 188-variable data set. This data set is the input of a data integration process that returns a structure with normalized variables and crisis indicators that facilitate a preliminary analysis and the application of a future Machine Learning analysis. This analysis shows how to detect some dependencies and reduce the number of variables after the data integration stage. Besides, data integration is carried out to complement the Young and Hamilton indexes on the mania and depression crisis (Llamocca, 2018; Junestrand, 2018) and to provide clinicians with additional information.

In addition to this introduction, section 2 shows the procedure of the research. Section 3 contains a description of the data and sources of information. In section 4, the integration of the data collected is explained. Section 5 is about the analysis and results and finally, section 6 shows the conclusions and the future work on the project.

 ISBN: 978-1-5262-0820-0

2. PROCEDURE

To accomplish the project and gather the prototype data, 12 patients from the partner hospital are involved in the project. The participation of these patients requires the compliancy with all privacy policies and prior informed consent. After getting all permissions and starting on September 2017, a period of collecting the data was scheduled. Data from patients were gathered using different methods and sources. The next section explains the most relevant sources. In addition, a process of cleaning and integration of the data was also developed by executing algorithms in R. One of the resulting data set, data2.csv, consist of 2328 observations of 188 variables. This data set contains the most relevant variables in the study and also several other indicators (calculated variables). The "crisis indicator" is a categorical variable that indicates the stage of the patient: Mania, Depression or Euthymia (the normal state of a patient). Only in a few observations this indicator is different than NA (not available) value. The set of observations in which the real value is known (set by the doctor during the consultations) will be used to train the system when the number of observations become large enough. Meanwhile the data set is used as a prototype to be utilized for experiments. Junestrand (2018) performed a comparison of the behaviour of several machine learning algorithms on the dataset. The following step is to model the euthymic state of each particular patient. Comparisons between the real state and the euthymic state of the patient are relevant for the prediction of crisis. This study facilitates the decision making for the clinician.

3. SOURCES OF INFORMATION

The most relevant sources of information in this project are described as follows.

- *Interviews.* The patients are observed periodically by a psychiatrist in an interview session. During the session, the psychiatrist registers a total of 38 variables. One of these variables is the "crisis indicator" according to his/her diagnosis. Also relevant HDRS scale (Hamilton Depression Rating Scale) and YMRS scale (Young Mania Rating Scale) are added in these interview sessions.
- *Smartwatches/bands.* Each patient involved in the project has its own medical smart band. This device gathers a set of important indicators (heart rate, physical activity, sleep quality…) and a total of 108 variables. Data are automatically recorded by the device and during the appointment with psychiatrist are periodically collected. (Bellivier, 2014) and (Anchiriaco, 2017) are publications on actigraphy as data source, the latter containing the R code that Bip4Cast has created to import data from the medical bands to the server.
- *Fill-in Forms.* Patients complete an electronic form daily: The Bip4Cast app (Martinez, 2016) or a Google form as alternative. This form was designed as part of the Bip4Cast project for gathering quantitative and qualitative data. The source provides them with 41 variables. Some of them are objective data, e.g. coffee consumption, tobacco or alcohol consumption, periods of menstruation, etc. Other data are subjective data, e.g. anxiety, concentration, comfort, etc.

4. INTEGRATION

Every source output yield data with different structure/formats. This phase deals with centralizing all the variables in a single structure. Some packages of *RStudio* as *tidyverse* are being used for data transformation. The following two goals are already achieved.

- *Building a single structure.* A single structure capable of storing any variable from any source. The new structure is also scalable for allowing to store variables coming from new sources (as voice recording, images and video). Currently, there are 187 variables extracted from sources already incorporated. Table 1 provides information about this structure. One record in this structure corresponds to one observation day for a specific patient.
- *Building an integration process.* This process is capable of treating all formats/structures from any incorporated source. To achieve this goal, several tasks are performed: cleaning, formatting, standardization, interpolation, etc. For categorical variables we use the factor data type in R which is a very useful way to make the analysis shorter in combination with the non-categorical variables.

 ISBN: 978-1-5262-0820-0

Table 1. *Integration Structure*

Group of Variables	Type	Nº of Variables	Source	Frequency
Patient	CHAR	1	All	Daily
Date	DATE	1	All	Daily
Fill-in Forms Parameters	NUMERIC	39	Fill-in Forms	Daily
Smartwatch Parameters	NUMERIC	108	Smartwatches	Daily
Interviews Parameters	NUMERIC	37	Interviews	Each 2 weeks
Crisis Indicator	NUMERIC	1	Interviews	Each 2 weeks
"New Source" Parameters	TBD	TBD	Voice, video	TBD

5. ANALYSIS AND RESULTS

In this study patient's data are individually analysed. As the structure contains all the data from all patients, filtering data per patient before analysing is necessary. One of the first steps is to build the correlation matrix. To remove noise, a Principal Component Analysis (PCA) iss carried out. The similarity between variables can be also visualized in graphics using "*ggplot*" R-package. For example, the relation between variables *animus* and *motivation* (Figure 1) is establish from the fill-in form data. The correlation coefficient between them reaches 85% average (for all patients). This result can be useful to discard one of the variables on the base of high correlation. In relation to the accelerometer features, Pearson's correlations are computed (Figure 2). The results show that the features related to sleep quality are the most relevant for crisis prediction: the maximum positive correlation (0.43) exists between features '*at least_5min_ wake_ at_ night*' and '*duration_day_minutes*'. Strong positive correlations exist between features '**dur_night_min**' and '*dur_day_min*'(0.39); and '*dur_night_sleep*' and '*ACC_tinday_min*' (0.37). Maximal values in negative correlation are between 'acc_onset' and 'during_day_LiG300_min' (-0.23); between 'during_day_OIN30_min' and 'acc_onset'(-0.23); and (-0.32) between '*during_day_OIN30_min*' and '*sleep_efficiency*'; even (-0.33) between '*at least 5 min wake at night*' and '*sleep_efficiency*'.

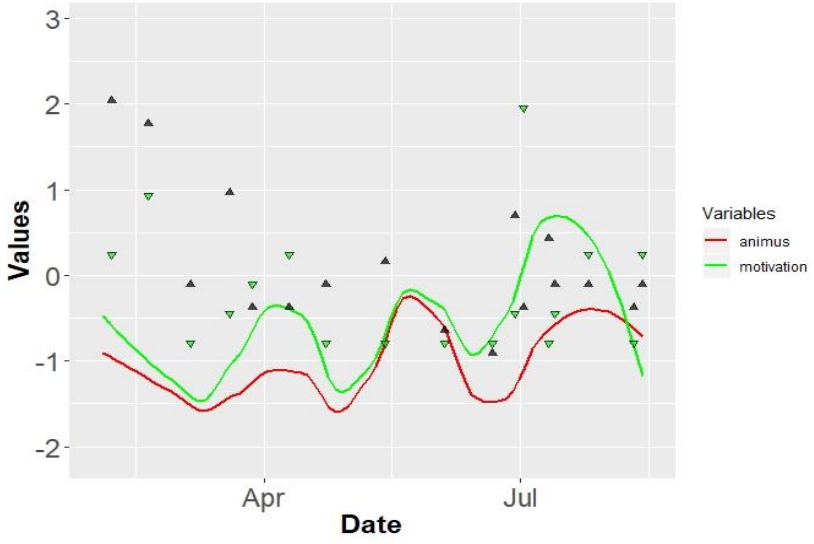


Figure 1. *Similarity between variables: animus (red) and motivation (green)*

Also that it is an implication for decorrelation of the features for further data mining, and especially a practical meaning for clinical psychiatrist. For instance, they could draw additional conclusions about the features most important to understand the dynamics in every particular patient. From data mining point of view, decorrelation is important to decrease the dimensionality of a problem leading to an easier further

 ISBN: 978-1-5262-0820-0

performance in machine learning sense. The data show an example of a good classification based on a smaller number of features once internal relations and connections are revealed. At this point, the study of the entropy measures of the data can be relevant. Cukic et al. (2018, 2019) relate presence of depression of mania crisis to a low entropy levels in de data of mood.

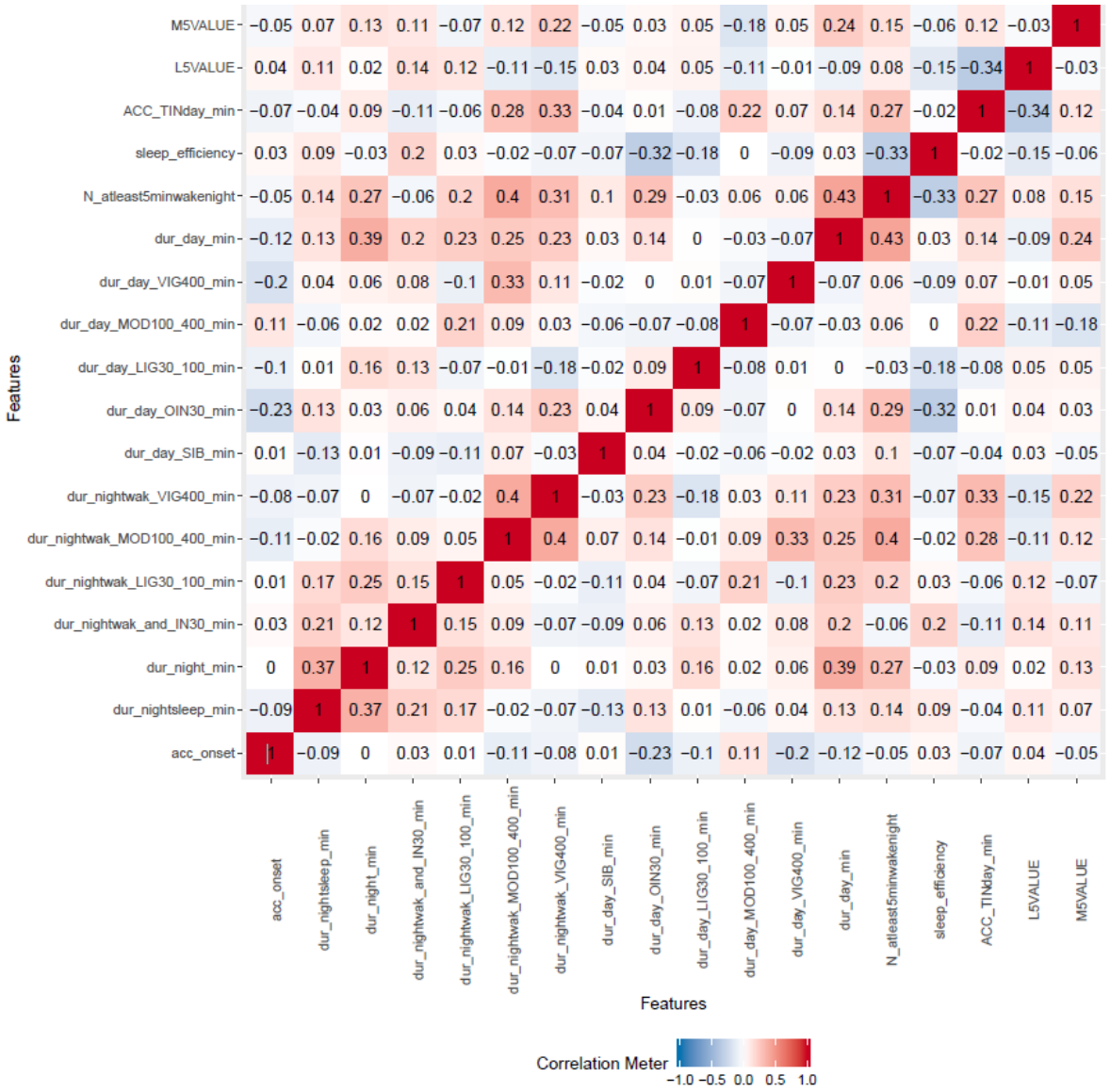


Figure 2. *Correlation between accelerometer features.*

6. CONCLUSIONS AND FUTURE WORK

This work describes the Bip4Cast project on mood disorders analytics. The project is in data analysis phase within real data. Several sources of data from patients are being collected and the behaviour and evolution of relevant variables are in process of analysis to provide predictions that supports the decision making in further medical treatments. As a future work, a prediction model will be developed and some machine learning algorithms will be used. In addition, Bayesian Analysis will be use to aggregate future data and debug the model where previous low mood fluctuations indicate the presence of depression or mania crisis.

Acknowledgements: This work is supported by the Project H2020-MSCA-RISE-2015-690874 (2016-2020).

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

7. REFERENCES

J. Anchiraico (2017), Design of a Big Data architecture for predicting crises in bipolar disorder, Master Thesis, *Eprints Complutense* (https://eprints.ucm.es/41633/)

F. Bellivier et al. (2014), Sleep in remitted bipolar disorder: A naturalistic case-control study using actigraphy, *Journal of Affective Disorders, 158, DOI: 10.1016/j.jad.2014.01.012,* pp. 1-7

M. Cukic et al. (2018), EEG machine learning with Higuchi fractal dimension and Sample Entropy as features for successful detection of depression, *Journal ArXiv Cornell University*, pp. 1-34

M. Čukić Radenković (2019). Novel Approaches in Treating Major Depressive Disorder (Depression), ISBN: 978-1-53614-382-9, NOVA Scientific Publishers ltd. July 2019.

M. Čukić Radenković and V. Lopez Lopez (2019). Machine Learning Approaches for Detecting the Depression from Resting-State Electroencephalogram (EEG): A Review Study. Cornell Repository Arxiv.org. https://arxiv.org/abs/1909.03115

A. Junestrand (2018) Application of Machine Learning Algorithms for Bipolar Disorder Crisis Prediction. Bachelor Thesis, *Eprints Complutense* (https://eprints.ucm.es/48866/)

P. Llamocca, A. Junestrand, M. Cukic, D. Urgeles, V. Lopez (2018). Data Source Analysis in mood disorder research, *Proceedings of the XVIII Conferencia de la Asociación Española para la Inteligencia Artificial* (CAEPIA), Granada, Spain, pp. 893-898.

V. Lopez, G. Valverde, J. Anchiraico and D. Urgeles (2016) Specification of a Cad Prediction System for Bipolar Disorder, *Uncertainty Modelling in Knowledge Engineering and Decision Making*, World Scientific Proceedings Series on Computer Engineering and Information Science, vol. 10, pp. 162-167

A. Martinez, (2016) System to predict bipolar disorder crises analysing massive data, Bachelor Thesis, *Eprints Complutense* (https://eprints.ucm.es/38722/)

N. Vanello et al. (2012), Speech analysis for mood state characterization in bipolar patients, *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Diego, CA, pp. 2104-2107.

ISBN: 978-1-5262-0820-0

The Alan Turing Institute

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Content-based Recommender Systems for Heritage: Developing a Personalised Museum Tour

Olga Loboda[1], Julianne Nyhan[2], Simon Mahony[3], Daniela M. Romano[4], Melissa Terras[5]

[1,2,3,4]Department of Information Studies, University College London,
Gower Street, London, UK

[5]College of Arts, Humanities and Social Sciences, University of Edinburgh,
George Square, Edinburgh, UK

[1] *olga.loboda.13@ucl.ac.uk*, [2]*j.nyhan@ucl.ac.uk*, [3]*s.mahony@ucl.ac.uk*,

[4]*d.romano@ucl.ac.uk*, [5]*m.terras@ed.ac.uk*

[1,2,3,4]*www.ucl.ac.uk*, [5]*www.ed.ac.uk*

ABSTRACT

How will a content-based recommender system be perceived by museum visitors? How will it transform visitor experience, and how can we adapt recommender systems to meet the needs of users in the museum domain? In this paper, we demonstrate the implementation of a content-based recommender system to generate personalised museum tours within the UCL Grant Museum of Zoology, London, UK. We also outline pilot usability tests that were carried out to collect initial feedback on the system performance in the wild. The findings help detect critical issues before the system is tested with museum visitors to explore the potential transformation in visitor experience that occurs with content-based recommender systems in physical museums.

1. INTRODUCTION

Museum recommender systems (RSs) have the potential to enhance visitor experience (VX) by providing a more personalised way to engage with museum collections. By focusing visitors' attention on a selection of exhibits, RSs could mitigate information overload associated with the overwhelming amounts of information that visitors have to process in a physical museum (Huang *et al.*, 2012). By tailoring recommendations to individual interests and needs, RSs can build a more personal connection between visitors and objects. By engaging visitors with a collection, RSs might also encourage exploration and stimulate learning and reflection (Kontiza *et al.*, 2018). In addition, the benefits of RSs can vary depending on user characteristics, as, for instance, they can assist visitors who are not familiar with the collection to identify their points of interest in the museum (Wang *et al.*, 2007; Fournier *et al.*, 2014; Bartolini *et al.*, 2016). However, little research has been done that provides solid evidence of enhanced VX with the help of RS.

Over the past few decades, researchers have been testing different approaches to generate museum recommendations: recent studies include Keller and Viennet, 2015; Rossi *et al.*, 2016; Cardoso *et al.*, 2017; Hashemi and Kamps, 2018; Kontiza *et al.*, 2018; Pavlidis, 2019. The studies often tend to be limited to offline evaluations of filtering methods, but they may not be able to reveal the efficiency of RSs in a real-world setting. The algorithms may theoretically be accurate, but the RS may not meet visitor needs because of many external or situated factors, such as a poorly designed interface and the position of points of interest in the exhibition (Hashemi and Kamps, 2018; Naudet and Deladiennée, 2019). Hence, it is necessary to carry out system evaluations in the wild. For instance, MyMuseum, a mobile guide at the National Arts Gallery of the Philippines, aimed to make gallery visits more informative and personalised by providing artwork recommendations based on personal information, art preferences, user location and item ratings (Alabastro *et al.*, 2010). By conducting online evaluations and collecting direct user experience (UX) feedback, the MyMuseum study revealed the difference in perception of the effectiveness of content-based and collaborative filtering approaches with and without contextual data. Their online evaluations indicated that the cumulative score of both accuracy and coverage was the highest for collaborative filtering without contextual information. From the user acceptance tests, the contextual recommendation approaches received more positive feedback, because the users felt more comfortable when the RS provided location-based suggestions. At the same time, it is also not enough to evaluate RSs with UX-related studies. Considering that the goal of the RSs is to enhance VX, it is necessary to take a step further and to learn about the system

 ISBN: 978-1-5262-0820-0

efficiency by analysing the RS-augmented interaction between the visitor and museum collection (Loboda *et al.*, 2018).

It is also important to acknowledge the idiosyncrasies of developing museum-specific RSs. For instance, Buder and Schwind (2012) suggested that, unlike in e-commerce, RSs in education should not strictly follow user preferences but rather challenge learners with opposite viewpoints. Moreover, the recommendations need to be assessed through an "item consumption" approach; while the evidence of RS's effectiveness in e-commerce is measured by purchasing rates, the same threshold may not be applicable to educational and, similarly, museum RSs (Buder and Schwind, 2012; Loepp and Ziegler, 2019). With a more defined scope of the requirements for museum RSs following their observed impact on VX, it is necessary to revisit the utility of complex RSs in order to make them suitable for the real-world museums. The development of museum RSs depends on data taken from the diverse museum databases which may not be appropriate to be used in recommenders, and this needs to be explicitly discussed to identify how to approach and transform the available datasets. Moreover, a user-item matrix with ratings for a collaborative filtering method or behavioural data for context-awareness may not be practicable to acquire and test in small-scale RS studies with limited museum and user data as well as restricted financial resources to build RSs with no commercial benefits. This may prevent researchers from replicating existing studies in other museum environments and enriching available knowledge about the efficiency of a specific RS approach across different museums.

This research aims to demonstrate the impact of museum RSs through a series of UX and VX-related field studies. We anticipate that the studies will reveal a set of requirements for the museum RSs that can enhance VX and be replicable at the same time. In this paper, we present reGrant, an RS with a traditional content-based filtering method, developed for the UCL Grant Museum of Zoology and Comparative Anatomy. We also discuss the pilot usability tests conducted to collect external feedback about the developed RS and to update the system before the main study that aims to explore the impact of RSs on VX.

2. SYSTEM OVERVIEW

Initially, we carried out front-end field evaluations in the Grant Museum to identify the possible areas for improvement for their visitors and to explore how an RS could meet visitors' needs. The Grant Museum was selected because of its role as an experimental museum space for research within a university (Macdonald and Nelson, 2012) where an RS could be tested with real-world visitors. The collected feedback indicated that visitors required more information about the displayed objects, while some also enquired about a museum map or a tour. With these aspects included, an RS could be used to generate a personalised tour with more information about the specimens relevant to the individual interests of the visitors and to highlight objects that they may otherwise have missed.

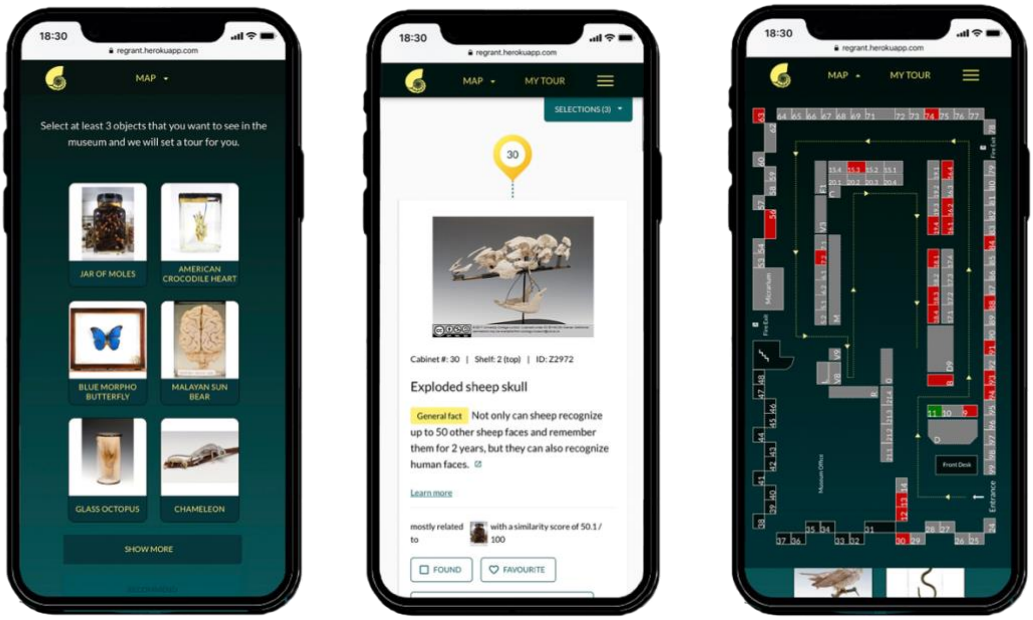


Figure 1. *Recommender system interface – home page with a user preference form (left), tour page with recommended objects (middle), map popup with recommended cabinets (right)*

 ISBN: 978-1-5262-0820-0

Our RS is being developed using Angular and Python Flask frameworks and has been deployed to a cloud service, Heroku, at regrant.herokuapp.com. Please note that a live version of our RS may differ from the system overview outlined in this paper due to periodic system updates that reflect our research findings and ongoing experimental studies. The first version, which was developed during April-May, 2019 (see Figure 1), was based on the following scenario: when the user launches the application, they first have to select at least three specimens of interest from a list of 100 featured museum objects so that the system could use these to generate a tour with 50 objects. On a tour page, the user can explore some interesting facts about recommended specimens, and they can learn more about a particular specimen by visiting the corresponding object pages. They can also use the map, where recommended cabinets are highlighted, to navigate in the museum. By tapping on a cabinet icon on the map, the user can see a list of objects showcased in the selected cabinet, and manually add preferred specimens to their tour. As their tour proceeds, the user can rate recommendations as well as mark specimens as favourite and/or found. At the end of the visit, the system provides a visit summary with some information about the user's tour, their preferences and 21 new recommendations for the next visit based on user input gathered during the tour.

3. RECOMMENDER ALGORITHM

In our RS, we employ a traditional content-based filtering method that relies on the similarities between item features and does not require any substantial prior user data to recommend a subset of items. The filtering method explores the similarities between ten object features, such as species, distribution, conservation status and object type. Figure 2 illustrates the executed recommendation process. More specifically, the data pre-processing step involves data cleaning to fill in missing values and to convert values into lowercase strings, data aggregation to gather all analysed strings in one column, tokenisation to break up strings into lists of words (tokens) and vectorisation to convert tokens into vectors. For our dataset with 500 museum objects, we use Count Vectorizer which counts how many times a token occurred in the description of each specimen. The returned term-document matrix is then used to calculate the similarity scores between all objects in the dataset using the cosine similarity algorithm. The algorithm evaluates the similarity between two vectors and is represented as a sum of multiplied tokens, A_i and B_i, from vectors A and B divided by the product of the two vectors' lengths:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

where A_i is a token from the vector A; B_i is a token from the vector B; n is a number of distinct tokens from the column with aggregated object features.

When the system receives a recommendation request, it constructs a subset of relevant objects based on the scores from the similarity matrix to then either display recommendations on individual object pages, list them on a tour page in a particular order, or provide new recommendations for the next visit on a visit summary page.

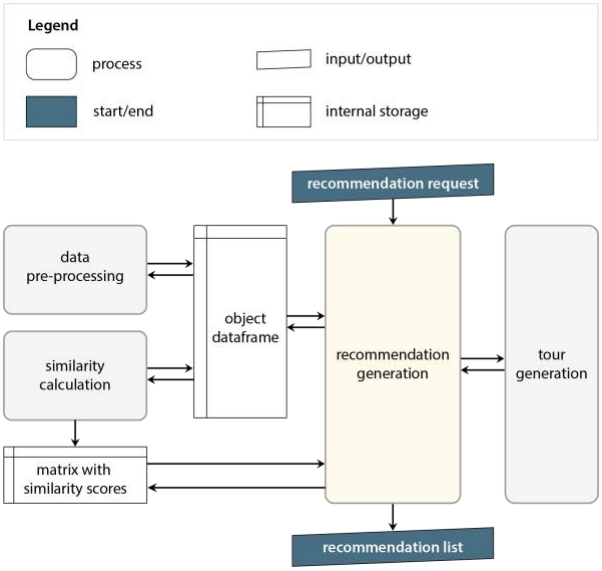


Figure 2. *Recommendation component diagram*

 ISBN: 978-1-5262-0820-0

4. PILOT USABILITY TESTING

4.1 Usability study evaluation procedure

Pilot usability tests were conducted at the museum and involved both quantitative and qualitative evaluations to collect some external feedback and identify critical issues with the system performance before the RS was tested with the Grant museum visitors in a large-scale study. Over the period of two weeks, 12 current and former UCL staff and students were recruited to take part in a usability test at the Grant museum (5 male and 7 female respondents aged between 22 and 44; 10 solo visitors and 1 group visit; 8 first-time and 4 returning visitors). Due to space constraints, in this paper we provide a brief summary of the study, but it will also be addressed in the following publications and the lead author's PhD thesis.

At the museum the participants were debriefed about the goal of the usability test and the reason why the RS was developed. However, we did not demonstrate how the system worked in order not to affect the perception of the system's ease of use, but rather to imitate a real-world scenario in which the visitors would not be able to participate in a demonstration session. The participants were then asked to explore the museum with the RS for around 20 minutes and then to return back for the post-test evaluations. The questionnaires were used for an overall evaluation of perceived system performance and encompassed 26 items (5-point Likert scales, 1 = disagree, 5 = agree) that were mostly adapted from the Pu et al.'s (2011) ResQue survey used to measure the RS qualities. The quantitative study was supported with semi-structured interviews to gather more in-depth feedback about encountered issues and potential system improvements.

Before reporting the findings, it should be noted that two participants had to be excluded from the quantitative study – one participant did not accurately follow the usability protocol, while the other encountered an error on their device which prevented them from testing the RS properly, but since they were with a companion, they followed their companion's tour and thus provided some feedback in the semi-structured interview.

4.2 Usability study results

Overall, the participants were satisfied with the recommender (M = 4.4, SD = .516, n = 10) and found the system easy to use (M = 4.5, SD = .707, n = 10) and easy to learn (M = 4.3, SD = .675, n = 10). During the interviews, 8 out of 12 participants mentioned potential educational benefits of the system. The respondents reported that the RS made their visits more structured, informative and helped them find interesting objects that they would otherwise have missed. One participant particularly liked the visit summary because it helped them learn not only about the collection but also about their museum preferences. Moreover, the users liked the idea of integrating recommenders into the museum environment ('I would use this recommender again': M = 4.4, SD = .516, n = 10; 'I would like to use similar systems in other museums': M = 4.6, SD = .699, n = 10).

From the quantitative findings, the recommendations received positive feedback ('The recommender gave me good suggestions': M = 4.1, SD = .316, n = 10), while the perceived accuracy score was adequate ('The recommended objects matched my interests': M = 4.0, SD = .471, n = 10). The participants, however, were not satisfied with the recommendation diversity ('The recommended objects were diverse': M = 3.5, SD = .527, n = 10). During the interviews, 3 system users suggested adding random objects to their tours to see how they would respond to less relevant content, while 2 participants were particularly curious about what their least relevant specimens were. In this regard, the participants were also curious to find out what criteria were used to generate the recommendations as the first RS version did not provide explanations.

It is also important to mention that the participants struggled to locate recommended objects in the museum ('It was easy to find the recommended objects in the museum': M = 3.5, SD = .845, n = 10). 5 respondents indicated that the cabinet numbers in the physical museum were confusing because they don't follow a sequential order and they were difficult to see because of their location and colour contrast. Moreover, since the RS was developed for a zoological museum where many objects can be found on the same shelf, it was challenging to locate individual objects in a cabinet. As a result, several interviewees reported that their attention decreased as their tour proceeded because 50 suggested specimens to find in one visit was overwhelming, and they skipped some objects if too many were recommended for the same museum cabinet.

4.3 Usability study discussion and limitations

Because the study was limited to the university students and staff, the sample is not necessarily representative of the Grant museum's wider and more heterogeneous audience to the extent that would allow

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

the findings to be generalised. In addition, the system was tested with only 500 objects and thus more unfavourable diversity results might have been observed by making use of content-based filtering with a larger object dataset. Nevertheless, the pilot usability evaluation helped identify the areas for improvement before the main study with the Grant museum visitors. The participants' enquiry about the least relevant and random objects revealed a substantial divergence from recommending only those objects that may match visitor's preferences. This suggests that diversity in museum RSs might be favoured over accuracy. Moreover, the feedback extended beyond the recommendation quality and indicated that an RS can also become a trigger of information overload and museum fatigue, because the users wanted to set custom tour lengths and asked to adjust external elements related to the physical museum.

Following the user feedback, we released the second version of our recommender where, for example, we updated the system to extract n_1 most similar objects as well as randomly pick n_2 least similar objects for the tour, where n_1 and n_2 are set dynamically in a ratio of 4:1 depending on a preferred tour length. More details about all system updates will be addressed in the next publications and will be comprehensively discussed in the PhD thesis.

5. CONCLUSION

In this paper, we demonstrate how we implemented the first version of our content-based RS. The suggested system is anticipated to be replicable in order to encourage other researchers to conduct similar evaluations in other museums, to aggregate and compare the findings and to elicit robust requirements for museum RSs. We emphasise the importance of field studies to collect real-world feedback about system performance, and thus we begin to analyse the UX with an early version of our RS system to learn how it should be further adapted for the targeted museum. The pilot usability testing presented here was necessary to also collect initial external feedback about the system before the main study is carried out where we explore the transformations of VX with the content-based RS in the physical museum.

Acknowledgements: We would like to thank the UCL Grant Museum of Zoology and Comparative Anatomy for providing information about their collection as well as allowing us to carry out our evaluations in the real-world museum environment. We would also like to thank all UCL staff and students who participated in our usability study for their invaluable feedback.

6. REFERENCES

Alabastro, P., Ang, M., DeGuzman, R., Muhi, M. and Suarez, M. (2010). 'MyMuseum: Integrating personalized recommendation and multimedia for enriched human-system interaction', in *DC '10: Proceedings of the 6th International Conference on Digital Content, Multimedia Technology and Its Applications*. IEEE, pp. 421–426.

Bartolini, I., Moscato, V., Pensa, R. G., Penta, A., Picariello, A., Sansone, C. and Sapino, M. L. (2016). 'Recommending multimedia visiting paths in cultural heritage applications', *Multimedia Tools and Applications*, Vol. 75, No. 7, pp. 3813–3842.

Buder, J. and Schwind, C. (2012). 'Learning with personalized recommender systems: A psychological view', *Computers in Human Behavior*, Vol. 28, No. 1, pp. 207–216.

Cardoso, P. J. S., Rodrigues, J. M. F., Pereira, J. A. R. and Sardo, J. D. P. (2017). 'An Object Visit Recommender Supported in Multiple Visitors and Museums', *International Conference on Universal Access in Human-Computer Interaction*. Edited by M. Antona and C. Stephanidis. Springer International Publishing, pp. 301–312.

Fournier, R., Viennet, E., Sean, S., Soulié Fogelman, F. and Bénaïche, M. (2014). 'AMMICO: social recommendation for museums', in *Proceedings of DI '14: Digital Intelligence – Intelligences Numériques*. Nantes, France.

Hashemi, S. H. and Kamps, J. (2018). 'Exploiting behavioral user models for point of interest recommendation in smart museums', *New Review of Hypermedia and Multimedia*. Taylor & Francis, Vol. 24, No. 3, pp. 228–261.

Huang, Y. M., Liu, C. H., Lee, C. Y. and Huang, Y. M. (2012). 'Designing a personalized guide recommendation system to mitigate information overload in museum learning', *Educational Technology and Society*, Vol. 15, No. 4, pp. 150–166.

Keller, I. and Viennet, E. (2015). 'Recommender Systems for Museums: Evaluation on a Real Dataset', in

 ISBN: 978-1-5262-0820-0

IMMM '15: The Fifth International Conference on Advances in Information Mining and Management Recommender, pp. 65–71.

Kontiza, K., Loboda, O., Deladiennée, L., Castagnos, S. and Naudet, Y. (2018). 'A museum app to trigger users' reflection', in *Mobile CH '18: Proceedings of the 2nd Workshop on Mobile Access to Cultural Heritage*. CEUR Workshop Proceedings.

Loboda, O., Nyhan, J., Mahony, S. and Romano, D. M. (2018). 'Towards evaluating the impact of recommender systems on visitor experience in physical museums', in *Mobile CH '18: Proceedings of the 2nd Workshop on Mobile Access to Cultural Heritage*. CEUR Workshop Proceedings.

Loepp, B. and Ziegler, J. (2019). 'Measuring the Impact of Recommender Systems – A Position Paper on Item Consumption in User Studies', in *ImpactRS '19: Proceedings of the 1st Workshop on Impact of Recommender Systems*.

Macdonald, S. and Nelson, T. (2012). 'A Space for Innovation and Experimentation: University Museums as Test Beds for New Digital Technologies', *A Handbook for Academic Museums: Beyond Exhibitions and Education.*, pp. 418–444.

Naudet, Y. and Deladiennée, L. (2019). 'Towards technology-mediated CH experiences: some open challenges', in *CI '19: Proceedings of the Workshop on Cultural Informatics*. CEUR Workshop Proceedings.

Pavlidis, G. (2018). 'Towards a Novel User Satisfaction Modelling for Museum Visit Recommender Systems', in *VR Technologies in Cultural Heritage*. Springer International Publishing, pp. 60–75.

Pu, P., Chen, L. and Hu, R. (2011). 'A User-Centric Evaluation Framework for Recommender Systems', in *RecSys '11: Proceedings of the Fifth ACM Conference on Recommender Systems*, pp. 157–164.

Rossi, S., Barile, F., Improta, D. and Russo, L. (2016). 'Towards a Collaborative Filtering Framework for Recommendation in Museums: From Preference Elicitation to Group's Visits', *Procedia Computer Science*. Elsevier Masson SAS, Vol. 98, pp. 431–436.

Wang, Y., Aroyo, L. M., Stash, N. and Rutledge, L. (2007). 'Interactive user modeling for personalized access to museum collections: The Rijksmuseum case study', in Conati, C., McCoy, K., and Paliouras, G. (eds) *International Conference on User Modeling*. Springer Berlin Heidelberg, pp. 385–389.

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Modeling a Decision-Maker in Goal Programming by means of Computational Rationality

Maura E Hunt, Manuel López-Ibáñez

Alliance Manchester Business School, University of Manchester,
Oxford Rd, Manchester UK

maura.hunt@postgrad.manchester.ac.uk, manuel.lopez-ibanez@manchester.ac.uk

ABSTRACT

This paper extends a simulation of cognitive mechanisms in the context of multi-criteria decision-making by using ideas from computational rationality. Specifically, this paper improves the simulation of a human decision-maker (DM) by considering how resource constraints impact their evaluation process in an interactive Goal Programming problem. Our analysis confirms and emphasizes a previous simulation study by showing key areas that could be effected by cognitive mechanisms. While the results are promising, the effects should be validated by future experiments with human DMs.

1. INTRODUCTION

Often, real-world problems contain multiple, conflicting criteria where no single alternative best satisfies all criteria simultaneously, causing alternatives that are mutually nondominated. The nature of nondominated alternatives requires preference information from a decision-maker (DM), which implies that the DM performs trade-off analysis for the criteria. In some cases, the number of alternatives to a decision problem is so large, that trade-off analysis would not be possible for a DM to perform. In these situations, an interactive approach is adopted to help a DM find their most preferred alternative.

A commonly used interactive algorithm for cases where the decision space contains many criteria and/or alternatives is Goal Programming (GP). GP aids a DM in trade-off analysis by finding the alternative that best satisfies the DM's aspiration for achievement on criteria. To begin the optimization process, the DM provides information about which criteria is most important and an initial achievement value is given to each criterion. The GP algorithm will then find the alternative that best satisfies the parameters the DM specified. Considering the alternative achieved, the DM will adjust the initial goals to continue exploring the decision space. Ideally, continuation of the GP process will allow the DM to perform adequate trade-off analysis and find an alternative that best satisfies their decision problem and underlying preferences.

Stewart (2005) proposed that the presence of cognitive biases could cause a lack of exploration of the decision space due to a premature termination of the GP process. Based on Tversky and Kahneman (1974), Stewart (2005) postulated that the readjustment of goals between iterations would be subject to avoidance of sure loss, anchoring and adjustment, and that the presence of these biases would result with alternatives that do not match the DM's underlying preferences. Stewart (2005) used these biases to simulate how a DM would adjust goals between iterations. His conclusion was that anchoring to the previous alternative observed by the DM (the GP alternative) had the biggest impact on the difference between the chosen alternative and the preferences of the DM. While anchoring to the previous goal and avoidance of sure loss have significant and similar effects, they were not as influential as anchoring to the GP alternative.

The conclusions of Stewart (2005) are informative, but the simulation of a DM could be improved with a newer concept from psychology known as computational rationality. Computational rationality is an explanatory extension of Simon's (1972) descriptive idea of bounded rationality. While Tversky and Kahneman (1974) continued Simon's (1972) ideas by describing different cognitive effects, they interpret the effects as biases, indicating that cognition fails to find the correct answers to problems by deviating away from normative theories. In contrast, computational rationality can provide an explanation to why cognitive effects occur by analyzing resource constraints (Lewis et al., 2014). A computational rationality perspective could potentially provide a more accurate representation of human cognition in analysis, which in turn, could lead to more generalizable and complete models (Lewis et al., 2014; Lieder et al., 2018).

In the current study, we compare two perspectives for simulating cognitive mechanisms with the replication and improvement models. First, we replicated the simulation from Stewart (2005) with a slight modification to how many datasets were used since Stewart (2005) only used one dataset per datatype. The

 ISBN: 978-1-5262-0820-0

reason for adding additional datasets was to produce a more percise simulation study. Then, we created the improved model by adapting resource rational methods from computational rationality to simulate the anchoring effect (Lieder et al., 2018; Vul et al., 2014). We expect that the improved model will provide a more accurate reflection of human decision-making processes and result in a lower error rate because of the simulation's ability to capture the variables contributing to the effect. We conclude our findings based on an ANOVA to determine which parameters had the strongest effect on deviating the chosen alternative away from the DM's actual preferred alternative in the simulation.

2. GOAL PROGRAMMING FORMULATION

Let decision space Z consist of n alternatives that are evaluated on m criteria, and z_j denote the value of criterion j for alternative $\boldsymbol{z} \in Z$. The preference structure that simulates the DM's underlying preferences is constructed using the following additive value function (AVF):

$$V(\boldsymbol{z}) = \sum_{j=1}^{m} w_j(M_j) \tag{1}$$

where the marginal value function $v_j(z_j)$ is assumed to have the following sigmoidal shape (Stewart, 2005):

$$M_j = \begin{cases} a) & \lambda_j \dfrac{e^{\alpha_j z_j} - 1}{e^{\alpha_j \tau_j} - 1}, \quad for\ 0 \le z_j < \tau_j \\ b) & \lambda_j + (1 - \lambda_j)\dfrac{1 - e^{-\beta_j(z_j - \tau_j)}}{1 - e^{-\beta_j(1 - \tau_j)}}, \quad for\ \tau_j \le z_j \le 1 \end{cases} \tag{2}$$

The reason for such a shape is to represent the DM's preferences following prospect theory which describes a DM's perception of alternatives as losses or gains based on their own internalized reference point (Kahneman and Tversky, 1979). The DM will then use the preference structure to choose between the nondominated alternatives and to determine if an alternative is perceived as a loss during the GP process.

In the GP algorithm, the DM first determines the goal or aspiration level for each criterion represented as g_j. Then, the GP algorithm uses $s_j = g_j - z_j$ to calculate how much the alternative deviates from the goal. The deviation from the goal is constrained to be non-negative and once the goal is achieved, s_j is set to 0. The GP process is initialized by setting g to the ideal value of 1 for all criteria. Following Stewart (2005), the deviation is then minimized using a linear combination of Archimedean with Chebyshev GP as shown in Equation 3, where w_j is the weight given by the DM for criterion j and ξ is a constant varied in analysis to determine if the different GP frameworks are more robust to human cognitive effects.

$$\boldsymbol{z}^* = argmin\left(\xi \sum_{j=1}^{m} w_j s_j + (1 - \xi)\, max_{j=1}^{m}\,(w_j s_j)\right) \tag{3}$$

3. SIMULATION OF THE DM'S GOAL ADJUSTMENT

3.1 Replicated Model

Stewart (2005) noted that when a DM receives the initial GP alternative, adjusting g could be difficult. For starters, the DM could lack the ability to process the entire decision space. The processing limitations of the DM was modeled by "perceiving" only a subset $V \subseteq Z$, such that $|V| = \theta \cdot n$, of the alternatives. From the perceived alternatives V, the DM could understand the decision space in a different manner than the analyst or available data itself. The incongruence for the decision space was modeled by multiplying each alternative in V with a log normally distributed value having a standard deviation of σ, where higher values correspond to greater variance, resulting in a new set of alternatives V'.

After transforming the decision space, the DM could use cognitive mechanisms when adjusting the goals. First, the DM could be loss aversive and perceive setting new goals as losing the alternative "won" from the first iteration. Simulating avoidance of sure loss was achieved by using the value function to compare how the DM valued each $\boldsymbol{v}' \in V'$ and $\boldsymbol{z}^*$. If the DM preferred $\boldsymbol{v}'$, $\boldsymbol{v}'$ is kept as the same value. However, if the DM preferred $\boldsymbol{z}^*$, four trials were used to improve $\boldsymbol{v}'$ by halving the distance between $\boldsymbol{v}'$ and $\boldsymbol{z}^*$ until the DM had a greater preference for $\boldsymbol{v}'$ or the trials expired. The resulting vector is stored in $\hat{\boldsymbol{z}}$, which represents the DM's

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

idealized adjusted goals. Then, the direction of the idealized adjustment, d, was found by subtracting $\boldsymbol{z}^*$ from $\hat{\boldsymbol{z}}$. If d was negative, the DM could perceive $\hat{\boldsymbol{z}}$ as a loss and be unwilling to accept a new goal. When d is negative, ρ is multiplied by d to represent the magnitude of the DM's aversion to losses which is stored in $\hat{d}$ along with the positive values of d that were unchanged.

Stewart (2005) further proposed that the anchoring effect could influence the DM's goal adjustment process. When creating the adjusted goals, the DM could potentially anchor to the previous goal, g_0, or to $\boldsymbol{z}^*$. The degree to which the DM anchors to g_0 is varied using the parameter γ, where higher values of γ indicates stronger anchoring. Additionally, the DM could anchor to $\boldsymbol{z}^*$ which is represented by varying φ. Opposite to γ, higher values of φ indicate less anchoring and as φ exceeds 1, the DM begins to explore alternatives beyond the bounds of the decision space and follow the direction of the idealized goals found in d. Equation 4 shows the formula used to calculate the adjusted goals between iterations.

$$g = \gamma g^0 + (1-\gamma)(\boldsymbol{z}^* + \varphi\hat{d}) \tag{4}$$

3.2 Improved Model

A few modifications were made to simulate the DM in the improved model using methods from computational rationality to represent anchoring and adjustment. The methods used were adapted from the studies done by Vul et al. (2014) and Lieder et al. (2018). Both studies proposed frameworks that modeled human decision-making based on resource constraints and sampling techniques that resemble machine learning. However, their proposals differ in the computational processing times and how the distributions were used for a decision problem. Vul et al. (2014) proposed that the human brain constructs an internalized posterior distribution and then randomly samples from this distribution to make inference and learning decisions. Instead, given the number of decisions a human makes in a given day, Lieder et al. (2018) believed human cognition would use the memorylessness properties of Monte Carlo Markov Chains (MCMC).

For traditional numeric decision problems and the methods described in Lieder et al. (2018), one anchor is altered to find one correct alternative. However, in MCDA problems, no "correct" alternative exists, only preferred alternatives specific to the DM. Thus, the most preferred alternative given by the preference function was used to replace the "correct" alternative and is denoted by K. Additionally, due to the complex nature of MCDA problems, the comparison of alternatives was not achieved by comparing an alternative's values to an anchor, but by finding the Euclidean distance between the anchor and the most preferred alternative.

To begin the simulation of how a DM performs adjustments, the proposed goal was created by randomly sampling a value, δ, from a normal distribution, and then added to g_0. If the probability of the Euclidean distance for the proposed goal, $P(g_0 + \delta \mid K)$, is greater than the probability of the previous goal, $P(g_0 \mid K)$, than the adjustment is accepted. If not, the adjustment can be accepted if $\frac{P(g0 + \delta \mid K)}{P(g0 \mid K)} < uniform\ (0, 1)$.

The process of adjusting an estimate does not typically continue until the correct alternative is found which is illustrated with the anchoring effect. Instead of describing the anchoring effect as occurring or not occurring, a resource rational explanation can be used to determine the optimal number of adjustments the brain should perform. Lieder et al. (2018) proposed that the optimal number of adjustments, denoted as t^*, is determined by considering the trade-off between error and time cost. As the number of adjustments increases, the error rate decreases as the time cost linearly increases at a rate of tc, where higher values indicate a harsher time cost. The more adjustments the DM performs, the less of an impact the anchoring effect has, and the "correct" or most preferred alternative becomes closer to the DM's chosen alternative. After using the computational rationality framework to explain anchoring, the actual adjustment of the goal originally shown with Equation 4 was replaced with Equation 5 for the improved model.

$$g = t^* + (1-t^*)(\boldsymbol{z}^* + \varphi\hat{d}) \tag{5}$$

4. RESULTS

The same synthetic data and parameters were used for both models. Data was randomly sampled from the surface of a hypersphere for either all positive or negative values. The results reported in the current study are the convex cases, or the data generated from positive values. After creation and transformation, data was standardized using a min-max scaling for each criterion. Twenty different datasets were generated with the dimensions of $m = 7$ and $n = 100$. Parameter values were created following Stewart (2005).

Analysis was performed separately for the models. A two-way analysis of variance (ANOVA) with first order interactions was performed by varying the three levels of ξ, θ, σ, ρ, φ, and γ to determine their effect on

 ISBN: 978-1-5262-0820-0

the chosen alternative which was scaled between the DM's most and least preferred alternatives. For the improved model, the parameter *tc* replaced γ and had the three levels of 0.03, 0.09, and 0.30.

All main effects were significant ($p < 0.01$) for the replicated model. The interaction effects varied in significance but followed the same patterns outlined in Stewart (2005). When examining the interaction effects, anchoring to $\boldsymbol{z}^*$ had the greatest effect on the chosen alternative, while the proportion of alternatives selected appeared to have the lowest impact. For these reasons, φ and θ were held constant for further analysis. While Stewart (2005) determined the findings based on the slight differences in averages, we will use Tukey Post-Hoc analysis to compare models and determine our overall findings.

Contrastingly, the improved model showed the main effects to be significant ($p < 0.000$) except σ which was not significant ($p > 0.05$). Patterns for the parameters were more difficult to detect and φ did not appear to contain as strong of an influence for interactions. However, θ appeared to have the least significant effects when interacting with the other parameters. For these reasons, σ and θ where excluded for further analysis to eliminate excess noise from the model.

Post-hoc analysis shows that only one average from γ was significantly different between level 1 and 3 ($p < 0.05$) in the replicated model. This suggests that only a low or high degree of anchoring impacted the DM's adjustment of goals. Interestingly, in the improved model all averages for anchoring were significantly different ($p < 0.05$) with low rates of error. This indicates that the computational rationality model was able to determine a stronger effect for anchoring that the replicated model could not capture. However, the improved model was not able to determine significantly different averages for ρ ($p > 0.05$). Since only the simulation of anchoring was altered, using two different frameworks could have negatively impacted the results for loss avoidance.

5. CONCLUSIONS

Our study extends the simulated DM proposed by Stewart (2005) by improving the simulation of human DMs with ideas from computational rationality. The findings of the current study show that anchoring to the previous goal has the highest impact on GP. Anchoring to the GP alternative also has a significant role in GP but potentially not as large of an effect that Stewart (2005) suggested. The proposed model is also useful for evaluating interactive multi-objective optimization algorithms (López-Ibáñez and Knowles, 2015). Moreover, our simulation results suggest trends and potential areas of future research. In particular, experiments with actual DMs must be conducted to validate our findings.

Acknowledgements: Dr. Paul Warren and Dr. George Farmer for their insights into psychological research.

6. REFERENCES

D Kahneman and A Tversky (1979), Prospect theory: An analysis of decision under risk, *Econometrica*, **47**, *2*, pp. 263–292.

R L Lewis, A Howes and S Singh (2014), Computational rationality: Linking mechanism and behavior through bounded utility maximization, *Topics in Cognitive Science*, **6**, *2*, pp. 279–311.

F Lieder et al. (2018), The anchoring bias reflects rational use of cognitive resources, *Psychonomic Bulletin & Review*, **25**, *1*, pp. 322–349.

M López-Ibáñez and J D Knowles (2015), Machine decision makers as a laboratory for interactive EMO. *Evolutionary Multi-criterion Optimization*, **9019**, pp. 295–309.

H Simon (1972), Theories of bounded rationality, *Decision and Organization*, **1**, *1*, pp. 161–176.

T Stewart (2005), Goal programming and cognitive biases in decision-making, *Journal of the Operational Research Society*, **56**, pp. 1166–1175.

A Tversky and D Kahneman (1974), Judgment under uncertainty: Heuristics and biases, *Science*, **185**, *4157*, pp. 1124–1131.

E Vul et al. (2014), One and done? Optimal decisions from very few samples, *Cognitive Science*, **38**, *4*, pp. 599–637.

Learning Spare Changes in Time-varying Markov Networks with Density Ratio Estimation and Its Application to fMRI

Y Zhang[1], C Langley[2], J Thai[3], S Liu[4]

[1]Faulty of Engineering, University of Bristol,
Queens Road, Clifton, Bristol, UK

[2]Department of Psychiatry, University of Cambridge,
Trinity Lane, Cambridge, UK

[3]Clinical Research and Imaging Centre, Bristol Medical School, University of Bristol,
St Michael's Hill, Bristol, UK

[4]School of Mathematics, University of Bristol,
Woodland Road, Clifton, Bristol, UK

[1]yulong.zhang@bristol.ac.uk, [2]cl798@medschl.cam.ac.uk, [3]jade.thai@bristol.ac.uk, [4]song.liu@bristol.ac.uk

[1,3,4]www.bristol.ac.uk, [2]www.cam.ac.uk

ABSTRACT

This paper proposes a method for estimating sparse changes in time-varying Markov networks. Rather than estimating two network structures separately and then obtaining the differences, we adopt a previously proposed method which directly estimates the ratio of the two Markov network densities. The sparse changes are tackled easily with a sparse inducing regularizer. Specifically, to consider the temporality of the networks, an importance weighting scheme is introduced. Moreover, an application to fMRI data demonstrates the potential of the proposed method.

1. INTRODUCTION

Changes in interactions between random variables are of great interest in numerous real-world applications, as they can provide insights of underlying mechanisms and help to make decisions. These interactions can be naturally encoded by networks. A rich literature has been developed under the assumption that the interactions are static in one state. With increasing availability of data sets that evolve over time, however, there is an emerging challenge to develop models for time varying networks.

In this paper we study the problem of estimating sparse changes in time-varying Markov networks (MNs), that is, MNs rewire over time. In particular, we focus on pairwise MNs, whose joint distribution can be factorized over single variables or pairs of variables (Koller, Friedman, and Bach 2009). Formally, given two sets of sequential samples $\{\mathrm{X}_t^p\}_{t=1}^{T_p}$ and $\{\mathrm{X}_t^q\}_{t=1}^{T_q}$ respectively drawn from two d-dimensional MNs with true parameters θ_p^* and θ_q^*, where $\theta_p^*, \theta_q^* \in \mathrm{R}^{d\times d}$, the goal is to estimate the sparse change $\delta\theta^* = \theta_p^* - \theta_q^*$. This problem can be encountered in many important real-world applications, including identifying the changes in the neural connectivity networks to identify features associated with different mental diseases, and the changes in the stock market dependency structures to spot trends.

To solve such a problem, one naive approach is first estimating θ_p^* and θ_q^* separately from two sets of samples and then obtain $\delta\theta^*$. We may utilize L1-norm regularization to produce sparse MNs and thus the change between MNs also becomes sparse (Banerjee, Ghaoui, and d'Aspremont 2008; Friedman, Hastie, and Tibshirani 2008; Lee, Ganapathi, and Koller 2007). However, this approach does not work if MNs are rather dense but change is sparse. One other approach is directly estimating $\delta\theta^*$ using the two sets of samples, without learning θ_p^* and θ_q^* separately (Zhao, Cai, and Li 2014; Liu et al. 2013; Fazayeli and Banerjee 2016). This approach results in better estimators of differential networks as they do not estimate nuisance parameters and require weaker assumptions (Kim, Liu, and Kolar 2019). However, all the previous works assume that the data generating process is time-invariant and that the relational structure is fixed, which is not suitable for time-varying MNs.

This paper proposes a direct method for estimating sparse changes in time-varying MNs. The method relies on the Density Ratio Estimation (DRE) method (Sugiyama, Suzuki, and Kanamori 2012), more specifically on

 ISBN: 978-1-5262-0820-0

Kullback-Leibler Importance Estimation Procedure (KLIEP) (Sugiyama et al. 2008). To tackle the time-varying MNs, we introduce an importance weighting scheme. The rest of this paper is organized as follows. In Section 2, we describe the proposed model for estimation of the changes in time-varying MNs. In Section 3, the method is applied to a real-world dataset. Conclusion is given in Section 4.

2. DIRECT LEARNING OF SPARE CHANGES IN TIME-VARYING MNs

In this Section, we first give a brief background on MN. Then, we review density ratio estimation and finally we describe how to consider time-varying networks and present the objective function for time-varying MNs.

2.1 MN Model

Let $\mathrm{X} = (x_1, x_2, \dots, x_d)$ denote a d-dimensional random vector. The pairwise MN over the random vector X is

$$\mathrm{P}(X;\theta) = \frac{1}{Z(\theta)}\exp\left(\sum_{i,j=1,i\geq j}^{d}\theta_{i,j}^{T}T_{i,j}(x_i,x_j)\right) = \frac{1}{Z(\theta)}\exp(\langle\theta,T(X)\rangle) \tag{1}$$

where $\mathrm{T}(X) = \mathrm{T}(x_i,x_j)_{i,j=1}^{d}$ are univariate and bivariate vector-valued basis functions, $\theta = \{\theta_{i,j}\}_{i,j=1}^{d}$ is parameters, and $\langle\cdot\rangle$ is the inner product operator. The partition function, $\mathrm{Z}(\theta)$, plays the role of a normalizing constant, ensuring that the probabilities add up to one.

2.2 Objective Function for Time-invariant MNs

Consider two MNs described by $\mathrm{p}(X;\theta_p^*)$ and $\mathrm{q}(X;\theta_q^*)$ respectively. The structure change $\delta\theta^* = \theta_p^* - \theta_q^*$ between these two MNs can be estimated directly based on density ratio $\mathrm{r}(X;\delta\theta^*)$, which is defined as follows (Liu et al. 2014).

$$\mathrm{r}(X;\delta\theta^*) = \frac{p(X;\theta_{\mathrm{P}}^*)}{q(X;\theta_{\mathrm{Q}}^*)} = \frac{\frac{1}{Z(\theta_p^*)}\exp(\langle\theta_p^*,T(X)\rangle)}{\frac{1}{Z(\theta_q^*)}\exp(\langle\theta_q^*,T(X)\rangle)} = \frac{1}{Z(\delta\theta)}\exp(\langle\delta\theta^*,T(X)\rangle) \tag{2}$$

To estimate ratio $\mathrm{r}(\mathrm{X}_p,X_q;\delta\theta^*)$, the method fits a modelled $\hat{p}(X;\theta_P) = \mathrm{r}(\mathrm{X}_p,X_q;\delta\theta^*)\mathrm{q}(X;\theta_Q^*)$ to the true distribution $\mathrm{p}(X;\theta_P^*)$ using the parameter $\delta\theta$. We use the Kullback-Leibler divergence as goodness of fit measure and the objective is to minimize

$$\begin{aligned} KL(p(X;\theta_P^*)\|\hat{p}(X;\theta_P)) &= KL\left(p(X;\theta_P^*)||\, r(X;\delta\theta^*)q(X;\theta_Q^*)\right) \\ &= \int p(X;\theta_P^*)\log\left(\frac{p(X;\theta_P^*)}{r(X;\delta\theta^*)q(X;\theta_Q^*)}\right)dX \\ &= \underbrace{KL\left(p(X;\theta_{\mathrm{P}}^*)\middle\|q(X;\theta_Q^*)\right)}_{\text{Constant}} - \int p(X;\theta_P^*)\log r(X;\delta\theta^*)dX \end{aligned} \tag{3}$$

As the first term is constant with respect to $\delta\theta$, the optimization objective is simplified as

$$\begin{aligned} -\int p(X)\log r\,(X;\delta\theta)dX &= -\frac{1}{n_p}\sum_{i=1}^{n_p}\log r\left(X_p,X_q;\delta\theta\right) \\ &= -\frac{1}{n_p}\sum_{i=1}^{n_p}\langle T(X_p),\delta\theta\rangle + \log\frac{1}{n_q}\sum_{i=1}^{n_q}\exp\langle T(X_q),\delta\theta\rangle \end{aligned} \tag{4}$$

2.3 Objective Function for Time-varying MNs

In this section we consider two time-varying MNs modelled by $p_{\mathrm{t}}(X;\theta_p^*)$ and $q_{t'}\ (X;\theta_q^*)$, which are changing with time t and t' respectively. For two given time points τ and τ', the corresponding distributions are described as $p_{\tau}(X;\theta_p^*)$ and $q_{\tau'}\ (X;\theta_q^*)$ respectively. In this case, the ratio we want to learn becomes $\mathrm{r}(\mathrm{X};\delta\theta^*) = \frac{p_{\tau}(\mathrm{X};\theta_{\mathrm{p}}^*)}{q_{\tau'}(\mathrm{X};\theta_{\mathrm{q}}^*)}$. However, there is normally only one sample for a distribution $p_{\tau}(X;\theta_p^*)$. To estimate the ratio, it is a common practice to make some assumption about $p_{\tau}(X;\theta_p^*)$. The assumption we make here, which is

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

important for the following work, is that $p_t(X;\theta_p^*)$ varies smoothly with time t. Specifically, the smoothness assumption means that samples drawn from $p_{\tau+\epsilon}(X;\theta_p^*)$ are increasingly similar to those drawn from $p_\tau(X;\theta_p^*)$ as $\epsilon \to 0$. This assumption allows us take advantage of the adjacent samples of X_τ, by assigning weights according to their time proximity to τ. In this study, we apply a well-known weighting function as given in Eq. (5).

$$k(\tau,t) = \exp\left(-\frac{(t-\tau)^2}{2\sigma_u^2}\right) \tag{5}$$

The value of $k(\tau,t)$ is maximized to 1 at $t=\tau$, and decreases towards 0 as t differs more from τ. By weighting samples based on this, $p_\tau(X;\theta_p^*)$ can be estimated with the full set of data. The same assumption is also applied to $q_{\tau'}(X;\theta_q^*)$.

Applying the weights to Eq. (5), the objective function becomes

$$\mathcal{L}_{\text{KLIEP}}(\delta\theta,\tau,\tau') = -\frac{1}{Z_\tau}\sum_{i=1}^{n_p} k(\tau,t)\langle T(X_p),\delta\theta\rangle + log\frac{1}{Z_{\tau'}}\sum_{i=1}^{n_q} k(\tau',t')\exp\{\langle T(X_q),\delta\theta\rangle\} \tag{6}$$

To consider sparse structure changes, we introduce L1 norm to the objective function. Finally, we consider the following optimization problem

$$\underset{\delta\theta}{\text{argmin}}\left(\mathcal{L}_{KLIEP}(\delta\theta) + \lambda_{n_p,n_q}\|\delta\theta\|_1\right) \tag{7}$$

3. EXPERIMENTAL RESULTS

3.1 fMRI Dataset

We use our method to analyse data from an fMRI study that aimed to find the neural correlates of cognitive impairment in a neurological disease. The data was collected from subject groups with Multiple Sclerosis (MS) and Healthy Controls (HC), respectively, with 0.906 second intervals from 90 regions of interest (ROI). There are three types of tasks in the experiment, i.e. sensorimotor task (T1), intrinsic alertness task (T2) and the extrinsic alertness task (T3). More details about the tasks can be found in (Kim, Liu, and Kolar 2019). Each subject was asked to go through four blocks. And each block contained all three types of tasks but arranged in different orders. The whole sequence is illustrated in Figure1.

T3	T1	T2	T2	T1	T3	T3	T1	T2	T2	T1	T3
Block 1			Block 2			Block 3			Block 4		

Figure 1. *Experiment design*

The valid recording time for a task is about 60s, which is believed to have best statistical properties. During the 60s, the event was presented continuously to try to keep the subject in the same mental state. We can simply assume the data follows identical distribution (MN) in the 60s. However, more and more evidence shows that the temporal network may contain important information (Preti, Bolton, and Van De Ville 2017; Damaraju et al. 2014). To this end, we assume that the distribution changed very slowly during this 60s, and set the parameter σ_u to 60.

3.2 Experimental Results

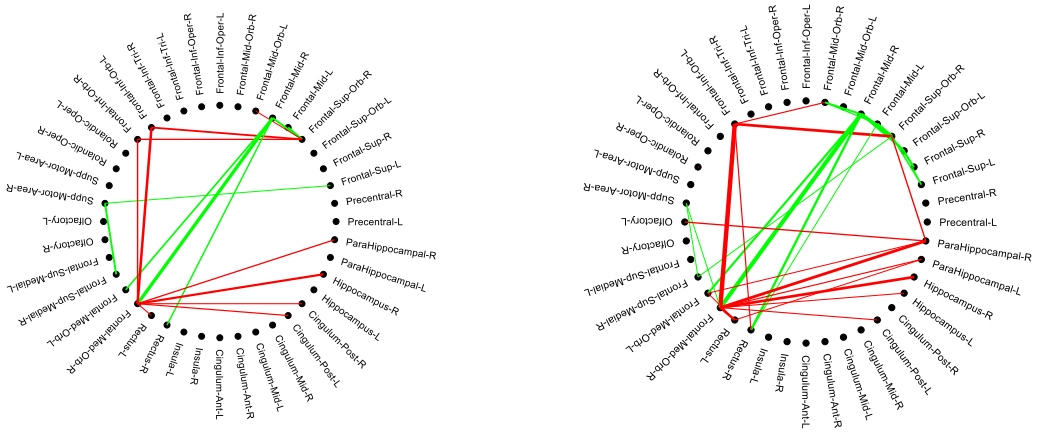
In this study, we compared the difference between HC-T2 and MS-T2, and the difference between HC-T3 and MS-T3. The samples for HC-T2 includes T2 readings from all four blocks of all HC subjects. The other sample sets are formed in the same way. In each experiment, the time point of interest is $\tau = \tau' = 30$. Applying the proposed method, the plots of the estimated difference graphs are given in Figure 2.

In Figure 2, we represent the results $\delta\theta$ with lines between two brain regions. The line width represents the change magnitude. Specifically, the green lines mean an increase of the dependence from MS to HC. The red lines indicate a decrease of the dependence from MS to HC. While the absence of line means there is no change happened.

These results suggest the neuroanatomical differences between MS and HC (van Antwerpen et al. 2018) may have functional consequences linked to fatigue and cognitive dysfunction in MS. However, the analysis

 ISBN: 978-1-5262-0820-0

we present here should be taken as a pilot study. A more delicate study would be carried out in cooperation with neuroscientists.



(a) *HC-Task 2 vs MS-Task 2* (b) *HC-Task 3 vs MS-Task 3*

Figure 2: *Estimated difference graph structures*

4. CONCLUSIONS

In this paper, we have developed a new method for direct estimation changes between two time-varying MNs. The method has extended the work in (Liu et al. 2014) to consider sequential data, by introducing an importance weighting scheme. The method was applied on a fMRI dataset, to help neuroscientists understand the neural mechanisms underlying cognitive impairments in MS disease. However, the results are preliminary and more inputs from neuroimaging specialists are required.

5. REFERENCES

Banerjee, Onureena, Laurent El Ghaoui, and Alexandre d'Aspremont. 2008. 'Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data', *Journal of machine learning research*, 9: 485-516.

Damaraju, Eswar, Elena A Allen, Aysenil Belger, Judith M Ford, S McEwen, DH Mathalon, BA Mueller, GD Pearlson, SG Potkin, and A Preda. 2014. 'Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia', *NeuroImage: Clinical*, 5: 298-308.

Fazayeli, Farideh, and Arindam Banerjee. 2016. "Generalized direct change estimation in ising model structure." In *International Conference on Machine Learning*, 2281-90.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2008. 'Sparse inverse covariance estimation with the graphical lasso', *Biostatistics*, 9: 432-41.

Kim, Byol, Song Liu, and Mladen Kolar. 2019. 'Two-sample inference for high-dimensional Markov networks', *arXiv preprint arXiv:.00466*.

Koller, Daphne, Nir Friedman, and Francis Bach. 2009. *Probabilistic graphical models: principles and techniques* (MIT press).

Lee, Su-In, Varun Ganapathi, and Daphne Koller. 2007. "Efficient structure learning of markov networks using $ l_1 $-regularization." In *Advances in neural Information processing systems*, 817-24.

Liu, Song, John A Quinn, Michael U Gutmann, Taiji Suzuki, and Masashi Sugiyama. 2014. 'Direct learning of sparse changes in Markov networks by density ratio estimation', *Neural computation*, 26: 1169-97.

Liu, Song, John A. Quinn, Michael U. Gutmann, and Masashi Sugiyama. 2013. "Direct Learning of Sparse Changes in Markov Networks by Density Ratio Estimation." In *Joint European conference on machine learning and knowledge discovery in databases*, 596-611.

Preti, Maria Giulia, Thomas AW Bolton, and Dimitri Van De Ville. 2017. 'The dynamic functional connectome: State-of-the-art and perspectives', *Neuroimage*, 160: 41-54.

Sugiyama, Masashi, Taiji Suzuki, and Takafumi Kanamori. 2012. 'Density Ratio Estimation in Machine Learning', *Cambridge University Press*.

ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Sugiyama, Masashi, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. 2008. 'Direct importance estimation for covariate shift adaptation', *Annals of the Institute of Statistical Mathematics*, 60: 699-746.

van Antwerpen, C, G Demarco, A Davies-Smith, R Jones, N Masuda, and NJ Thai. 2018. "Neural correlates of fatigue: a voxel-based morphometric MRI study in multiple sclerosis." In *ECTRIMS Online Library*, P1139.

Zhao, Sihai Dave, T Tony Cai, and Hongzhe Li. 2014. 'Direct estimation of differential networks', *Biometrika*, 101: 253-68.

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Extracting Emerging Patterns with Change Detection in Time for Data Streams

C.J. Carmona[1], A.M. Garcia-Vico[1], P. Gonzalez[1], M.J. del Jesus[1]

[1]Andalusian Research Institute on Data Science and Computational Intelligence, University of Jaen, 23071, Jaen SPAIN

[1]{*ccarmona | agvico | pglez | mjjesus*}*@ujaen.es*

ABSTRACT

Currently, many sources such as the Internet, sensor networks, the Internet of Things and others, generate information on a continuous basis, in a big data context. The processing of these data is well-known as data stream analysis, where data is processed as soon as it arrives in the system, not being stored. There is a large number of techniques for processing these data throughout the literature. In particular, emerging pattern mining attemps the search for discriminative relationships in data in order to describe differentiating properties between classes. Its main objective is the description of interesting knowledge in data with respect to a target variable. In this contribution, we raised the possibility to modify the extraction of emerging patterns in data streaming through the use of the time as target variable. This is well-known into the community as change mining. The benefits of extracting a precise and interpretable knowledge in data streaming could be very relevant for the community, because emerging patterns with change detection can support recommender systems by means of the extraction of alarms based on interpretable and precise patterns.

1. INTRODUCTION

Every day the amount of data generated has suffered an exponential growth since the last decades due to the Internet, new storage systems, mobile devices, social networks, and so on. These huge amounts of data are well-known as big data which is a hot topic on enterprises. Within big data, data streaming is a complex environment where data arrive into the system in an unbounded and ordered sequence throughout time at a variable speed. Latter, the machine learning technologies must be adapted because these must be able to work with data generated continuously and from different sources.

One of the most successful and widespread application of machine learning technologies in business are recommender systems (Lu et al., 2015). These can be applied in different scenarios where many users interact with many items such as retail, video on demand, music streaming, medical apps, e-commerce websites, amongst others. Within recommender systems two main categories of algorithms can be found (Burke, 2007): content based and collaborative filtering, although new recommender systems combine both types.

In data mining, there is a set of techniques halfway between predictive and inductive learning, the supervised descriptive rule discovery. This concept groups different techniques such as subgroup discovery (Herrera et al., 2011) or emerging pattern mining (Garcia-Vico et al., 2018), amongst others. In addition, a variation of the extraction of emerging patterns (EPs) for change detection in time was presented in (Liu et al., 2000). The main objective is to search for EPs throughout time.

The main objective of this proposal is to show the capabilities and benefits that EPs can bring to decision and recommendation systems, especially when the variable of interest in the analysis is the time in data streaming environments. It can be carried out thanks to the differentiating characteristics extracted from this type of algorithms, which are able to discover trends in data.

The document is organised as follows: Firstly, the main concepts of EP are briefly described in Section 2. Next, Section 3 presents change detection in time with concept drift which is a key component in data streams environment. Section 4 shows the properties and possible benefits of the use of EPs with change detection in time for data streams. Finally, the conclusions are outlined.

 ISBN: 978-1-5262-0820-0

2. EMERGING PATTERNS

The EP concept appears in (Dong & Li, 2005) and its goal is to search for itemsets whose support increases significantly from one dataset to another. Specifically, the concept is defined as (Dong & Li, 1999):

> *An EP is an itemset whose growth rate are larger than a given threshold p. When applied to timestamped databases, an EP can capture emerging trends, when applied to datasets with classes, the EP can capture useful contrasts between classes.*

$$GR(P) = \begin{cases} 0 & if\ Sup_{C1}(P) = Sup_{C2}(P) = 0 \\ \infty & if\ Sup_{C1}(P) \neq 0\ \wedge\ Sup_{C2}(P) = 0 \\ \dfrac{Sup_{C1}(P)}{Sup_{C2}(P)} & another\ case \end{cases}$$

The growth rate (*GR*) is the quality measure employed for analysing the patterns and it is calculated through the ratio of the support obtained for one class (Sup_{C1}) between the support obtained for another class (Sup_{C2}). A pattern is considered emerging if the ratio is upper than the threshold (*p*) which is considered to 1, generally. Throughout the literature a high number of restrictions with respect to *p* and *GR* have been presented in order to optimise the value of this quality measure.

This data mining technique has been widely employed in order to obtain good classifiers, however a recent study about EPs has been presented in (Garcia-Vico et al., 2018) analysing their properties within supervised descriptive rule discovery (Carmona et al., 2018) and not so much from the predictive perspective. The contribution presents the benefits in the use of this type of patterns in order to describe complex problems for supervised learning.

3. CHANGE MINING WITH CONCEPT DRIFT

The change detection in time is well-known into the literature as change mining. It appears in (Liu et al., 2000) associated to the decision tree models in order to give the characteristic descriptions of changes with respect to the error rate difference obtained in this type of models discovered on two different datasets. The concept is employed to discover the changes in new data incorporated with respect to old data. This concept is related with others terms appeared into the literature such as change patterns or changing domains (Chen et al., 2005).

It was defined as (Song et al., 2001):

> *The change detection problem consists of finding all emerging patterns, unexpected changes and added/perished rules between datasets which are collected at a different time and ranking the changed rules in each type by the degree of change.*

As can be observed in this definition, the main objective is to extract EPs which describe changes throughout time, i.e., for this problem class one (C1) are data at a given time, as opposed to data at another time which would be class two (C2). In fact, the majority of papers related to change mining are employed in e-commerce environments in order to analyse trends, data streams, sequences, change detection for decision support systems (Huang et al., 2018), and so on.

The interpretation of changes in trends is closely related to concept drift which occurs when a non-stationary target concept is modified in the environment. Formally, a real drift is defined by a change in time *t* with respect to a time *t + delta* on the conditional probability of the classes given an instance *x* (Gama et al., 2014). Considering both cause and effect of the probability distribution associated to data we can observe two types of drift (Krawczyk et al., 2017):

- Real drift. The probability distribution is modified with respect to the class but the incoming data probability could not be modified.

 ISBN: 978-1-5262-0820-0

- Virtual drift. Incomplete data representation or changes in data distribution without affecting the concept, for example.

In data streams environments, the concept drift must be analysed and taken into account for the design and development of the algorithms, regardless to the data mining induction performed.

4. EMERGING PATTERNS WITH CHANGE DETECTION IN TIME FOR DATA STREAMS

A data stream is, by definition, an unbounded, ordered sequence of instances that arrives at the system throughout time at a variable speed (Gaber, 2012). This simple definition of a data stream together drifting concepts leads to a high amount of differences with respect to classical static datasets and data mining. Some important aspects must be considered in order to create learning algorithms in this environment data: structures, constraints, methodology with respect to the analysis of data, informed mechanisms, and so on.

These aspects lead to the online adaptive learning where data are arriving to the system, and an output value is obtained using the current model. In a latter stage, the estimation of this prediction is analysed and the model can be modified or confirmed. In this way, data stream mining algorithms should determine how the instances will be processed, how they will learn from these instances in order to handle the concept drift and how the quality of the model will be determined. Thus, four important elements were defined in (Gama et al., 2014):

- memory,
- change detection,
- learning, and
- loss precision.

The memory component is very relevant because it is the element in charge of processing new information and forgetting the old one. On the other hand, change detection and learning methodology are components with a wide number of techniques and mechanisms in order to detect drift or changes in distributions, and to update structures employed. Finally, the loss estimation is related to the quality of the knowledge extracted.

In this contribution, we present the adaptation from EP to change mining in data streams environments. The objective is to extract rules in order to describe emerging behaviour between different time instants. Specifically, the main idea is to convert the time into the variable of interest for EP, batch t is C1 and batch t-1 is C2. In this way, the most relevant elements into the design are the memory component and loss precision. The algorithms must analyse the *GR* for new rules extracted in new data with respect to data of previous stages as can be observed in Fig. 1.

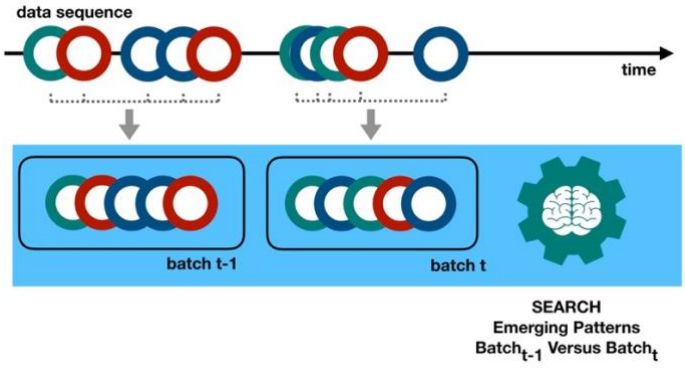


Figure 1. *Operation scheme in order to extract emerging patterns with change detection in time for data stream environments*

 ISBN: 978-1-5262-0820-0

Data are considered by means of batches where they are collected until a block of a predetermined number of instances is obtained and then the whole block is processed. The forgetting mechanism is related to the outdated observations for the batch t-1. On the other hand, we are able to detect possible change into the distribution thanks to the change detection in time.

The use of this methodology for real recommender and decision system allows us to discover alarms, also making us able to monitor the evolution of the systems and users over time. Finally, it is important to highlight that the use of EPs allows the consideration of different quality measures in order to show/analyse the quality of the systems in order to support the decision making in complex systems.

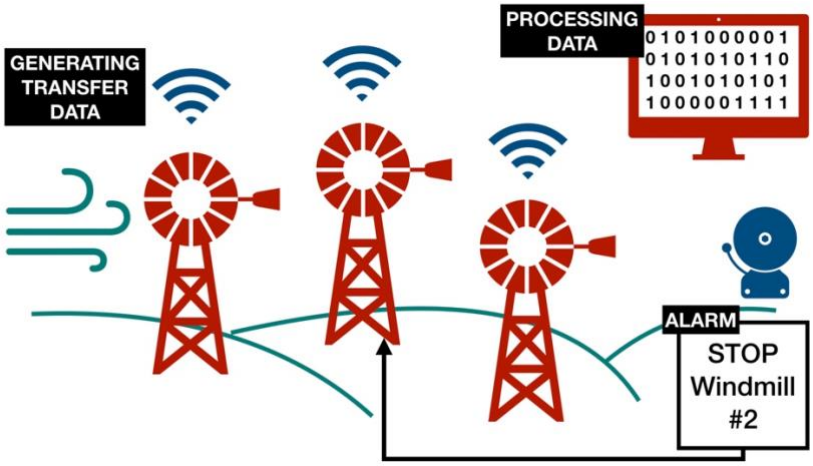


Figure 2. *Example of a real recommender system based on change mining for data streaming*

An example of a recommender system for data streaming can be observed in Fig. 2, where different devices in a wind turbine farm are generating data in a stream, and the data center is analysing and processing these data as they arrive to the system. In this way, the system is able to analyse data with respect to time in order to search for change detection, and system launches alarms considering the knowledge extracted. As you can see in Fig. 2, the windmill number 2 must be stopped due to a possible error in the mechanism with respect to the meteorological conditions and the performance of the remaining windmills.

Definitely, this technique allows us to integrate multiple measures simultaneously in the evaluation process, thereby evaluating and measuring multiple dimensions. Moreover, it could help and complement the recommender systems with interesting and unusual knowledge in real time.

5. CONCLUSIONS

The advance in technology and the current dynamic environments in which we are immersed cause the need to adapt and develop new data analysis models. In this contribution, we present the adaptation of EPs to change detection in time for data streaming environments. Their capacities and benefits for describing data from different perspectives (interpretability, generality and precision) make them in interesting for decision support and recommender systems, where experts need to understand the knowledge, especially in complex problems such as data streaming environments.

Acknowledgements: This work was supported by the Spanish Ministry of Economy and Competitiveness under the project TIN2015-68454-R (FEDER Founds) and FPI 2016 Scholarship reference BES-2016-077738 (FEDER Founds).

6. REFERENCES

R. Burke (2007). Hybrid web recommender systems. In: The Adaptive Web, Pages: 377–408. Springer.

C.J. Carmona, M.J. del Jesus, F. Herrera (2018). A Unifying Analysis for the Supervised Descriptive Rule Discovery via the Weighted Relative Accuracy. Knowledge-Based Systems, Volume 139, Pages: 89-100.

M.C. Chen, A.L. Chiu, H.H. Chang (2005). Mining changes in customer behavior in retail marketing. Expert Systems with Applications. Volume 28, Pages: 773–781.

G.Z. Dong, J.Y. Li (1999). Ecient Mining of Emerging Patterns: Discovering Trends and Dierences, in: Proc. of the 5th ACM SIGKDD. International Conference on Knowledge Discovery and Data Mining, ACM Press. Pages: 43–52.

G.Z. Dong, J.Y. Li (2005). Mining border descriptions of emerging patterns from dataset pairs. Knowledge and Information Systems. Volume 8, Pages: 178–202.

M.M. Gaber (2012). Advances in data stream mining, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 2, Issue 1, Pages: 79–85.

J.A. Gama, I. Zliobaite, A. Bifet, M. Pechenizkiy, A. Bouchachia, (2014). A survey on concept drift adaptation, ACM Computer Survey, Volume 46, Issue 4, Pages: 44:1–44:37.

A.M. García-Vico, C.J. Carmona, D. Martín, M. García-Borroto, M.J. del Jesus (2018). An Overview of Emerging Pattern Mining in Supervised Descriptive Rule Discovery: Taxonomy, Empirical Study, Trends and Prospects, WIREs Data Mining and Knowledge Discovery, Volume 8, Issue 1.

F. Herrera, C.J. Carmona, P. González, M.J. del Jesus (2011). An overview on Subgroup Discovery: Foundations and Applications, Knowledge and Information Systems, Volume 29, Issue 3. Pages: 495-525.

T.C.K. Huang, P.T. Yang, J.H. Heng. (2018). Change detection model for sequential cause-and-effect relationships. *Decision Support Systems*, Volume 106, Pages: 30-43.

B. Krawczyk, L. L. Minku, J.A. Gama, J. Stefanowski, M. Wozniak (2017). Ensemble learning for data stream analysis: A survey, Information Fusion, Volume 37, Pages 132–156.

B. Liu, W. Hsu, H.S. Han, Y. Xia (2000). Mining Changes for Real-Life Applications, in: Proc. on the 2nd Data Warehousing and Knowledge Discovery Conference, Pages: 337–346.

J. Lu, D. Wu, M. Mao, W. Wang, G. Zhang. (2015). Recommender system application developments: a survey. Decision Support Systems. Volume 74, Pages: 12-32.

H.S. Song, J.K. Kim S.H. Kim (2001). Mining the change of costumer behavior in an internet shopping mal. Expert Systems with Applications. Volume 21, Pages: 157–168.

ISBN: 978-1-5262-0820-0

The Alan Turing Institute

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Personalised Playlist Prediction

Lewis Bell[1], Carlos Del Rey[2], Eimear Cosgrave[3]

[1]Department of Computer Science, University of Bristol,
Bristol, UNITED KINGDOM

[2]Department of Computer Science, Universidad Carlos III de Madrid,
Madrid, SPAIN

[3]Department of Computer Science, Trinity College Dublin,
Dublin, IRELAND

lb16618@bristol.ac.uk, 2cdelreyg@ucsc.edu, 3ecosgrav@tcd.ie

[1]*cs.bris.ac.uk*, [2]*inf.uc3m.es*, [3]*scss.tcd.ie*

ABSTRACT

In recent years, the class of recommendation problems have been increasingly solved using Collaborative Filtering (CF) methods due to the increased viability of artificial neural networks in practical applications. We present a new novel method for Item-Based-CF recommendation using unsupervised clustering in Embedding Spaces of abstract song objects such as those produced by the Word2Vec methods detailed in Mikolov et al.'s seminal papers on the topic [1][2], abstracted away from the original use of finding natural language semantics to build embeddings of generalised items (So called 'Item2Vec' methods). We give an application of this concept in the form of recommending playlists to a user who has built playlists that they enjoy, based upon a dataset of 1.4 million unique songs contained within 11,000 playlists from other users to perform collaborative filtering, and detail the various other models and techniques we attempted on the way to our final solution. The results of these methods are successful by both empirical and subjective metrics, with our solution being suited to performing prediction on arbitrarily sized inputs, and being able to predict arbitrarily sized outputs.

1. INTRODUCTION

This project exists to investigate various contemporary methods of developing recommender systems using Item-Based Collaborative Filtering[3]. This is opposed to the more classical style of Content Based Filtering, which was preferred for many years until the modern developments of deep neural networks allowed for embedding spaces to be viable and performant options. We aim to develop an example of one such systems. This problem has been well explored in the past, being interpreted as both a supervised and unsupervised learning task in different implementations. For example, Huang's[4] implementation of using a Recurrent Neural Network with items mapped into an Embedding Space to learn sequences of items, with the aim of learning links between them. Alternatively, Gong et al.'s [5] technique views the problem as an unsupervised learning task, using clustering to group different users into clusters and recommending items to users in the same clusters. Our approach becomes novel in the fact that we take some inspiration from both of the sides presented above. We present an approach that does an unsupervised clustering of items **in** the embedding space. Then building a 'taste vector' of each user, assigning them to a cluster and recommending songs from the user's cluster.

 ISBN: 978-1-5262-0820-0

2. METHODS

2.1 Dataset

We begin with defining the terms *playlist* and *song*. A song is the combination of the song's name and the song's artist, this combination is assumed to be unique within the dataset. A playlist is an ordered list of songs designed to be listened to together. We selected a dataset from Kaggle named "950K music playlists: huge dataset of vk.com users playlists" (no longer available). Each element in the dataset has a playlist ID, a song title and the name of the artist. Therefore, each ordered list of songs with the same playlist ID constitutes one playlist. As we believe the order of the songs within a playlist is informative in and of itself, we did not shuffle the playlists.

The data was cleaned to comply with memory restrictions by first removing all datapoints that didn't match a simple regex of Latin characters and basic punctuation, then all playlists with fewer than 115 unique songs in them were also removed. After this, the dataset was formed by 4,151,989 songs in 11,682 different playlists. In a further implementation, the dataset could be expanded to include songs of all alphabets to decrease the bias towards songs with titles/lyrics in Latin-based languages.To perform feature selection, a `sklearn LabelEncoder` was used to assign a unique identification number to each unique *songArtist* pair. We finished this process with 1,455,005 unique songs. A simple one-hot encoding of each song in a playlist, or a composite many-hot encoding of the entire playlist was insufficient for our use-case of training a neural network, as the data was very highly dimensional. This inspired the use of an Variational AutoEncoder to construct a latent, lower-dimensional embedding space in which each song occupies a unique *m*-dimensional vector, and a playlist is an ordered list of these vectors.

2.2 Models and Algorithms

The objective of all models in this project was given a playlist of songs a user is known to like, recommend new songs not in the playlist they would also like.

Primarily, we attempted to train the most simple model we could imagine in order to get early results and judge the viability of our overarching idea. We attempted to use a simple two layer, fully connected, feed forward network. It would take a many-hot encoded vector (A vector with 1 in any position corresponding to that song being present) representing the playlist, and using a categorical-cross-entropy loss function, would put the input vector in a class representing the song closest to it. To achieve this, we treated each unique song as a category, and built a space of binary features with $|S|$ dimensions, where S is the set of all songs in the dataset. We defined a playlist as being a vector in this space, which would have a 1 in the component corresponding to the dimension of each song in the playlist, and 0s in all other positions.

The shortcomings of this model are twofold. First, there isn't really enough semantic information between the large groups of categorical data for a simple model to really converge any kind of relationship between input or output. Secondly, and more pressingly this model was almost impossible to train due to it's excessive memory requirements. As illustrated in **fig.1**, this model had $|S|$ neurons in both the input and output layers. We did not have access to a machine with sufficient memory space to be able to train the network as is.

We needed to compress our instance space somehow, the high dimensionality and sparseness of the instance vectors was simply too much of a problem. It was at this point we began to contemplate dimensionality reduction techniques.

 ISBN: 978-1-5262-0820-0

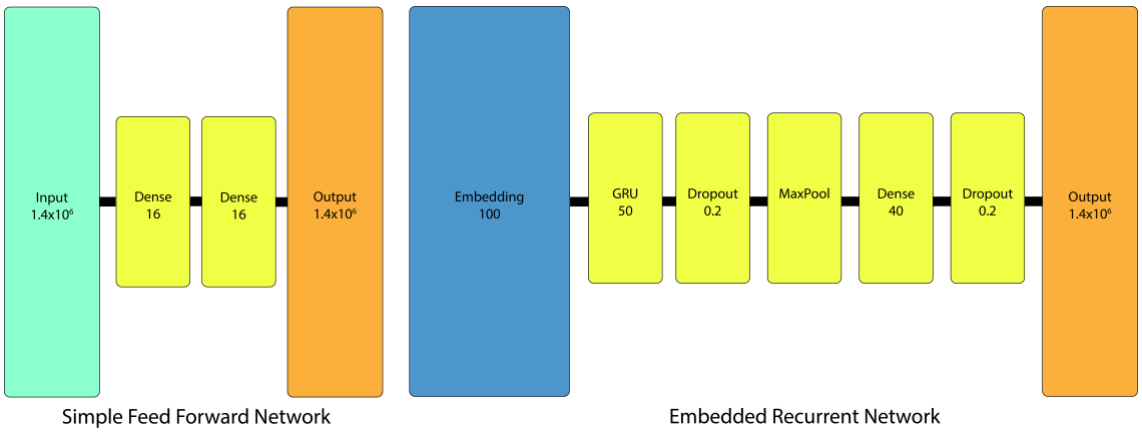
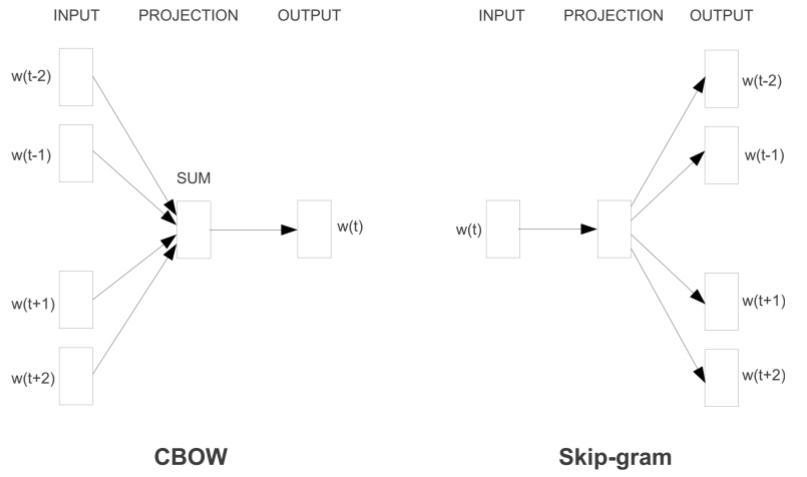


Figure 1. *The original two neural network architectures attempted. Left: Simple Fully Connected Neural Network. Right: Huang's Embedded Recurrent Neural Network*

Embedding Spaces are a concept routed in the field of Natural Language Processing. Firth's 1957 principal of Distributional Semantics[6] raised the concept of the meaning of a word emerging from the context in which it is used. It's important to mention here that only angle-based distance metrics such as cosine similarity have any semantic meaning in these spaces. Furthermore, there exists a linearity among the semantics of objects in an embedding space. While being theorised and conceptually sound for many years, like many concepts in Deep Learning, there were few practical applications of Word Embeddings until the early 2010s, when Mikolov et al. published their seminal paper on the Word2Vec models[1]. These models allowed for corpra



of text to be trained into word embeddings in a very efficient manner using a selection of simple neural network architectures.

Figure 2. *The Skip-Gram and Continuous Bag Of Words (CBOW) model architectures proposed by Mikolov et al. in [1]*

In recent years, the contemporary idea of generalising the concept of word embeddings away from their original domain of corpora of text can be seen in examples like [7]. The underlying concept is still the same, items gain their semantics from the other items they are frequently encountered with. For example, if *User A* purchases an alarm clock and some batteries, then *User B* also purchasing an alarm clock would also probably like some batteries as well.

 ISBN: 978-1-5262-0820-0

We trained a Word2Vec skip-gram network to build an item-embedding of our unique song ids. This embedding allowed us to represent any song from our originally *1,400,000* dimensional space in a much lower dimension (in our implementation. *D=100*), and we could even represent a whole playlist as the mean vector of all it's constituent songs. To reiterate, this embedding was built entirely on a processed corpora of lists of song IDs; nothing about the words of the song titles or artist names were trained over. Following some research, we decided to make use of the Recurrent Neural Network architecture proposed by Huang[4] for the purpose of doing collaborative-filtering based recommendation in embedding spaces. The Recurrent architecture is suited to the sequential nature of the playlists we were training over, and the architecture also includes dropout and dense layers to increase generality (see **fig.1** for full details of the architecture). While this improved the training time of our model (down to about 5 minutes per epoch), we found that the results were still not very good. Getting the network to converge was difficult and after trying a range of optimisers (RMSProp, Adagrad, Adam), none gave us satisfactory results.

It was at this stage that changed our approach to the problem. A range of classical supervised classification approaches had failed to provide any satisfactory results, so we decided to instead treat the problem as unsupervised clustering. The general idea here is to segment the instance space into a number of clusters, with each cluster containing a certain semantic group of items.

It is here that we present the novel concept of our research. Generally, a number of features are extracted from the data to form the dimensions of the instance space, with each item pertaining to a certain vector in the space, and the clustering is applied here. However, in our Collaborative-Filtering style of recommendation, we needed to be able to perform this clustering without knowing any intrinsic properties of the items themselves. As such, the clustering algorithms were applied *in* the embedding space, where the Word2Vec model has automatically determined a space of features for us to use.

Initially, we made use of a simple K-Nearest-Neighbours(KNN) model to recommend a list of items, given a list of inputs. The basic algorithm was to determine the vector representation of each item in the input playlist, and find the 'playlist vector' by taking the average of it's components. Note that we can do this due to the semantic linearity of vectors in embedding spaces[2], the concepts of adding and preforming scalar multiples have well defined semantics. Once we have determined the average playlist vector, we find the *n* vectors in the space that have the highest cosine similarity with the playlist vector. The model was then evaluated by computing the average vector of the predictions, and measuring it's cosine similarity with the average vector of a held-out validation set of *n* songs from the original playlist. The objective is to maximize this metric, as a cosine similarity of *1* represents that the two vectors are the same.

This technique gave good results, but we wanted to further investigate the potential of clustering methods. We implemented a simple K-Means clustering algorithm that given a number of clusters *k*, automatically segments the instance space. This method does not scale well, and was not suited to our desired application as it introduced very long train times. We then pivoted to an implementation of the MiniBatch K-Means Algorithm as presented in [8]. An important note to make is that the K-Means algorithm and it's derivatives rely on a euclidean distance metric being used to calculate the distances between cluster centeroids, but our embedding space relies on a cosine similarity distance metric. To solve this issue, it is enough to simply normalize the vectors in the embedding space before applying the k-means algorithm. This is due to cosine similarity being linearly related to euclidean distance for normalized vectors.

After experimenting with various different batch sizes and numbers of clusters, we applied the same evaluation strategy as with the KNN algorithm presented above. We were very pleased with the results.

3. RESULTS, ANALYSIS AND CONCLUSIONS

Here our results are presented for the two experiments which produced measurable results. None of the supervised-learning based models we attempted to use finished, either due to the memory issues detailed above, or because of difficulties getting the model to converge during training.

The first viable model was performing K-Nearest-Neighbours on items in the embedded space produced by the Word2Vec model. The following clustering models require very little training time before predictions are ready to be made, but as a trade-off the prediction time involves significantly more computation compared to the supervised methods used earlier.

An important decision to be made was the evaluation metric by which we judge success of our solutions, as there is no measurement intrinsic to the model (Unlike in a neural network for example, which features a loss function). We decided on using the average cosine similarity between the songs in the predicted playlist and

 ISBN: 978-1-5262-0820-0

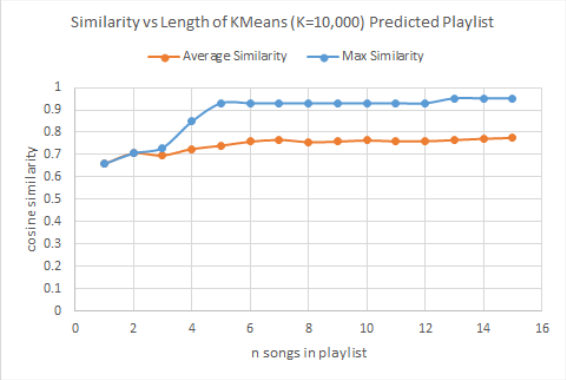
the songs contained in a held-out test set from each playlist, as the cosine similarity has a strong semantic meaning, and a higher cosine similarity means the two sets of songs are more similar.

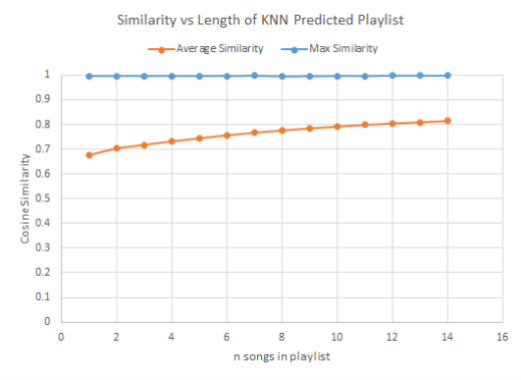
For all subsequent experiments, We measured how the cosine similarity changed as we varied the number of songs being predicted per playlist. Generally, the more songs in the output playlist, the better the cosine similarity, up to a point where there are diminishing returns.

Our first experiment began with using KNN (fig.\ref{fig:knn}), this method was a lot more efficient to use than the previous neural network based approaches, with each prediction taking about 90 seconds to complete. The results here were pleasing, as we saw a peak average cosine similarity of *0.81*.

In an attempt to improve on these scores, we began making use of a K-Means based model. Along with the parameter of the length of the predicted playlist, we also could vary the number of clusters to group into. Hence the final 3 graphs in fig.\ref{fig:results} show the results of K-Means with varying Ks. The results show that increasing the number of clusters definitely increased the performance of our solution, though again with diminishing returns. We settled on a final implementation of *20000* clusters, the best performing in our tests. It is worth mentioning that the K-Means method also requires more significant training time than the KNN model, with the *20000* cluster model taking approximately 45 minutes to train on our machine, however each prediction now takes *~0.68* seconds.

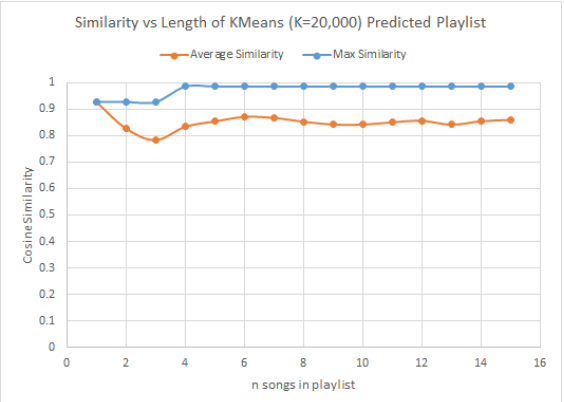
It is obvious from the graphs that the *20000* cluster K-Means model is the highest performing of our tested models. We believe this is due to there being approximately 20,000 distinct user groups visible in our data, and at this level we can maximally assign one user to a taste cluster they fit in the best. In terms of future work, we believe we could improve the model by using variations on K-Means such as Gaussian Mixture Models to allow for songs to belong to more than one cluster, and by decreasing the biases in our model by

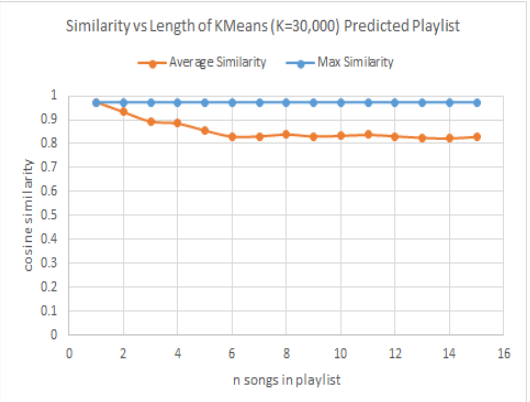




expanding the variety of songs in our dataset.

Acknowledgements: Professor Narges Norouzi (JBSOE, UCSC).





 ISBN: 978-1-5262-0820-0

4. REFERENCES

[1] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[2] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Advances in neural information processing systems, pp. 3111 3119, 2013.

[3] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl, et al., "Item-based collaborative filtering recommendation algorithms.," Www, vol. 1, pp. 285–295, 2001.

[4] S. Huang, "Introduction to recommender system. part 2 (neural network approach)," Feb 2018.

[5] S. Gong, "A collaborative filtering recommendation algorithm based on user clustering and item clustering.," JSW, vol. 5, no. 7, pp. 745–752, 2010.

[6] J. R. Firth, "A synopsis of linguistic theory 1930-55.," vol. 1952-59, pp. 1–32, 1957.

[7] F. Vasile, E. Smirnova, and A. Conneau, "Meta-prod2vec: Product embeddings using side-information for recommendation," in Proceedings of the 10th ACM Conference on Recommender Systems, pp. 225–232, ACM, 2016.

[8] D. Sculley, "Web-scale k-means clustering," in Proceedings of the 19th international conference on World wide web, pp. 1177–1178, ACM, 2010.

ISBN: 978-1-5262-0820-0

The Alan Turing Institute

User-centric design of a clinical decision support system for critical care treatment optimisation.

CJ McWilliams[1*], IG Gilchrist[1], MJ Thomas[2], T Gould[2], RS Rodriguez[1], CP Bourdeaux[2]

[1] University of Bristol, Bristol, UK
[2] University Hospitals Bristol NHS Foundation Trust, Bristol, UK

**Corresponding author: chris.mcwilliams@bristol.ac.uk*

ABSTRACT

In this paper we present the concept for an interactive clinical decision support system that will replace the intensive care unit dashboards currently deployed at the Bristol Royal Infirmary. The proposed system is intended to promote compliance with treatment guidelines and will improve on the pre-existing dashboards by introducing predictive modelling, capturing usage data for algorithm development and providing an enhanced user interface that is co-created by our interdisciplinary research team and the clinical users of the system. The intention is to design both the software and the algorithms such that the system could be deployed across multiple NHS trusts in the future.

1. INTRODUCTION

The intensive care unit (ICU) is a high-tech and data-rich environment where critically ill patients undergo sophisticated monitoring and organ support. Increasingly ICUs are collecting and storing these data in clinical information systems (CIS) as part of routine care. The data are used in real-time to inform clinical decision making and retrospectively for audit and research [Johnson 2016, McWilliams 2019(1), Shillan 2019]. At University Hospitals Bristol the CIS product (*IntelliSpace Critical Care and Anaesthesia)* is provided by Philips Healthcare and is used across the four intensive care units in the trust. This product provides a graphical user interface that aims to replace traditional paper charting. It is a high-quality piece of commercial software that is used routinely to capture a data-rich description of each patient's stay on ICU. The data collected include medications and interventions, vital signs, laboratory results and free-text medical notes. Some fields are entered manually by clinical staff while others are captured via live data-streams from monitoring and treatment devices. Although the CIS is a valuable tool, the volume of data collected can make it difficult to manually review all the information that is relevant to any given clinical decision and there has been some resistance to adoption because of perceived limitations and usability issues.

Across medicine the translation of evidence-based interventions into clinical practice is poor [Morris 2011] and in the ICU this problem is particularly pronounced because of high patient heterogeneity and the complexity of individual cases. Intensive care patients present with complex co-morbidities and there is a significant cognitive burden on clinical staff, with intensivists delivering up to 200 interventions per patient per day [Pronovost 2013]. Treatment guidelines are regularly not met and clinical practice is often variable. These problems are not the result of individual decision 'errors', but rather are a feature of the challenging decision-making environment in which the ability of the individual to make decisions is constrained by a range of social and organisational factors [Tversky 1981, Starbuck 2015]. Under these conditions the rational model of decision-making [Simon 1955], whereby the individual integrates all relevant information to make the optimal decision, is not applicable. In this context there is extensive scope for decision support systems to aid clinical staff in making better decisions by performing algorithmic data integration and reducing cognitive burden wherever possible.

We have previously demonstrated that behavioural interventions targeted at clinical staff can reduce variability in care and improve compliance with clinical guidelines [Bourdeaux 2014, 2015 and 2016]. These interventions were intended to alter clinician behaviour using 'nudge' principals. However, the goal was not to override clinician autonomy, but rather to reduce the need for cognition in actions that can be safely automated and to focus clinical attention on tasks where human expert deliberation is needed. For example, there is no need for a doctor to actively prescribe chlorhexidine mouthwash to a mechanically ventilated patient. All such patients should receive this treatment so it can be safely automated via 'opt-out' prescribing. Similarly, there is no sense in a human having to divide breath volume by a patient's ideal body weight and check that the result does not exceed 6ml/kg. It is more effective to get a computer to perform the calculation

 ISBN: 978-1-5262-0820-0

and notify the clinician when this value is exceeded. The clinician may still decide to proceed to ventilate the patient with over-sized tidal volumes, but this results from a conscious decision rather than a failure to integrate information.

The main avenue for these behavioural interventions has been our clinical dashboards, which add a basic decision support layer on top of the Philips CIS by condensing and presenting key patient data in an easily interpretable format (see section 2). We are now looking to develop the next generation of these dashboards, which will incorporate machine learning and predictive models in order to support improved decision making and to avoid deviations from treatment guidelines before they occur. For example, we are currently developing a regression model to predict patient fluid balance at the end of each day. This prediction will help nurses at the bedside adjust hourly diuretic dosages in real-time in order to hit midnight fluid balance targets. Similarly, our recent work [McWilliams 2019(2)] has laid the foundations for a discharge planning tool which, complemented by other algorithms [Badawi 2012, Desautels 2017], could help clinicians evaluate risks during discharge decisions. With the increasing availability of critical care datasets [Harris 2018, Johnson 2016] the rate of development of new algorithms and methods is going to increase. A successful decision support system will need to be able to deploy these new algorithms in a clinical environment as and when they are proven safe and effective. We intend to develop a system with this extendibility, beginning with pre-existing algorithms and building on previous successes in deploying simple but effective clinical decision support.

2. SYSTEM DESIGN

2.1 The current clinical dashboards.

The dashboards currently in use on our intensive care unit (Figure 1) show selected variables in a table format. Each row corresponds to a patient and each column to a different variable of interest. For example, the third column in Figure 1 represents the quality of sleep of each patient over the last 24 hours and the fourth column gives the tidal volumes (*VT kg*), which are only relevant for mechanically ventilated patients. In general, variables are colour coded with values of concern illustrated in orange or red and acceptable values in green or blue. The dashboards are mounted in four public locations in the unit, where they are visible to all staff.

Despite their proven efficacy the current dashboards have limited functionality. They provide a static view on data which is extracted from the Philips CIS with conditional formatting to highlight problematic or abnormal variable values. The technology behind these dashboards (*SQL Server Reporting Services*) would not easily support the introduction of more complex data analysis, algorithmic decision support and user-interactivity. We feel that the inclusion of these features will maximise the utility and efficacy of the system.

Bed Number	Consultant/Nurse	Sleep Wake Flag	VT kg	p O2	PAC Flag	Nursing Tasks	Medical Tasks	CUR Review Decision	Level	Transfer Ward
01	Dr		6.17	10		VAP bundle	VAP bundle	Q		
02	Dr			9		VAP bundle	VAP bundle	Q		
03			9.82	12		VAP bundle	VAP bundle	Q		
04	D		9.44	5				Q		
05			11.44	8		VAP bundle	VAP bundle	Q		
06	Dr			10		VAP bundle	VAP bundle	Q	2	
07	D					VAP bundle	VAP bundle	Q		
08	Dr			8				Q		
09	Dr		5.15	10		VAP bundle	VAP bundle	Q		

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Figure 1. *Snapshot of the current dashboard on the general intensive care unit at the Bristol Royal Infirmary. Each column represents a physiological or treatment variable of interest, and abnormal values are coloured orange or red according to severity. This simple dashboard has proven effective in altering clinical practice (see main text). Consultant and nurse names have been masked for data protection.*

2.2 The new clinical dashboards.

We propose replacing the current dashboards with an interactive system. The preliminary design, sketched in Figure 2, has emerged from discussions with clinical staff who use the current system. The intention is to retain a simple and intuitive visual interface whilst extending the functionality. The next stage of the project will consist of a user-centric design process [Brunner 2017] in which we will hold a series of co-creation workshops with representatives from all clinical user groups to fully specify functional requirements and to iteratively improve the design. We envisage adhering to the following four design principles for the new system:

- *Smart* – to include machine learning models which compute personalised risks and predict problems before they occur.
- *Interactive* – to allow users to engage with interpretable algorithm outputs and to query possible treatment decisions.
- *Responsive* – customisable and adaptable to respond to current clinical and operational concerns on the unit.
- *Extendible* – designed to facilitate continuous integration of new features and models derived from ongoing research.

2.3 The technology behind the new dashboards.

The system will be a Flask web application that is hosted on a hospital server – the 'application server' shown in Figure 3. The application will interact with the Philips CIS database, which is stored on another server, via a bespoke SQL adaptor to extract live patient data. To do this the application will periodically query the SQL reporting database of the CIS, ensuring that the displayed data is up to date but minimising the number of individual queries run. Users will be able to access the system via secure log-in from any trust device, which will enable usage on ward rounds, during office meetings and outside of ICU (e.g. for bed planning). Notifications will be provided as visual alerts on the public dashboard touchscreens displayed around the ICU, and users will be able to click-through for more information about each notification (Figure 2). Alerts will also be sent to *Careflow*, which is a mobile application used by clinical staff to manage tasks relating to each of their patients. This type of integration with pre-existing technologies and workflows will maximise the uptake and efficacy of the system. As shown in Figure 3, the intention is that the system could also connect to data feeds from external NHS trusts. We believe that a real-time comparison of treatment delivery between multiple intensive care units could promote changes in clinical practice that would benefit many patients.

In the first instance the application will only extract data from the CIS and any data produced by the application will be stored locally on the application server (see Figure 3). These data will include outputs of algorithms, user-interactions with the system and manually input data fields, all of which could be used for learning to improve the system. In the future there is scope for a two-way data feed such that locally generated data could be fed back into the CIS database where relevant. We are also keen to extend the application to interact with databases from the wider hospital such as *Medway* (admissions, demographics etc) and *VitalPac* (electronic ward observations), which would increase the volume and the type of data available to drive decision support. Initially the application stack would be full-Python and as such the machine learning components would make use of libraries such as Scikit-learn, PANDAS and KERAS. Future iterations should incorporate functionality from R, which is preferred by many clinical staff and biomedical researchers, via the Python libraries PypeR or RPy2. The inclusion of R would allow clinical staff to develop their own decision support algorithms and statistical models in-house in accordance with our extendible and user-centric design principles. The success of this approach would require a well-documented API and clear tutorials.

 ISBN: 978-1-5262-0820-0

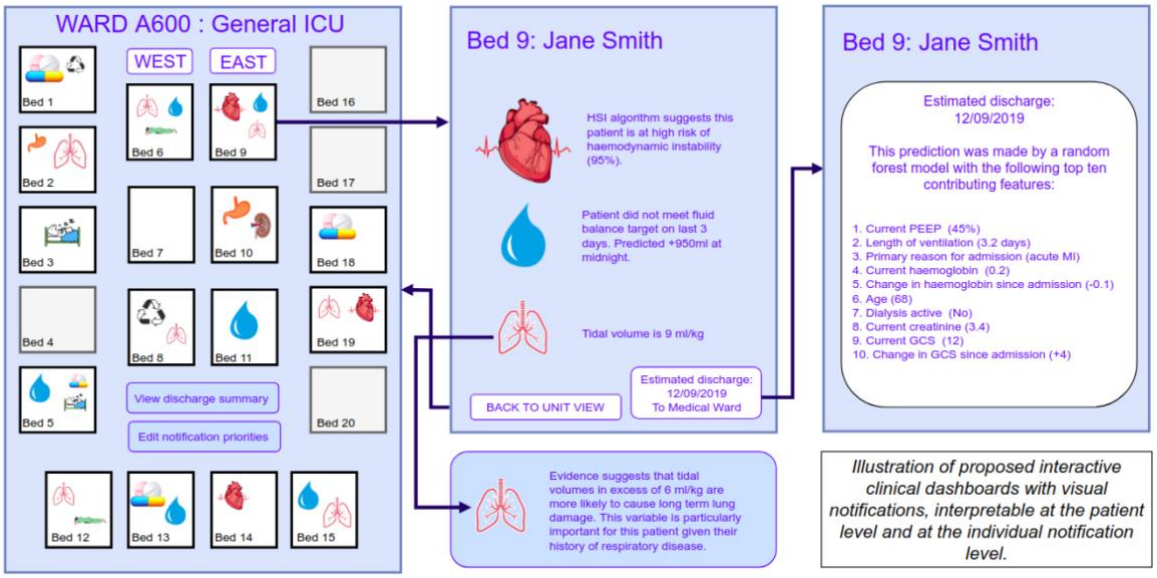


Figure 2. *Sketch of the proposed interactive dashboard interface. Notifications are given via a graphical icon in the relevant bed space. Arrows indicate touchscreen transitions between views.*

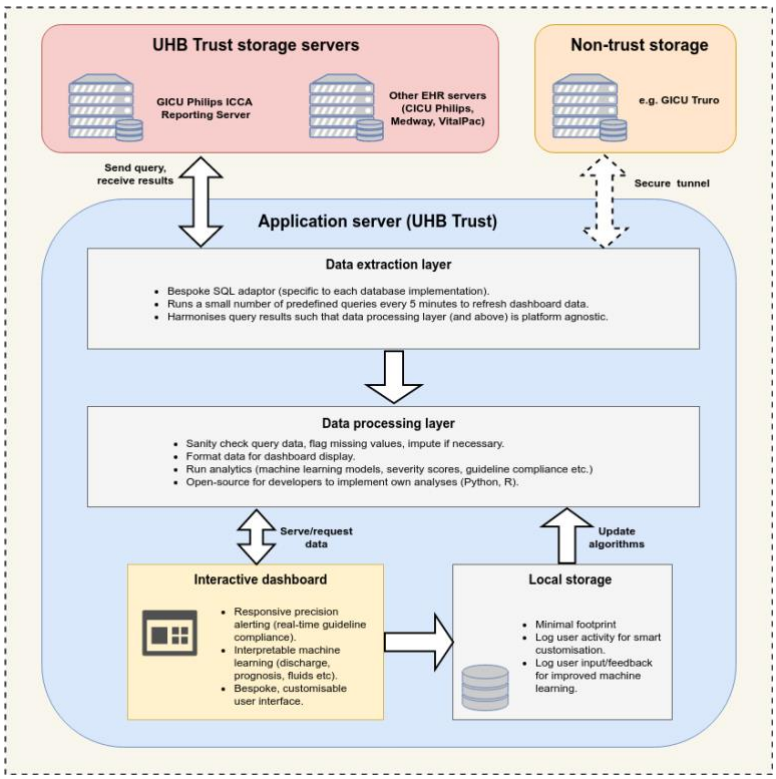


Figure 3. *Schematic of the Python technology stack supporting the proposed interactive dashboards. The dashboard is a Flask application and the SQL adaptor uses PyODBC to connect to a Microsoft SQL Server instance. UHB: University Hospitals Bristol; GICU: General Intensive Care Unit; CICU: Cardiac Intensive Care unit; ICCA: IntelliSpace Critical Care and Anaesthesia (clinical information system).*

 ISBN: 978-1-5262-0820-0

3. CONCLUSION

In this extended abstract we have laid out the motivation, design considerations and proposed functionality for a clinical decision support system for critical care. Extensive research will be required order to arrive at a safe and effective system, and this work will be carried out in collaboration between the Bristol Royal Infirmary, the University of Bristol and with input from several partner ICUs from across the regional intensive care network. The clinician-led user-centric design process will ensure the system is well aligned with clinical need and that it is embedded in the cognitive ecology [Hutchins 2017] of the intensive care unit and the multidisciplinary decision-making processes that underpin patient care. The system will also be designed with a user-centric approach to evaluation [Pu 2011] according to which iterative improvements will be made that maximise uptake and efficacy. Finally, a modular design will ensure the extendibility of the system to benefit from the ongoing development of algorithms in the field of critical care.

Acknowledgements: We would like to thank all UHB staff who have supported the use of clinical dashboards for ICU and the re-use of data to improve patient care. We also thank the *Elizabeth Blackwell Institute*, the *EPSRC IAA* and *Above and Beyond* for their contributions to funding this work.

4. REFERENCES

Badawi, O., & Breslow, M. J. (2012). Readmissions and death after ICU discharge: development and validation of two predictive models. *PloS one*, *7*(11), e48758.

Bourdeaux, C. P., Davies, K. J., Thomas, M. J., Bewley, J. S., & Gould, T. H. (2014). Using 'nudge'principles for order set design: a before and after evaluation of an electronic prescribing template in critical care. *BMJ Qual Saf*, *23*(5), 382-388.

Bourdeaux, C. P., Birnie, K., Trickey, A., Thomas, M. J. C., Sterne, J., Donovan, J. L., ... & Gould, T. H. (2015). Evaluation of an intervention to reduce tidal volumes in ventilated ICU patients. *British journal of anaesthesia*, *115*(2), 244-251.

Bourdeaux, C. P., Thomas, M. J., Gould, T. H., Malhotra, G., Jarvstad, A., Jones, T., & Gilchrist, I. D. (2016). Increasing compliance with low tidal volume ventilation in the ICU with two nudge-based interventions: evaluation through intervention time-series analyses. *BMJ open*, *6*(5), e010129.

Brunner, J., Chuang, E., Goldzweig, C., Cain, C. L., Sugar, C., & Yano, E. M. (2017). User-centered design to improve clinical decision support in primary care. *International journal of medical informatics*, *104*, 56-64.

Desautels, T., Das, R., Calvert, J., Trivedi, M., Summers, C., Wales, D. J., & Ercole, A. (2017). Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. *BMJ open*, *7*(9), e017199.

Drew, B. J., Harris, P., Zègre-Hemsey, J. K., Mammone, T., Schindler, D., Salas-Boni, R., ... & Hu, X. (2014). Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PloS one*, *9*(10), e110274.

Harris, S., Shi, S., Brealey, D., MacCallum, N. S., Denaxas, S., Perez-Suarez, D., ... & Beale, R. (2018). Critical Care Health Informatics Collaborative (CCHIC): Data, tools and methods for reproducible research: A multi-centre UK intensive care database. *International journal of medical informatics*, *112*, 82-89.

Hutchins, E. (2017). Cognitive ecology. In *Introduction to Vygotsky* (pp. 226-236). Routledge.

Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*, 160035.

McWilliams, C., Inoue, J., Wadey, P., Palmer, G., Santos-Rodriguez, R., & Bourdeaux, C. (2019). Curation of an intensive care research dataset from routinely collected patient data in an NHS trust. *F1000Research*, *8*, 1460.

McWilliams, C. J., Lawson, D. J., Santos-Rodriguez, R., Gilchrist, I. D., Champneys, A., Gould, T. H., ... & Bourdeaux, C. P. (2019). Towards a decision support tool for intensive care discharge: machine learning algorithm development using electronic healthcare data from MIMIC-III and Bristol, UK. *BMJ open*, *9*(3), e025925.

Morris, Z. S., Wooding, S., & Grant, J. (2011). The answer is 17 years, what is the question: understanding time lags in translational research. *Journal of the Royal Society of Medicine*, *104*(12), 510-520.

Pronovost, P. J. (2013). Enhancing physicians' use of clinical guidelines. *Jama*, *310*(23), 2501-2502.

 ISBN: 978-1-5262-0820-0

Pu, P., Chen, L., & Hu, R. (2011). A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems* (pp. 157-164). ACM.

Shillan, D., Sterne, J. A., Champneys, A., & Gibbison, B. (2019). Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Critical Care*, *23*(1), 1-11.

Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, *69*(1), 99-118.

Starbuck, W. H. (2015). Karl E. Weick and the dawning awareness of organized cognition. *Management Decision*, *53*(6), 1287-1299.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

A hybrid decision making system using image analysis by deep learning and IoT sensor data to detect human falls

Pingfan Wang[1], Sean McGrath[2], Nanlin Jin[3]

[1,3] Faculty of Engineering and Environment, Northumbria University, Newcastle upon Tyne, Newcastle, United Kingdom

[2] Department of Electronic and Engineering, University of Limerick, Limerick, Ireland

[1]pingfan.wang@northumbria.ac.uk, [2]Sean.Mcgrath@ul.ie, [3]nanlin.jin@northumbria.ac.uk

[1,3]www.northumbria.ac.uk/, [2]https://ece.ul.ie/

ABSTRACT

Falling is the leading cause of injury and accidental death for those who are older or reduced disability. In recent years, the problem of population aging is becoming more and more serious, and the number of disabled people is also growing. There is evidence that people with disabilities are at higher risk of injury from falls than a person without a disability. Therefore, the high-efficiency and accurate fall detection are highly demanded, to protect the health and safety of the elderly, at the same time, it helps to reduce the burden on individuals and society. This paper proposes a method that combines image processing and IoT to analyze and process input image and sensor data, then predict the possible physical state of the people, which could be used as the basis for subsequent decision making of health care. In this method, image data is processed and analyzed by deep learning network and feature matching algorithm, the sensor data is collected and processed by multiple sensors and cloud platform. We first get the analysis results of image input and sensor data respectively, then the hybrid decision making system will be employed to combine two results to obtain the most similar matching action, finally, the result whether the elderly fall is obtained. Compared to the use of image or sensor only, the result of the hybrid decision making method is more accurate and less delayed. The result of the experiment has shown the hybrid decision method based on data fusion has higher accuracy and low latency.

1. INTRODUCTION

In the modern life, the living environment of people has been improved by the emergence of the new technology, but for some people who in the poor body situation or mental illness, is difficult for them to ask for help and timely treatment if there is falling or other accidents. There are three mainstream methods are employed to this problem, sensors, vision and data fusion respectively. The sensors-based approach mainly relays on the wearable sensors and accelerator to detect the posture of the body, A two-axis accelerometer with a posture sensor was used in [1] for fall detection. [2] propose a fall detection system consisting of an inertial unit that includes triaxial accelerometer, gyroscope, and magnetometer with efficient data fusion and fall detection algorithms, its result shows excellent accuracy by placing the wearable sensor on the waist of subject. However, the intrusion and the fixed relative relations with the object, which could cause the device to be easily disconnected.

Furthermore, for the vision-based approach, machine learning and deep learning are the mainly used methods, which has higher accuracy and less intrusion. In [3] a neural network system was incorporated in the fall detection computation algorithm, which can successfully identify the falls in specificity and sensitivity perspective. [4] proposed a vision-based solution using the convolutional neural network, decide if a sequence of frames contains a person falling. This method requires a lot of computing resources and strict requirements for the applicable scenarios.

Moreover, Data fusion technology is increasingly used in fall detection, [5] proposes an improved framework by fusing the Doppler radar sensor result with a motion sensor network, to offset false alarms

 ISBN: 978-1-5262-0820-0

from visitors. [6] have developed an algorithm that combines both features extracted from the signal of a tri-axial accelerometer and one ECG lead, several different kinds of data are merged to classify body postures. [7] adopts several heterogeneous sensors to collect data and experiment accuracy has been improved compared to the application of individual sensors. A convolutional neural network-based sensor fusion system is proposed in [8], to monitor falls in healthcare application, which simultaneously using deep image input and inertial signal, the result indicates the system is effective to monitor the fall as well as the transition movements. However, there are a few ways to combine visual data and perceptron data for data fusion. Many previous studies have been based on different sensor data fusions. Data fusion results in different dimensions will make the prediction results more accurate and reliable.

2. Hybrid decision making system architecture

The hybrid decision making system consists three parts, semantic segmentation and analysis, IoT (Internet of Things) sensors data collection and analysis, and the integrated decision-making, respectively, as seen in Figure 1. The originality of our work is that the decision is jointly made by both the results of image processing and the sensor data collected via the IoT platform, the result of the image processing is used as the main basis for decision making, the result of the IoT data as a supplementary basis for the unsure judgement.

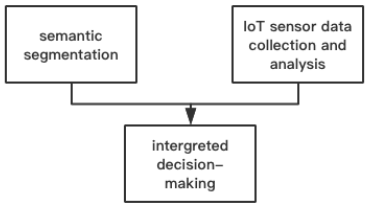


Figure 1. *The procedure of the semantic segmentation for the people indoor.*

2.1 Image analysis: semantic segmentation and analysis based on deep neural network

DNN (Deep Neural network), compare to simple neural network, the number of the hidden layers is much more than the simple neural network, usually it will reach dozens of layers. FCN (Fully Convolutional Network), indicates that the neural network is composed of convolutional layers without any fully-connected layers usually found at the end of the network. To our knowledge, the idea of extending a convert to arbitrary-sized inputs first appeared in [9], which extended the classic LeNet to recognize strings of digits, because their net was limited to one-dimensional input strings, used Viterbi decoding to obtain their outputs.

2.1.1 Image data process and body part extraction

ResNet (Residual Neural Network) was proposed by Kaiming He of Microsoft Research Institute [10]. It has several different network forms depending on the depth of the network, the ResNet-101 is adopted based on the consideration of top-1 and top-5 error rate.

The trained dataset is the Ade20k, which is compressive dataset that consists of 20,210 annotated images, compared to other similar datasets, it contains more indoor images that facilitate to the improve the accuracy of the network. The trained network has been employed to process the image input, the body part is extracted, as seen in Figure 2.

Figure 2. *The procedure of the semantic segmentation for the people indoor.*

2.1.2 Action dataset building

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

we have developed an IoT system which collects the following dataset, which has multiple kinds of actions. It contains most of the cations in everyday life, including normal standing, bending, squatting, walking, sitting, and other unusual movements such as kneeling, sitting on the floor, lying down, and other abnormal actions. The actions with intermediate state between normal and abnormal actions is also included, which aims to increase sample integrity and diversity, also guarantee the reliability. The part of the dataset is seen in Figure 3.

Figure 3. *Part of the action dataset.*

2.1.3 Feature matching and similarity calculation

The feature matching is mainly applied based on the SIFT (Scale-invariant feature transform) [9], which is invariance to image scale and rotation,

The image inputs will be divided into three parts, the first is the normal action, include but not limited to the standing, kneeling, walking, sitting; the second part is the abnormal action, include but not limited to falling, lying down; the last part is the confusion group, which is the transition state of normal action to abnormal action. Originally, each test group has 30 images, after rotating 90 degrees, 180 degrees and 270 degrees respectively, the total number of samples is increased to 360.

After feature matching and similarity calculation with the action dataset, the list of the similarity result will be calculated, we define the action in the action dataset with the highest similarity to the test image as the final result.

2.2 IoT sensors data collection and analysis

The physiological indicators of the human body are closely related to the state of the human body. When an accident occurs, people's heartbeat will speed up, body temperature will rise, and people will sweat more. To this end, several physiological sensors such as heart rate sensors, body temperature sensors and humidity sensors are recommended for use in wearable bracelets. At the same time, heart rate acceleration is the most significant parameter when compared to other physiological manifestations.

We have used the bracelet in our experiments, which include BME680 environment sensor, DHT11 temperature & humidity sensors, and Heart rate sensor developed by the University of Limerick. It is the integrated framework that consists of the three types of sensors, the data will be collected by the different type of sensors, then be transmitted to the raspberry pi, then the final data operation will be carried on in the cloud platform, as seen in Figure 3.

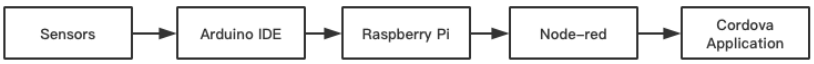


Figure 3. *The procedure of the IoT sensors data collection and analysis.*

The bracelets have been deployed in the human body and monitor the multiple physical indicators in real-time, the physiological state human body is divided into two states, a normal state and a special state. Generally speaking, the physical state of the fallen people will be considered a special state.

2.3 Integrated decision-making

Our proposed integrated solution will fill the gap in integrated decision making. As seen before, the result obtained in the image semantic segmentation phase is not ideal, especially for the confusion group, its correct rate is much smaller than other groups. It cannot be used alone as a basis for decision-making, on the other hand, the result from the IoT data analysis could be the strong evidence for the final decision generation. Generally, people in an unexpected state will inevitably undergo physiological changes, heartbeat acceleration may be the most obvious feature. But such changes are not only associated with falls.

 ISBN: 978-1-5262-0820-0

In terms of the comprehensive consideration for the decision-making, the result of the semantic segmentation will be marked separately, the state belongs to the normal group is marked as M_n, the confusion group corresponds to M_c, the abnormal group corresponds to M_a. For the purpose of the calculation, $M_n = 1$, $M_c = 0$, $M_a = -1$.

For the result of the IoT data analysis, the normal state is marked as I_n, and the special state is marked as I_s. $I_n = 1$, $I_s = -1$.

The evaluation result of the whole system is represented by score S,

$$S = w_m * M_i + w_i * I_j \tag{1}$$

The w_m and w_i are the weights of the semantic segmentation result and IoT data analysis result, respectively, they were initialized to 0.5. M_i represents the result of semantic analysis, and the I_j represents the result of the IoT data analysis.

$$S = \begin{cases} 1, & s > 0 \\ 0, & s = 0 \\ -1, & s < 0 \end{cases} \tag{2}$$

If the value of S = 1, which means the people not fell; S = 0, which means the states of the people are uncertain; S = -1, which means people fall.

2.4 Test and evaluation

According to section 2.1.3, the test data is divided into 3 groups, each has 120 test data. The first step of the test aims to get the result of the image input processing, at the same time, the second step is to collect and analyze relevant sensor data from the wearable bracelet on the people.

***Table 1.** detection result only applied the semantic segmentation.*

	Correct number	*Error number*	*Correct rate*
Normal group	*104*	*16*	*87%*
Confusion group	*64*	*56*	*55%*
Abnormal group	*92*	*28*	*77%*
Total	*260*	*100*	*72%*

We have conducted experiments on three different types of groups, as seen in Table 1. The total correct rate of the testing is 72%, which is acceptable, a large part of the input image can be recognized and matched. Both the normal and non-normal groups have relatively high accuracy rates since their results are relatively certain. Then for the confusion group, because the action of the intermediate state is very uncertain, it leads to a correction rate of only around 50%.

The above shows that the result of the normal and abnormal action is relatively certain, which can be considered as reliable evidence for the decision-making. However, for the confusion group, its result will need additional information for decision making. After testing, the vast majority of results are 1 or -1, meaning that people are falling or not. Only a few results are 0, which means the image detection results show that the person did not fall, but the analysis of the Internet of Things data indicates that the person is in a state of falling, this is a contradictory conclusion.

3. CONCLUSIONS

Our proposed new method combines the image semantic segmentation and data from IoT devices, which can effectively classify whether the human body falls. The result of the image semantic segmentation can be the first evidence for the decision-making, but the result of confusion group in it cannot be reliable evidence. Therefore, the analysis data of the IoT as supplementary evidence can strengthen the verification of normal and abnormal results, especially for the confusion group, its results are decisive. From the experimental results, the result is showed that the method has practical effects for fall detection, it can make a realistic

 ISBN: 978-1-5262-0820-0

judgement. Especially for the older people or people who live along, its decision-making results can be used as a processing condition after an accident, and an in-depth analysis of the result may also have a preventive effect on such an accident.

Acknowledgements: *This work was based on my Master dissertation, was supervised by Dr. Sean McGrath, and the encouragement and strong support from Dr. Nanlin Jin.*

4. REFERENCES

A Núñez-Marcos et. al (2017), Vision-Based Fall Detection with Convolutional Neural Networks, wireless communications and mobile computing.

D Curone et. al (2010), Heart Rate and Accelerometer Data Fusion for Activity Assessment of Rescuers During Emergency Interventions, IEEE Trans. pp.702 - 710.

D G. Lowe. (2004), Distinctive Image Features from Scale-Invariant Keypoints, International Journal of Computer Vision, 60(2), 91-110.

F Bagala et. al (2012), Evaluation of Accelerometer-Based Fall Detection Algorithms on Real-World Falls. PLoS One 2012, 7: e37062. 10.1371/journal.pone.0037062.

H Gjoreski, M Lustrek, and M Gams (2011), Accelerometer Placement for Posture Recognition and Fall Detection, 2011 Seventh International Conference on Intelligent Environments, Nottingham, pp. 47-54.

H Li et. al (2017), Multisensor data fusion for human activities classification and fall detection, 2017 IEEE SENSORS, Glasgow, UK, , pp. 1-3, doi: 10.1109/ICSENS.2017.8234179

K He, X Zhang, S Ren, and J Sun. (2015), Deep Residual Learning for Image Recognition.

L Alhimale,H Zedan, and A Al-Bayatti (2014), The implementation of an intelligent and video-based fall detection system using a neural network, Applied Soft Computing, 18, pp. 59-69.

L Liu et. al (2014), An automatic fall detection framework using data fusion of Doppler radar and motion sensor network, IEEE, pp. 5940-5943.

N Dawar, N Kehtarnavaz. (2018), A Convolutional Neural Network-Based Sensor Fusion System for Monitoring Transition Movements in Healthcare Applications, 2018 IEEE 14th International Conference on Control and Automation (ICCA), Anchorage, AK, USA, DOI: 10.1109/ICCA.2018.8444326.

P Pierleoni et. al (2015), A High Reliability Wearable Device for Elderly Fall Detection, IEEE Sensors Journal, 15(8), pp. 4544-4553.

 ISBN: 978-1-5262-0820-0

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

On Tour: Harnessing Social Tourism Data for City and Point of Interest Recommendation

Tom Bewley[1] and Iván Palomares Carrascosa[2]

Department of Engineering Mathematics[1] and Department of Computer Science[2],
University of Bristol, Bristol, UNITED KINGDOM

{ *tom.bewley, i.palomares* } *@bristol.ac.uk*

ABSTRACT

We introduce various data-driven models for recommending both city destinations and within-city points of interest to tourists. The models are implemented with a novel dataset of travel histories, derived from social media data, which is larger by size and scope than in prior work. All proposed models outperform simple baselines in cross-validation experiments, with the strongest variants reliably including tourists' true movements among their top recommendations.

1. INTRODUCTION AND BACKGROUND

Tourism is a popular activity for millions of people worldwide, with city breaks being a common form of holiday. Travellers face complex choices between potential destinations, then between specific points of interest (POIs) to visit upon arrival. In the existing literature, recommender systems have been developed to suggest both high-level destinations [4,6,7] and specific POIs [2,5,10,11] to individual tourists, informed by their explicitly- or implicitly-defined preferences and past visitors' experiences. In some works [1,8], the problem has been expanded to consider tourist groups with diverse preferences, which much be aggregated into a unified preference model or a list of 'compromise' recommendations that balances their needs.

Several recent tourism recommenders [9,10,13] operate by exploiting trends in large social datasets, such as those provided by the *Flickr* photo sharing platform. However, the reported results have been limited to only a small number of target cities, and no attempt has yet been made to perform both city- and POI-level recommendation using the same underlying dataset. In addition, relatively little effort has been dedicated to developing robust techniques for inferring accurate travel histories from the raw social data, and the accuracy of the derived datasets has not been questioned or investigated.

The overall aim of this project is to develop a set of recommendation tools to assist groups of travellers in the planning of city holidays, using both personalised data and population-wide statistics from a novel dataset of travel histories. This dataset is constructed by combining data from *OpenStreetMap* (OSM) with that from *YFCC100M* [12], the world's largest publicly-available media collection, which features accurate time and location tags for tens of millions of photos uploaded to *Flickr*. Two specific problems are addressed: city recommendation for groups of travellers, and within-city POI recommendation for a single traveller. In both cases, the output is a ranking of all available options, from most- to least-recommended.

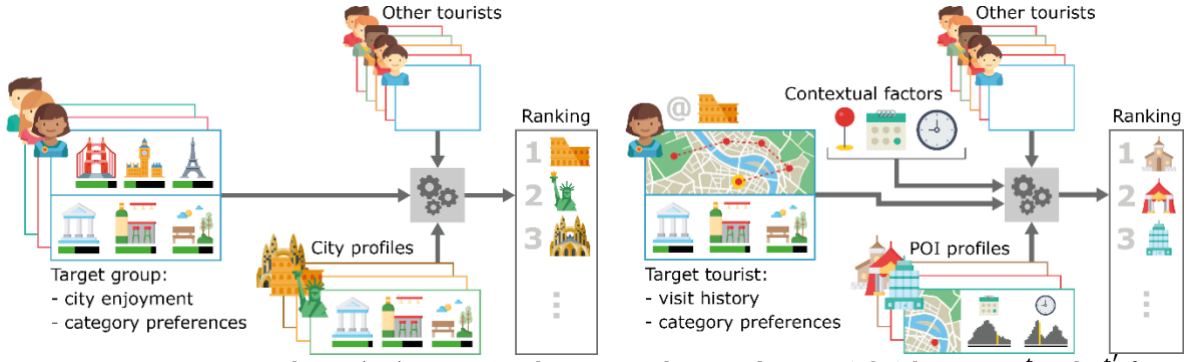
2. RECOMMENDATION PROBLEMS AND MODELS

2.1 City Recommendation

The first problem addressed in this project is that of recommending new cities for groups of tourists to visit. Fig. 1 (left) illustrates the key elements of the city recommendation problem. T is a set of tourists, C is a set of cities and X is a set of POI categories. Each $t \in T$ has visited cities $V^t = \{c_1, \dots, c_l\} \subseteq C$, and a vector of city enjoyment values $e^t = [e^t_c \forall c \in C]$, which is non-zero only for the cities in V^t. t also has a vector of preferences over POI categories (e.g. churches, parks) $p^t = [p^t_x \forall x \in X]$, which sums to 1. Each city c has a distribution $i^c = [i^c_x \forall x \in X]$, which also sums to 1 and can be interpreted as the *importance* of each category in that city (e.g. churches are highly prevalent in Rome; parks comprise a large part of Vienna). A tourist group G is a subset of T which is the target of recommendation. The goal is to produce a ranking R of all unvisited cities $U^G = \bigcap_{t \in G} U^t$ (where $U^t = C - V^t$), ordered from most- to least-recommended for G. Considering first the single-tourist case ($G = \{t\}$) we propose three ranking models:

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute



- *Tourist-tourist similarity* (TT). Compute the Jensen-Shannon distance (JSD) between e^t and $e^{t'}$ for each $t' \neq t$ and collect the 50 most similar tourists into a neighbourhood N^t. For each $c \in U^t$, define the recommendation score $S_{TT}^t(c)$ as the mean of $e_c^{t'} \cdot \left(1 - JSD\left(e^t, e^{t'}\right)\right)$ across all $t' \in N^t$. Produce the ranking R by sorting the cities in U^t by score. This model follows the conventional rationale of collaborative filtering: similarly-travelled tourists should continue to enjoy similar places in future.

Figure 1. *Key elements of the city (left) and POI (right) recommendation problems.*

- *City-city similarity* (CC). For each $c \in U^t$, compute its similarity to each visited city $c' \in V^t$ in terms of their POI category importance distributions i^c and $i^{c'}$, again using JSD. Define the recommendation score for c as the mean of $e_{c'}^t \cdot \left(1 - JSD\left(i^c, i^{c'}\right)\right)$ across all $c' \in V^t$, and again rank by score. The rationale for this model is that tourists tend to visit cities with features in common.
- *Tourist-city similarity* (TC). For each $c \in U^t$, simply define the score for ranking as $1 - JSD(p^t, i^c)$. This model most closely resembles a conventional content-based recommender system: cities are recommended that align closely with the tourist's category preferences.

Multi-tourist groups ($|G| > 1$) add complexity since multiple preferences and visitation histories must be considered. We consider two techniques for producing aggregate recommendations for a group:

- *Aggregation-of-scores* (AoS). Complete scoring and ranking independently for each $t \in G$, then take the mean value of the per-tourist scores for each city.
- *Aggregation-of-preferences* (AoP). Define $V^G = \cup_{t \in G} V^t$ and p^G as the elementwise mean of the category preference vectors of the tourists in the group. Complete scoring and ranking with this group-level information, as if it represented a single tourist.

2.2 POI Recommendation

The second problem addressed is that of recommending POIs to visit during an ongoing trip to a specific city. Only the single-tourist case is considered due to project time constraints and foreseeable difficulties with group evaluation. Fig. 1 (right) illustrates the key elements of this problem. In addition to the notation introduced above, $P^c = \{p_1, \ldots, p_o\}$ is the set of POIs in city c and $H^{t,c} = (v_1, \ldots, v_q)$ is the chronologically-ordered history of *visits* by tourist t in city c, each defined by its start time, end time and associated POI. When finding a POI for t to visit next in c, the goal is to produce a ranking R containing all POIs in P^c, ordered from most- to least-recommended. The POI of the most recent visit $H_q^{t,c}$ is taken to be t's current location. We propose a hybrid model that uses the following six features to quantify the suitability of each POI p in this circumstance:

- $fPop$. The overall popularity of p, defined as $1 - 2^{-visitors/\alpha}$ where $visitors$ is the number of tourists that have been to p. Formally, $visitors = |\{t \in T : p \in H^{t,c}\}|$. In our implementation, $\alpha = 100$.
- $fCat$. *t's* expected preference for p, as represented by p_x^t where x is the category of p.
- $fProx$. A measure of proximity / travelling convenience, defined as $1 - 2^{d/\beta}$ where d is the Euclidian distance in metres to p from t's current location. In our implementation, $\beta = 2000$.
- $fTime$. The appropriateness of p given the present hour of the day. This is computed by finding the proportion of recorded visits that lie within this hour, both to p itself (h_p), and also to p's category x at large across all cities (h_x). The two proportions are weighted by the number of visits to p, $visits$, using the formula $2^{\varepsilon} \cdot h_x + (1 - 2^{\varepsilon}) \cdot h_p$ where $\varepsilon = visits/\gamma$. In our implementation, $\gamma = 100$.

 ISBN: 978-1-5262-0820-0

- $fDate$. The appropriateness of p given the present month, defined in an analogous fashion to $fTime$.
- $fHist$. A measure of the coincidence of p and the previously-visited POIs in $H^{t,c}$, within the visit histories of all other tourists $t' \neq t$. For each $p' \in H^{t,c}$, count the number of times p and p' both occur in another tourist's history, and weight each coincidence according to the time in hours Δ between the two visits, using $w = \zeta + (1 - \zeta)\exp(-\Delta / \kappa)$. The weighted values are summed across all p' and t', and normalised by p's total visit count, $visits$. In our implementation, $\zeta = 0.1$ and $\kappa = 24$.

We consider various techniques for mapping the six features into a single recommendation score for ranking: a trivial summing operation Sum, a linear regression model Lin and two small neural network topologies NN_3 (one 3-neuron hidden layer) and $NN_{6,6}$ (two 6-neuron hidden layers), both with logistic activation functions. In each case, we first z-normalise the features across a large bank of recommendation scenarios, since this improves learning stability.

3. TRAVEL HISTORIES DATASET

We synthesise a novel dataset of travel histories, consisting of approximately 812,000 POI-level visits by 65,000 tourists across 200 cities worldwide, using location-tagged photos from *YFCC100M* [12]. We group the raw photos first by city (this information is available in the *Places* expansion pack) then by user, and sort chronologically. Consecutive photos taken within a 10 metre radius and 1 hour timeframe are combined into a single *visit*. These values are rather conservative, chosen so that bursts of photos taken at exactly the same location are grouped, but visits to two POIs on the same street or town square are not.

For each visit, we assemble a set of *visit words* from the user-provided titles, descriptions and tags, which can be compared with metadata for nearby POIs on OSM. Where a sufficiently strong match is found, the visit is labelled with the POI. Each POI can easily be assigned a category based on OSM's taxonomy of entity types. Consecutive visits to the same POI on the same day are combined into one, thereby preventing the aforementioned conservative visit creation parameters from yielding erroneous duplicate visits.

In a second 'bootstrapping' pass through the data, we use the visit words for already-labelled visits to assemble a distribution of word frequencies for each POI (e.g. *game* might be a high-frequency word for a stadium, and *lion* a high-frequency word for a zoo). The words for unlabelled visits are reassessed with respect to these distributions, and those that pass a fixed total summed probability threshold for a nearby POI are labelled with that POI. The bootstrapping pass boosts the number of labelled visits by approximately 30%.

A manual assessment of the dataset (viewing the underlying photos for one visit per city and looking for the presence of the POI) indicates that labelling accuracy is on the order of 75-85%, depending on how minor errors (e.g. labelling with the wrong building of the correct University) are penalised. This compares favourably with a dataset from prior work [9], which only uses location data for POI labelling, and for which we estimate an accuracy value around 45-60%. The 200-city coverage also far exceeds the eight included in the prior dataset. The statistics of our dataset align well with lists of popular POIs according to *TripAdvisor*, and reflect intuitive trends in the diurnal and seasonal variations in per-category POI visitation (e.g. restaurants are popular at mealtimes, and gardens in spring).

This dataset has the potential to serve as a widely-applicable resource of real-world tourist behaviour, complete with a measure of POI enjoyment through the proxy of number of photos taken (though it does lack an indicator of negative opinion, which may be sought at a later date through sentiment analysis of user-provided text). In light of its wide applicability, and its derivation from public-domain resources, We have published the dataset at https://github.com/tombewley/OnTour-TourismRecommendation/tree/master/dataset.

It exists as a single 65MB JSON file whose core elements are individual visits. Each visit entry contains the POI (as represented by the unique OpenStreetMap identifier), the start and end timestamps and the number of photos taken. Visits are categorised at the highest level by user, then by city, and ordered chronologically. The dataset also includes two separate subschemas containing the total number of photos taken by each user in each city, and the name, category and coordinates of each POI.

4. IMPLEMENTATION, EVALUATION AND OPTIMISATION

We implement our models in Python to work with the travel histories dataset. Preferences are defined in terms of photo counts and metadata: city enjoyments e^t and category preferences p^t are defined by the fraction of t's photos taken in each $c \in C$ or $x \in X$. Similarly, per-city POI category importance values i^c are the fraction of photos taken at each category. Tourist t's history $H^{t,c}$ is defined directly as their visits in c within the dataset.

Following previous work [9,13], our evaluation method is one of cross-validation, which takes the general form of *forgetting* an item from the dataset and assessing the models' ability to predict it using what remains.

 ISBN: 978-1-5262-0820-0

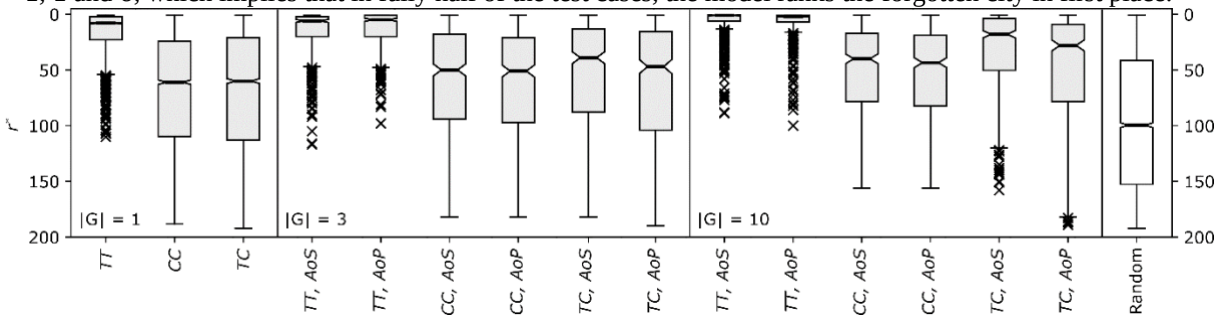
For city recommendation: a group of tourists is assembled who have all visited a common city c^* (as well as at least two others), this city is removed from their histories and the remaining data used as input to the models. For POI recommendation (single-tourist only): a tourist's history in one city is cut short at a point that ensures at least 20 visits remain, and the time gap between the visits immediately before and after (v and v^*) is at most 8 hours. The POI of v is taken as the tourist's current location and that of v^* is the forgotten POI p^*. In both city and POI recommendation, performance is measured via r^*, the position of c^*or p^* in the ranking. For the POI recommendation problem, this is reported as a fraction of the number of POIs in the city.

For the POI recommender, the scoring methods Lin and NN_3 and $NN_{6,6}$ need to be trained on the dataset. To do this, we create a bank of all valid recommendation scenarios (15,326 in total), partition into training, validation and test sets (60: 20: 20 split) and run the training set through the model. In each scenario, we compute r^*, sample a random POI placed higher in the ranking, and define the error as the difference between the two scores. We employ a stochastic gradient descent optimiser, and perform a secondary set of weight updates to non-aggressively push the score magnitudes into the range $[0,1]$ for stability purposes. Training is stopped when validation set performance begins to decrease, which prevents overfitting.

5. RESULTS

5.1 City Recommendation

The box plots in fig. 2 summarise the performance of the three models on 500-case test sets with a group size $|G|$ of 1, 3 and 10. Cases are sampled uniformly across cities to avoid biasing to popular destinations. A clear result is that all three models perform markedly better than random. However, the collaborative filtering model TT is the strongest by a wide margin, with a median r^* of 8 out of 200 with $|G| = 1$. Performance improves as the group size increases. The best results are attained with $|G| = 10$ using AoS; here the r^* quartiles lie at 1, 1 and 6, which implies that in fully half of the test cases, the model ranks the forgotten city in first place.



With $|G| = 1$, the median r^* for the CC and TC models is 61 and 60 respectively; the two are almost inseparable. A gap does open up for larger group sizes, with TC ranking the forgotten city around 20 places higher on average for $|G| = 10$, most notably when the AoS aggregation method is used. With the other two models, the aggregation method has no consistent effect. An investigation of the weaker performance of the CC and TC models reveals few strong results, though it is clear that more visits to the forgotten city afford a more accurate estimation of its POI category importance vector, and in turn a higher ranking.

Figure 2. *City recommendation performance with each model and various group sizes.*

5.2 POI Recommendation

Fig. 3 shows the performance of our model with each scoring method on the 3,055-sample test set, alongside baseline results from using each feature in isolation, as well as random ranking. All methods outperform all baselines, and are remarkably similar despite differences in complexity. This suggests that even a trivial summation of features can reach close to the maximum attainable performance on the dataset, given the set of features. That said, the larger neural network topology ($NN_{6,6}$) is the strongest, placing the forgotten POI at a median position of 6% down the ranking of candidate POIs. The other methods perform around 1-2% worse by rank percentage. On a case-by-case level, $NN_{6,6}$ gives the highest or joint-highest r^* 58% of the time.

In surrounding analyses, we find that visits in popular and Anglophone cities are predicted best, likely due to better data quality in these locations. Static POI categories (e.g. monuments) are also better predicted than those that are weather- or event-sensitive (e.g. theatres). Correlation analysis of the four scoring methods shows a high degree of similarity in their r^* values across the test set; if one method performs well, the others are likely to do the same. Further correlation analysis of each method with respect to each feature shows that the learning models place greater weight on $fProx$, $fPop$ and $fHist$ than the other three. This aligns with the

 ISBN: 978-1-5262-0820-0

greater predictive efficacy of the baseline rankings generated by these features. A final set of feature-to-feature correlations shows that they are largely independent, which is desirable since this maximises informativeness.

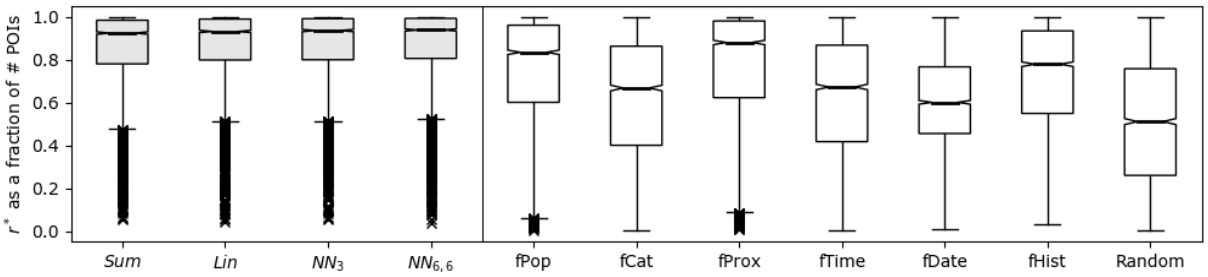


Figure 3. *POI recommendation performance with each scoring method, and single-feature and random baselines.*

6. CONCLUSION

We have presented models for the data-driven recommendation city destinations and POIs and evaluated them by cross-validation on a novel dataset. In all cases, performance exceeds that of simple baselines. For city recommendation, the collaborative filtering model *TT* performs strongest by a wide margin. For POI recommendation, even a simple summation of features is effective, though small gains can be attained by using a neural network for scoring. Future work could involve model refinement (new features and optimised parameters), testing via user trials, and deployment as an interactive web application.

Distinct from our specific choice of recommender models, the travel histories dataset itself provides a large (812,000 entries; 65,000 tourists; 200 cities) and high-accuracy resource of POI-level touristic visits. It is synthesised from two freely-available data sources – YFCC100M and OpenStreetMap – thus can itself can be freely used as part of future work in the field of tourism recommendation.

More details on this project can be found in the full MSc thesis paper [3], which is currently being graded but will be made available at https://tombewley.com/msc.

Acknowledgements: Many thanks to Ercan Ezin, who was co-supervisor for this MSc thesis project.

7. REFERENCES

[1] L Ardissono et al. (2003), Intrigue: personalized recommendation of tourist attractions for desktop and hand held devices, *Applied artificial intelligence 17.8-9: 687-714.*

[2] M Batet et al. (2012), Turist@: Agent-based personalised recommendation of tourist activities, *Expert Systems with Applications 39.8: 7319-7329.*

[3] T Bewley (2019), On Tour: Harnessing Social Tourism Data for City and Point-of-Interest Recommendation, *MSc Thesis, University of Bristol.*

[4] L Cao et al. (2010), A worldwide tourism recommendation system based on geotagged web photos, *IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE.*

[5] L Castillo et al. (2008), Samap: An user-oriented adaptive system for planning tourist visits, *Expert Systems with Applications 34.2: 1318-1332.*

[6] J A Delgado and R Davidson (2002), Knowledge bases and user profiling in travel and hospitality recommender systems, *Proceedings of the Enter 2002 Conference. Springer Verlag, Wien, NY, 1–16.*

[7] M Goossen et al. (2009), My ideal tourism destination: Personalized destination recommendation system combining individual preferences and GIS data, *Information Technology & Tourism 11.1: 17-30.*

[8] A Jameson (2004), More than the sum of its members: challenges for group recommender systems, *Proceedings of the working conference on Advanced visual interfaces. ACM.*

[9] K H Lim et al. (2015), Personalized tour recommendation based on user interests and points of interest visit durations, *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

[10] K H Lim et al. (2017), Personalized itinerary recommendation with queuing time awareness, *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

 ISBN: 978-1-5262-0820-0

[11] R Logesh et al. (2019), Efficient user profiling based intelligent travel recommender system for individual and group of users, *Mobile Networks and Applications 24.3: 1018-1033.*
[12] B Thomee et al. (2015), YFCC100M: The new data in multimedia research, *arXiv:1503.01817*.
[13] L Wan et al. (2018), A hybrid ensemble learning method for tourist route recommendations based on geo-tagged social networks, *International Journal of Geographical Information Science 32.11: 2225-2246*.

 ISBN: 978-1-5262-0820-0

Decision making model based on expert evaluations extracted with sentiment analysis

Cristina Zuheros[1], Eugenio Martínez-Cámara[2], Enrique Herrera-Viedma[3], Francisco Herrera[4]

[1,2,3,4]Andalusian Research Institute in Data Science and Computational Intelligence,
University of Granada,
C/Periodista Daniel Saucedo Aranda S/N 18071 Granada, Spain

[1]czuheros@ugr.es, [2]emcamara@decsai.ugr.es, [3]viedma@decsai.ugr.es, [4]herrera@decsai.ugr.es

ABSTRACT

Decision making requires the evaluation of a set of alternatives by a group of experts. Those alternatives are usually assessed with a set of pre-defined linguistic terms. However, the evaluation and the experts are constrained by those linguistic terms. Hence, we propose that experts express their views in natural language, extract the aspects of interest of the problem and inferring the opinion meaning about them using Aspect-based Sentiment Analysis methods. We will apply our model to a real case of study for evaluating restaurants.

1. INTRODUCTION

Humans face up the problem of decision making in their daily live, hence a large number of models have emerged to solve those decision-making problems. These models tackle the decision situation by selecting a single alternative, a set of alternatives or ranking a set of alternatives from a set of possible ones [4]. Those alternatives are evaluated by experts according to one or more criteria, which may be expressed by numerical values or set of linguistic terms, which are more interpretable for experts.

The evaluation of the alternatives addressed by a set of pre-defined linguistic terms limits the evaluative expressiveness of the experts, and it thus constrains the quality of the decision-making (DM) model. In other words, traditional DM limits the quality of the solutions of DM problems by constraining the set of linguistic terms with which experts can make their assessments. The acquisition of assessments from experts should not be limited by a set of linguistic terms. Accordingly, we propose that experts can express their views in natural language, inferring the meaning of their evaluations and categorising them in the set of criteria of a specific DM problem. We propose the development of a DM model built upon an aspect-based sentiment analysis method (ABSA) for extracting and inferring the opinions of the experts.

The ABSA system will have the ability of extracting the aspects associated to the criteria for each alternative of the DM problem, and will infer the opinion meaning about each aspect. Unlike other existing DM models based on counting opinion keywords [6], our proposal will conduct a semantic understanding process of the evaluation of each expert. Consequently, the experts will provide higher quality evaluations, because they will express them in natural language, and our model will automatically understand the meaning of these evaluations.

The interpretation of the views of experts from natural language will provide more flexibility to DM models, because it will allow to use those models in the context of social networks, especially in e-commerce sites based on user opinions. We plan to evaluate our proposal in the restaurant selection problem, hence we plan to compile a dataset of restaurant reviews, to develop an ABSA model for inferring the user evaluations at aspect-level and to develop a DM model.

This paper is structured as follows. In Section 2, the basis of DM and ABSA are presented. The description of our model is shown in Section 3. A real case of study of a restaurant selection DM problem is described in Section 4. Finally, Section 5 points out some conclusions.

2. BACKGROUND

This section is focused on the basics behind our proposal. Section 2.1 defines the DM task, and Section 2.2 presents the ABSA task.

 ISBN: 978-1-5262-0820-0

2.1 Decision Making

DM is defined through a set of alternatives $X = \{x_1, \dots, x_n\}$, which must be evaluated by a set of experts $E = \{e_1, \dots, e_m\}$. Each alternative is usually assessed following a set of criteria $C = \{c_1, \dots c_l\}$. The workflow of a DM model is described as following [1, 7]:

- *Getting individual evaluations.* Experts provide their evaluations about a set of alternatives according to the stablished criteria.
- *Computing collective evaluation.* All the expert evaluations are aggregated into a collective evaluation, for instance using the Ordered Weighting Averaging operators.
- *Selecting the alternatives*. The alternatives are selected according to the collective evaluation. Depending on the DM problem, a single, a set or a ranking of alternatives may be selected.

Experts sometimes disagree with the solution obtained in the previous step, which requires the calculation of the level of agreement with consensus measures. The solution is satisfactory when the consensus exceeds a threshold. Otherwise, experts should debate and modify their preferences. The end goal is to find out the solution which arouses a higher agreement among the experts, allowing a certain flexibility determined by the threshold.

2.2 Aspect-based Sentiment Analysis

Natural Language Processing (NLP) is the area that integrates knowledge from computer science and linguistics, and its aim is two-fold: (1) the use of computational techniques for understanding human language, and (2) the definition of the right representation of knowledge for the generation of language. The aim of this paper is to understand the position of an expert on the domain of interest, hence we will only focus on the goal of understanding of language goal.

The point of view of an expert is its private state in relation to a topic, in our case a DM problem. Private states are conveyed in opinions, evaluations, emotions and speculations [9], and those utterances are expressed with subjective language [10]. Accordingly, the understanding of the view of an expert requires the computational treatment of subjective language. The task centred on subjective and especially in opinions and evaluations is Sentiment Analysis (SA) [3]. SA encompasses three levels of analysis [2], which the aspect-based sentiment analysis (ABSA) is the one that matches with the requirements of our proposal. ABSA attempts to fully label an opinion, because it aims at extracting all the entities and their related aspects mentioned in a text and classifying the opinion expressed about them. For instance, given a restaurant review "I like the somosas, chai, and the chole, but the dhosas and dhal were kinda dissapointing", an ABSA system extracts the aspect terms "samosas", "chai", "chole", "dhosas" and "dhal", categorises them as the criteria "food" of the restaurant, and labels their opinion meaning, which is positive for the first three aspect terms and negative for the last two ones. Likewise, an ABSA model may define the numerical representation of the opinion meaning, which is essential for the building the input of a DM model.

3. A DECISION MAKING MODEL BASED ON EXPERT KNOWLEDGE ACQUIRED WITH SENTIMENT ANALYSIS

E-commerce sites based on user opinions assist the decision-making process of their users in relation to the commercial service that they provide. Those commercial services may be considered as the alternatives on which to decide, and user opinions as the evaluations of those alternatives. Accordingly, an e-commerce site may be defined as a large number of alternatives evaluated by a large number of experts following in most cases a set of pre-defined criteria. Also, the experts may not evaluate all the criteria of the alternatives or they may not even evaluate all the alternatives. Consequently, we propose to tackle the problem of DM in e-commerce sites as a large-scale DM model, which we detail in the following sections and expose in Figure 1.

 ISBN: 978-1-5262-0820-0

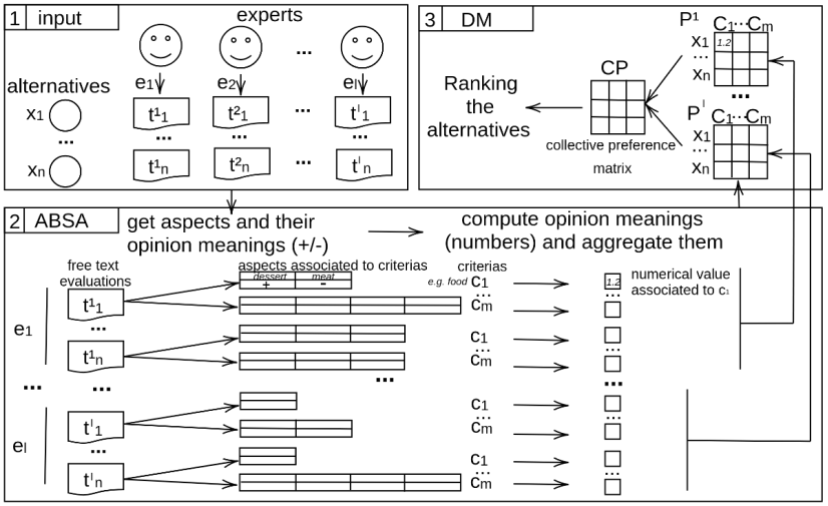


Figure 1: Outline of our proposal, which is composed of three elements: the evaluation of the experts in natural language (1), the classification of the evaluations with an ABSA model (2) and the ranking of the decision alternatives with a DM model (3).

3.1 Defining the DM problem

Given a DM problem, the alternatives and the set of expert must be defined. Since we aim to work with opinions from e-commerce sites, the experts will be the users of those sites.

The set of alternatives $X = \{x_1, \dots, x_n\}$ are the set of entities to be evaluated by the users. Each alternative is related to a pre-defined set of evaluative criteria $C = \{c_1, \dots c_m\}$ linked to a set of weights $\omega = \{\omega_1, \dots \omega_m\}$. For example, users could evaluate restaurants according to their quality of *products*, *price* and *atmosphere*.

The set of experts $E = \{e_1, \dots, e_l\}$ are the set of users that evaluate the alternatives in the e-commerce site. Experts may be previously fixed, so we only seek the comments of this group of experts related to the alternatives. However, if experts are not fixed, we will only look for any comment about the alternatives.

3.2 Getting individual evaluations

After the setting of the alternatives and the experts, the next sept is the extraction of the expert knowledge from the evaluations, which is composed of the following stages:

Extracting the evaluations. Experts evaluate the alternatives using natural language. We search for comments that refer to the alternatives of the problem, which allow us to define for each expert $e_k, k = 1, \dots, l$, a vector T^k such that each element t_i refers to the evaluation to the alternative $x_i,\ i = 1, \dots, n$. Therefore, each element t_i^k is the original comment of the user e_k to the alternative x_i.

In those scenarios where the e-commerce sites also allow experts to provide numerical assessments on certain criteria, our model allows to take into account such assessments. Let consider that experts may make their numerical evaluations in a scale of $2\tau + 1$ levels of opinion intensity. In these cases, for each expert e_k a matrix with numerical evaluations NE^k is defined such that each element ne_{ij}^k refers to the evaluation to the alternative x_i according to the criterion $c_j, j = 1, \dots, m$.

Getting experts' preferences. The first step to interpret the view of the expert is to identify all the references to the pre-defined criteria. That extraction will not build upon of a set of pre-defined set of terms in order to avoid to constraint the model, and it will be conducted by a deep learning method with language understanding capacity in the line of the state of the art in ABSA [5]. Given the comment of an expert about an alternative (t_i^k), the ABSA system will have the ability of (1) identifying the aspect terms related to the criteria of the problem, (2) categorising those aspect-terms in their corresponding criteria, (3) classifying the opinion meaning related to each aspect-term and (4) calculating the opinion meaning value of each criterion according to the opinion meaning of all its aspect-terms. Formally, our model obtains for each expert e_k a matrix P^k such that each element p_{ij} refers to the processed evaluation to the alternative x_i according to the criterion c_j. Each element p_{ij}^k is a numerical value in the interval $[-\tau, \tau]$ which represents the numerical value of the opinion meaning generated by the ABSA model.

 ISBN: 978-1-5262-0820-0

At this point, each expert e_k has associated the matrices NE^k and P^k. Our model linearly combines both matrices from each expert into a single matrix MP^k. The combination is a pairwise weighted sum of the elements of the matrices NE and P, in which the weight ω_{ne} represents the relevance of the numerical evaluation provided by the expert, and the weight ω_p represent the importance of the evaluation in natural language, such that $\omega_{ne} + \omega_p = 1$. Formally, each element mp_{ij} is defined by $mp_{ij} = \omega_{ne} \cdot ne_{ij} + \omega_p \cdot p_{ij}$. If the DM problem does not specify otherwise, it will be considered $\omega_{ne} = \omega_p = 0.5$. When experts do not provide numerical assessments to certain criteria, the mp_{ij}value is just the p_{ij} value.

3.3 Computing collective evaluation

Once each expert e_k has linked to its evaluation matrix MP^k, our model aggregates all these evaluations into a collective evaluation represented by the matrix *CP*. Each element cp_{ij} from CP is defined by $\phi(mp_{ij}^k)$, where ϕ refers to the mean operator that integrates the evaluation of the k experts. Thus, the element cp_{ij} is a numerical value in the interval $[-\tau, \tau]$.

3.4 Selecting the alternatives

For each alternative, our model combines the evaluation obtained for each of its criteria. The collective evaluation ce_i for the alternative x_i is calculated by $ce_i = \omega_1 \cdot cp_{i1} + \cdots + \omega_m \cdot cp_{im}$. This means that each alternative has an associated value in the interval $[-\tau, \tau]$ based on the expert evaluations of all the criteria. The ranking of the alternatives can be obtained just by ordering these values.

4. THE CASE STUDY: DECIDING AMONG RESTAURANTS

We will evaluate our theoretical proposal for DM in the problem of restaurant selection. We will conduct the DM problem taking into account the reviews of some restaurants written by a set of users of an e-commerce site of restaurants. According to our theoretical proposal we need an e-commerce site with a set of restaurants, which they will be the target of the decision, and a set of reviews about those restaurants, which will be the source of knowledge for the DM model. Since the e-commerce site TripAdvisor matches those two requirements, we will use it as our source of alternatives and evaluations for our DM theoretical proposal. Moreover, we will first evaluate our proposal with restaurant reviews written in English, and then we will assess it with reviews written in Spanish.

Concerning the different elements of our proposal described in Section 3, we will detail each of them in the specific scenario of the restaurant selection problem as follows.

Defining the DM problem. The set of alternatives is composed of a set of restaurants from TripAdvisor. Since, we will first evaluate our proposal with reviews written in English and then with reviews written in Spanish, we will first work with restaurants from the United Kingdom and them with restaurants from Spain. As the evaluation criteria, we will consider the set $C = \{Food, Service, Drinks, Ambience, Location\}$, because TripAdvisor considers "*Food*" and "*Service*" as numerical evaluation criteria, and "Drinks", "Ambience" and "Location" are evaluation criteria usually mentioned by TripAdvisor users. Each criterion has an associated weight representing its importance compared to the rest of criteria. For example, if we consider the criteria *Food* has a greater impact than the other criteria, the associated weights could be provided by the vector $\omega = \{0.4, 0.15, 0.15, 0.15, 0.15\}$.

The set of experts will be composed by those users that have been posted a review about the restaurants that composed the set of alternatives of our problem. Users can publish their evaluation as a free text review about the five criteria, and they can also publish a numerical evaluation in a five-scale opinion intensity of the criteria *Food* and *Service*.

Extracting the evaluations. The extraction of the evaluations consists of obtaining the reviews that experts provide for evaluating each alternative. Since each user write a free review evaluating each restaurant in TripAdvisor, we have to extract this information from this e-commerce site. Then, this step focus on extracting the free reviews provided by users from TripAdvisor. To achieve it, we will develop a web crawler that allows to download the reviews.

We will develop the web crawler to obtain both the evaluations in natural language and the numerical evaluations for the criteria *Food* and *Service*, as experts are allowed to provide their reviews through natural language texts and numerical ratings. All reviews expressed in natural language provided by an expert will be collected into a vector as exposed in Section 3.2, i.e., vector T^k collects all restaurants evaluations expressed in natural language for e_k user. Similarly, all numerical evaluations provided by an expert will be collected into a matrix. This matrix, known as NE^k, will have specific values for just the criteria which are numerically

 ISBN: 978-1-5262-0820-0

evaluated by the user. In this case study, this matrix will have numerical evaluations for the criteria *Food* and *Service* and the rest of its values will be defined as missing data.

Getting experts' preferences. Once we get the evaluations of the users expressed in natural language, we have to extract the expert knowledge. For this purpose, an ABSA system will be developed. This model must extract all aspects associated with each criterion and its corresponding polarity for each natural language review t_i^k. Like supervised system, the ABSA model needs training data. Training set must contain a wide set of restaurants reviews which aspects (and its polarities) associated to the criteria are pointed out. We could compile a corpus of restaurant reviews to build this training data, but we will consider existing high-quality corpus such as SemEval-2016 [8]. This corpus provides an extended collection of restaurants reviews with annotated aspects associated to criteria and annotated polarity based on each aspect. SemEval-2016 is available for English and Spanish reviews of restaurants. This annotated training set allows to train the ABSA model to extract aspects associated to criteria and to set up the polarity of each aspect. After training the ABSA model, it would be applied to the test set. Test set is just the collection of restaurants evaluations expressed in natural language, that is, the collection of vectors T^k. We will apply the ABSA model to this test corpus to extract the aspects associated to the criteria as well as the polarity of each aspect for each natural language review concerning restaurants.

After inferring the polarity of aspects associated to each criteria thanks to the ABSA model, we will transform such polarities into a numerical evaluations on a five-scale as the user numerical evaluations of the matrix NE^k. Fusing all aspects related to each criterion by means of a linear combination, we will get the matrix P^k which collects the inferred evaluations of the e_k expert for each alternative and each criterion. This matrix is combined with the matrix NE^k to get the matrix MP^k as exposed in Section 3.2. The matrix MP^k, which represents the numerical evaluations combining the inferred evaluations obtained at the ABSA model and the specific numerical evaluations of the e_k user for each alternative according to each criterion, is the input of the DM model.

Computing collective evaluation. After inferring the knowledge of each expert through the ABSA model, we have to merge the knowledge of all of them. We will consider that all experts have the same importance. Then, the collective matrix CP is computed by means of the mean operator considering all the MP^k matrices as shown in Section 3.2. The matrix CP represents the collective evaluation for each alternative according to each criterion.

Selecting the alternatives. To get the final evaluation for each alternative, the model combines the evaluation obtained for all criteria of each restaurant by a weighted average as shown in Section 3.4. By ordering these values, the ranking of the restaurants is obtained.

In order to analyse a larger set of restaurants, it is necessary to keep in mind that experts may not evaluate all of them. For this challenge, we will study two alternatives:

1. Meta-expert groups. We will create group of similar experts taking into account their views. Each group evaluates the whole set of alternatives. Then, we will treat each group as a unique expert.
2. Experts evaluate a minimum set of alternatives. We will only consider experts that evaluate a minimum number of alternatives. In this case, we will study the development of a decision-making model with the ability of working with unknown information.

5. CONCLUDING REMARKS

This proposal has the aim of using NLP for getting experts evaluations from opinions, as a common-sense scenario for decision making where experts use the language for expressing their opinions.

Acknowledgements: this work was partially supported by the Spanish Ministry of Science and Technology under the project TIN2017-89517-P. Cristina Zuheros was supported by the FPI Programme (PRE2018-083884) from the Spanish Government and Eugenio Martínez Cámara was supported by the Juan de la Cierva Formación Programme (FJCI-2016-28353) from the Spanish Government.

 ISBN: 978-1-5262-0820-0

6. REFERENCES

[1] B Liu, Q Zhou, RX Ding, I Palomares and F Herrera (2019), Large-scale group decision making model based on social network analysis: Trust relationship-based conflict detection and elimination, *European Journal of Operational Research* 275 (2), 737-754.

[2] B Liu (2015), *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge University Press.

[3] B Pang and L. Lee (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1-135.

[4] C Zuheros, C-C Li, F J Cabrerizo, Y Dong, E Herrera-Viedma and F Herrera (2018), Computing with words: Revisiting the qualitative scale, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 26, 127-143.

[5] D Ma, S Li, F Wu, X Xie and H Wang (2019). Exploring Sequence-to-Sequence Learning in Aspect Term Extraction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pp. 3538-3547.

[6] J A Morente-Molinera, G Kou, Y Peng and C Torres-Alberto (2018), Analysing discussions in social networks using group decision making methods and sentiment analysis, *Information Sciences* 447, 157-168.

[7] L Martínez, R M Rodríguez and F Herrera (2015), *The 2-Tuple Linguistic Model: Computing with Words in Decision Making*. Springer Publishing Company, Incorporated.

[8] M Pontiki, D. Galanis, H. Papageorgiou,…,G. Eryigit (2016), SemEval-2016 Task 5: Aspect Based Sentiment Analysis. Association for Computational Linguistics. Proceedings of the 10th International Workshop on Semantic Evaluation, 19-30

[9] R Quirk, S Greenbaum, G Leech and J. Svartvik (1985), *A comprehensive grammar of the English language*. Longman, London.

[10] J Wiebe, T Wilson, R Bruce, M Bell and M Martin (2004), Learning subjective language. *Computational linguistics*, 30(3), 277-308.

ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Realising the Potential for ML from Electronic Health Records

Haoyuan Zhang[1], D. William R. Marsh[2], Norman Fenton[3], Martin Neil[4]

[1, 2, 3, 4]School of Electronic Engineering and Computer Science, Mile End Road, London, United Kingdom

[1]*haoyuan.zhang@ucl.ac.uk*, [2]*d.w.r.marsh@qmul.ac.uk*, [3]*n.fenton@qmul.ac.uk*, [4]*m.neil@qmul.ac.uk*

ABSTRACT

The potential for applying Machine Learning (ML) to Electronic Health Records (EHRs) has been widely agreed but practical progress has been slow. One reason why EHR data are not immediately usable for ML is lack of information about the meaning of the data. An improved description of the data would help to close this gap. However, the description needed is of the data journey from the original data capture, not just of data in the final form needed for ML. We use a simplified example to show how typical EHR data has to be transformed in a series of steps to prepare data for analysis or modelling building. We outline some of the typical transformations and argue that the data transformation needs to be visible to the users of the data. Finally, we suggest that synthetic data could be used to accelerate the interaction between medical practitioners and the ML community.

1. ML AND EHR

An Electronic Health Record (EHR) system contains a collection of digitised patient and population health information that has been collected as part of routine clinical care, including the management and financial functions. The records include various types of data, such as patient demographics, medical history, administrative information, laboratory tests, radiology images, and billing information. There are millions of patient records in EHRs with billions of data points that potentially can help people make better-informed decisions. Machine Learning (ML) techniques can potentially use this vast data to improve medical decision-making and assist with research goals such as disease prediction, biomarker discovery, phenotype identification and quantification of intervention effect (Shickel et al., 2017).

Yet surprisingly, machine learning has, in practice, had little impact in medical decision-making (Rajkomar et al., 2019, McLachlan et al., 2019). Generally, the engagement of the research community with data has been the key to the success of ML development. For example, the UCI repository[1] provides a range of benchmark datasets that is freely accessible to everyone, and competitions such as Kaggle[2], encourage many researchers to tackle various practical problems using data science and ML techniques. These platforms help shape the popularity and the development of ML algorithms and have inspired novel applications in many fields. However, because of the confidentiality of healthcare data most researchers have no direct access to health data and there is less interaction between the health and ML communities.

Efforts have been made to bridge this gap by sharing some of the anonymised health data for research purposes. The MIMIC III database[3] is one example, with more than 60,000 intensive care unit stays spanning from 2001 to 2012 in the US. The database contains data such as demographics, vital signs, and laboratory tests, and has been used for studies using ML techniques. In the UK, an initiative called CLOSER[4] was established in 2012 to provide access to data from several longitudinal studies. The data within the repository are provided with descriptive statistics on each variable and are openly available under licences. The project shows the type of information needed about data for it be widely usable: before requesting data from an EHR an ML researcher would need an understanding of the shape and statistics of the data. However, the CLOSER data was not collected as part of routine care; instead, major resources were committed to these studies, and each of them has their own aims and objectives, which have influenced the designs of the extracted data.

This paper proposes a research direction to realise the potential for ML from EHRs. Section 2 describes the typical steps undertaken to transform raw EHR data for analysis by clinical researchers. Using a simplified example based on a case study we explain why this process hampers the use of ML modelling

[1] https://archive.ics.uci.edu/ml/index.php

[2] https://www.kaggle.com

[3] https://mimic.physionet.org

[4] https://www.closer.ac.uk/

 ISBN: 978-1-5262-0820-0

techniques. In Section 3, we propose a way forward. We argue that it is necessary to improve the visibility of these transformations with descriptions of both data and the process that generates the analysis data from the raw data. We argue that synthetic data could be used to do this, increasing the effectiveness of the interaction between medical practitioners and ML researchers. Section 4 concludes this paper.

2. UNDERSTANDING EHR DATA

In the following, we introduce an example to show how data currently travels and is processed before analysis; we outline how it is possible to improve this situation.

2.1 The Data Journey: Collection, Linkage and Transformation

In England, health providers (e.g. hospitals and clinics) submit health data to a data warehouse called Secondary Uses Service (SUS), linking records from Admitted Patient Care (APC), Outpatient (OP) appointments to Accident and Emergency (A&E). This data warehouse is primarily used by commissioners, such as Clinical Commissioning Group (CCG), to keep track of treatment and care activities of the service providers. At pre-arranged dates during each financial year, data in SUS undergoes cleaning, quality checks and then is further compiled by Commissioning Support Unit (CSU) as Hospital Episode Statistics (HES) to a wider community. In the financial year 2018/19 (April to March), around 168 million hospital episodes from 558 NHS providers and 1426 independent providers were recorded in HES.

Apart from commissioning of services and tariff reimbursement purposes, health data in SUS or HES are often further transformed and used for secondary purposes including research and healthcare planning. One example of transformation is the aggregation of individual diagnostic categories into broader categories. Further, the data from one source become more useful when linked with data from other sources: for example, primary and secondary data can be linked.

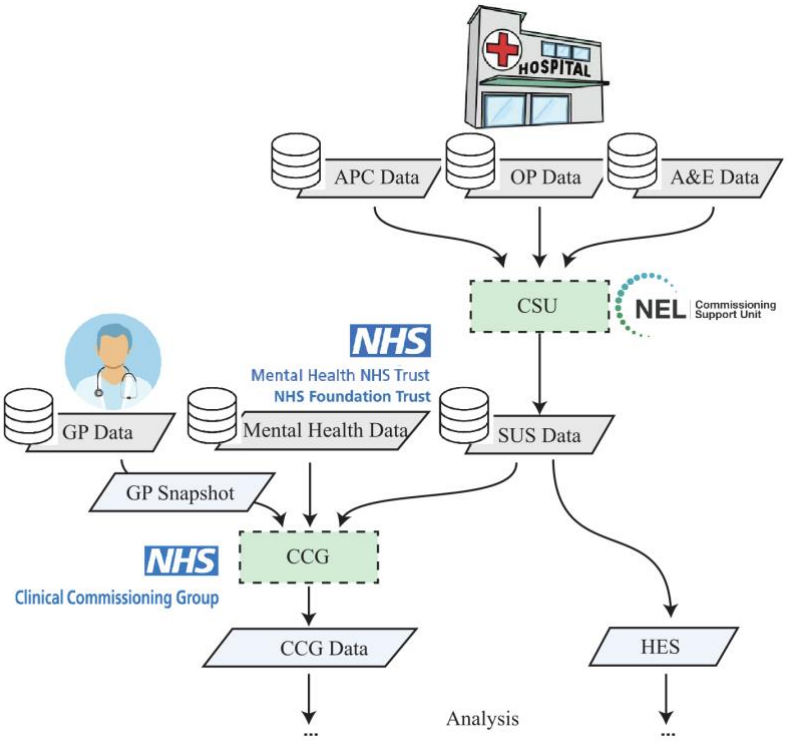


Figure 1. *Health Data Journey to Local CCG.*

2.2 An Example Data Analysis

We experienced the data journey as collaborators on a project with a local CCG. The project aim was to investigate the impact of mental health service availability on patient A&E usage, covering both mental health conditions and 'physical' co-morbidities (such as diabetes and heart disease). The project planned to use a mix of data analytic methods and ML, with the aim of creating tools to help plan expenditure. Figure 1 summarises how the data used in this project was collected, linked and transformed.

 ISBN: 978-1-5262-0820-0

The SUS dataset is collated at CSU and flows to CCG, which links the SUS dataset with GP data and data from other service providers (e.g. mental health service providers) through unique patient identifiers. The linked EHRs consist of a wide variety of data fields and these data fields are structured following the national Commissioning Data Sets (CDS) standard. For example, the CCG data has demographic information such as age, gender, and ethnicity.

A range of additional fields that are derived by the CSUs. Figure 2 gives an example of a common transformation made within a medical organisation. *Read codes*, a clinical terminology system that encodes patient conditions such as clinical signs, symptoms, and diagnoses, are used in the GP data. However, codes are too numerous to be used directly in modelling, so flags are derived from these codes to tag whether a patient has conditions of interest. Each flag is defined by a set of codes. For example, in 2017, *Patient 10001* was assigned with a *1BT...11* Read code and a *Eu34114* Read code from two separate visits. These two visits are merged into one record in the GP snapshot in the financial year 2017/18 record. Three flags are raised for this patient: *low mood*, *depression*, and *anxiety*. The snapshot is further transformed at CCG for research purpose. The flags are aggregated into a variable by counting the number of mental health conditions.

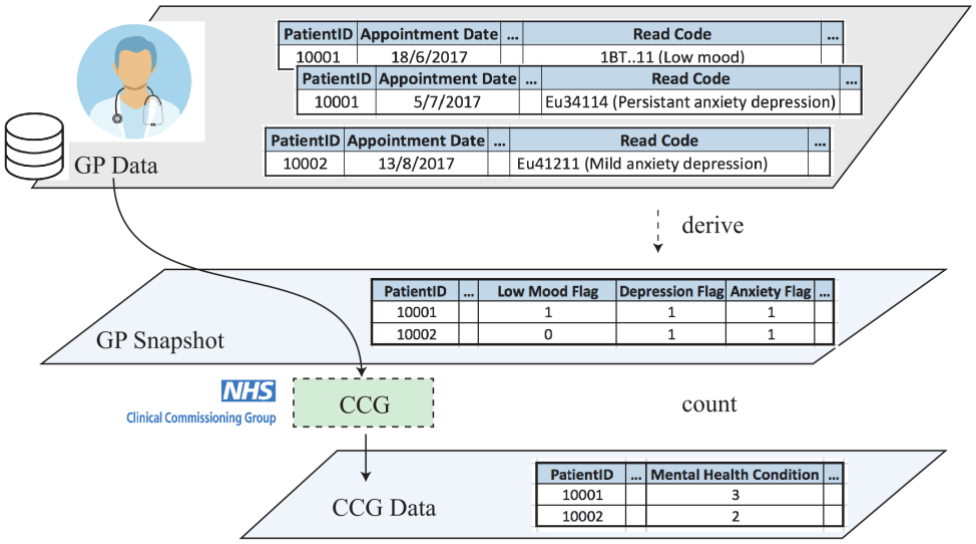


Figure 2. *Data Transformation.*

2.3 Data Challenges

Although data analysis always faces challenges, the complexity of the data journey creates specific challenges. Most importantly, the contents of the data are not well documented, but it is not clear how this should be achieved. The meaning of the data derives both from its collection ('how is this field used in the GP's EHR?') and from the transformations that occur in the data's journey from their origin. The recipients of the processed dataset may lack information about the transformations that have occurred on this journey.

In the mental health study, it was hard to establish at the outset if the data were sufficient to support the intended modelling. The sufficiency of the data depends on the proposed used: for example, training a decision support system requires data on clinical outcomes whereas predicting length of stay in hospital does not. Our challenge was obtaining sufficient detail about the use of A&E resource. The A&E data came from one of the linking steps, with uncertainty about both how it was linked and the original contents.

Several specific challenges were apparent in the mental health study:

1. Data is held by the CCG in a relational database management system, but there is no direct access to this for the analysis team. The overall structure of the original data is a sequence of encounters (GP appointments or hospital stays). It is assumed that analysis will require a flat dataset, but this limits the ways that encounters can be combined. For example, suppose a GP patient had two appointments and was tagged with a low mood indicator in visit 1 and persistent anxiety depression in visit 2. The sequence of these events will be missed in the derived data: they are both considered as events happening in a given year.

 ISBN: 978-1-5262-0820-0

2. Issues occurred in the aggregation of diagnostic codes to create flags. The information about the sets of codes used was not immediately available; when found, not all the decisions made were considered valid. In addition, some information can be lost during the transformation. In our example, codes *Eu34114* and *Eu41211* are both flagged as 'depression' and 'anxiety' in the snapshot, losing the descriptive information 'persistent' and 'mild'.
3. A comparison of drug prescriptions with recorded diagnoses provided evidence of under-reporting of certain mental health conditions in the GP data. Discussions with practicing GPs suggested several mechanisms for this, some relating to possible cultural biases. However, having only aggregated data made investigation of this difficult.
4. Just as the uncertainty created in the original data collection may be disguised in the transformations, these transformations may create additional uncertainty. As pointed out by Goldstein et al., 2012, when linking data from different sources, matches are often treated as perfect, whereas in fact uncertainty may be introduce in the data linkage process.

The context of the challenges is the need for confidentiality and information governance. It is not possible for EHR data simply to be made available in its entirety. Instead, at least the data linkage and much data transformation must take place in a secure environment. It may be possible for anonymised data to be transferred to another (still secure) environment for modelling building, but this requires the precise data needed to be requested. These constraints limit the interaction between the EHR data controllers and the ML community who wish to experiment with novel algorithms and models.

3. MAKING THE DATA JOURNEY VISIBLE

We propose that the interaction of the ML community with EHR data can be improved using a combination of existing techniques applied across the data journey. Although data cannot be made public, documentation of the contents of datasets does not need to be confidential. This documentation needs to be available at all stages of the data journey, covering both the original data, its linkage and transformation. Provided that the form of the documentation is sufficiently rigorous, it could be used to achieve two types of automation. Firstly, some transformations could be automated, allowing them to be tailored to the specific needs of an analysis or modelling project. Secondly, the documentation can be made executable so that synthetic data can be generated, using available knowledge about the uncertainties introduced at each stage of the data journey. The synthetic data can be made available for experimentation and preparation, accelerating the process of model building with real, confidential, data in the secure environment.

3.1 Data Dictionaries and the Data Schema

Techniques for documenting data are well known. Clinical Practice Research Datalink (CPRD)[5] is an example of a project that summarises a list of data dictionaries across various datasets, with each containing data field information such as type, format, and source of data, valid range and field description.

Several tools are available to automatically capture the data dictionary information from the metadata. For example, SchemaSpy is a Java-based tool that analyses the metadata and generates an XML file corresponding to the schema in a database and a graphical representation of it in an HTML site and textual document. SchemaSpy can automatically reverse engineer the Entity-Relationship (ER) diagrams of the database and allows us to click through the hierarchy of tables by both HTML links and ER diagrams. It also identifies a list of potential anomalies in the database that fail to meet constraints between keys.

3.2 Documenting Data Transformations

Different projects may require different transformation procedures. Hence, we need transparent documented procedures. A project like CALIBER[6] is an example that aims to do this by sharing coding lists and programming scripts used to extract both data and clinical coding to researchers. This approach can be extended to other transformations, represented in a library of parameterised transformations. Setting the parameters of the transformation will serve both to show clearly what has been done and to allow transformation in a secure environment to be automated to suit the needs of a particular project.

[5] https://www.cprd.com/home

[6] https://www.ucl.ac.uk/health-informatics/caliber

 ISBN: 978-1-5262-0820-0

3.3 Synthetic Data and Automating Transformation

Even with detailed documentation, requesting data for analysis still requires interaction between data controllers and ML researchers. This interaction can be expensive if the data extraction is repeatedly refined. More likely, on many projects such refinement is not possible because of resource limitations. A further step is to allow the researchers to play with the data while preserving the confidentially using synthetic EHR data.

There has been much research on generating synthetic populations. However, methods either lack validation or can only handle very limited variables (Baowaly et al., 2018). McLachlan et al. (2016) developed a methodology to generate EHRs from health incidence statistics and clinical practice guidelines. Park et al. (2013) proposed generating synthetic data from an algorithm that learn the statistical characteristics of a real EHR, but their methods only work on low dimensional binary data. Choi et al. (2017) developed an approach called medical Generative Adversarial Network which learns from real patient records – the synthetic data are statistically sound but only works with discrete variables such as binary flags and counts.

Probabilistic methods focusing on estimating the joint probability distribution of data can be used to model more detailed population synthesis. Sun and Erath (2015) proposed learning the conditional dependencies between variables through a scoring approach in the form of a Bayesian Network (BN) and sample synthetic data from the joint distribution. This method has been extended into a hierarchical mixture modelling framework in Sun et al. (2018), where the model can generalize the associations of individual variables as well as the relationships between cluster members. Unfortunately, their study is restricted to discrete data. Key EHR variables are continuous (e.g. spending, blood pressure). However, inference algorithms for BNs with both continuous and discrete variables (e.g. dynamic discretisation in Neil et al. (2008)) make it possible to learn the statistical features of EHRs with both continuous and discrete variables. With the learned probabilistic models, we can sample the population statistical distributions to generate realistic synthetic EHRs.

In the mental health case study, we used a probabilistic model to generate synthetic data. This allowed the model building code to be created outside the secure environment, speeding up the development cycle. The probabilistic model generated data that corresponded in form but only approximated what we knew about the distributions of the data. It did not need to be an accurate sample of the real data. However, the synthetic data was limited to the final stage of the data journey: the future goal is to extend this across the data journey.

Executable data documentation would combine the different elements: data dictionary, schema, transformations and probabilistic models. The first goal is to be able to generate data with the correct 'shape' – covering, fields and types, so that the generated data can be used to develop and debug models, before they are applied to real data. For this goal, conditional probability distributions with only a few parents could be learnt from data or estimated from knowledge, since it is not necessary for the synthetic data to be an accurate sample of the real data. However, a second goal is also possible. As shown in our case study, uncertainty is introduced when data capture is imperfect, such as under-recording of diagnoses that carry a social stigma. This cannot be detected in the data but useful information can be elicited from practitioners and documented using a probabilistic model.

4. CONCLUSION

To increase the application of ML modelling to EHR data we must improve the understanding and accessibility of medical data, communication between the medical practitioners who originate data, those who control it and ML researchers is simpler and more efficient. Using an example, we have illustrated how health data travels across organisations and is transformed, emphasizing the importance of transparent documentation and data description. We propose that the documentation of data should be executable so that it is possible to generate and share synthetic data that captures the precise form of the real data and approximates its statistics. Machine learning researchers would be able to exploit such data and hence help achieve the goal of true 'learning health systems'. Our project aims to build a website that allows users to explore data fields and relationships captured from metadata. When users select variables, we would generate synthetic data through sampling from a pre-trained probabilistic model learned from real EHRs.

Acknowledgements: The authors acknowledge funding support from the Alan Turing Institute (R-QMU-005) and EPSRC (EP/P009964/1: PAMBAYESIAN).

 ISBN: 978-1-5262-0820-0

4. REFERENCES

Baowaly, M. K., Lin, C.-C., Liu, C.-L. & Chen, K.-T. 2018. Synthesizing electronic health records using improved generative adversarial networks. *J AM Med Inform ASSN*, 26, 228-241.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F. & Sun, J. 2017. Generating multi-label discrete patient records using generative adversarial networks. *arXiv*:1703.06490.

Goldstein, H., Harron, K. & Wade, A. 2012. The analysis of record-linked data using multiple imputation with data value priors. *Stat Med*, 31, 3481-3493.

Mclachlan, S., Dube, K. & Gallagher, T. Using the Caremap with health incidents statistics for generating the realistic synthetic electronic healthcare record. *Proc. IEEE Intl. Conf. ICHI, 2016*, 439-448.

McLachlan, S., Dube, K., Johnson, O., Buchanan, D., Potts, H. W. W., Gallagher, T., Marsh, D.W., Fenton, N. E. 2019. A framework for analysing learning health systems: are we removing the most impactful barriers?. *Learn Health Syst*, e10189.

Neil, M., Tailor, M., Marquez, D., Fenton, N. & Hearty, P. 2008. Modelling dependable systems using hybrid Bayesian networks. *Reliab Eng Syst Safe*, 93, 933-939.

Park, Y., Ghosh, J. & Shankar, M. Perturbed Gibbs samplers for generating large-scale privacy-safe synthetic health data. *Proc. IEEE Intl. Conf. ICHI, 2013*, 493-498.

Rajkomar, A., Dean, J. & Kohane, I. 2019. Machine learning in medicine. *N Engl J Med*, 380, 1347-1358.

Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *Proc. IEEE Intl. Conf. Biomed Health Inform*, 22, 1589-1604.

Sun, L. & Erath, A. 2015. A Bayesian network approach for population synthesis. *Transp. Res. Part C Emerg*, 61, 49-62.

Sun, L., Erath, A. & Cai, M. 2018. A hierarchical mixture modeling framework for population synthesis. *Transport Res B-Meth*, 114, 199-212.

ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Vectology – exploring biomedical variable relationships using sentence embedding and vectors

Benjamin Elsworth[1], Yi Liu[2], Tom R Gaunt[3]

MRC Integrative Epidemiology Unit, University of Bristol,
Bristol, United Kingdom

[1]ben.elswoth@bristol.ac.uk, [2]yi6240.liu@bristol.ac.uk, [3]Tom.Gaunt@bristol.ac.uk

ABSTRACT

Many biomedical data sets contain variables that are identified by simple, and often short, descriptions. Traditionally these would either be manually annotated and/or assigned to ontologies using expert knowledge, facilitating interactions with other data sets and gaining an understanding of where these variables lie in the biomedical knowledge space. An alternative approach is to utilise sentence embedding methods and convert these variables into vectors, calculated from precomputed models derived from biomedical literature. This provides a data-driven alternative to manual expert annotation, automatically harnessing the expert knowledge captured in the existing literature. These vectors, representing the biomedical space embodied by each specific piece of text, enable us to apply methods for exploring relationships between variables in vector space, notably comparing distances between vectors. From here, it is possible to recommend a set of variables as the most conceptually similar to a given piece of text or existing vector, whilst also gaining insight into how a group of variables are related. Vectology is made available via an API (http://vectology-api.mrcieu.ac.uk/) and basic usage can be explored via a web application (http://vectology.mrcieu.ac.uk).

1. INTRODUCTION

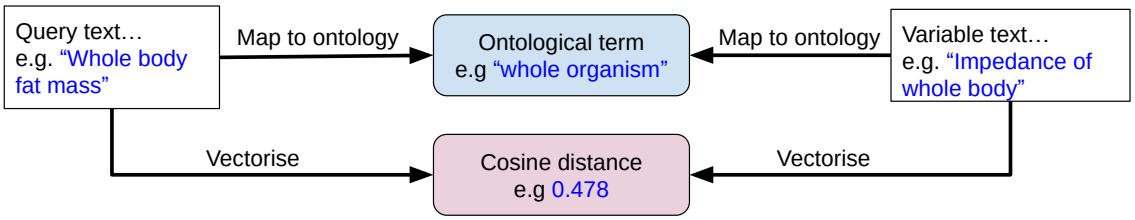
Recent studies (Blagec *et al*., 2019; Duong *et al*., 2019; Karadeniz & Özgür, 2019) have shown the benefits of going beyond manually curated ontology systems, embracing sentence embedding methods to tackle the challenges associated with biomedical text, ontologies and semantic similarity. However, so far, no efforts have been made to apply these ideas to methods focused around exploring the distances between sets of established biomedical entities, e.g. human phenotypes and diseases in clinical and cohort studies.

With the recent advances in Natural Language Processing (NLP) (Bojanowski et al., 2017; Peters et al., 2018; Devlin et al., 2018) where their architectures allow for a multi-stage training through transfer learning, domain specific language models can be trained from domain specific corpora on well-established state-of-the-art pretrained models with relative ease and have showed to sufficiently improve the accuracy in domain specific text embeddings (Chen et al., 2018; Lee et al., 2019; Zhang et al., 2019).

Vectology provides a novel approach in assisting researchers in the biomedical field to better understand the interconnectedness of entities, converting text into vectors and creating distances between them. These distances represent the similarity of these text entities based on their context, and can be used to identify most similar traits. This method avoids the dependence on an intermediate value that is necessary when relating text terms through a common ontology or vocabulary, and the potential errors associated with two such ontology mappings (Figure 1). In contrast to an ontology mapping it also provides a quantitative estimate of the distance between a pair of text terms.

 ISBN: 978-1-5262-0820-0

Figure 1. *Vectors and ontologies*



2. THE VECTOLOGY PLATFORM

Vectology offers the following functionalities:

1. For a list of query texts, derive their text embeddings from biomedical-oriented state-of-the-art language models (word2vec, sent2vec, BERT).
2. Compare any list of text to itself using precomputed models and numerous methods.

For any queried entities, Vectology can suggest the most similar entities from well-established variable data sets. For example, by comparing the text embeddings of the queried texts with pre-computed text embeddings of UK Biobank variables (https://www.ukbiobank.ac.uk). These are indexed and searched using Elasticsearch (https://www.elastic.co/).

Figure 2 demonstrates the implementation of Vectology and its major components. Figure 3 illustrates the web application interface of Vectology and its application programming interface (API) for programmatic use

Figure 2. *Recommending the most similar variables to a given text or vector*

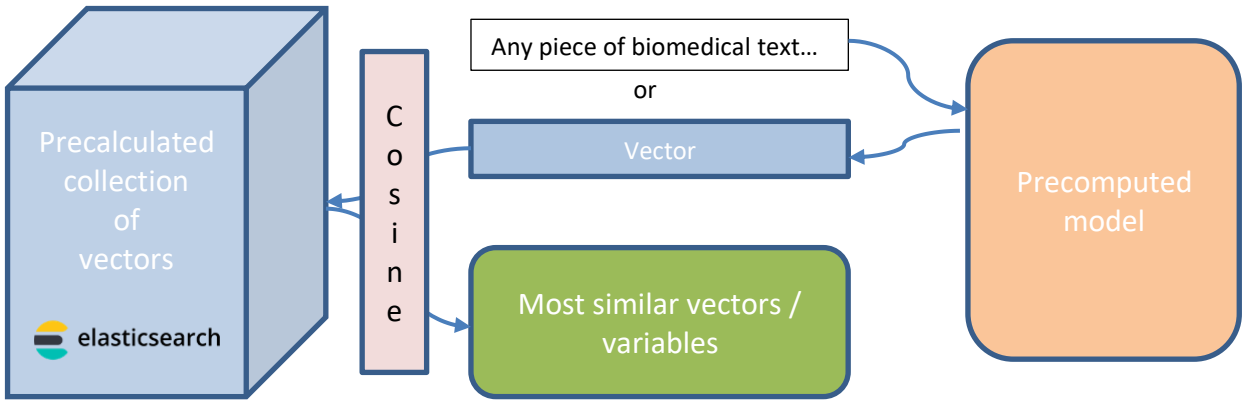
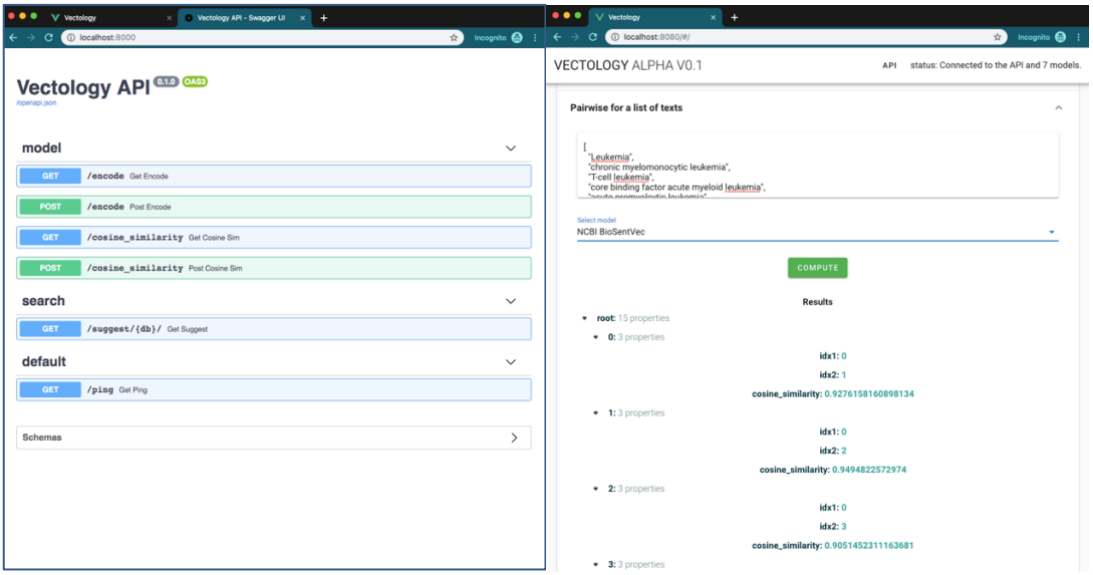


Figure 3. *API and example web application*

 ISBN: 978-1-5262-0820-0



3. USE CASE

Large population studies such as UK Biobank (https://www.ukbiobank.ac.uk/) include tens of thousands of variables measured on tens or hundreds of thousands of individuals. Many such population studies exist around the world. A researcher with a specific question about a disease or risk factor needs to be able to identify the variables relevant to that question. Vectology provides a system to recommend relevant variables and estimate their conceptual distance from the risk factor/disease of interest based on knowledge derived from the medical literature.

Any set of biomedical variables can be used to demonstrate this approach, in this case we have taken 17 traits from the UK Biobank covering a range of human phenotypes, diseases and drugs. Figure 4 is a simple hierarchical clustering of the cosine distances between trait vectors, obtained using the BioSent2Vec model. With no additional training, the vectors cluster accurately into groups representing their biomedical origins. In particular, the two clusters of drugs group with their associated phenotypes, Tamoxifen and Docetaxel being drugs commonly used to treat cancer, and Gemfibozil, Ezetimibe, Atorvastatin and Fluvastatin grouping with traits associated with heart disease and cholesterol. Doing an exercise of this kind manually would involve expert knowledge across a wide domain of biomedical disciplines.

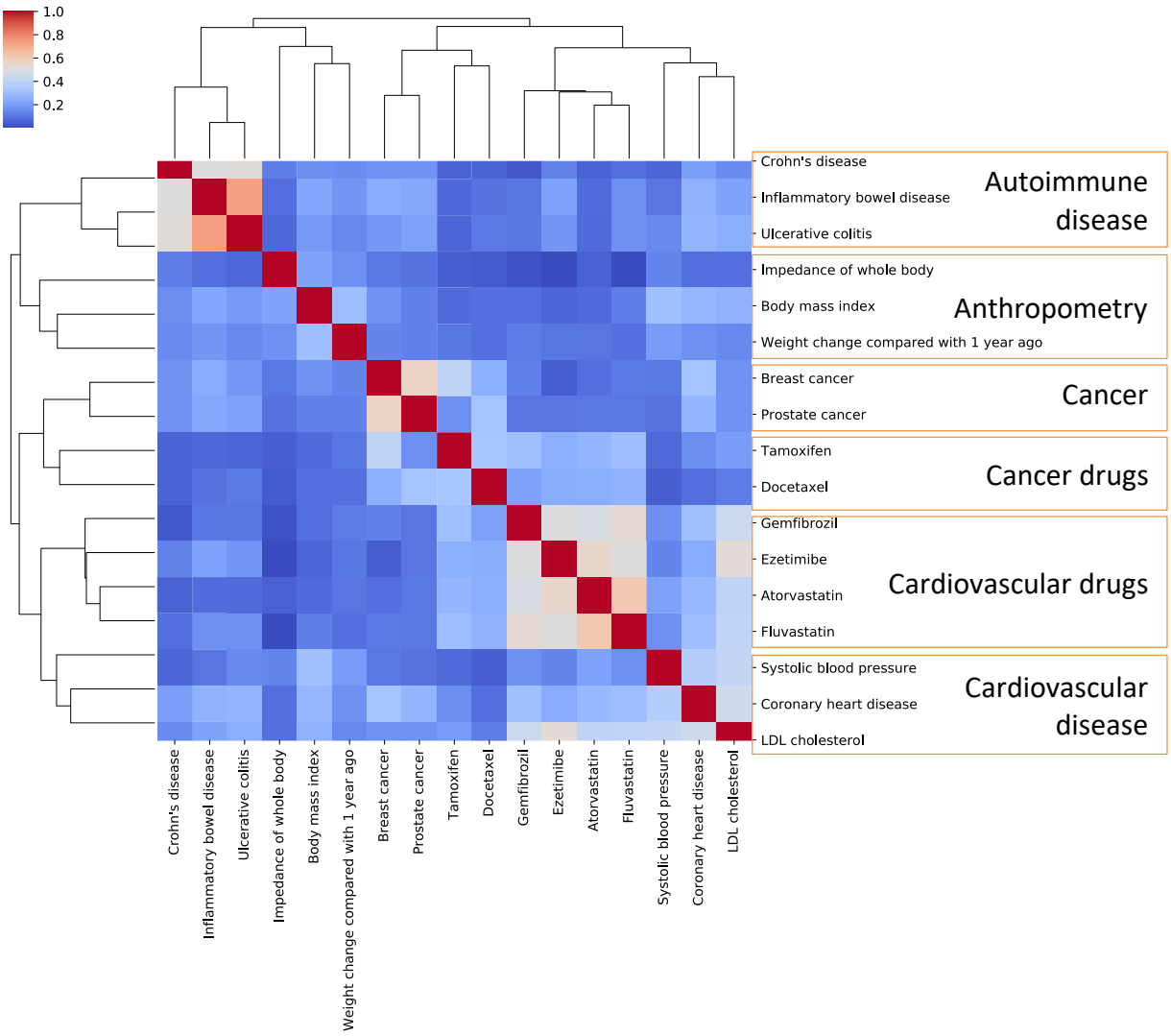
There are a number of practical applications for this type of analysis:

1. **Variable queries:** Being able to search variables using vectors derived from the biomedical literature provides a sophisticated and powerful alternative to standard search methods.
2. **Dimensionality reduction:** By identifying clusters of related variables we can reduce the number of variables to use in a hypothesis free analysis, e.g. by selecting only one representative variable from each cluster.
3. **Evidence synthesis:** Identification of related variables in different publications offers the potential to expand systematic review and meta-analysis to include a broader array of variables related to the same core concepts.
4. **Identifying bias in the literature:** Comparison of variable clusters/distances with clusters/distances derived by other methods (e.g. trait correlation) may identify differences that highlight where published human knowledge disagrees with the data, resulting in new knowledge and a better understanding of biases in scientific reporting.
5. **Systematic analysis:** The ability to map concepts in a research question to appropriate variables in studies with tens of thousands of variables makes these large and complex datasets more amenable to systematic analysis.

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Figure 4. *Hierarchical clustering of example biomedical trait distances*



4. CASE STUDIES

UK Biobank variables

Using over 50,000 variables from UK Biobank, we have created vector indices using 8 NLP methods and models. We have also created some simple pre-processing rules to help remove uninformative text, for example, 'Treatment/medication code: tamoxifen' becomes just 'tamoxifen'. This is something that needs to be carefully considered, as common words across variables will invariably make them more similar in vector space, so efforts made to decrease the number of uninformative words are important.

MRC-IEU GWAS Database

A more practical applications of this now exists in the MRC-IEU GWAS Database (https://gwas.mrcieu.ac.uk/). Here, again we have created indices using the same method, but this time based on the GWAS traits present in this database. For each GWAS, the top ten most similar GWAS traits are

 ISBN: 978-1-5262-0820-0

returned based on cosine similarity to all other traits. For example Table 1 lists the ten GWAS trait vectors with the closest cosine similarity score to the cholesterol lowering drug 'atorvastatin' (https://gwas.mrcieu.ac.uk/datasets/UKB-b:10008/). We can clearly see that similar statins are returned with highest similarity, but next are other cholesterol lowering drugs (ezetimibe, fenofibrate, gemfibrozil) that contain no text in common with the original search term (the prefix 'Treatment/medication code:' is removed prior to embedding and searching). This demonstrates the biomedical knowledge captured in the underlying model. A query will return traits that are related biomedically, not simply by text similarity.

Table 1. *Top 10 most similar GWAS trait names to xxx*

GWAS Trait Name	Similarity Score
Treatment/medication code: atorvastatin	1
Treatment/medication code: simvastatin	0.74
Treatment/medication code: simvastatin	0.74
Treatment/medication code: rosuvastatin	0.69
Treatment/medication code: pravastatin	0.62
Treatment/medication code: fluvastatin	0.62
Treatment/medication code: ezetimibe	0.55
Treatment/medication code: ezetimibe	0.55
Treatment/medication code: fenofibrate	0.50
Treatment/medication code: gemfibrozil	0.48

5. CONCLUSIONS

Here we present a method and platform to explore and compare biomedical text using sentence embedding methods, with the aim of improving our understanding of human derived variable descriptions, as well as enhancing the usability and understanding of complex data sets. In our case study we illustrate how Vectology offers the potential to support analytical decisions in health research through the recommendation and prioritisation of variables.

6. FUNDING

This research was funded by the UK Medical Research Council (MC_UU_00011/4).

7. REFERENCES

Blagec, K., Xu, H., Agibetov, A. and Samwald, M., 2019. Neural sentence embedding models for semantic similarity estimation in the biomedical domain. *BMC bioinformatics*, *20*(1), p.178.

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information., 2017 *Transactions of the Association for Computational Linguistics* **5**, 135–146.

Chen, Q., Peng, Y. and Lu, Z., 2018. BioSentVec: creating sentence embeddings for biomedical texts. *arXiv preprint arXiv:1810.09302*.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

 ISBN: 978-1-5262-0820-0

Duong, D., Ahmad, W.U., Eskin, E., Chang, K.W. and Li, J.J., 2019. Word and sentence embedding tools to measure semantic similarity of Gene Ontology terms by their definitions. *Journal of Computational Biology*, *26*(1), pp.38-52.

Hemani, G., Bowden, J., Haycock, P.C., Zheng, J., Davis, O., Flach, P., Gaunt, T.R. and Smith, G.D., 2017. Automating Mendelian randomization through machine learning to construct a putative causal map of the human phenome. *BioRxiv*, p.173682.

Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R. and Tan, V.Y., 2018. The MR-Base platform supports systematic causal inference across the human phenome. *Elife*, *7*, p.e34408.

Karadeniz, I. and Özgür, A., 2019. Linking entities through an ontology using word embeddings and syntactic re-ranking. *BMC bioinformatics*, *20*(1), p.156.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H. and Kang, J., 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L., 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Zhang, Y., Chen, Q., Yang, Z., Lin, H. and Lu, Z., 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific data*, *6*(1), p.52.

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

From Pictures to Touristic Profiles: A Deep-Learning Based Approach

Mete Sertkan, Julia Neidhardt, Hannes Werthner

Research Unit of E-Commerce, TU Wien,
Favoritenstrasse 9-11/194-04, Vienna, Austria

{mete.sertkan, julia.neidhardt, hannes.werthner}@tuwien.ac.at

https://www.ec.tuwien.ac.at

ABSTRACT

Tourism products are typically very complex and strongly tied to emotional experiences. Thus, for many people it is hard to state their preferences and needs explicitly. To overcome such difficulties, we follow the idiom "A Picture is worth a thousand words". Thus, in this work tourists as well as tourism destinations are profiled within the Seven-Factor Model of travel behavioural patterns by using picture collections. Pre-labelled pictures are used in order to train convolutional neural networks (CNNs) with the goal to determine the Seven-Factor representation of a given picture. We demonstrate that touristic characteristics can be extracted out of the pictures. Furthermore, we show that those characteristics can be aggregated for a collection of pictures such that a Seven-Factor representation of a tourist or tourism destination respectively can be determined.

1. INTRODUCTION

In essence, recommender systems (RS) can be defined as "software tools and techniques providing users with suggestions for items a user may wish to utilize" (Ricci et al., 2015). Items in the tourism domain are usually complex (i.e., they typically combine accommodation, transportation, activities, food, etc.), mostly intangible, and highly related to emotional experiences (Werthner and Klein, 1999; Werthner and Ricci, 2004). Thus, consumers have difficulties to explicitly express their preferences, needs, and interests, especially in the early phase of travel decision making (Zins, 2007).

Previous research showed that a reasonable way to counteract such difficulties is to elicit a user's preferences, needs, and/or personality implicitly by using pictures as a pivotal tool. For example, in (Neidhardt et al., 2014, 2015) preferences and personality of a traveller are determined by a gamified and simple picture selection process, where the user has just to select three to seven pictures out of a predefined static set of 63 pictures. Other approaches are leveraging low-level and/or high-level features of user generated pictures (e.g., Instagram, Flickr, etc.) to determine a user's personality traits or to classify a user into basic tourist classes (Ferwerda et al., 2015, 2018; Figueredo et al. 2018). However, previous research has mainly focused on characterizing the user. In an ideal case, picture-based methods should be applicable universally and thus also on pictures of the recommended item, such that users and items are characterized in the same manner and therefore are easily comparable.

We present a novel concept to characterize users and recommendation items (i.e., tourism destinations) by utilizing any kind of picture collections (e.g., pictures of a user's social media stream, pictures of a tourism destination provided by a destination management organization, etc.). Thus, our work is not limited to a dedicated set of pictures. In this work tourists as well as tourism destinations are described by the Seven-Factor Model of travel behavioural patterns (Neidhardt et al., 2014, 2015). In this way, in addition to the personality we also consider tourism related factors. Since tourists as well as destinations are rather complex entities and thus can have many different characteristics, we avoid assigning them onto one factor. Rather, we model both as a mixture of the Seven-Factors to capture multiple aspects. To determine the Seven-Factors of a particular picture, we trained and evaluated CNNs with pre-labelled pictures.

In this paper we show that, given a picture, the Seven-Factors can be extracted. Furthermore, we conceptually demonstrate that those factors can be aggregated for a collection of pictures to obtain a Seven-Factor representation of a tourist or a tourism destination respectively.

Iván Palomares (Ed.): Proc. 1st International 'Alan Turing' Conference on Decision Support and Recommender Systems (DSRS-Turing'19) 75
The Alan Turing Institute, London, United Kingdom, 21-22nd November 2019
 ISBN: 978-1-5262-0820-0

2. MATERIALS AND METHODS

The main idea of this work is, given any picture collection (either of a user or of an item), to determine a Seven-Factor representation of the respective collection. The Seven-Factors can be summarized as following:

Sun and Chill-Out **(F1)** a neurotic sun lover, who likes warm weather and sun bathing and does not like cold, rainy or crowded places; ***Knowledge and Travel*** **(F2)** an open minded, educational and well-organized mass tourist, who likes travelling in groups and gaining knowledge, rather than being lazy; ***Independence and History (F3)*** an independent mass tourist, who is searching for the meaning of life, is interested in history and tradition, and likes to travel independently, rather than organized tours and travels; ***Culture and Indulgence (F4)*** an extroverted, culture and history loving high-class tourist, who is also a connoisseur of good food and wine; ***Social and Sports (F5)*** an open minded sportive traveller, who loves to socialize with locals and does not like areas of intense tourism; ***Action and Fun (F6)*** a jet setting thrill seeker, who loves action, party, and exclusiveness and avoids quiet and peaceful places; ***Nature and Recreation (F7)*** a nature and silence lover, who wants to escape from everyday life and avoids crowded places and large cities (Neidhardt et al., 2014, 2015).

We propose two steps to determine a Seven-Factor representation of a given picture collection, namely *Classification* and *Aggregation*. The purpose of the *Classification* step is to determine a Seven-Factor representation of each picture of a given collection. Therefore, seven CNNs are trained as binary classifiers, one for each factor, and the outputs (i.e., class probabilities) are combined into a seven-dimensional vector (i.e., Seven-Factor representation). The data set in use consists of 300 pictures for each factor with 150 positive examples and 150 negative examples. For example, a picture of sunbathing people on the beach is a positive example for the factor *F1* and a picture of a rainy day in an urban area is a negative one. Considering the limited amount of data, a pretrained ResNet50 (He et al., 2016) implementation is adapted as binary classifier and only finetuned with the training data (i.e., Transfer Learning) (Karpathy, 2019). Each model is trained with 200 pictures and validated with 100 pictures and additionally the training data is enriched by using data augmentation techniques (i.e., random crop and horizontal flip). Furthermore, stochastic gradient descent (SGD) is used as an optimizer in combination with cross entropy loss as the loss function (Karpathy, 2019).

For each picture in a given collection the *Classification* step returns a seven-dimensional vector f^p of Seven-Factor scores (i.e., Seven-Factor representation). Hence, the main role of the *Aggregation* step is to aggregate the individual Seven-Factor representations f_i^p of a collection with $i = 1 \dots N$ pictures into one representation, which characterizes the whole collection. In this work the *Aggregation* step is modelled naively as a simple mean (see Equation 1), but in future work different approaches (e.g., ordered weighted mean, etc.) will be explored and compared.

$$\frac{1}{N}\sum_{i=1}^{N} f_i^p \tag{1}$$

3. RESULTS

In Figure 1 one can see that the pre-trained ResNet is enabling a proper start performance with validation accuracies between 53% and up to 90% already after one epoch of training.

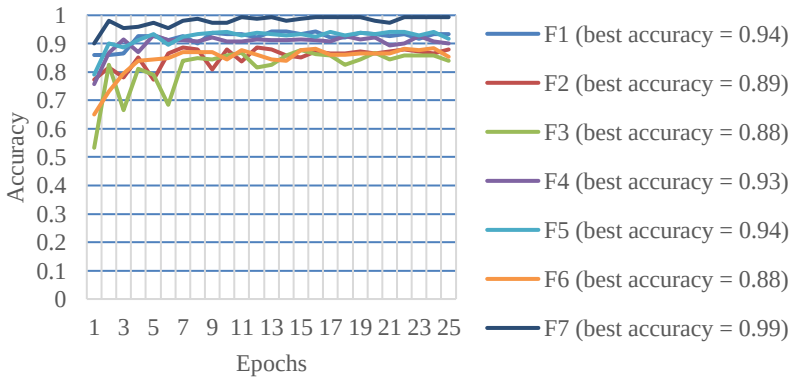


Figure 1. *Validation accuracies of the CNNs.*

 ISBN: 978-1-5262-0820-0

The Alan Turing Institute

Overall, all seven CNNs show a good performance with validation accuracies of greater or equal to 88%, especially the CNN of the factor *F7*, where 99% of the pictures in the validation set are classified correctly.

Figure 2. *The picture collection Action and it's corresponding Seven-Factor representation. In red the proposed approach and in blue the picture-based approach introduced in (Neidhardt et al., 2014, 2015).*

The proposed *Profiler* (i.e., *Classification* plus *Aggregation*) is exemplary demonstrated and compared against a baseline, namely the picture-based approach introduced in (Neidhardt et al., 2014, 2015). In (Neidhardt et al., 2014, 2015) a user has to select three to seven pictures out of a predefined set of 63 pictures, in order to determine the respective user's profile (i.e., Seven-Factor representation). Hence, a profile is created using this approach and the picture collection *Action* (see Figure 2) as input and compared against the output of the proposed CNN approach using the same collection (i.e., *Action*) as input. As expected, the factor *F6* scores the best in both approaches, since the input contains pictures related to partying and thrill-seeking. Also, the factor *F5* shows an increased score in both, since the considered collections contains two different kinds of sports. Furthermore, the CNN approach captures the nature aspect (i.e., mountains) of the pictures with a *F7* score of 0.36, whereas the corresponding score in the baseline is 0.08. On the other hand, mass touristic aspects (i.e., carnival) are not covered compared to the baseline, where the factors *F2* and *F3* are scoring with 0.45 and 0.4 respectively.

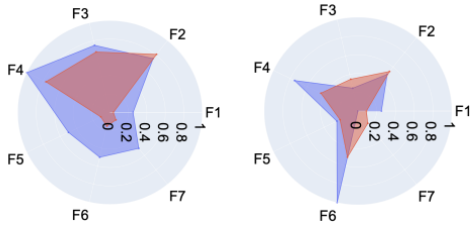


Figure 3. *Seven-Factor representation of Vienna (left) and Las Vegas (right). In red the proposed approach and in blue the experts' opinion collected in (Sertkan et al., 2018, 2019).*

Next, Seven-Factor representations of two tourism destinations, namely *Vienna* and *Las Vegas*, are created using the proposed CNN approach. The resulting profiles are compared against expert knowledge, which was collected in (Sertkan et al., 2018, 2019). Pictures of the Google travel guide of each destination are used as input. The resulting profiles are presented in Figure 3, where the experts' opinion is shown in blue and the results of our approach in red. In case of *Vienna*, one can see that the experts' opinion is more comprehensive, knowing that there is a lot of nature, possibilities for sports, and also places to go out. In case of *Las Vegas*, the experts' opinion and the CNN output have similar shapes, but relevant characteristics are clearly more highlighted by the experts. Factors positively associated with high-class, exclusivity, luxury, thrill-seeking like *F6* and *F4* are scoring relatively higher in the experts rating compared to the CNN results. In both cases, the shortcomings of the introduced approach are rather caused by the respective picture collections than by the CNNs. The used collections might not reflect the actual characteristics of the destinations, and the CNN approach heavily relies on the collection of pictures.

Finally, based on a distance measure a list of top-N recommendation can be given. For example, one can use the Euclidean Distance between the user's and the destination's Seven-Factor representations:

$$\sqrt{\sum_{i=1}^{7}(F_i^{User} - F_i^{Destination})^2} \tag{2}$$

Based on this and the Seven-Factor representations obtained from the proposed approach, the collection *Action* (Figure 2), i.e., the user, is closer to *Las Vegas* than to *Vienna* with distances of 0.87 and 1.43 respectively.

We conceptually demonstrated our approach and we are aware that the evaluation is not yet fully fletched. Thus, currently we are developing a web-application in order to conduct a user study.

 ISBN: 978-1-5262-0820-0

4. CONCLUSIONS

In this paper we introduced a novel way to characterize tourists and tourism destinations out of a collection of pictures. The main aim was, given a picture to determine its Seven-Factors scores and furthermore, to aggregate the individual Seven-Factor scores of a set of pictures, into one Seven-Factor representation, which then can be accounted as the profile of the respective picture collection (i.e., depending on the source of the collection, either user profile or destination profile). To do so, we trained seven convolutional neural network (CNNs), each for one factor of the Seven-Factor Model. Overall, the trained CNNs showed promising results with validation accuracies between 88% up to 99% depending on the factor. Finally, we conceptually demonstrated preference elicitation, destination characterization, and recommendation and compared the proposed approach against previous work (Neidhardt et al., 2014, 2015; Sertkan et al., 2018, 2019).

To conclude, we demonstrated that the concept introduced in this work is a feasible way of automatically characterizing tourists and tourism destinations in a comparable way. However, we are aware of the limitations and challenges of the followed approach. Future work will consider a more thorough and comprehensive evaluation of the resulting user and destination profiles, since in this work we only conceptually demonstrated this idea. Furthermore, the data acquisition will be conducted more systematically by developing a taxonomy of tourism related pictures or tourism products, and by following this taxonomy during training and test data collection. Work in this direction has already begun (Grossmann et al., 2019) and will be continued and improved in the future. Finally, it is planned to combine and compare different sources (i.e., other than Google travel guide) in order to obtain a more comprehensive view of the tourism destinations and thus minimize bias.

5. REFERENCES

Ferwerda B., Schedl M. and Tkalcic, M., (2015), Predicting personality traits with instagram pictures, *Proc. ACM Workshop EMPIRE*, Vienna, pp. 7-10

Ferwerda B. and Tkalcic M., (2018), Predicting Users' Personality from Instagram Pictures: Using Visual and/or Content Features?, *Proc. ACM UMAP*, Singapore, pp. 157-161.

Figueredo M., Ribeiro J., Cacho N., Thome A., Cacho A., Lopes F. and Araujo V. (2018), From Photos to Travel Itinerary: A Tourism Recommender System for Smart Tourism Destination. *Proc. IEEE BigDataService*, Bamberg, pp. 85-92

Grossmann W., Sertkan M, Neidhardt J. and Wethner H. (2019), Pictures as a tool for matching tourist preferences with destinations. In *Personalized Human-Computer Interaction* (M. Augstein & E. Herder & W. Wörndl, Eds), De Gruyter, Berlin

He K., Zhang X., Ren S. and Sun J. (2016), Deep residual learning for image recognition. *Proc IEEE Conf. CVPR*, Las Vegas, pp. 770-778

Karpathy A. (2019), CS231n Convolutional Neural Networks for Visual Recognition. http://cs231n.github.io/, online, accessed 20-June-2019

Neidhardt J., Schuster R., Seyfang L. and Werthner H. (2014), Eliciting the users' unknown preferences, *Proc. ACM RecSys*, New York, pp. 309-312

Neidhardt J., Schuster R., Seyfang L. and Werthner H. (2015), A picture-based approach to recommender systems, *Intl. J. Information Technology & Tourism*, 15, 1, pp. 49-69

Ricci F., Rokach L. and Shapira B. (2015), Recommender systems: Introduction and challenges. In *Recommender Systems Handbook* (Ricci F. & Rokach L. & Shapira B., Eds), Springer, Boston, pp. 1–34

Sertkan M., Neidhardt J. and Werthner H. (2018), Mapping of tourism destinations to travel behavioural patterns. *Proc. Conf. ENTER*, Jönköping, pp. 422-434

Sertkan M., Neidhardt J. and Werthner H. (2019), What is the personality of a tourism destination? *Intl. J. Information Technology & Tourism*, 21, 1, pp. 105-133

Werthner H. and Klein S (1999), *Information technology and tourism: a challenging relationship*, Springer-Verlag, Wien

Werthner H. and Ricci F. (2004), E-commerce and tourism, *Communications of the ACM*, 47, 12, pp.101-105

Zins A.H (2007), Exploring travel information search behavior beyond common frontiers *Intl. J. Information Technology & Tourism*, 9, 3-1, pp. 149-164

ISBN: 978-1-5262-0820-0

The Alan Turing Institute