FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Framework for Multi-Agent Simulation of User Behaviour in E-Commerce Sites

Duarte Duarte

DISSERTATION PLANNING

Mestrado Integrado em Engenharia Informática e Computação

Supervisor: Hugo Sereno Ferreira

Second Supervisor: João Azevedo

February 1, 2016

Framework for Multi-Agent Simulation of User Behaviour in E-Commerce Sites

Duarte Duarte

Mestrado Integrado em Engenharia Informática e Computação

February 1, 2016

Abstract

Customers interact with e-commerce websites in multiple ways and the companies operating them rely on optimizing success metrics for profit. Changing what, how and when content such as product recommendations and ads are displayed can influence customers' actions.

Multiple algorithms and techniques in data mining and machine learning have been applied in this context. Summarizing and analyzing user behaviour can be expensive and tricky since it's hard to extrapolate patterns that never occurred before and the causality aspects of the system are not usually taken into consideration. Commonly used online techniques have the down side of having a high operational cost. However, there has been studies about characterizing user behaviour and interactions in e-commerce websites that could be used to improve this process.

The goal of this dissertation is to create a framework capable of running a multi-agent simulation, by regarding users in an e-commerce website that react to stimuli that influence their actions. Furthermore, some probabilistic models can be used to guide how these agents interact with the system. By taking input from web mining, which includes both static and dynamic content of websites as well as user personas, the simulation should collect success metrics so that the experimentation being run can be evaluated.

"Quidquid latine dictum sit,
altum videtur"

Latin proverb

Contents

List of Figures

LIST OF FIGURES

List of Tables

LIST OF TABLES

Abbreviations

ABS	Agent Based Simulation
API	Application Programming Interface
CBMG	Customer behavior model graph
CPC	Cost per Conversion
CTR	Click through Rate
DAG	Directed acyclic graph
DES	Discrete Event Simulation
HMM	Hidden Markov model
JPD	Joint probability distribution
LTV	Lifetime Value
MOOC	Massive Open Online Course
PGM	Probabilistic Graphical Model
SD	System Dynamics
SWOT	Strengths, Weaknesses, Opportunities and Threats
WCM	Web content mining
WSM	Web structure mining
WUM	Web usage mining
WWW	*World Wide Web*

Chapter 1

Introduction

In this chapter we intend to introduce the report, starting by describing its context, motivations and objectives that will drive the dissertation. It ends with a description of the report structure.

1.1 Context

Customers interact with e-commerce websites in multiple ways and the companies operating them rely on optimizing success metrics such as CTR (Click through Rate), CPC (Cost per Conversion), Basket and Lifetime Value and User Engagement for profit. Changing what, how and when content such as product recommendations and ads are displayed can influence customers' actions.

Multiple algorithms and techniques in data mining and machine learning have been applied in this context.

1.2 Motivation and Goals

Modelling user behaviour on the web is not a new problem. It has been applied with different objectives, from improving the performance of cache servers, to the improvement of search engine, influencing purchase patters or recommending related pages or products [DK01, J S00]. However, all these approaches were done with a machine learning mindset – *predicting* which page the user or customer will browse next. This requires extensive use of existing and historical training datasets which might not expose all the causality aspects of the system. What if the data (or the time needed to gather it) is simply not available?

Let's imagine that we developed a new recommendation engine algorithm. One of the most common ways to evaluate it is by testing the engine with *A/B testing*[1], which is a randomized experiment where a group of users are presented with one version of the engine (control) and the other group is shown the improved version of the engine. By analysing how the two groups

[1] formally, two-sample hypothesis testing

behave differently, it's possible to assess the quality of the two versions, comparatively. However, this approach may not be feasible in all situations. For the experiment to be statistically significant, the number of users shown the two versions of the product must be enough. The experiment also takes time to run and the metrics used to compare both versions might not be easy to choose. [Ama15].

The goal of this dissertation is to create a framework capable of running a multi-agent simulation (chapter 3), by regarding users in an e-commerce website and react to stimuli that influence their actions (chapter 2). Furthermore, some statistical constructs such as Baysian networks, Markov chains or probability distributions (chapter 4) can be used to guide how these agents interact with the system. By taking input from web mining (Web structure mining (WSM), Web usage mining (WUM) and Web content mining (WCM)), which includes both static and dynamic content of websites as well as user personas, the simulation should collect success metrics so that the experimentation being run can be evaluated.

1.3 Report Structure

Besides this introduction, this report has 5 more chapters.

In chapters 2, 3 and 4, we describe the literature review and state of the art in regard to e-commerce, simulation systems and probabilistic models, respectively. The chapter 2 focuses on e-commerce background, what metrics can be used on e-commerce websites and the customer life cycle, an important part in the simulation. The chapter 3 describes three main topics regarding simulating systems: agent based, discrete event simulation and some hybrid approaches. Finally, the chapter 4 deals with describing some probabilistic models, with emphasis on graphical models and Bayesian statistics.

Chapter 5 is concerned with describing the methodology and proposes a work plan for the dissertation.

The final chapter, chapter 6, concludes the work realized and describes a SWOT[2] analysis applied to the project.

[2]Strengths, weaknesses, opportunities and threats

Chapter 2

E-commerce background

In this chapter we discuss some key concepts related to e-commerce, for the purpose of giving context to the dissertation. We discuss the typical customer life cycle in an e-commerce website, some metrics that might be used and some ways on how the customer interaction with the website might be influenced and improved.

2.1 Introduction

E-commerce, or electronic commerce, can be described by the trading of products or services over the Internet (or other computer networks). The type of e-commerce businesses we are interested are those who sell their goods directly to the customer, e.g online shopping, using an online store or catalog of products. Some popular online stores[1] are Amazon[2], Ebay[3] and Alibaba[4].

2.2 Customer life cycle

An important concept to understand the customer is by describing its life cycle, as presented by [SC00, Section 6] in figure 2.1.

It starts by reaching the target audience or market up to an established customer base, not forgetting about those that drop mid way, due to abandonment or attrition.

- Reach happens outside of the website and refers to the number of potential customers. For example, if the online store is advertised on a social network, the reach is the number of users who were served the ad in that other website, they may or may not ignore it.

[1] http://www.alexa.com/topsites/category/Top/Shopping
[2] http://www.amazon.com/
[3] http://www.ebay.com/
[4] http://www.alibaba.com/

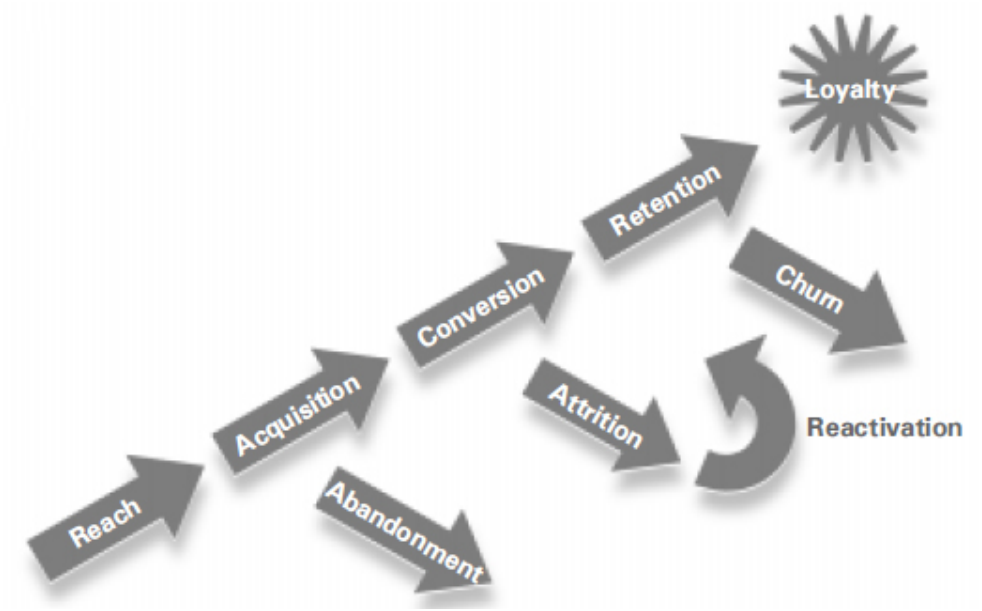


Figure 2.1: Customer lifecycle [SC00]

- Acquisition is the next stage, where the user decides to act on and visits the website (or some other action like subscribing to a newsletter).

- Conversion is the stage where a visitor stops being a user and starts being a customer. It usually means that the user made a purchase but some companies might consider a sign up or registration in the website as a conversion.

- Retention focuses on making existing customers, that made at least one purchase before, repeat purchases.

- Loyalty is a stronger form of retention, which represents a greater trust level of the customer in the store.

- Abandonment is defined by the customers that started the buying process but do not finish it. For example, a customer may add items to the online shopping cart but instead of moving to the next step, e.g. enter credit card details, they exit the website or go elsewhere. This may happen in any store with a multi-step buying process, which is very common.

- Attrition happens when a retained customer ceases buying from the store and starts using a competitor store.

- Churn is defined by the number of customer that attrited during a certain period divided by the total number of customers at the end of that period. It measures how much of the customer base "rolls over" in a certain time period.

2.3 Customer Behavior Model Graph (CBMG)

A state transition graph named Customer Behavior Model Graph (CBMG) can be used to describe the behaviour of customers browsing a website. The nodes represent the possible states or pages, e.g home page, product page, search, and a probability is associated with each transition. An example of such a CBMG is shown in figure 2.2.

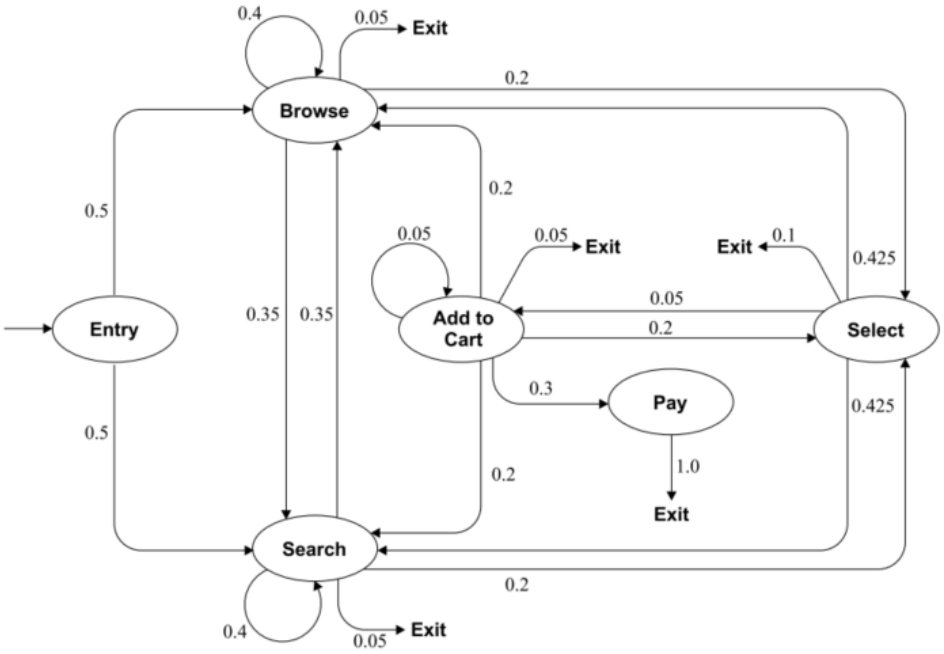


Figure 2.2: Example of a customer behaviour model graph [MAFM99]

[MAFM99] describes how CBMGs can be used to analyse the workload of an e-commerce store server and how metrics can be derived directly from the CBMG alone.

2.4 E-commerce metrics

Metrics are a common way to quantify, measure, benchmark or evaluate some process. In an e-commerce setting, businesses are interested in optimizing, mostly, for profit. Different businesses prioritize metrics in different ways, adapted to each use case. Here we present some common used metrics, but this list is by no means exhaustive. [SC00, MAFM99]

- *Conversion Rate* (CR) is the percentage of visitors that buy a product or a service;
- *Shopping Cart Abandonment* is the percentage of visitors that added a product to the online cart but did not complete the process;
- *Average Order Value* (AOV) is the average size of an order;

- *Customer Lifetime Value* (LTV) is the projected value that a customer will spend on the store;
- *Clicks to Buy* (CTB) is the average number of clicks a visitor has to do to complete a buy order;
- *Churn Rate* the percentage of customers that do not make a repeated purchase;
- *Bounce Rate* is the percentage of visitors that arrive at the homepage of the online store but leave immediately, without clicking anything or visiting a different page.

There are other common metrics such as *Acquisition Cost*, *Cost Per Conversion*, *Net Yield* or *Connection Rate* however they are associated with promotion campaigns that happen outside of the store website, therefore they are not interesting in the context of our work.

2.5 Influencing user behaviour

Tracking a bunch of statistics and metrics about an online store is no good if there are no actionable changes that can be done using that information. [Con04] describes functionality, psychological and content factors that can influence the visitor experience, represented in the table 2.1.

Table 2.1: Main building blocks of Web experience and their sub-categories [Con04]

Functionality factors	
Usability	**Interactivity**
Convenience	Customer service/after sales
Site navigation	Interaction with company personnel
Information architecture	Customization
Ordering/payment process	Network effects
Search facilities and process	
Site speed	
Findability/accessibility	

Psychological factors	**Content factors**	
Trust	**Aesthetics**	**Marketing mix**
Transaction security	Design	Communication
Customer data misuse	Presentation quality	Product
Customer data safety	Design elements	Fulfillment
Uncertainty reducing elements	Style/atmosphere	Price
Guarantees/return policies		Promotion
		Characteristics

Regarding usability of the online store, providing a personalized experience to each customer can be very beneficial for both the customer and the business. A common way to do this is by recommending products that the customer might be interested it [AT05]. For example, if we know that a customer buys mostly football related products, recommending her more products in the same category might increase sales.

2.6 Summary

In this chapter we covered a brief overview of e-commerce, starting with the customer lifecycle, how to measure it using metrics and presenting a common way to model the users' behaviour, the CBMG.

Chapter 3

System Simulation

This chapter intends to introduce some approaches to computational simulation systems and engines, namely agent based and discrete event simulation. To finish the chapter, we show some novel approaches to simulation.

3.1 Introduction

Simulations are used to reproduce the behaviour of a system. They have been applied to different areas like physics, weather, biology, economics and many others. There are many types of simulations: stochastic or deterministic, steady-state or dynamic, continuous or discrete and local or distributed [Wik15]. These categories are not exhaustive nor exclusive.

In this literature review, we are particularly interested in studying simulations which can model stochastic processes and not dynamic (dynamic systems are usually described by differential equations and are continuous).

3.2 Agent Based Simulation (ABS)

In agent based simulation (ABS), sometimes described as agent based computing [Woo98, Jen99], the individual entities in the model are represented discretely and maintain a set of behaviours, beliefs or rules that determine how their state is updated. [Nia11] lists three different approaches to agent based modelling and simulation:

- *Agent-oriented programming* which puts emphasis on developing complex individual agents rather than a large set of agents;
- *Multi-agent oriented programming* focus on adding *some* intelligence to agents and observe their interactions;

- *Agent-based or massively multi-agent modelling* where the main idea is to build simple models for the agents which interact with a large population of other agents to observe the global behaviour.

[SA10] describes ABS as "well suited to modelling systems with heterogeneous, autonomous and pro-active actors, such as human-centred systems.", which make them a good candidate to be used in the development of this dissertation. However, existing literature is quite confusing and broad, using different terms to refer to the same concepts, without clear distinctions between different agent based approaches and without consensus [Nia11, Bra14].

Many platforms and frameworks were developed to support agent-based modelling and similation. Some notable examples include Repast [Col03], NetLogo [WE99], StarLogo [Res96] or MASON [PL05]. An updated list is maintained at OpenABM [Ope16].

Agents have been applied to e-commerce context mostly in two distinct areas: recommendation systems [XB07, WBS08] and negotiation [RKP02, MGM99]. No relevant literature was found regarding simulating user behaviour in websites with agents.

3.3 Discrete Event Simulation (DES)

A discrete event simulation (DES) models a process as a series of discrete events, where the state of the system changes only at well defined points in time [SA10]. It was originally proposed by Kiviat in 1969 [Kiv69] and there is extensive research in this simulation technique. Banks et al. [BCNN04] provides a comprehensive description and analysis of DES. The algorithm 1 is a possible implementation of a very simple and single-threaded DES.

Algorithm 1 Basic DES algorithm

```
EndCondition ← false
Clock ← 0
EventList ← initialEvent
while EndCondition = false do
    CurrentEvent ← POP(EventList)
    Clock ← TIME(CurrentEvent)
    EXECUTE(CurrentEvent)                ▷ might put new events in EventList
    UPDATESTATISTICS()
end while
GENERATEREPORT()
```

The major concepts in DES are as follows [BCNN04]:

- Entity, objects explicitly represented in the model (e.g, a customer);
- Event, an occurrence that changes the state of the system (e.g, a customer enters the website);

- Event list (or future event list or pending event set), a list of future events, ordered by time of occurrence;
- Clock, used to keep track of the current simulation time.

Event list is one of the fundamental parts of the system and it has been widely researched [HOP+86, Jon86, TT00, DGW13].

Pidd [Pid98] proposes a three-phased approach that consists of: jump to the next chronological event, executing all the unconditional events (or type B) that happen that moment and then executing all the conditional events (or type C). This approach has advantages in terms of less usage of resources compared to other simplistic approaches. Also, there has been studies on how to scale DES to distributed and parallel (PDES) executions [Mis86, Fuj90].

[SMG+10] states that "DES is useful for problems (...) in which the processes can be well defined and their emphasis is on representing uncertainty through stochastic distributions", which makes DES a good candidate to model the problem at hand.

3.4 Hybrid and novel approaches

In recent years, there has been research which proposes a marriage between agent based model and simulation with discrete event simulation, however, this concept is not widely recognized [Bra14]. Brailsford states that the line that divides agent based models (and simulation) and DES is spurious and that common distinctions between the two approaches are artificial. Casas et al. [FRJ11] describe a method where multi agent system components have been added to an existing discrete event simulation implemented in OMNeT++[1] [Var01]. Onggo [Ong07] shows how agent based models can be ran on top of a DES engine. Kurve et al. [KKK13] describes an agent based performance model of a PDES kernel. Regarding existing software, AnyLogic claims to be "the only simulation tool that supports Discrete Event, Agent Based, and System Dynamics Simulation" [Any00]. AnyLogic was first shown in 2000 at the Winter Simulation Conference.

3.5 Summary

This literature reviews shows that there is vast research regarding simulation, either agent based or DES, however not everyone is speaking the same language. The extensions to DES seen above are particularly interesting since they can be used to scale the simulation to a greater number of entities as well as modelling real world processes with more fidelity.

[1] C++ based discrete event simulation toolkit

Chapter 4

Probabilistic Models

4.1 Introduction

Probabilistic or statistical models represent explicit assumptions about a problem domain, in the form of a model. This model usually encompasses random variables[1], in the form of probability distributions, and the relation and dependence between the variables. [WB13]

In the following sections we describe a common way to represent probabilistic models, probabilistic graphical models (PGM) or, simply, graphical models.

4.2 Probabilistic Graphical Models

A PGM is a graph based model where the nodes represent random variables and the (directed or undirected) edges represent a conditional dependence between variables. An example is shown in figure 4.1.

PGMs and their extensions, where we show some examples of them in the following sections, are exceptionally well suited for reasoning and to reach conclusions based on available information (both domain expert and data), even in the presence of uncertainty. PGMs provide a general framework that allows representation, inference and learning on these models. [KF09]

There is extensive research and available literature in this area. Some notable examples include, but are not limited to, the books *"Probabilistic Graphical Models: Principles and Techniques"* by Daphne Koller and Nir Friedman [KF09] and *"Pattern Recognition and Machine Learning"* (Chapter 8: Graphical Models) by Christopher Bishop [Bis06]. It is also worth mentioning that there is a MOOC [2] named *"Probabilistic Graphical Models"*, also by Daphne Koller (Stanford), freely available on Coursera [3].

[1] Variable whose value is given by a probability distribution, commonly represented by Θ.
[2] Massive Open Online Course
[3] https://www.coursera.org/course/pgm

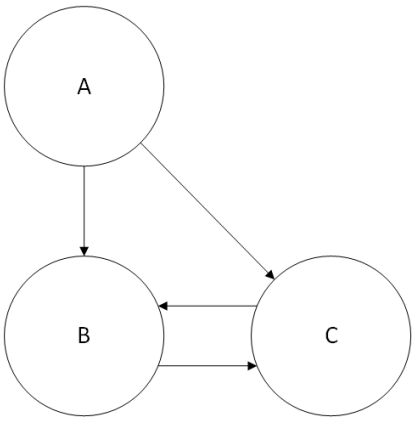


Figure 4.1: Example of a PGM: B and C depend on A, B depends on C and C depends on B.

In the following sections, we describe three important categories of graphical models: Bayesian networks, Markov random fields and its extension to hidden Markov models. There are plenty of other graphical models however they were deemed not relevant enough to be included in this literature review.

4.3 Bayesian Networks

Bayesian networks, also named directed graphical models, is a type of PGM where the edges in the graph representation are directed and represent causal relationships between random variables or group of random variables (see figure 4.1). This concept was first introduced by Pearl in 1985 [Pea85], which uses Bayes' conditioning [Bay63] as the basis for updating information.

Bayesian networks follow the Bayesian approach to statistics and probabilities. In contrast to classical or physical probability, Bayesian probability (of an event) is a person's *degree of belief* in that event [Hec96]. While it may seen that a degree of belief is somewhat arbitrary or may lack precision and accuracy, multiple authors [Ram31, TK74, Sha88] argue that small variations in probability do not have a big influence in the decision making process and that measuring beliefs lead to the same rules of probability (which can be summarized with the product rule 4.1 and the sum rule 4.2 [Mac05]).

$$P(x,y \mid \mathscr{H}) = P(y \mid x,\mathscr{H})P(x \mid \mathscr{H})^{4} \tag{4.1}$$

$$P(x,\mathscr{H}) = \sum_{y} P(x \mid y,\mathscr{H})P(y \mid \mathscr{H}) \tag{4.2}$$

Formally [Pea88], a Bayesian network B represents a joint probability distribution (JPD) over a set of variables $\mathbf{U}$ and can be defined by a pair $B = \langle G, \Theta \rangle$. B is a DAG (directed acyclic graph) where the vertices represent the random variables $X_1, ..., X_n$. Θ represents the set of parameters

[4] $\mathscr{H}$: hypotesis or assumptions the probabilities are based

that quantify the network. For each possible value x_i of X_i, and $\prod_{x_i}$ of $\prod_{X_i}$ (set of parents of X_i in G), it contains a parameter $\theta_{x_i|\prod_{x_i}} = P_B(x_i \mid \prod_{x_i})$. Therefore, the JPD can be defined as

$$P_B(X_1,...,X_n) = \prod_{i=1}^{n} P_B(X_i \mid \prod_{X_i}) = \prod_{i=1}^{n} \theta_{X_i|\prod_{X_i}} \tag{4.3}$$

which expresses the factorization properties of the JPD. [Bis06, section 8.1.] goes in detail on how to apply the eq 4.3.

These properties of Bayesian networks make it an excellent tool for expressing causal relationships. Heckerman [Hec96] lists multiple advantages of Bayesian networks on modelling and data analysis: "readily handles situations where some data entries are missing", "gain understanding about a problem domain and to predict the consequences of intervention", "ideal representation for combining prior knowledge and data" and "efficient and principled approach for avoiding the overfitting of data".

Regarding the area of e-commerce specifically, some research has been done where Bayesian networks are applied. [NMK14] is an attempt at predicting sales in e-commerce using social media data. [MCGM02] also proposes a Bayesian based model to predict online purchasing behaviour using navigational clickstream data.

4.4 Markov Random Fields

Markov random fields (MRF) or Markov networks are undirected graphical models [Kin80] (in contrast to Bayesian networks which are directed and acyclic). The nodes still represent variables or group of variables however the links do not carry arrows. The concept was originally proposed as the general setting for the Ising model[5] [Kin80]. Again, Bishop [Bis06] provides a very good overview of this topic.

MRFs factorize as

$$p(x_1,...,x_n) = \frac{1}{Z} \prod_{C \subset \mathfrak{C}} \psi_C(x_C) \tag{4.4}$$

where C is a clique[6] of the graph and x_C is the set of variables in that clique, Z is a constant used to normalize the distribution (might be defined for each x), ψ_C is a compatibility or potential function [WJ08, section 2.1.2] [Bis06, section 8.3]. The equation 4.4 highlights an important property of MRFs: the Markov property or memoryless property. That is, the conditional probability distribution of future states depends only on the present state.

Markov models were shown to be well suited for modelling and predicting e-commerce purchasing and user's browsing behaviour [DK01]. The same article states "there will be cases in which a user's Web site browsing process is not Markovian and, in these cases, such an assumption will lead to inaccurate modeling" however this claim is groundless.

[5] Ising model: mathematical model of ferromagnetism in statistical mechanics

[6] clique: fully connected subset of vertices

4.5 Hidden Markov Models

Hidden Markov models (HMMs) are a PGM with unobserved or hidden states. They are considered a dynamic Bayesian network[7]. They have been originally defined in the 60s by Baum and colleagues [BP66]. [Rab89] defines HMMs as "the resulting model (...) is a doubly embeded stochastic process that is not observable, but can only be observed though another set of stochastic processes that produce the sequence of observations.".

A common example found in literature is the Coin Toss Model [Rab89]: imagine someone on one side of a curtain performing a coin (or multiple coin) tossing experiment. The other person will not tell us about what she is doing, only the outcome of each coin flip (heads or tails). Multiple HMMs can be built to explain the coin toss outcomes, i.e, assuming that one, two or more biased coins are being used in the experiment. The figure 4.2 is a possible model that can account to 3 coins being tossed.

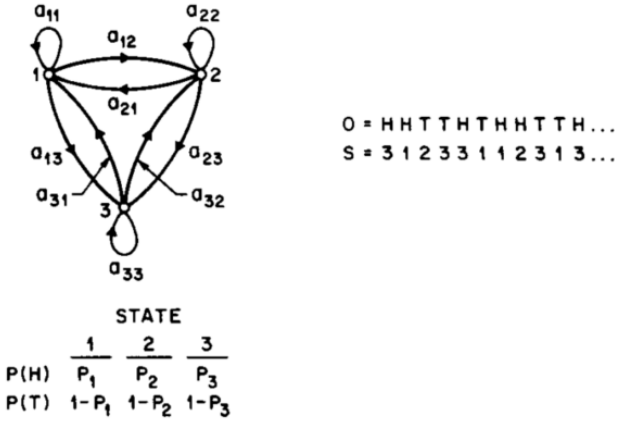


Figure 4.2: Example of a 3-coin model [Rab89]

A HMM is characterized by the following:

- N which is the number of states in the model where individual states are represented by $S = \{S_1, ..., S_N\}$ and the state at time t is q_t;
- M which is the number of distinct observation symbols per state (individual symbols are represented by $V = \{V_1, ..., V_M\}$);
- $A = \{a_{i,j}\}$, the state transition probability distribution where

$$a_{ij} = p(q_{t+1} = S_j \mid q_t = S_i), 1 \leq i, j \leq N \tag{4.5}$$

- $B = \{b_j(k)\}$, the observation symbol probability distribution in state j:

$$b_j(k) = p(v_k \; at \mid q_t = S_j), 1 \leq j \leq N, 1 \leq k \leq M \tag{4.6}$$

[7] dynamic Bayesian network: Bayesian networks adapted with time steps

- Finally, $\pi = \{\pi_i\}$, the initial state distribution:

$$\pi_i = p(q_1 = S_i), 1 \leq i \leq N \tag{4.7}$$

The formal model can be summarized as $\lambda = (A, B, \pi)$ [Rab89].

Multiple algorithms have been studied and applied to HMMs: for inference, the forward algorithm, forward-backward algorithm [BE+67] or the Viterbi algorithm [FJ05, MJ00]. Regarding learning, the algorithm Baum-Welch [BP66, BE+67] can be used.

Regarding e-commerce and web user behaviour there is some research done. [XY09] explains how to use a hidden semi-Markov model to detect anomalies on user browsing behaviour. [ADW02] describes very briefly a relational hidden Markov model for the behaviour of web site users, in order to improve predictions and personalization of websites.

4.6 Summary

In this section we reviewed the literature for graphical models. They provide a tool of excellence to model real world phenomena, enabling decision making under uncertainty and noisy observations.

There are multiple categories of graphical models however we focused on Bayesian and Markov networks and hidden Markov models, due to their applicability in the work at hand (in chapter 5 we will define how PGMs can be applied).

Chapter 5

Methodology

In this chapter we intend to describe the methodology to be followed on this dissertation and propose a work plan.

5.1 Methodology

Like any software development project, a simulation project also has a life cycle. In this section we describe the steps to apply in the simulation methodology, based on Ulgen et al. [UBJK94] and Banks et al. [BCNN04, section 1.11], which can be summarized as follows:

1. *Problem formulation*: Clear statement of the problem by the analyst and stakeholders;
2. *Setting of objectives and overall project plan*: Questions to be answered by the simulation, plans for the study, cost and number of days for each phase, with the results expected at each stage;
3. *Model conceptualization*: Select, modify and iterate over the assumptions that characterize the system;
4. *Data collection*: Collect the necessary data to run and validate the model, assuming that required data will change with the increasing complexity of the system;
5. *Model translation*: Materialization of the system in a program;
6. *Verification*: Making sure that the program behaves correctly accordingly to its inputs;
7. *Validation*: Calibration of the model, comparing the model against an actual system;
8. *Experimental design*: Tweak the experiments, comparing alternative designs;
9. *Production runs and analysis*: Estimate measures of performance for the systems that are being simulated;

10. *Documentation and reporting*: Document both the program and the progress of the study;

11. *Implementation*: End result of the study, including the entire simulation process.

This process can be visualized in figure 5.1.

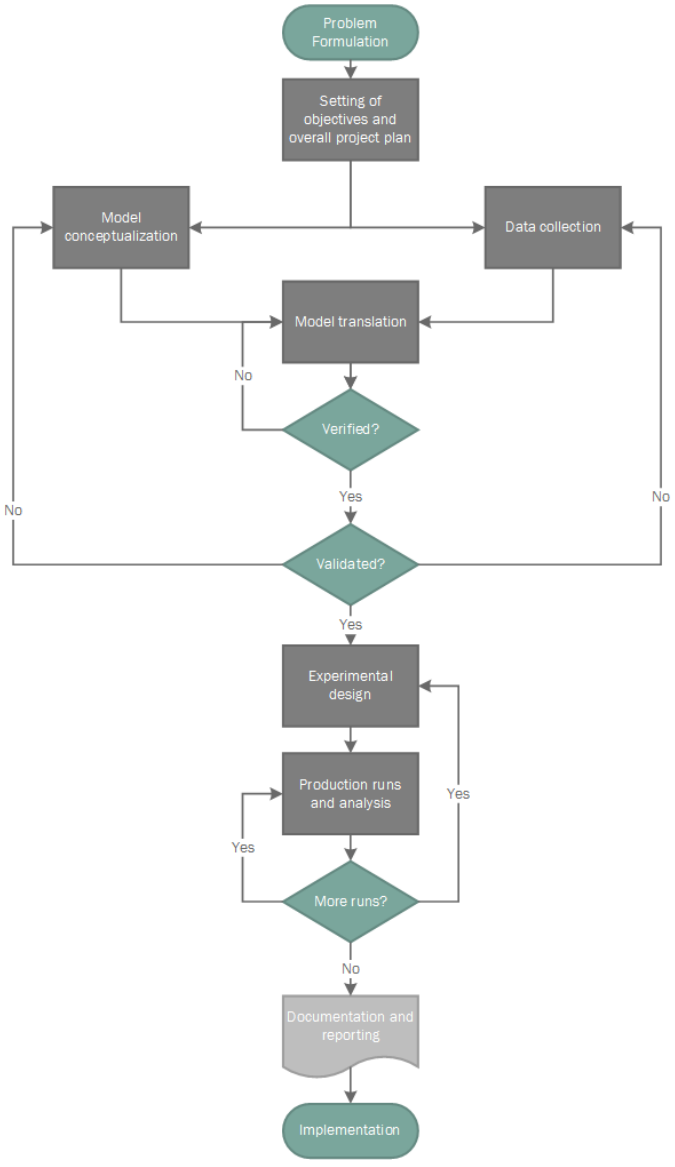


Figure 5.1: Steps in a simulation study [BCNN04]

5.2 Planning

Taking into account the steps described in section 5.1, a Gantt diagram was developed to define the tasks and their duration. This diagram is shown in appendix A.

We now present the details of the most important planned tasks. The name scheme follows [BCNN04].

5.2.1 Discovery

- **Description**: Consists of the steps *Problem formulation* (1) and *Setting of objectives and overall project plan* (2) presented above. In this stage the initial statement of the problem and objectives are fine-tuned and clarified.

- **Input**: Literature review on simulation, e-commerce and user behaviour modelling and reunions with the supervisor.

- **Result**: Further clarification on the problem, overview and analysis of the literature related with the dissertation and this report.

- **Time frame**: 01/10/2015 - 19/02/2016 (20 weeks).

5.2.2 Model building and data collection

- **Description**: Corresponds to the steps 3 to 7 presented above. In this stage, the fundamental part of the work is done and iterated multiple times. Emphasis should be given to the reproducibility and validation of the process. This stage will be further detailed during the dissertation period.

- **Input**: This report.

- **Result**: Runnable and testable simulation system, accompanied by plausible scenarios (data) to feed the simulation.

- **Time frame**: 22/02/2016 - 15/04/2016 (8 weeks).

5.2.3 Running the model

- **Description**: Involves the steps *Experimental design* (8) and *Production runs and analysis* (9). In this stage, the model/system is tweaked and statistical analysis is done on the experiment results.

- **Input**: Runnable simulation system.

- **Result**: Measurements of performance of the system being evaluated.

- **Time frame**: 18/04/2016 - 06/05/2016 (3 weeks).

5.2.4 Implementation

- **Description**: Includes the steps *Documentation and reporting* (10) and *Implementation* (11). This final stage should conclude the dissertation, by delivering the implementation of the framework/simulation engine and the final dissertation report.
- **Input**: All the deliverables produced up to this point.
- **Result**: Documentation, dissertation report, concrete simulation system implementation and verifiable results.
- **Time frame**: 09/05/2016 - 04/07/2016 (8 weeks).

Architecture of the solution, implementation details (e.g. technologies) and on what actually consists a *framework* were purposely left out of this planning dissertation report.

Chapter 6

Conclusion

This chapter concludes the work realized for the dissertation planning. It presents the conclusions including a SWOT analysis.

6.1 Conclusion

The state of the art literature review focused on three main areas, e-commerce, simulation and probabilistic models. There was an attempt at showing the classic techniques and approaches but also in showing extensions and improvements to those.

In order to better assess the work required for the dissertation, a SWOT analysis is in order.

6.1.1 SWOT Analysis

Regarding strengths, there is a wide and vast research in the area of simulation and modelling. There is also a strong confidence that the methods, approaches and techniques reviewed in the state of the art can be used to model user browsing behaviour successfully.

Concerning weaknesses, the proposed framework relates to many different areas, which may make the dissertation too broad in scope, making it hard to focus on something specific ("do one thing and do it well"[1]).

In regards to opportunities, no other similar framework or tool was found, which means that the work being done might be novel and may provide extra value to third parties (e.g the scientific community as tool to validate and test others models like recommendation systems or to companies focusing on e-commerce).

Regarding threats, testing and validating the framework may be problematic, in terms of corroborating results with real data or even finding a suitable approach to validation itself.

[1] Unix philosophy

Conclusion

References

[ADW02] Corin R Anderson, Pedro Domingos, and Daniel S Weld. Relational Markov Models and their Application to Adaptive Web Navigation. pages 143–152, 2002.

[Ama15] Xavier Amatriain. How do you measure and evaluate the quality of recommendation engines?, 2015. Available on http://qr.ae/RUNcIK, accessed last time at 14 of February 2016.

[Any00] AnyLogic. Anylogic simulation software, 2000. Available on http://www.anylogic.com/, accessed last time at 14 of February 2016.

[AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[Bay63] Mr Bayes. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfrs philosophical transactions, 53: 370–418. *URL http://rstl. royalsocietypublishing. org/content/53/370. short*, 1763.

[BCNN04] Jerry Banks, John Carson, Barry L Nelson, and David Nicol. Discrete-Event System Simulation. *PrenticeHall international series in industrial and systems engineering*, page 624, 2004.

[BE+67] Leonard E Baum, John Alonzo Eagon, et al. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bull. Amer. Math. Soc*, 73(3):360–363, 1967.

[Bis06] Christopher M Bishop. Pattern recognition. *Machine Learning*, 2006.

[BP66] Leonard E. Baum and Ted Petrie. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[Bra14] Sally Brailsford. Modeling Human Behaviour - An (ID)entity Crisis? *Proceedings of the 2014 Winter Simulation Conference*, (Id):1539–1548, 2014.

[Col03] Nick Collier. Repast: An extensible framework for agent simulation. *The University of Chicago's Social Science Research*, 36:371–375, 2003.

[Con04] Efthymios Constantinides. Influencing the online consumer's behavior: the Web experience. *Internet Research*, 14(2):111–126, 2004.

REFERENCES

[DGW13] Tom Dickman, Sounak Gupta, and Philip A. Wilsey. Event pool structures for pdes on many-core beowulf clusters. In *Proceedings of the 1st ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, SIGSIM PADS '13, pages 103–114, New York, NY, USA, 2013. ACM.

[DK01] M Deshpande and G Karypis. Selective Markov Models for Predicting Web Page Access. *Proc. of First SIAM Intl Conf on Data Mining*, 4(2):163–184, 2001.

[FJ05] G David Forney Jr. The viterbi algorithm: A personal history. *arXiv preprint cs/0504020*, 2005.

[FRJ11] Pau Fonseca i Casas, Miquel Ramo Nñerola, and Angel A. Juan. Using specification and description language to represent users' profiles in OMNET++ simulations. *Proceedings of the 2011 Symposium on Theory of Modeling & Simulation: DEVS Integrative M&S Symposium*, 2011.

[Fuj90] Richard M. Fujimoto. Parallel discrete event simulation. *Commun. ACM*, 33:30–53, 1990.

[Hec96] David Heckerman. A Tutorial on Learning With Bayesian Networks. *Innovations in Bayesian Networks*, 1995(November):33–82, 1996.

[HOP+86] James O Henriksen, Robert M O'Keefe, C Dennis Pegden, Robert G Sargent, Brian W Unger, and Douglas W Jones. Implementations of time (panel). *Proceedings of the 18th conference on Winter simulation*, pages 409–416, 1986.

[J S00] M Deshpande J Srivastava, R Cooley. Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data. *ACM SIGKDD Explorations Newsletter 1.2*, 1(2):12–23, 2000.

[Jen99] Nicholas R Jennings. Agent-based computing: Promise and perils. 1999.

[Jon86] Douglas W. Jones. An empirical comparison of priority-queue and event-set implementations. *Commun. ACM*, 29(4):300–311, April 1986.

[KF09] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[Kin80] Markov random fields and their applications, 1980.

[Kiv69] P.J. Kiviat. Simscript ii programming language (automatic computation), 1969.

[KKK13] Aditya Kurve, Khashayar Kotobi, and George Kesidis. An agent-based framework for performance modeling of an optimistic parallel discrete event simulator. *Complex Adaptive Systems Modeling*, 1(1):12, 2013.

[Mac05] David J C MacKay. *Information Theory, Inference, and Learning Algorithms David J.C. MacKay*, volume 100. 2005.

[MAFM99] Daniel A. Menascé, Virgilio A. F. Almeida, Rodrigo Fonseca, and Marco A. Mendes. A Methodology for Workload Characterization of E-commerce Sites. *Proceedings of the 1st ACM conference on Electronic commerce - EC '99*, pages 119–128, 1999.

REFERENCES

[MCGM02] Wendy M Moe, Hugh Chipman, Edward I George, and Robert E McCulloch. A Bayesian Treed Model of Online Purchasing Behavior Using In-Store Navigational Clickstream. (April), 2002.

[MGM99] Pattie Maes, Robert H Guttman, and Alexandros G Moukas. Agents that buy and sell. *Communications of the ACM*, 42(3):81–ff, 1999.

[Mis86] Jayadev Misra. Distributed discrete-event simulation. *ACM Comput. Surv.*, 18(1):39–65, March 1986.

[MJ00] James H Martin and Daniel Jurafsky. Speech and language processing. *International Edition*, 2000.

[Nia11] Muaz Ahmed Khan Niazi. Towards A Novel Unified Framework for Developing Formal , Network and Validated Agent-Based Simulation Models of Complex Adaptive Systems. page 275, 2011.

[NMK14] Wamukekhe Everlyne Nasambu, Waweru Mwangi, and Michael Kimwele. Predicting Sales In E-commerce Using Bayesian Network Model. 11(6):144–152, 2014.

[Ong07] Bhakti S S Onggo. Running agent-based models on a discrete-event simulator. 2007.

[Ope16] Openabm modeling platforms, 2016. Available on https://www.openabm.org/page/modeling-platforms, accessed last time at 9 of February 2016.

[Pea85] Judea Pearl. Bayesian Networks A Model of Self-Activated Memory for Evidential Reasoning, 1985.

[Pea88] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[Pid98] Michael Pidd. Computer simulation in management science. 1998.

[PL05] Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, 2005.

[Rab89] Lawrence R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition, 1989.

[Ram31] Frank P Ramsey. Truth and Probability. *The Foundations of Mathematics and Other Logical Essays*, Ch. VII(1926):156–198, 1931.

[Res96] Mitchel Resnick. Starlogo: An environment for decentralized modeling and decentralized thinking. In *Conference companion on Human factors in computing systems*, pages 11–12. ACM, 1996.

[RKP02] Iyad Rahwan, Ryszard Kowalczyk, and Ha Hai Pham. Intelligent agents for automated one-to-many e-commerce negotiation. In *Australian Computer Science Communications*, volume 24, pages 197–204. Australian Computer Society, Inc., 2002.

[SA10] Peer-Olaf Siebers and Uwe Aickelin. Introduction to Multi-Agent Simulation. *Ecology*, pages 1–25, 2010.

[SC00] Jim Sterne and Matt Cutler. E-Metrics: Business Metrics for the New Economy. page 61, 2000.

REFERENCES

[Sha88] Ross D Shachter. Decision Making Using Probabilistic Inference Methods. 1988.

[SMG+10] P. O. Siebers, Charles M. Macal, J. Garnett, D. Buxton, and M. Pidd. Discrete-event simulation is dead, long live agent-based simulation! *Journal of Simulation*, 4(3):204–210, 2010.

[TK74] Amos; Tversky and Daniel Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science (New York, N.Y.)*, 185(4157 (Sept. 27, 1974)):1124–31, 1974.

[TT00] Kah Leong Tan and Li-Jin Thng. Snoopy calendar queue. In *Proceedings of the 32Nd Conference on Winter Simulation*, WSC '00, pages 487–495, San Diego, CA, USA, 2000. Society for Computer Simulation International.

[UBJK94] Onur M. Ulgen, John J. Black, Betty Johnsonbaugh, and Roger Klungle. Simulation Methodology - a Practitioner'S Perspective. *International Journal of Industrial Engineering, Applications and Practice*, 1(2):16, 1994.

[Var01] Andras Varga. The OMNeT++ Discrete Event Simulation System. *Proceedings of the European Simulation Multiconference*, pages 319–324, 2001.

[WB13] John Winn and Christopher Bishop. Model-based machine learning, 2013. Available on http://www.mbmlbook.com/, accessed last time at 9 of February 2016.

[WBS08] Frank Edward Walter, Stefano Battiston, and Frank Schweitzer. A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems*, 16(1):57–74, 2008.

[WE99] Uri Wilensky and I Evanston. Netlogo: Center for connected learning and computer-based modeling. *Northwestern University, Evanston, IL*, pages 49–52, 1999.

[Wik15] Wikipedia. Computer simulation — wikipedia, the free encyclopedia, 2015. [Online; accessed 14-February-2016].

[WJ08] Martin J. Wainwright and M I Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

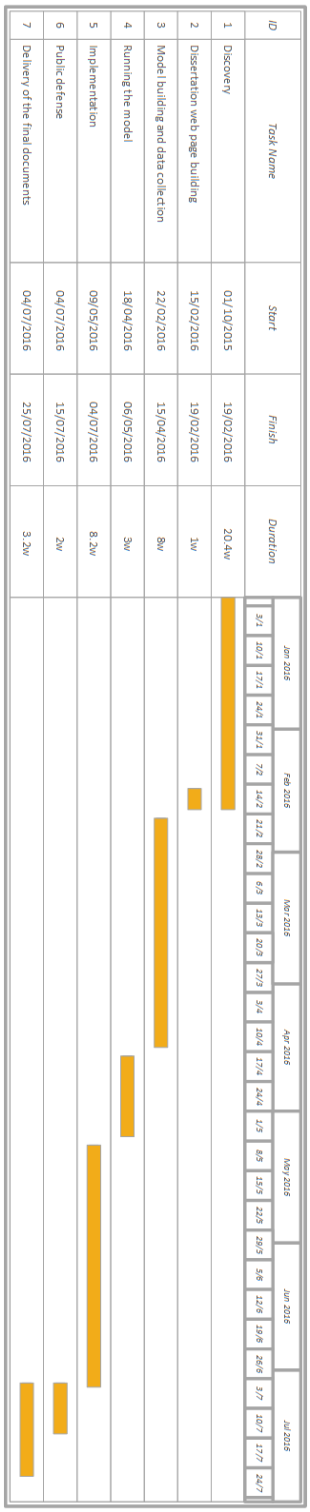
[Woo98] Michael Wooldridge. Agent-based computing. *Interoperable Communication Networks*, 1:71–98, 1998.

[XB07] Bo Xiao and Izak Benbasat. E-commerce product recommendation agents: Use, characteristics, and impact. *Mis Quarterly*, 31(1):137–209, 2007.

[XY09] Yi Xie and Shun-zheng Yu. A Large-Scale Hidden Semi-Markov Model for Anomaly Detection on User Browsing Behaviors. *IEEE/ACM Transactions on Networking*, 17(1):54–65, 2009.

Appendix A

Work Plan



ID	Task Name	Start	Finish	Duration
1	Discovery	01/10/2015	19/02/2016	20.4w
2	Dissertation web page building	15/02/2016	19/02/2016	1w
3	Model building and data collection	22/02/2016	15/04/2016	8w
4	Running the model	18/04/2016	06/05/2016	3w
5	Implementation	09/05/2016	04/07/2016	8.2w
6	Public defense	04/07/2016	15/07/2016	2w
7	Delivery of the final documents	04/07/2016	25/07/2016	3.2w

Figure A.1: Gantt Diagram