THE KNOWLEDGE DISCOVERY PROCESS

Professor: OSCAR CORCHO GARCIA
E-Mail: ocorcho@faculty.ie.edu

Academic Background
Full Professor in Computer Science, Facultad de Informática, Universidad Politécnica de Madrid (2007-now)
Visiting Professor, IRSTEA, France (2014)
Marie Curie Research Fellow, University of Manchester (2005-2007)
Visiting Researcher, Stanford University (2002)
Adjunct Professor, Universidad Pontificia de Salamanca (2000-2004)
PhD in Artificial Intelligence, Universidad Politécnica de Madrid (2004)
MSc in Knowledge Engineering, Universidad Politécnica de Madrid (2001)
BSc in Computer Science, Universidad Politécnica de Madrid (2000)

Corporate Experience
Co-founder, LocaliData S.L. (Madrid, 2013-now)
OBJECTIVES

Motivation: Digital data are changing the way we think and how we behave. With information always at our fingertips, we are not necessarily better informed. We live in a world of instant information; far too much information for us to cope with.

One of the goals of professionals taking this course and master (be them information strategists, decision makers, practitioners, students, researchers) is to accelerate and facilitate knowledge discovery from such growing wealth of data. Many of the courses in this master are focused on providing students with working knowledge of the tools and techniques that can be used in order to deal effectively with the most typical knowledge discovery problems that are arising in today’s digital world.

However, while many of the courses in the master will be focused on specific data mining or machine learning techniques and tools, there are many other issues that need to be considered in any knowledge discovery project. Every step from finding and gaining access to the data, through information extraction to knowledge delivery must be technically and methodologically well supported. For instance, it is not uncommon to see, in business contexts, presentations of the results of knowledge discovery processes where results are presented without questioning their veracity, when they have only used a minuscule percentage of the data available or when the techniques that have been applied may not be applicable for the type of data used as a source. The precision of knowledge discovery process results must be trustworthy and well understood, and the obtained knowledge must be delivered to each person in a form that suits their needs. It is also important to consider the socioeconomic issues related to data ownership and how people react to requests for projects to use “their” data.

Focus and objectives: In this course our focus is on the methodology that can be followed for typical knowledge discovery processes, taking into account existing methodologies that provide guidance during the knowledge discovery process (from well established ones to more recent agile processes). As such, the course will cover the following learning objectives, so that by the end of the course students will be able to:

1. Understand and put into practice common reference methodologies that provide guidance during the knowledge discovery process (e.g., CRISP-DM, SEMMA, Agile Data Analytics).
2. Apply techniques and common tools for the understanding of available data sources and for their pre-processing.
3. Design sound evaluation methods to assess the main characteristics of the models that have been learned during the knowledge discovery process (e.g., quality, performance, etc.).
4. Adapt existing knowledge discovery processes inside an organisation to new situations coming from changes in the organization or in its objectives, in data sources, or in available technology.

It will not be the focus of this course to enter into the details of specific algorithms or techniques that can be used for the learning of models. Similarly, tools that can be used to deal with small and Big Data sources will be out of the scope of this course. All of these will be covered in detail by other courses in the master.

METHODOLOGY

The course is structured in 10 sessions, with one of them devoted to the presentation of the hands-on work that will have been done individually and another one devoted to doing an exercise in groups. In general, sessions will be structured in three main blocks:

- Group discussion on the material reviewed by students as part of their preparation for the session.
• Presentation of additional advanced material related to the core topic of the session.
• A final set of problem-based learning activities focused on solving a real use case.

BIBLIOGRAPHY

The materials that will be the main source for the course will be provided via the campus online. These materials include chapters from popular textbooks addressing diverse areas of the knowledge discovery process, selected articles from scientific journals that go into depth of some of the relevant aspects for this course (e.g., the complete process, evaluation design), and blog entries and short and long videos that will inspire discussions during the course. Some of these materials and proposed activities will be compulsory, while others will be optional.

Below we provide the main references to books from which chapters will be selected:


PROGRAM

SESSIONS 1 & 2

Introduction to the Knowledge Discovery Process

In this session we will review and discuss about the main activities involved in the knowledge discovery process. We will review several processes and methodologies proposed in the state of the art, and analyse the activities and tasks that are normally performed in each of them.

R.A.: A survey of data mining and knowledge discovery process models and methodologies (The Knowledge Engineering Review)
B.C.: Chapter 5. Data-Intensive Analysis (The DATA Bonanza)
B.C.: Chapter 6. Problem Solving in Data-Intensive Knowledge Discovery (The DATA Bonanza)

Required reading or viewing

• Book chapter (DataBonanza_Chapter5): The Data Bonanza. Chapter 5. Data-intensive analysis. Pages 125-144.
• Book chapter (FourthParadigm_Chapter1): The Fourth Paradigm. Chapter 1: Jim Gray on eScience: A Transformed Scientific Method. Pages xix-xxxiii

Optional reading or viewing

SESSION 3

Data Acquisition and Pre-processing: Strategies to Use Data APIs

Other: How to use APIs from Twitter, Google and Facebook to find data, ideas (The Poynter Institute)

In this session we will discuss about some of the tasks involved in the data acquisition and pre-processing steps of the knowledge discovery process. As a hands-on exercise, the lecturer will expose in the classroom a problem that requires performing data acquisition through complementary REST APIs and we will analyse and propose alternatives for data acquisition and pre-processing with each of these APIs.

The students are encouraged to read the news post that is proposed in the materials section, and review the main characteristics of the APIs from Google Places, Yelp, and FourSquare, to prepare for the discussion to be held during the lecture.

SESSIONS 4 & 5

Do the Data Confirm my initial Hypothesis? Guidelines to an effective evaluation

Other: Hypothesis Testing (Laerd Statistics)

In these two sessions we will explore a number of evaluation methods that can be applied to determine whether our hypotheses in the initial steps of the knowledge discovery process hold. The lecturer will initially propose a set of problems, and students will need to propose initial hypotheses that may be derived from the available data and from the problem descriptions. Once the hypotheses are discussed and agreed upon, the students will work in groups, on the second session, proposing different types of evaluation methods to assess the validity of the results that may have been obtained after the modelling activity.
SESSION 6

Social Computing and Collective Intelligence for Knowledge Discovery

In this session we will discuss about one of the alternative methods that are being used currently for some of the phases in the knowledge discovery process, consisting on the usage of social computing and collective intelligence for data acquisition, pre-processing and curation, and for evaluation purposes.

Video: Introduction to Social Computing (Video_1)
Video: Programming the Social Computer (Video_2)


Required reading or viewing

- Video (Video_1): Introduction to Social Computing (9minutes) https://www.youtube.com/watch?v=JQL28IpwpYM

Optional reading or viewing

- Video (Video_2): Programming the Social Computer (Dave Robertson, 1 hour) https://www.youtube.com/watch?v=27L970r8J9U

SESSION 7

A Case study of Knowledge Discovery in the Marketing domain

In this session we will go through a whole knowledge discovery process coming from a real-world case from a company specialised in the marketing domain. We will start with the analysis of the problems to be solved by such a company, and we will then go through the whole process. As a result of the session, the students will have constructed their first complete project description for knowledge discovery.

R.A.: A classification of user-generated content into consumer decision journey stages (Neural Networks)

Additional material for this session will be made available when we approach this session. The recommended reading is the paper that is indicated in the session materials.

SESSION 8

Group work session

In this session students will prepare, in groups, for one hour, the design of a knowledge discovery task that will be given at the beginning of the class. This will be evaluated offline by the lecturer.
SESSION 9

Avoiding common knowledge discovery mistakes

This session will be devoted to describing common knowledge discovery mistakes, especially for early practitioners, which get much more important in the world of Big Data. We will discuss, based on examples, on how to do a proper data audit, how to determine whether a method is applicable for a specific problem, etc.

No material will be given in advance for this session, but will be based mostly on the results from the evaluation done over the group work presented in the previous session, plus some other examples of where common mistakes are usually found in knowledge discovery projects.

SESSION 10

Individual Presentations of Knowledge Discovery Projects

In this session students will make short presentations, followed by questions and discussions from the whole group, on their individual knowledge discovery projects.

EVALUATION METHOD

Final grades will be determined by combining three main evaluation metrics, based on the class participation, the quality of the teamwork and presentation, and finally the quality of the individual report and presentation.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Score %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual presentation</td>
<td>60%</td>
</tr>
<tr>
<td>Class Participation</td>
<td>15%</td>
</tr>
<tr>
<td>Workgroups</td>
<td>25%</td>
</tr>
</tbody>
</table>

CLASS PARTICIPATION

Active student participation in class is essential to the learning process and to the success of the class. Please note that you will be evaluated on the quality of your interaction, not necessarily the quantity. However, some level of quantity is also necessary. Quality participation moves the discussion along and contributes to collective learning – it adds value to the discussion. Please note that thorough preparation for classes and case discussions is key for constructive participation.

Because almost all learning is done in the classroom, attendance is required at all class sessions. A missed session will result in a zero grade for that day. Because of the limited number of meetings and the interactive nature of the course it is not possible to make up missed sessions. Please notify the professor in advance if you are unable to attend class due to illness or any other emergency.

TEAM WORK AND PRESENTATION

All students will belong to a group (number of members to be determined during the course) and, during the course, there will be a session where they will develop an assignment together, which will be delivered in the class and evaluated. A more detailed format with the specific topics to be covered in the presentation will be shared with the students at a latter stage. All team members are expected to participate, although the whole group will receive a single grade. The grading will be based on the presentation done.
INDIVIDUAL PRESENTATION

Students are expected to prepare one individual presentation on a case of their choice, and present it during the final session. The presentation will be due at midnight on the day before the class when these will be presented and they have to be submitted via the campus online tool.

The grading criteria for these presentations will include the following points:

- Quality of the analysis. A clear problem statement identifying the main challenges covered by the work done, and the main hypotheses to be tested.
- Quality of the proposed design and evaluation. A clear design and selection of the steps and techniques to be applied throughout the knowledge discovery process, together with an appropriate evaluation of the results.
- Quality of the synthesis. A clear description of the obtained results and conclusions from the knowledge discovery process.
- Presentation style and organization - logical structure, clarity, and conciseness, and quality of the answers provided on the questions.