

“Ecosystem for European Education Mobility as a Service: Model with Portal Demo (eMEDIATOR)”

Project No. 2021-1-LV01-KA220-HED-000027571

INTERNAL REPORT 3:

Pozition	
Document type	Internal Report
Responsible Partner	Transport and Telecommunication Institute (LATVIA)
Editors	B.Misnevs, I.Kabashkins, O.Zervina
Period	3
Dissemination level	Confidential
Organizations	TTI, UL, UM, UoI, AU.
Submission date	31.10.2022
Number of pages	



DOCUMENT HISTORY

Version #	Submission date	Responsible Person	E-mail	Reviewer	E-mail	Reviewer organization	Date of review submission
1.0	31.10.22	Igor Kabashkin	kiv@tsi.lv	Boriss Misnevs	Misnevs.B@tsi.lv	TTI	02.11.2022

The European Commission support for the production of this publication does not constitute an endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein

Disclaimer

Responsibility for the information and views set out in this report lies entirely with the authors.

Reproduction is authorized provided the source is acknowledged.

Information contained in our published works has been obtained by the authors from sources believed to be reliable. However, neither this consortium nor its authors guarantee the accuracy or completeness of any information published herein and this consortium nor its authors shall be responsible for any errors, omissions, or claims for damages with regard to the accuracy or sufficiency of the information contained in this consortium publications.

Abbreviations and Acronyms:

Content

1. A3.1. Development of the functional architecture of the portal. (TTI).....	8
1.1 INTRODUCTION.....	8
1.2 GENERAL REQUIREMENTS TO FUNCTIONAL ARCHITECTURE OF THE INFORMATION PORTAL.....	8
1.3 PROVIDING THE KEY INTEROPERABILITY ENABLER ARCHITECTURAL BUILDING BLOCKS	11
1.4 PRINCIPLES OF eMEDIATOR PORTAL DEVELOPMENT ON THE BASE OF EUROPEAN INTEROPERABILITY FRAMEWORK	12
1.5 SERVICE BASED FUNCTIONAL STRUCTURE OF THE MODEL	15
1.6 CONCLUSION	18
REFERENCES	18
LIST OF AUTHORS.....	18
2 . A3.2 Development of the learning delivery model of the portal (UL)	19
2.1 INTRODUCTION.....	19
2.2 LEARNING DELIVERY MODEL BASED ON THE EDUCATIONAL PRINCIPLES.....	19
2.3 LEARNING DELIVERY BY THE RECENT METHODOLOGICAL SOLUTIONS	24
2.4 LEARNING DELIVERY BY THE MOST COMMON TECHNICAL SOLUTIONS	26
2.5 CONCLUSIONS AND FURTHER IMPLICATIONS.....	28
2.6 eMEDIATOR – PRACTICAL CONSIDERATIONS FOR DESIGN.....	29
REFERENCES.....	32
LIST OF AUTHORS.....	35
3 . A3.3. Development of the model of academic and non-academic resources (TTI)	36
3.1 INTRODUCTION.....	36
3.2 OER ARCHITECTURE FRAMEWORK	37
3.3 KNOWLEDGE AND INFORMATION BASE.....	39
3.4 CONCLUSION	40
REFERENCES	40

LIST OF AUTHORS.....	41
4 . A3.4 Development of the model for job application support. (UMU)	42
4.1 CURRENT SCENARIO.....	42
4.2 MODEL VALIDATION.....	43
4.2.1 VALIDATION STEPS	45
4.2.2 MODEL REFINEMENT	49
4.3 COMPETENCE CATEGORISATION	51
4.4 COVERAGE PERCENTAGE	52
4.5 CONCLUSIONS	54
REFERENCES	55
LIST OF AUTHORS.....	56
5 . A3.5 Design of the Search Engine (UoI)	57
5.1 INTRODUCTION.....	57
5.2 SEARCH ENGINES.....	58
WHAT IS A SEARCH ENGINE.....	58
WHAT ARE SEARCH ENGINES FOR?	59
THE HISTORY OF SEARCH ENGINES.....	59
OPERATION OF SEARCH ENGINES	65
TYPES OF SEARCH ENGINES	70
EXAMPLES OF SEARCH ENGINES.....	72
PROS AND CONS.....	78
THE FUTURE OF SEARCH ENGINES.....	78
TOP 10 SEARCH ENGINES IN THE WORLD IN 2022	80
5.3 MACHINE LEARNING	81
TYPES OF MACHINE LEARNING	82
APPLICATIONS OF MACHINE LEARNING IN SEARCH ENGINES.....	85
5.4 DATA MINING	88

DATA MINING AND KNOWLEDGE DISCOVERY.....	89
METADATA	90
PRACTICAL DIFFICULTIES	90
CATEGORIES OF DATA MINING SYSTEMS.....	91
DATA MINING METRICS.....	92
DATA	93
DATA MINING PROCESS	100
DATA MINING AND COMBINATION WITH OTHER FIELDS.....	103
DATA MINING TECHNIQUES AND ALGORITHMS	105
5.5 eMEDIATOR SEARCH ENGINE.....	123
A CUSTOMIZED SEARCH ENGINE	123
APPLYING MACHINE LEARNING TECHNIQUES FOR CLASSIFYING COMPETENCES/COURSES	127
PROVIDE COMPETENCES RECOMMENDATION BASED ON REPUTATION SYSTEMS RATING.....	127
REFERENCES	130
LIST OF AUTHORS	132
6 . A.3.6. Development of the mock-up testing procedure and test case requirements (AU)	133
6.1 eMEDIATOR Mock-Up FUNCTIONALITIES.....	133
6.2 TEST CASE REQUIREMENTS.....	137
6.3 TESTING PROCEDURE / ROADMAP	137
LIST OF AUTHORS.....	138
APPENDIX 1. Examples of Functional Architecture Requirements testing procedure and test case requirements (TTI).....	139
APPENDIX 2. Examples of Learning Delivery Model Requirements testing procedure and test case requirements (UL).....	139
APPENDIX 3. Examples of Education Data Structuring and Storage Experience testing procedure and test case requirements (UM).....	139



APPENDIX 4. Examples of Technologies and API used for Digital Education System testing procedure and test case requirements (UoI).....	139
APPENDIX 5. Examples of Digital Platforms for Education Service Delivery testing procedure and test case requirements (AU).....	140
APPENDIX 6. Examples Miscellaneous	140



1. A3.1. Development of the functional architecture of the portal. (TTI)

1.1 INTRODUCTION

The development of information technologies leads to the creation of many various kinds of information systems within the European information space. Many of these systems, especially those operating within the framework of state and public institutions, require the mutual exchange of information. This exchange of information is required both between different services at the same territorial level (horizontal compatibility) and at the level of compatibility at the European, national, regional, or local level (vertical compatibility). This puts forward special requirements for the functional and information compatibility of existing and developed information systems and portals, as well as for their functional interaction.

To solve this problem, the European Union has developed a document that defines the requirements for the European Interoperability Reference Architecture (EIRA) [1].

EIRA is a reference architecture focused on the interaction of existing and developing IT systems, primarily providing government and public services.

The recommendations of the document [1] can be used to develop the functional structure of the eMEDIATOR portal.

1.2 GENERAL REQUIREMENTS TO FUNCTIONAL ARCHITECTURE OF THE INFORMATION PORTAL

For mutual understanding of all interested parties about the functionality of the system at all stages of their life cycle, a special document is developed that describes the behavior of the system in various modes of its application. This description should be equally clear to developers of portals and IT systems, as well as to users and other interested parties.

Table 1.1 lists the main functional requirements that are most often used in information systems [2].

The functional requirements may appear in the following forms, that are listed in Table 1.2.

The definition of non-functional requirements is quality attributes that describe ways your product should behave. The list of basic non-functional requirements is shown in Table 1.1.

Table 1.1: The main requirements to information systems

Main functional requirements	Main non-functional requirements
<ul style="list-style-type: none"> • Business Rules • Transaction corrections, adjustments, and cancellations • Administrative functions • Authentication • Authorization levels • Audit Tracking • External Interfaces • Certification Requirements • Reporting Requirements • Historical Data 	<ul style="list-style-type: none"> • Usability. Usability is the degree of ease with which the user will interact with your products to achieve required goals effectively and efficiently. • Legal or Regulatory Requirements. Legal or regulatory requirements describe product adherence to laws. If your product violates these regulations, it may result in legal punishment, including federal fines. • Reliability. Such a metric shows the possibility of your solution to fail. To achieve high reliability, your team should eliminate all bugs that may influence the code safety and issues with system components. • Performance. Performance describes how your solution behaves when users interact with it in various scenarios. Poor performance may lead to a negative user experience and jeopardize system safety.

Table 1.2: Forms of functional requirements

Forms of functional requirements
<p><input type="checkbox"/> Functional requirements specification document. The documentation includes detailed descriptions of the product's functions and capabilities. These could be a single functional requirements document or other documents, such as user stories and use cases. As well as the form, the specification document must consist of the following sections:</p> <ul style="list-style-type: none"> • Purpose. This section includes background, definitions, and system overview. • Overall description. The description document consists of product vision, business rules, and assumptions. • Specific requirements. The requirements might be database requirements, system attributes, and functional requirements. • Use cases. Use cases describe the interaction between the system and external users that leads to achieving particular goals. Each use case includes three main elements: <ul style="list-style-type: none"> ○ Actors are the users who will interact with your product. ○ System functional requirements describe the intended behavior of the product. ○ Goals describe all interactions between the users and the system. <p><input type="checkbox"/> User stories. User stories are documented descriptions of software features from the end-user perspective. The document describes scenarios of how the user engages with the solution.</p> <p><input type="checkbox"/> Functional decomposition. A functional decomposition or work breakdown structure (WBS) illustrates how complex processes and features break into simpler components. By using the WBS approach, the team can analyze each part of the project while capturing the full project picture.</p>

Non-functional requirements are essential during the development of information systems and largely determine the success or failure of a project. Some methods for determining non-functional requirements by the developer at the design stage of IT systems are given in Table 1.3.

Table 1.3: Main methods for determining non-functional requirements

Main methods for determining non-functional requirements	
•	Use a defined classification and classify them into three groups: operation, revision, and transition. In this way, the stakeholders and the development team build a consistent language for discussing non-functional needs.
•	With a list of pre-defined elicitation questions, you may increase the development teams productivity. Besides, you can save time when preparing for elicitation interviews and workshops.
•	Engage with the development team during the requirements definition to ensure that you are on the same page with the development team.
•	Use 'Invented Wheels' and reuse the requirements written for other systems, since software systems have a lot in common when comparing nonfunctional requirements.
•	Use automated testing tools. Such tools will help to check your product performance faster and reveal more non-functional requirements.

During the design phase, it is important to define both the functional and non-functional requirements for the IT system. The more accurately and fully this will be done, the less time will be spent on their clarification in the development process, and, consequently, the development costs will be lower.

Functional requirements are determined by the idea of the project. They include the basic functions of the IT system and how users use it.

Non-functional requirements depend on the experience of the developers and their understanding of the use of the IT system. These requirements are formed due to the accumulated experience of using the original product or similar IT systems.

The parametric differences between functional and non-functional requirements are described in Table 1.4. This distinction is based on the understanding that the first of the requirements describe what the system does, while the other requirements describe how the system works.

Table 1.4: Main parameters of functional and non-functional requirements

Parameters	Functional Requirements	Non-Functional Requirements
Requirement	It is mandatory	It is non-mandatory
Capturing type	It is captured in the use case	It is captured as a quality attribute
End-result	Product feature	Product properties
Objective	Helps you verify the functionality of the software	Helps you to verify the performance of the software
Area of focus	Focuses on user requirement	Concentrates on the user's expectation and experience
Documentation	Describe what the product does	Describes how the product works
Product Info	Product Features	Product Properties

1.3 PROVIDING THE KEY INTEROPERABILITY ENABLER ARCHITECTURAL BUILDING BLOCKS

When developing ICT systems, the main factors of partnership in the European Union at the horizontal and vertical levels are interoperable solutions [2]. At the same time, interoperability issues are key when integrating various subsystems within a single ecosystem. The key interoperability should be provided in the areas shown in the Table 1.5 [1].

Table 1.5: Key Areas of IT systems interoperability

Main Architecture Building Blocks	Main components of Architecture Building Blocks
Key Sharing and Reuse readiness Architecture Building Blocks. These blocks are key interoperability enablers for sharing/provisioning and reusing/consuming.	<ul style="list-style-type: none"> Legislation catalogue; an inventory of legal documents. This ABB is a key interoperability enabler for sharing/provisioning and reusing/consuming legal documents. Public service catalogue; a collection of descriptions of active public services that are provided by public administrations at any administrative level (i.e. local, regional, national or pan-European). All public service descriptions published in a catalogue of public services conform to a common data model for representing public services. This ABB is a key interoperability enabler for sharing/provisioning and reusing/consuming of front-office public services. Data Set catalogue; a curated collection of datasets. This ABB is a key interoperability enabler for sharing/provisioning and reusing/consuming Data. Service registry; Implements the functionality of registering the system service within a catalogue to be discovered by other services. This ABB is a key interoperability enabler for sharing/provisioning and reusing/consuming back-office services.
Key Exchange readiness Architecture Building Blocks. These ABBs are key interoperability enablers for assessing compatibility.	<ul style="list-style-type: none"> Public Policy Formulation and Implementation Instrument; Techniques or means for the development of pertinent and acceptable proposed courses of action for dealing with public problems and carrying out of a policy decision. This ABB is a key interoperability enabler for assessing the compatibility of legal/juridical certainty in exchanged information. Exchange of Business Information; communication of business information by a business capability. This ABB is a key interoperability enabler for assessing the compatibility of interaction in exchanged information. Representation; the perceptible form of the information carried by a business object. This ABB is a key interoperability enabler for assessing compatible interpretations of Data. Machine to Machine Interface; a boundary set of means enabling the exchange of data between a service and other services. This ABB is a key interoperability enabler for assessing compatible interfaces. Human Interface; a boundary set of means enabling the exchange of data between an individual and a service. This ABB is a key interoperability enabler for assessing compatible interfaces.

Conceptual model of a reusable Interoperable European Solution is shown at the Fig. 1.1 The main component of this model is listed in the Table 1.6 [1].

Table 1.6: The main components of the IT system conceptual model

The main components of the IT system conceptual model
<ul style="list-style-type: none"> One or more (integrated) public services; One or more software components provide application services that are public service neutral (application component); One or more interfaces (human interface or machine-to-machine interface); One orchestration service specific to the supported public service; ☐ One choreography service specific to the supported public service. One or more IES services (such as application mediation enablers, workflow enablers) as well as external and internal information sources and services; One or more DSI services (such as collaboration enablers and infrastructure mediation enablers); One or more catalogues that document the interoperability solutions.

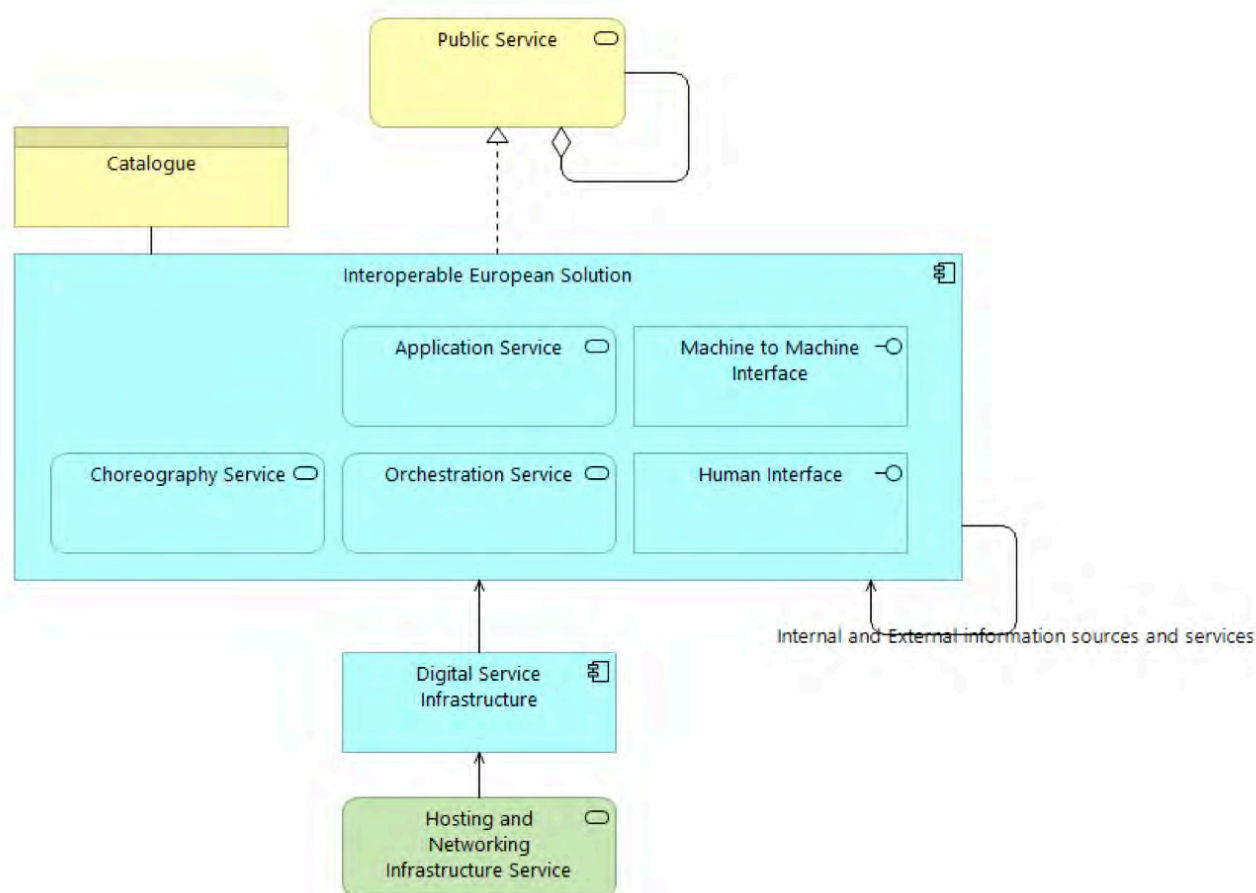


Figure 1.1: Conceptual model of a reusable Interoperable European Solution

1.4 PRINCIPLES OF eMEDIATOR PORTAL DEVELOPMENT ON THE BASE OF EUROPEAN INTEROPERABILITY FRAMEWORK

In order to achieve horizontal and vertical compatibility of IT systems developed in the European ICT space, the twelve principles of interoperability are proposed in [1]. They describe the content of the services provided and shown in Fig. 1.2.

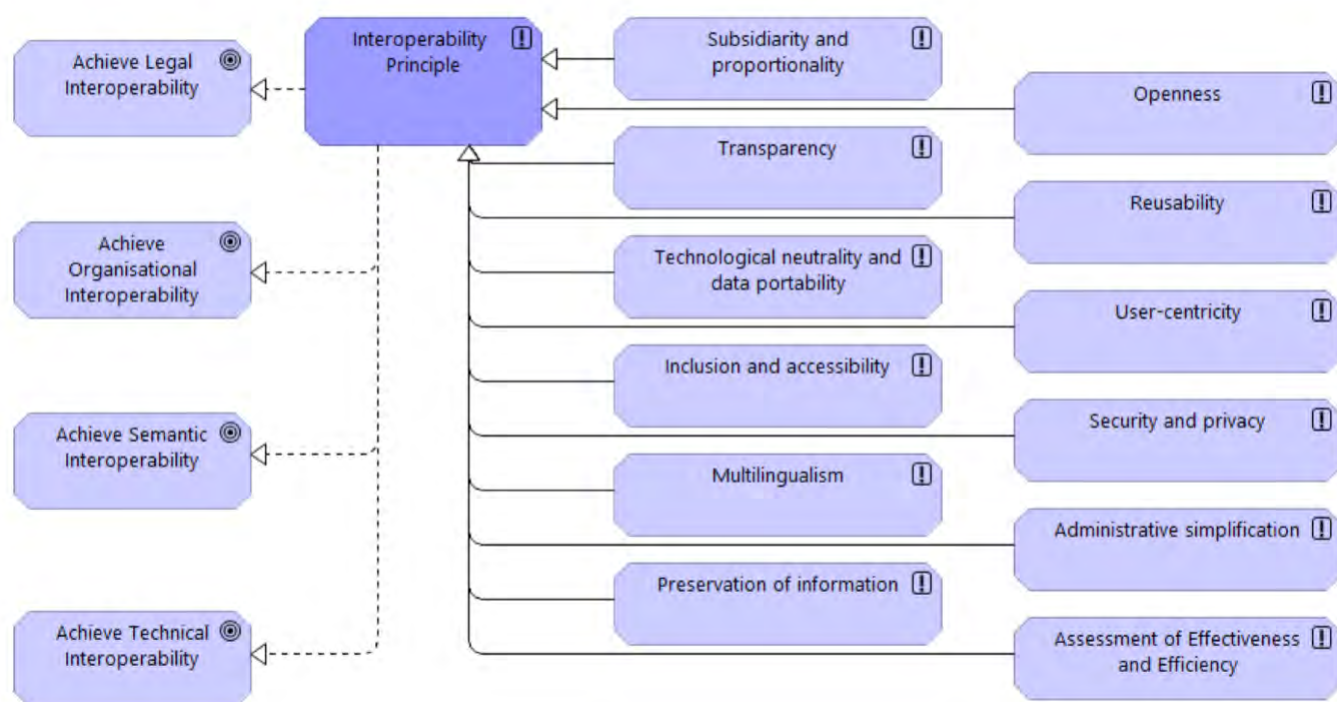


Figure 1.2: Principles of European Interoperability Framework

In order to adapt the portal model to the rapid development of modern networks and meet the diverse requirements of customers, the architecture of the portal should be based on cloud technologies with the separation of functionality between the client and server parts, as well as offline maintenance after the installation period.

Three-layer structures [3, 4] correspond to these conditions in the form of a presentation layer, a functional structure layer and a data layer. The presentation layer is on the client, the functional structure layer and the data layer are on the server. The presentation layer implements the functions of registering / logging in to the client system and online asking questions, and also receives knowledge from the e-learning system; the key to developing an e-learning system is the level of functional structure, this level mainly provides the client with knowledge related to problem solving, and performs customized training for the client; the design of the data layer is based on all the information and data of the functional structure layer that is necessary to achieve the respective functional modules.

The general view of the portal model with a three-layer structure is shown in Fig.1.3. The main functionality of the each level of this structure is described in the Table 1.7.

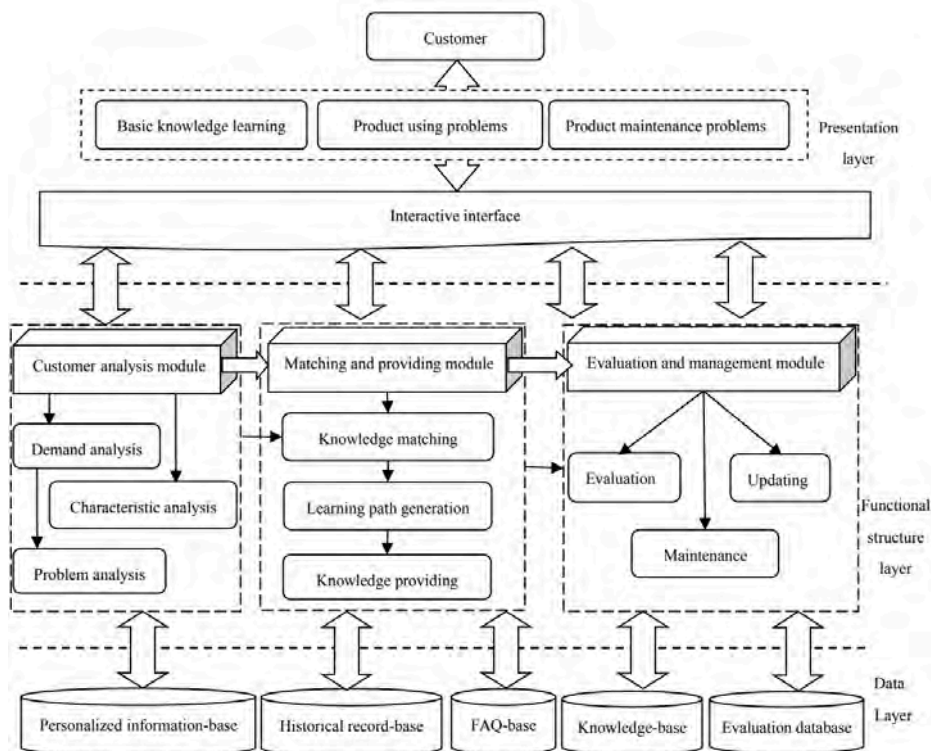


Figure 1.3: The model of portal with a three-level structure

Table 1.7: The main functionality of the three-layer structure of portal model

Structure layers of the portal model	Description
Presentation layer	Presentation layer receives customer's requests through interactive interface which gives customers the right to entrance e-learning system to fulfill problem-oriented online learning. This layer transfers customer's related information such as demands, problems and characteristics to the functional structure layer and intelligently presents knowledge to customers for problem solving. Presentation layer is not only the entrance for customers to problem-oriented e-learning system for product after-sales service but also the window of providing a variety of knowledge for customers.
Functional structure layer	<p>Functional structure layer is the main layer for problem-oriented e-learning system for product after-sales service. This layer receives customer's demands information which come from presentation layer, achieves customer's demands through interacting with data layer and sends back related knowledge to presentation layer.</p> <ul style="list-style-type: none"> Customer analysis module mainly obtains and analyzes personalized information, such as customer's characteristics, demands and specific problems, saves the analysis results into personalized information-base which is convenient to provide personalized knowledge service. Customer analysis module includes characteristic analysis sub-module, demand analysis sub-module and problem analysis sub-module which is the basis and premise for achieving personalized knowledge service. Knowledge matching and providing module achieves matching, collecting and providing of personalized solving knowledge which is the core of e-learning system. The system collects knowledge from knowledge-base, FAQ-base to match the problem solving knowledge according to personalized information and generate learning paths. Self-collecting problem solving knowledge or system recommendation knowledge based-on knowledge-base which achieves adaptive learning mode is high real-time while cuts down service costs, but largely relies on knowledge-base; Instruction learning mode solves customer's problems with high quality, but the real-time and efficiency are low and the service costs are high; Service personnel matches problem solving knowledge or system recommendation knowledge based-on FAQ-base which meets customer's real-time requests, but cannot provide personalized knowledge. Process evaluation and management module evaluates and analyzes the whole e-learning knowledge process, stores the results into learning evaluation database and forwards customer's new problems to all kinds of service departments in firms which much better fulfills internal department-oriented functions. This module is also beneficial for maintaining and knowledge updating of e-learning system so that continuously improving the after-sales service efficiency and service quality.
Data layer	Data layer locates at the bottom of architecture which is used to store all resources for achieving the process and functions of problem-oriented e-learning system for product after-sales service. Data layer mainly includes customer personalized information-base, knowledge-base, FAQ-base, learning historical record-base, and learning process evaluation database. Data layer saves and manages all types of information and data which support to achieve the problem-oriented e-learning system functions for product after-sales service.

All levels of the portal architecture are both mutually independent and interconnected. However, the connection between neighboring levels is rather weak. This three-layer architecture offers good scalability combined with ease of use and maintenance.

1.5 SERVICE BASED FUNCTIONAL STRUCTURE OF THE MODEL

The service delivery model provides the framework within which users receive services. The arrangement or configuration of time, resources, location of services, and collaboration among all actors makes up the service delivery model selected that will best meet individual user needs.

The general functional structure of the platform services for all users is represented by taxonomy in Fig. 1.4. This figure shows a basic set of services that can be refined and detailed during platform development and testing of its capabilities.

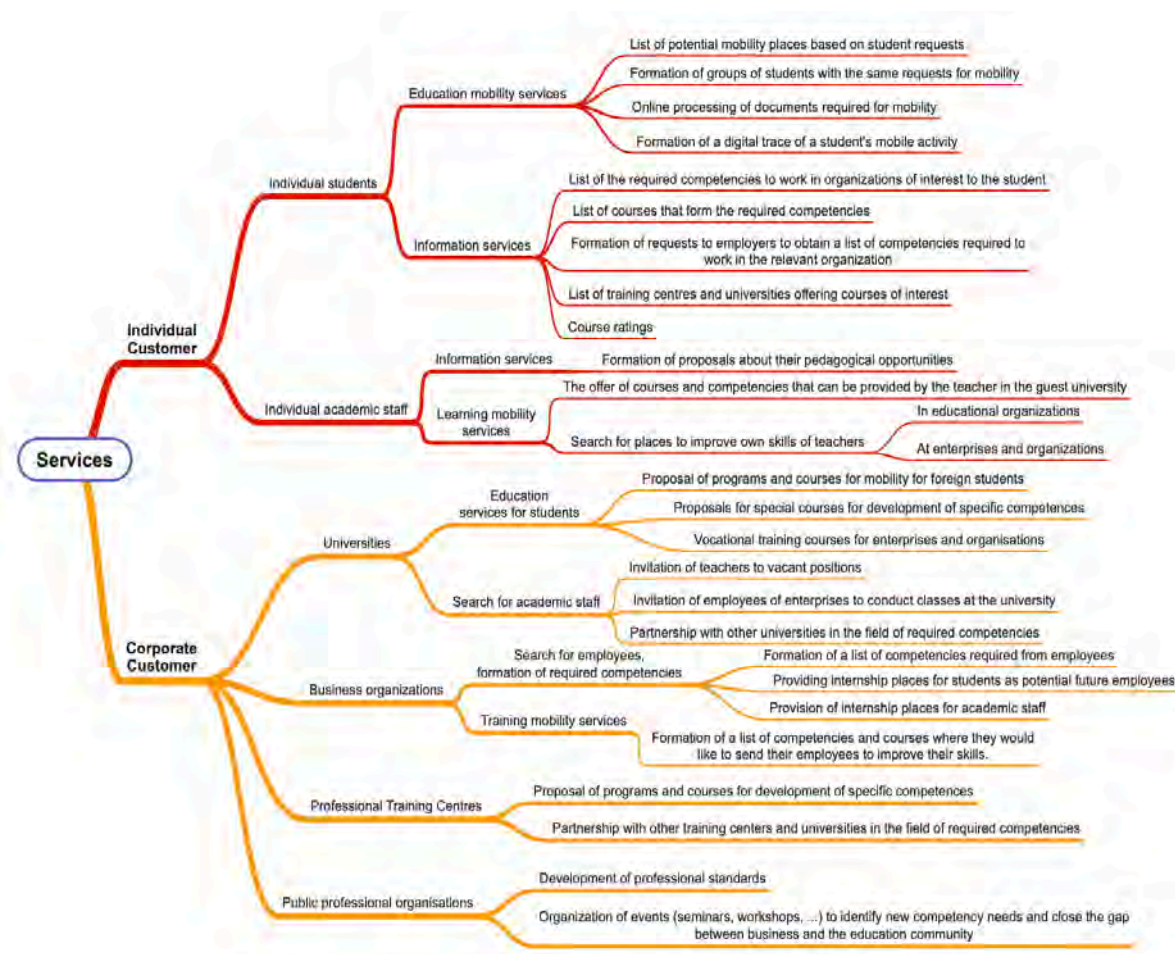


Figure 1.4: The general structure of EaaS framework services

Considering the above, we can state that the platform being developed (and, accordingly, the framework itself) as a complex system should have a multi-level structure with distributed functionality and many horizontal links on demand within the framework. At the same time, for the end user, most of the intermediate services should be transparent (invisible). Obviously, when solving this problem, mainly as a problem of integrating a variety of existing educational services, it is necessary to solve the problem of creating a universal interface within a certain architecture. At the same time, this interface should be based on some entity that connects the interests of all stakeholders in education. All communications with users of the described framework should be related to certain competencies (requested or supplied). The user will be provided (if available) with the required educational service (external to the framework or internal) for the formation of the required competence (Fig. 1.5). In the absence of the required service, the framework will be able to organize the search or creation of the required service for a specific competency request. Thus, another important requirement for the framework appears - it must be open to expanding the set of services.

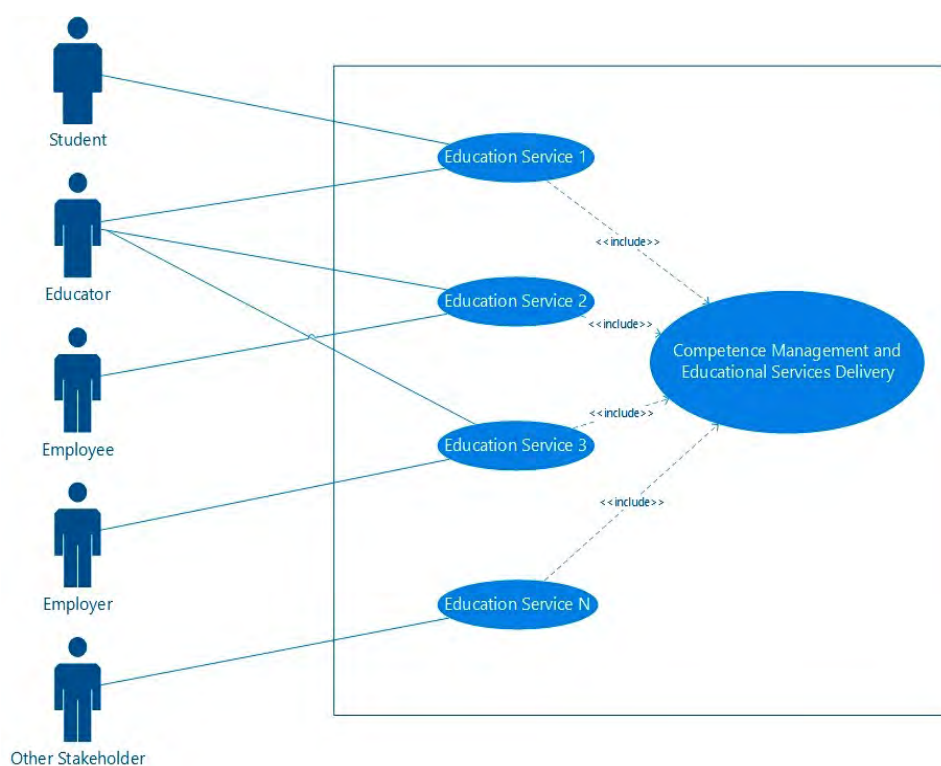


Figure 1.5: General view of the competence-based digital framework for Education as a Service.

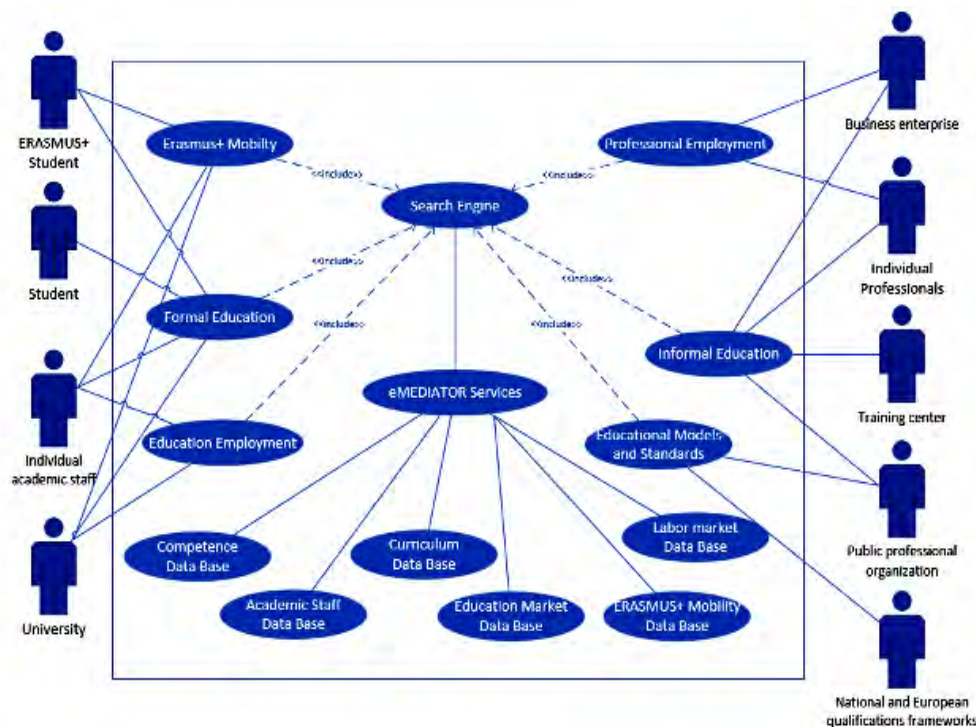


Figure 1.6: Use case diagram of EaaS information portal

A use case diagram (Fig.1.6) is a graphical depiction of a user's possible interactions with a system and shows various use cases and different types of users the EaaS eco-system has.

As a result of the analysis, it was shown that the platform being designed has a clear focus on processing large amounts of data. Therefore, we proposed another model that generally reflects the expected data processing processes in the eco-system. This architectural model is presented below (Fig.1.7).

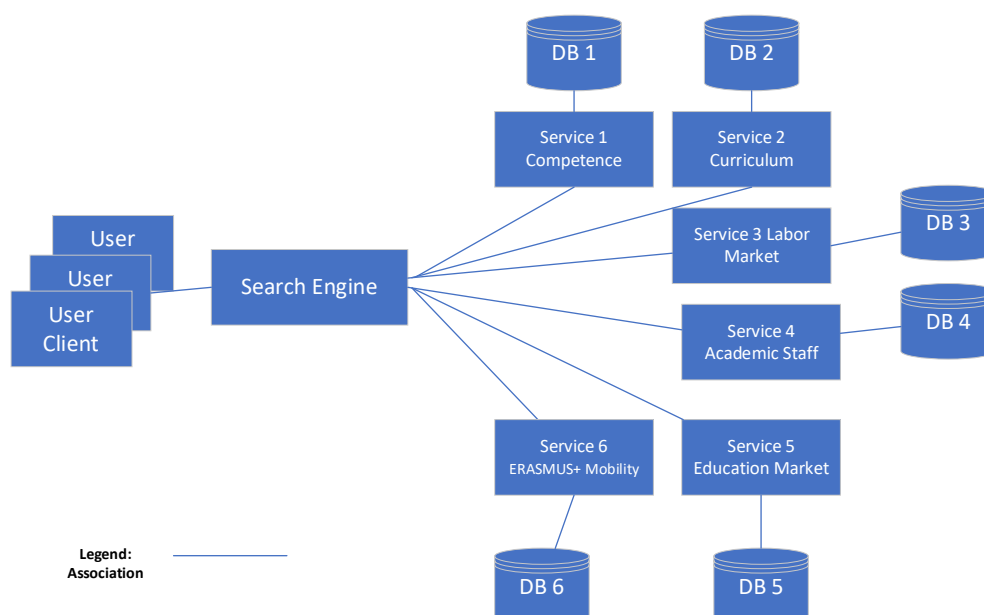


Figure 1.7: Data flow processing architecture diagram of EaaS information portal

The main element of the platform is a search engine, which provides information for each of the potential users within the framework of the services provided.

The complexity of communications both with users and between services within the system places special requirements on the development of interfaces (both GUI and API).

1.6 CONCLUSION

The section of report describes the basic approach for development of functional architecture and digital framework for creating platform for higher and professional education based on Education as a Service (EaaS) model. The complexity of the problem being solved are noted.

The section of report identified main directions of the framework and out-lined possible solutions.

The basic principles of interoperability and functional structure for the development of a digital platform that implements the ecosystem of the EaaS framework using a competency-based approach are formulated, and the functional architecture for services delivery model of the framework are described.

REFERENCES

1. An introduction to the European Interoperability Reference Architecture (EIRA©) v2.1.0.
https://ec.europa.eu/isa2/eif_en/
2. DECISION (EU) 2015/2240 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 25 November 2015 Establishing a programme on interoperability solutions and common frameworks for European public administrations, businesses and citizens (ISA2 programme) as a means for modernising the public sector.
3. Leina Deng, Zhigao Chen. An Architecture of Problem-oriented E-learning System for Product After-sales Service: Design and Application. Proceedings of the Wuhan International Conference on e-Business, 600-608.
4. Functional architecture of multimedia content delivery networks. Recommendation ITU-T H.644.3, 2020.

LIST OF AUTHORS

1. Boris Misnevs
2. Igor Kabashkin
3. Olga Zervina

2 . A3.2 Development of the learning delivery model of the portal (UL)

2.1 INTRODUCTION

The pioneers in using educational platforms are Australia and USA. Due to the vast areas and distances to be travelled, students learned under the supervision of a teacher, but staying at home. Currently, educational platforms are used all over the world, and their popularization took place during the pandemic SARS-CoV-2. The additional factor in the development of educational platforms is ecology and economy. No need for daily trips of students and teachers to schools and to other educational institutions means huge energy and financial savings globally.

Educational platforms are perceived as a tool for the effective transfer of knowledge without personal (face-to-face) contact between the teacher and the student and for independent adult learning. However, users of educational platforms still experience problems and challenges. Educational platforms – despite similar functions – differ in structure, interface, operating principles, task creation technique, way of recording or authorization. The experience gained while working with one platform helps to understand the general principles of others, but it is not enough to switch to another without additional training. Another problem arises when migrating resources, entire courses, users accounts to other applications due to the lack of full compliance despite existing standards, for example, SCORM.

Another challenge to the development of educational platforms are not technical possibilities but social norms and perception of this type of learning. Still in many societies, it is believed that this type of education has poor quality. Educational platforms as any breakthrough innovation that is significantly different from the established model initially causes resentment and fear. Therefore, scientific research and practical activities are necessary that would deflect the social perception of educational platforms and ensure high quality of education carried out in this way.

It seems particularly important to develop principles that regulate and define the educational processes on the platforms. Only education based on pedagogical and psychological principles can be of high quality and credible in the eyes of platform users.

2.2 LEARNING DELIVERY MODEL BASED ON THE EDUCATIONAL PRINCIPLES

On-line education – like all human activities aimed at achieving a given goal – cannot be chaotic.

The principles of education (teaching - learning) are defined as general norms of teaching behavior of a teacher and student activity. They can be understood in two ways:

- as claims based on scientific laws governing education
- as norms of behavior recognized as binding in education

They arise from basic regularities of the educational process, human ability to learn and currently implemented educational processes. Adherence to the principles of education increases the likelihood of achieving reliable learning

goals. The principle is a didactic standard that should be followed when organizing and implementing the process teaching-learning.

There are different classifications and typologies of educational principles. In this report, we present those that may be applicable in organizing education on the eMEDIATOR platform. In our opinion, these are the following principles: the principle of conscious and active learning, the principle of systematic learning, the principle of individuality and teamwork, the principle of combining theory with practice and the principle of developing learning skills.

2.2.1. The principle of conscious and active learning

Many studies show that learners generally perceive active learning as the most favorable process in education [1, 2] and increasing their self-efficacy [3]. Additionally, the use of active learning in STEM (science, technology, engineering, and mathematics) education has been linked to improvements in student retention and learning, particularly among students from some underrepresented groups [4-6].

According to Charles Bonwell and James Eison [7]:

“Using an active learning environment can enhance the integration of practice and theory in the classroom. We think of active learning as using instructional activities involving students doing things and thinking about what they are doing. Some characteristics of active learning are:

- Students are involved in more than listening;
- Less emphasis is placed on transmitting information and more on development of students’ skills;
- Students are involved in higher order thinking (analysis, synthesis, evaluation);
- Students are engaged in activities (such as writing, reading, discussing, and observing);
- Greater emphasis is placed on students’ exploration of their attitudes and values”.

Being active is therefore a necessary condition for effective learning. Therefore, the platform users should be aware of the goals of their activity. There are three types of activity in the education process:

- intellectual activity
- emotional activity
- practical activity

Practical implications for eMEDIATOR platform:

- referring to the life experiences, needs and cognitive interests of users
- developing positive motivation to learn, encourages and mobilizes users to effort
- applying educational methods that enable users to acquire knowledge and acting independently
- systematically making the users aware of their progress

2.2.2. The principle of systematic learning

This principle emphasizes the need to implement the teaching process - learning in logical order.

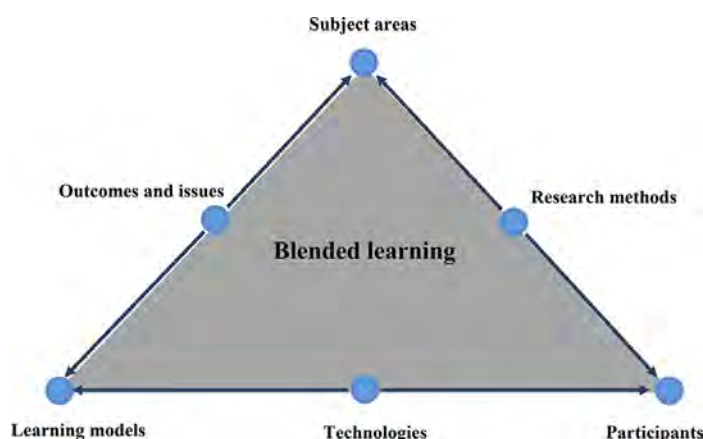


Figure 2.1. Learning logical order,

Source: https://www.dovepress.com/cr_data/article_fulltext/s331000/331741/img/PRBM_A_331741_O_F0001g.jpg

In the learning process, the users should systematically assimilate knowledge and skills and consolidate them.

Practical implications for eMEDIATOR platform:

- referring to the mastered material, bind its individual parts in whole
- emphasizing (with color, tone) the main and important issues
- showing logical relationships
- dividing the content into meaningful fragments, but in constant reference to the whole
- systematize and generalize at the end of the topic and at the end of the course
- offering users additional activities that require a longer and systematic effort

2.2.3. The principle of individuality and teamwork

According to Kim Carlotta von Schönfeld, Wendy Tan, Carina Wiekens & Leonie Janssen-Jansen [8]:

“Learning is a process of adaptation to one’s environment, in which an experience in one moment leads to alterations in (implicit) knowledge structures¹ and eventually is likely to impact behaviour. In psychology, four types of learning are usually identified: classical conditioning, operant conditioning, cognitive learning and social learning. Classical conditioning works through the gradual association of a representation of something with the thing itself. Classical conditioning might lead an individual to learn that pink is a ballet colour through the continuous appearance of the colour pink in ballet shoes. Operant conditioning works through perceived consequences of voluntary actions. Operant conditioning might induce an aspiring dancer to learn that an intense warm-up is unpleasant but leads to better results during practice. Cognitive learning is learning through reading or other internal activities, such as thinking to oneself. A dancer might read about human mechanics and then relate this gained knowledge to the way she can perfect a certain movement. *Social* learning is understood as imitation or other forms of learning *through a social context* (e.g. direct instruction). In this case, the dancer might learn by observing and copying the movements of other dancers, or through the discussion with others of how certain movements could work. These types of learning are not mutually exclusive; for instance, cognitive learning can occur through social learning. The only point at which any of the other types of learning

exclude social learning is when individuals learn cognitively by reading an informational text or experimenting by themselves”.

These researchers highlight the meanings of both individual and team learning experiences.

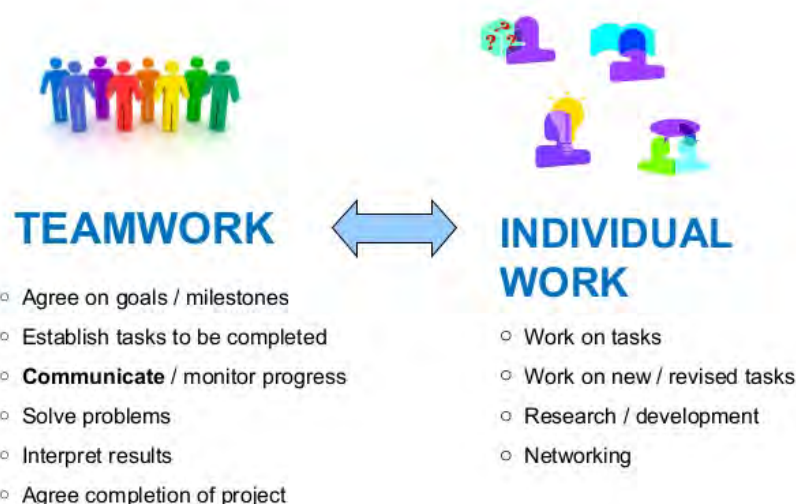


Figure 2.2. Team work and individual work,

source: <https://www.consultshol.com/leadership-management-business-and-economics.php>

The principle of individuality and teamwork requires to organize learning on the platform in such a way as to take into account both the individual capabilities of each user and the potential of cooperation between users.

Practical implications for eMEDIATOR platform:

The platform should include both individual activities and group tasks, but preferably the users themselves decide which group they want to join or create these groups themselves and invite others.

2.2.4. The principle of combining theory with practice

Learning by doing is seen as the most effective. Therefore, in recent years in education, methods of combining theory with practice have become more and more popular, which improve skills useful in everyday life, e.g., logical thinking, discussing, creativity, information selection, etc.

Developing skills in action has many benefits for both students and teachers. Children who actively participate during the lessons are definitely more open to knowledge and achieve better learning results. Frequent use of activating methods in the teaching process contributes to the integration of the team and building a proper teacher-student relationship based on mutual respect and trust.

Practical methods develop creative and critical thinking, broaden interests, and develop the ability to communicate in a group, argue and draw conclusions. By working with activating methods, we also support the emotional development of our students.

The use of practical methods requires the teacher to properly prepare the lessons in such a way as to create as many educational situations as possible for students in which they can experience and experience. The teacher should act as an adviser who, instead of providing information, teaches the student how to find it. He organizes work in such a

way that the student can: search, act and cooperate in a team, ask questions, discuss, search, and select information. In activating methods, the role of the student is the most important, without his active participation, the teaching process is incomplete.



Figure 2.3. Relations between theory and practice,

source: <https://aquetzalsaved.wordpress.com/2014/04/29/combining-theory-and-practice/>

Combining theory with practice develops the users' conviction about the usefulness of knowledge, evokes positive emotions and motivates (especially adults) to use the platform.

Practical implications for eMEDIATOR platform:

- connecting new knowledge with natural and social reality, technical, cultural
- practical activities should be preceded by theoretical knowledge
- the rules, principles, definitions, and laws underlying the activity of users of activities should be a product of their own activity

This principle should be considered where possible. We cannot forcibly link theory with practice where there are no such links.

2.2.5. The principle of developing learning skills

Formal education is most intense in childhood and adolescence, but man has the ability to learn throughout life. Learning is the process of acquiring knowledge, skills and habits strongly related to memory. What people learn is first stored in short-term memory, and through revision and retention, it passes into long-term memory, where it becomes knowledge. What was written there is almost entirely for the rest of your life.

Positive feelings are conducive to effective learning: optimism, moderate joy, self-esteem, and adequate motivation. Adequate motivation promotes concentration and perception and makes you resistant to stress and fatigue. It is very important to recognize your own learning style. Everyone uses all their senses to acquire knowledge and skills. It is important to determine which sense is dominant and which is supporting.

The development of mental abilities covers three groups of topics:

- speed reading skills,
- ability to process information effectively (including proper notetaking),
- ability to memorize large amounts of material quickly and permanently.

Due to the rapid and unpredictable changes in the labor market, the learning occurs throughout life. It should also be assumed that the eMoodle platform will not be the only and last educational experience of its users. It is therefore important that the platform develops positive associations with learning, so that its users will want to continue learning in the future.

Practical implications for eMEDIATOR platform:

- clear and fair rules of using the platform
- appreciating even no smaller achievements
- no comparisons with other users
- technical, psychological, and pedagogical support at every stage of using the platform

2.2.6. Putting principles into practice

The creators of educational platforms should consider the complexities, possibilities and needs of implementing the principles of education discussed here. Thanks to these principles, education will be more effective and friendly to learners and educators. These principles should be implemented by fully exploring the technical possibilities of platforms, using modern teaching methods and teaching aids.

The implementation of the principles of education on educational platforms is a bit more demanding than in stationary education. This is mainly due to the anonymity of platform users and the lack of knowledge of educators about their previous educational experiences. Therefore, it is important that the users of the eMEDIATOR platform have pedagogical and psychological support in the form of the possibility of consultation with the educator. Users' awareness that they are not left alone with technology can reduce fears related to educational failures and improve the quality of education.

2.3 LEARNING DELIVERY BY THE RECENT METHODOLOGICAL SOLUTIONS

2.3.1. The Four-Dimensional Instructional Design – new delivery model

A conceptual framework for designing educational activities is essential when applying state-of-the-art solutions, such as informatics or virtual reality tools. This topic incorporates human-computer interaction (HCI) aspects and pedagogical and psychological theories, concepts, and models. We propose a new approach to instructional design (ID) for online courses and virtual reality training.

2.3.1.1. Introductory Remarks: Instructional Design and Human-Computer Interactions Design

The e-Mediator project can be designed according to a user-centred design (UCD), which is among the most popular design philosophies, according to which users must take centre stage in the design of any computer system. The requirements and limitations of a given solution are articulated by users, designers, and technical practitioners. Work on user interfaces must adhere to specific principles regarding display design and methodology. Perceptual, mental,

attention, and memory principles are described in An Introduction to Human Factors Engineering [9]. Several perception principles can be applied to displays, such as legibility or audibility, top-down processing, redundancy gain, and the use of discriminable elements. The mental model principle applies to attention and memory; it involves realistic representations and moving elements according to the user's mental model. There are a number of principles based on attention, including the principle of proximity compatibility, the principle of multiple resources, and the principle of minimizing information access and interaction costs. Memory can be replaced with visual information using three principles: knowledge of the world, predictive aids, and memory consistency. The principles of HCI design are applied to ID to design intuitive and understandable interactions. The HCI principles can be incorporated into ID models if the digital resources are developed from the beginning.

An essential aspect of HCI is that it considers not only perceptual and cognitive factors of interactions but also their motor components. Based on a focus on perceptual and motor aspects of the interaction [10], we can develop an ID process that recognizes a psychomotor dimension of learning. To accomplish this, it is crucial to investigate a learning theory which intends to provide a comprehensive and integrated understanding of learning.

2.3.1.2. Cognitive, Emotional, Social and Psychomotor Dimensions of Instructional Design

Contemporary learning theories assume that learning is a three-dimensional process involving cognition, emotions, and social interactions [11]. Cognition refers to knowledge and skills, emotion to feelings and motivations, and social to communication and cooperation. According to the concept, one dimension may predominate in the learning process, but the other two are always present. Cognition, emotions, and environment are placed at the top of Illeris's triangle inverted.

Since HCI focuses on motor-related aspects of interactions, we can propose an extended version of the approach [12], which focuses on physical movement and coordination in designing learning activities. In Bloom's taxonomy of educational objectives (1956), the psychomotor domain was included in addition to the cognitive and affective domains [13]. Romiszowski (1993) (Message Design for Psychomotor Task Instruction) [14] adapted it to e-learning activities based on theories developed by Dave (1970) [15], Simpson (1972) [16], and Harrow (1972) [17].

As the central nervous system controls motor learning and knowledge, Illeris includes them in the cognition dimension [11]. Because of the specific nature of learning in a virtual environment, a diversity of learning activities requires taking into account a separate dimension, such as touching or moving. The goal of extending this theory to instructional design in virtual reality is to maintain a balance in the design of activities. This approach may also serve as an evolved version of instructional design, a Learning Experience DesignTM (LX or LX Design), a synthesis of instructional design, educational pedagogy, neuroscience, social sciences, design thinking, and user experience (UE) or user interface (UX) design. Table 1 presents examples of instructional design based on four dimensions.

As an advocate for users, an instructional designer creates intuitive solutions for learners to accomplish tasks effectively. These solutions are applicable to e-learning courses created in 2D, 3D, or other integrated learning environments, like VR.

Table 2.1. Exemplary tips for instructional design in four dimensions [18]

Cognitive dimension	Emotional dimension	Social dimension	Psychomotor dimension
<p>1) The content should be presented in small doses - bite-sized knowledge pills.</p> <p>2) Do not use unnecessary words that do not add value to the content.</p> <p>3) Utilize bookmarks containing short snippets of text, pop-ups displaying content. 4) Provide examples to illustrate the content. 5) Make multimedia meaningful.</p> <p>6) The course difficulty should be graded.</p>	<p>1) A friendly interface and visual appeal will ensure the course's visual appeal and intuitiveness</p> <p>2) Using intuitive and graphically refined templates, analyze the visual aspect of the course based on user experience design</p> <p>3) Provide illustrative examples to maintain interest in the content via visual aids.</p>	<p>1) Support and facilitate social communication.</p> <p>2) Use the chat box or discussion forum to interact with other course participants.</p> <p>3) Establish contact rules and use friendly messages - ensuring a friendly relationship.</p> <p>4) Provide technical support to course participants.</p> <p>5) Provide opportunities for individual and group learning.</p>	<p>1) The use of technology, such as Kinect or smart accessories, can be used to provide an adequate number of repetitions of an issue.</p> <p>2) Design a task so that it differentiates between activity in a virtual space, e.g. drag and drop items, find words or objects, select words or objects, type words, grab objects in VR, move your position, run, etc.</p>

2.4 LEARNING DELIVERY BY THE MOST COMMON TECHNICAL SOLUTIONS

2.4.1. 2D e-Learning Platforms

An e-learning platform is an online system for developing and distributing content; a Learning Management System (LMS) or Learning Content Management System (LCMS) is an e-learning platform. It provides learners access to suitable courses and resources through e-learning management facilities. By contrast, Learning Content Management Systems (LCMS) provide course developers with appropriate tools for creating e-learning content. As one of the most popular 2D platforms, Moodle (Modular Object-Oriented Dynamic Learning Environment) provides educators, administrators, and learners with a robust, secure, and integrated system for creating personalized learning environments [19]. As another example of an LMS that supports learning content management, collaboration,

communication, evaluation, and assessment functions, ILIAS (Integrated Learning, Information, and Work Cooperation System) is based on the German language.

Creating short, informative chunks of knowledge and aggregating them into smaller parts based on the rule of grading difficulty and highlighting important information while hiding less critical information can be used to achieve a cognitive dimension through designing activities. A click on an interactive picture switch to more detailed information after the highlighted words serve as visual anchors. A second dimension is emotions, which involves liking or disliking the course's visual properties (colours, shapes, video elements, interactive elements). Information gathered through a consistent template must be presented attractively and engagingly. The social dimension, which involves building collaborative interactions, may be challenging to implement. A flat 2D platform does not allow students to interact naturally. Social interactions occur through the system's built-in communication tools, which are often text-based and text-heavy instead of problem-solving. A forum or chat will therefore be primarily responsible for the social dimension. In most cases, they are separate from course objects, which a tutor moderates. Because of the specificity of the learning environment, simplified tasks such as drag-and-drop exercises or object movement can be designed to realize the last dimension.

2.4.2. Virtual Worlds – 3D Platforms

Virtual worlds, or three-dimensional (3D) social platforms, are communication systems that permit multiple interactants to share a three-dimensional digital space, regardless of their physical location [20]. They are computer-simulated worlds that provide perceptual stimulation to the user, who can manipulate elements of the modelled world and experience some level of presence. These environments are engaging, collaborative, participatory and conducive to learning [21]. Various factors may be incorporated into virtual worlds to enhance learners' engagement in a course. Nevertheless, this kind of solution depends on learners' preferences and ages. Immersive technologies provide alternative environments for situated learning [22], based on the idea that students are more involved in learning when they interact with real-world problems or situations. As part of the immersive experience, various sensory inputs are employed (graphics, sounds, visual perceptions of moving through the environment, the ability to touch objects, and maps with location clues) and social communication layers [23]. This relates to the appearance of avatars in virtual worlds and the communication between objects via language and text. Second Life (SL) or Sansar, created by Linden Lab, are examples of 3D virtual islands.

In such complex environments, ID is challenging; it requires setting goals and finding a balance between the types of activities. For the cognitive dimension, exemplary activities include conducting virtual experiments, which allow students to learn through mistakes or constructing the meaning of concepts through social interactions (which also relates to ID's social dimension). A meaningful animation or exploration of fantasy items with knowledge pills can provide the emotional component. Colours, shapes, blinking interactive objects, and in-world residency are the tools that stimulate learners' motivation; background music also encourages attention and interest in the course [24]. In this dimension, instructional designers must consider visual and aural aspects. In the social size, interactions involve design tools for in-world communication, such as group chat, instant messaging (IM), and voice communication. The interaction between 3D platforms and external social media can sometimes be transferred outside the platform. The interaction between students

and teachers increases motivation and reduces procrastination [25]. The SL interactions can be fully diversified in the ID process, which allows students' motivation to be easily controlled. In the last psychomotor dimension, the avatar moves (walks, runs, flies), or objects are constructed using the mouse. In addition, instructional designers must consider kinesthetic activities such as exploring new places and teleporting avatars over holodecks; these kinesthetic activities provide visual and auditory stimuli to learners.

These dimensions, if balanced, provide balance in virtual world instructional design.

2.4.3. Virtual Reality and VR Accessories

A new approach to instructional design can be applied in training through virtual reality and appropriate accessories. Using VR goggles or simulators, users can move through virtual worlds as if they were real. Various genres of educational games may be used for their VR content. Other than that, VR can also be used as a vest to make immersive games. Even VR gym facilities can either make exercising more enjoyable or less like work [26]. There seems to be a dominant psychomotor component to the activities. When VR tools are combined with tools to practice specific skills, the psychomotor aspect is usually emphasized.

As VR accessories are used more intensively, the psychomotor dimension appears to dominate. Thus, instructional designers must make sure that other tasks are assigned to other dimensions in order to maintain the balance between activities. As an example, the cognitive dimension could involve repeating tasks, grading the difficulty level, presenting content in 3D scenery, and exploring information through specific activities based on VR capabilities. Using appealing scenery, applying intuitive 3D objects to perform tasks, or illustrating the action with sound effects will achieve the emotive dimension. Immersion in a realistic setting increases attention and engagement in task completion. As part of the social dimension, virtual humans should interact with each other. For the psychomotor dimension, it is essential to emphasize that state-of-the-art solutions should be incorporated into the design of learning activities. In the synchrony condition, participants reported significantly higher levels of social closeness than in the non-synchrony condition [27]. E-Mediator should not ignore the findings when building its platform.

2.5 CONCLUSIONS AND FURTHER IMPLICATIONS

Learning scenarios based on the four-dimensional ID approach, which assumes a balance between cognitive, emotional, social, and psychomotor dimensions of learning, are more engaging for learners.

When planning activities, designers may be able to make them more engaging and attractive by thinking about them through the lens of four-dimensional learning. In designing a lesson, for example, it is essential to ensure that there are concise instructions, a sufficient number of repetitions, a level of difficulty of the task, illustrations, sound, social learning opportunities, and kinesthetic activities. To design the course,

keeping the balance between these dimensions is essential. The choice of activities will depend on several conditions, such as the tool's functionalities, the course's purpose, the training content, and the group's needs.

Designing activities in cyberspace will gain momentum in the future. It's not just about reaching a broader audience or saving money that entirely virtual university courses are on the rise. Implementing education in virtual spaces aims to create new interdisciplinary learning pathways since virtual resources can be updated, processed, modified, and merged quickly. In terms of virtual resource design and development, older and new technologies are typically integrated, and hybrid solutions are incorporated into the didactic process. A novel breed of software endless innovations is powered by immersive 3D virtual learning environments (supported by artificial intelligence) [28], which go beyond traditional e-learning platforms. When adapted into educational processes, these innovations require new approaches. It is possible to harmonize the design of courses with the author's approach to instructional design (or even its developed version, LX Design), which addresses the four dimensions of learning.

2.6 eMEDIATOR – PRACTICAL CONSIDERATIONS FOR DESIGN

From the teaching point of view, the optimal platform is an invisible one. Where all the space is dedicated exclusively to didactics. The platform should be a cost-effective environment for students, academics, scientists, and business as well. The platform should enable effective content administration and management, supported by the latest technologies. eMEDIATOR should also ensure comfortable work even at low Internet speeds. It may be strange, but high-speed Internet is not a standard. It should also be remembered that the excess of content interferes with the teaching process! Sometimes less is better.

A few practical comments can be made that reflect the essence of cost-effective solutions in the field of a modern educational platform:

- eMEDIATOR should provide a friendly and cost-effective environment for all participants: students, teachers, and business.
- eMEDIATOR should be open to new technologies in education and constitute a kind of HUB of the best solutions in this area.
- One should strive to individualize teaching according to the student's preferences, supporting this AI process.
- Automating platform management processes should be a priority. This will allow you to focus on the teaching process.
- eMEDIATOR should provide a comprehensive solution for the preparation of the teaching process, its implementation, monitoring and reporting.

The platform should also support security. In addition to data security, it should support the student in choosing a university (Erasmus) and be a source of reliable information. Therefore, eMEDIATOR should work hard in the background so that participants in the teaching process can concentrate on the educational goal!

Therefore, eMEDIATOR should enable correct operation also without Internet access. For example, a student at a university synchronizes his software with the eMEDIATOR: downloads current tasks and materials. Outside the university, he or she works without the Internet, and after returning he sends completed tasks (re-synchronization). This solution is currently not available on other platforms, but very much needed by an Erasmus student abroad. This will

enable not only cost reduction, but above all, continuous access to the platform and teaching materials. So, studying comfortably all the time. Such functionality will effectively solve many problems and allow you to use time efficiently, e.g., when traveling to university and returning.

Therefore, when creating the platform, we should use the latest achievements of new technologies. New technologies for teaching are developing very quickly and are becoming more and more user-friendly. The eMEDIATOR platform should be open to these technological achievements and use them skilfully and moderately. It is not about chasing technological novelties and fireworks but using cost-effective solutions. Importantly, all technologies should be compatible with the platform, maintaining the standards. This will ensure the continued use of already produced teaching materials and their updating. New technologies should also automate the process of information exchange, teaching monitoring and reporting. Artificial intelligence should support the individualization of teaching according to the student's preferences. Currently, new technologies are developing very intensively, and some of them can be used to build an educational platform:

- Artificial Intelligence/ Machine Learning
- Learning Analytics
- Big data
- Virtual Reality /VR, AR, MR/
- Learning Simulations
- Video Streaming
- Interactive Video
- The Cloud
- Hybrid Learning
- Adaptive Learning Platforms
- Automation /Zapier, Webhooks, Triggers/
- App-Based Learning
- FlipGrid
- Hybrid Learning
- Blockchain

The platform should support the didactic process by “intelligently” assuming time-consuming duties. The teacher will then have more time for teaching. New technologies have the potential to create tailor-made learning. The following technologies can be used for this: Artificial Intelligence/Machine Learning, Learning Analytics, Adaptive Learning Platforms, Big data.

The student must feel comfortable despite the intensive work. Moreover, the acquired knowledge and competences should be effectively preserved. It's not about bombarding the student with news and replays. A better solution is to select the content of interest to the student, in accordance with the priorities of the lecturer. The platform should monitor the teaching process within the subject, build optimal teaching models. On this basis, the individualization of teaching should be based, in accordance with the student's time preferences (i.e. taking into account his daily schedule). The student should be skilfully motivated to work. All these tasks should be performed autonomously by the platform. The didactic goal should be achieved by the way, the most important thing is to awaken the student's passion and desire for self-development. Each of us is aware that the teaching content is far behind science which is developing very quickly. Therefore, it is very important to skilfully graduate the complexity of the teaching content so as not to

discourage or discourage the student. It is best to arouse his scientific curiosity when solving research problems that are important, e.g., in business (in cooperation with businesspeople)

Selected tasks for which new technologies in this field can be used:

- Analyzing the educational process on a general and individual level. Preparing and providing students with small portions of information;
- Stimulating the effective repetition of the material in order to consolidate knowledge and improve competences;
- Individualization of teaching according to the student's preferences, in particular taking into account the student's day plan;
- Motivating the student to work, reminding about important deadlines;
- Monitoring student works in terms of their originality - fight against plagiarism;
- Proposing educational content to the student, according to his / her interests;
- Supporting the content management process;
- Support for the platform management process.

New technologies also offer the possibility of being immersed in a digital environment. The following technologies can be used for this: Virtual Reality /VR, AR, MR/, Learning Simulations. It is a completely different environment that affects almost all senses. The digital environment offers opportunities not available in classical education or prohibited (by law, safety regulations). Thanks to this, the student can learn without consequences from his own mistakes. It is not always painless learning, because there are now also interfaces for pain sensing! Learning in a digital environment is extremely engaging for the student.

The potential uses of virtual reality in education are endless. The teacher's creativity is paramount. You can prototype solutions, conduct scientific experiments or create your own digital avatar. You can enter a virtual nuclear reactor and observe the reactions taking place in it. The role of the MEDIATOR will be to combine these technologies. It will also be important to create repositories of tools for the teacher and student. This includes access to software, multimedia materials, databases, and instructions and guides as needed.

Currently, there are many teaching materials available, but with commitment you can create them yourself. It is not about creating complex applications, but a visual representation of the taught content. The following technologies can be used for this: Video Streaming, Interactive Video. After all, a short video can explain the essence of the problem better than long and complex scientific arguments. On the other hand, advanced, dedicated teaching software can be created as part of didactics at IT faculties - implemented as part of team student projects.

The e-Mediator platform should be a didactic HUB that integrates new technologies for education. It is not about all the novelties, but only effective and useful technologies: Automation /Zapier, Webhooks, Triggers/, The Cloud, Hybrid Learning, App-Based Learning, FlipGrid.

There should be an educational software repository with dedicated instructions for use, educational materials, studies, and multimedia materials. In short, everything that is needed in the didactic process. Teachers should have access to user-friendly and proven software so that they can quickly and efficiently prepare didactic materials on their own, or modify previously developed ones.

The eMEDIATOR platform should relate to the University's management system, with the timetable. The platform should enable efficient communication using social media. It should be a place of distribution of educational materials on various types of devices in formats adapted to the recipient's Internet speed. Many of these activities should be automatic so as not to involve the teacher who should focus on teaching.

Motivation and a personal example are very important in teaching. Therefore, close cooperation with business on solving important practical and scientific problems will be very motivating. The eMEDIATOR platform will enable cooperation with the use of an intelligent search engine, which will allow the business to find people with the desired competencies. On the other hand, students will be able to search for a suitable placement, internship and contact the POC in the enterprise. The platform will enable efficient contact between universities and business in projects related to solving current problems in business. Working in these projects will enable students to acquire specialized competences and contact business. It will be a big attribute when looking for a job in the industry after graduation. Businesses interested in acquiring graduates with specific competences will also have an impact on the content of education. This will enable, for example, the implementation of diploma theses on utilitarian problems with the support of business specialists.

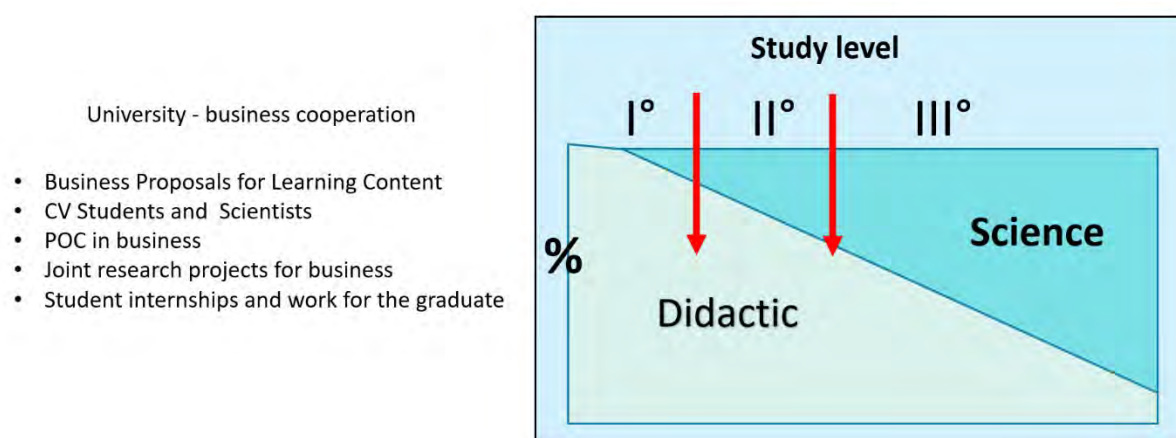


Figure 2.4. A model of cooperation with businessmen at subsequent levels of education at a university.

The last and most important thing is safety. It is about data security as well as data reliability. Data security can be supported by the latest technologies. Teachers should be aware of the dangers, for example, during data transfer, and take appropriate preventive actions beforehand, e.g. make the appropriate software configuration during a video conference. The security of repositories and processed data is important. Under no circumstances can the emitter be a security hole through which hackers can break in. Meeting places in virtual reality should be available only to authorized persons, e.g. with an individual password.

Security in the real world is even more important. Here, the platform will also be helpful. When planning a trip, a student, an Erasmus program participant, must choose the best university for him / her based on reliable information. It is about comprehensive knowledge that allows not only to choose a university, but also to prepare for departure, travel, stay and return.

REFERENCES

- [1] P. L Machemer, P. Crawford, Student perceptions of active learning in a large cross-disciplinary classroom. *Active Learning in Higher Education*, 8(1), 9–30. <https://doi.org/10.1177/1469787407074008>. 2007.

- [2] L. E. Patrick, L. A. Howell, W. Wischusen, Perceptions of active learning between faculty and undergraduates: Differing views among departments. *Journal of STEM Education: Innovations and Research*, 17(3), 55 <https://www.jstem.org/jstem/index.php/JSTEM/article/view/2121/1776>, 2016.
- [3] G. S. Stump, J. Husman, M. Corby, Engineering students' intelligence beliefs and learning. *Journal of Engineering Education*, 103(3), 369–387. <https://doi.org/10.1002/jee.20051>, 2014.
- [4] M. T. Chi, R. Wylie, The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219–243. <https://doi.org/10.1080/00461520.2014.965823>, 2014.
- [5] S. Freeman, S. L. Eddy, M. McDonough, M. K. Smith, N. Okoroafor, H. Jordt, M. P. Wenderoth, Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, 111(23), 8410–8415. <https://doi.org/10.1073/pnas.1319030111>, 2014
- [6] M. Prince, Does active learning work? A review of the research. *Journal of Engineering Education*, 93, 223–232. <https://doi.org/10.1002/j.2168-9830.2004.tb00809.x>, 2014
- [7] C. C. Bonwell, J. A. Eison, Active learning: Creating excitement in the classroom. Washington, DC: Eric Clearinghouse on Higher Education, 1991.
- [8] K.C. von Schönfeld, W. Tan, C. Wilkens, L. Janssen-Jansen, Unpacking social learning in planning: who learns what from whom?, *Urban Research & Practice*, 13:4, 411-433, DOI: 10.1080/17535069.2019.1576216, 2020.
- [9] C. D. Wickens, J. D. Lee, Y. Liu, S. E. G. Becker, An Introduction to Human Factors Engineering. Second ed. Pearson Prentice Hall, Upper Saddle River, NJ, 2004.
- [10] P. Haazebroek, B. Hommel, Towards a Computational Model of Perception and Action in Human-Computer Interaction. International Conference on Digital Human Modeling. Springer, Berlin, Heidelberg, 2009, 247-256.
- [11] K. Illeris, The Three Dimensions of Learning: Contemporary Learning Theory in the Tension. Field between the Cognitive, the Emotional and the Social. Roskilde University Press/Leicester, Roskilde, Denmark, 2002.
- [12] M. Gawlik-Kobylińska, Cztery wymiary uczenia się w projektowaniu scenariusza kursu e-learningowego [The Four Dimensions of Learning in e-Learning Course Design], *Scientific Quarterly of the National Defence University* 105 (2016), 39-52.
- [13] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, D. R. Krathwohl. Taxonomy of Educational Objectives, Handbook I: The Cognitive Domain. David McKay Co Inc, New York, 1956.
- [14] A. J. Romiszowski, Psychomotor principles. In M. Fleming and W. H. Levie (Eds.), Instructional message design. Educational Technology Publications, Englewood Cliffs, 1993.
- [15] R. H. Dave, Psychomotor Levels in Developing and Writing Behavioral Objectives, Educational Innovators Press, Arizona, 1970.
- [16] E. J. Simpson, The Classification of Educational Objectives in the Psychomotor Domain. Gryphon House, Washington DC, 1972.
- [17] A. Harrow, A Taxonomy of Psychomotor Domain: A Guide for Developing Behavioral Objectives. David McKay, New York, 1972.
- [18] Gawlik-Kobylińska, M. (2018, December). The Four-Dimensional Instructional Design Approach in the Perspective of Human-Computer Interactions. In Petkov, N., Strisciuglio, N., & Travieso-González, C. M. (Eds.). (2018). Applications of Intelligent Systems: Proceedings of the 1st International APPIS Conference 2018 (Vol. 310). IOS Press. 146--156. DOI= <http://dx.doi.org/10.3233/978-1-61499-929-4-146>.

- [19] T. Mayes, S. de Freitas, Learning and e-learning. Rethinking pedagogy for a digital age, Taylor & Francis Group, Routledge, 2007, 13-25.
- [20] P. Alahuhta, E. Nordbäck, A. Sivunen, T. Surakka, Fostering team creativity in virtual worlds. Journal For Virtual Worlds Research, 7(2014).
- [21] J. Lebień, M. Szwoch, Virtual Sightseeing in Immersive 3D Visualization Lab. Proceedings of the Federated Conference on Computer Science and Information Systems (4th Conference on Multimedia, Interaction, Design and Innovation), Gdańsk 2016, Annals of Computer Science and Information Systems, Vol. 8, 2016, pp. 1641–1645.
- [22] L. Dawley, C. Dede, C. Situated learning in virtual worlds and immersive simulations. In Handbook of research on educational communications and technology, Springer, New York, NY, 2014, 723-734.
- [23] S. Warburton, Second Life in higher education: Assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. British Journal of Educational Technology, 40(2009), 414–426.
- [24] L. Ferreri, L. Verga, Benefits of Music on Verbal Learning and Memory: How and When Does It Work? Music Perception: An Interdisciplinary Journal, 34(2016), 167-182.
- [25] J. K. Kim, The Effects of Teaching Image, Communal Sense, and Class Environment on Academic Procrastination in a University E-Learning Setting. International Information Institute (Tokyo). Information, 20(2017), 55.
- [26] L. Painter, Best VR accessories of 2018 [02 Jan, 2018], Tech Advisor, <https://www.techadvisor.co.uk/buying-advice/game/best-vr-accessories-of-2018-3643903/> (30.03.2018).
- [27] B. Tarr, M. Slater, E. Cohen, Synchrony and social connection in immersive Virtual Reality. Scientific reports, 8(2018), 3693.
- [28] F. Herpich, G. B. Voss, F. B. Nunes, R. R. Jardim, R. D. Medina, R. D., Immersive virtual environment and artificial intelligence: A proposal of context-aware virtual environment. In UBICOMM 2014: The Eighth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 68-71.

BIBLIOGRAPHY

- W. C. Allen, Overview and evolution of the ADDIE training system. Advances in Developing Human Resources, 8(2006), 430-441.
- M. Bettoni, E. Obeng, W. Bernhard, N. Bittel, V. Mirata, The Importance of Space in Knowledge Sharing Online: The QUBE Approach. 18th European Conference on Knowledge Management (ECKM 2017). Academic Conferences and publishing limited, 2017, 122.
- P. Hardré, The motivating opportunities model for Performance SUCCESS: Design, Development, and Instructional Implications. Performance Improvement Quarterly, 22 (2009). doi:10.1002/piq.20043.
- T. Hewett, R. Baecker, S. Card, T. Carey, J. Gasen, M. Mantei, G. Perlman, G. Strong, W. Verplank, ACM SIGCHI Curricula for Human-Computer Interaction, 2004. Available at: <http://sigchi.org/cdg/cdg2.html>
- ILIAS Platform of the War Studies University
https://ilias.akademia.mil.pl/ilias.php?ref_id=21933&cmdClass=ilplistofoobjectsgui&cmd=showObjectSummary&cmdNode=gg:ny:74:79&baseClass=ilrepositorygui#after_sub_tabs (20.07.2018)
- J. M. Keller, Development and use of the ARCS model of motivational design. Journal of Instructional Development, 10(1987), 2-10.

E. L. C. Law, M. Hassenzahl, E. Karapanos, M. Obrist, V. Roto, Tracing links between UX frameworks and design practices: dual carriageway. Proceedings of HCI Korea, Hanbit Media, Inc., 2014, 188-195.

M. K. O'Malley, Principles of human-machine interfaces and interactions. In Life Science Automation: Fundamentals

O. E. Osuagwu, Learning objects: The Nerve centre of learning content management systems (LCMS) for e-learning in the WWW, Journal of Mathematics and Technology, 1 (2010), 109–125.

S. Porter, Human Cognition and Aesthetic Design in Pedagogy and Online Learning. Georgia International Conference on Information Literacy 57 (2016),

<https://digitalcommons.georgiasouthern.edu/gaintlit/2016/2016/57>.

R. A. Reiser, J. Dempsey, Trends and issues in instructional design and technology. Second ed. Pearson, Upper Saddle River, 2007

A. P. Rovai, M. J. Weighting, J. D. Baker, L. D. Grooms. Development of an instrument to measure perceived cognitive, affective, and psychomotor learning in traditional and virtual classroom higher education settings. The Internet and Higher Education, 12 (2009), 7-13.

N. M. Seel, T. Lehmann, P. Blumschein, O. A. Podolsky, Instructional Design for Learning Theoretical Foundations. Sense Publishers, Rotterdam, 2017.

LIST OF AUTHORS

1. Małgorzata Gawlik-Kobylińska
2. Paweł Maciejewski
3. Marcin Rojek
4. Joanna Leek

3 . A3.3. Development of the model of academic and non-academic resources (TTI)

3.1 INTRODUCTION

The effective functioning of the developed model of the educational ecosystem is possible only through the use of open information resources.

Open educational resources (OER) as a term have been actively used since its first use at the UNESCO conference in 2002. As a rule, the term open educational resources means the free use of academic and scientific information, including its distribution and editing (D'Antoni, 2009; Hilton, Wiley, Stein, & Johnson, 2009; Plotkin, 2010; Wiley, 2009).

UNESCO (2007, 2010) describes the concept of open educational resources as all kinds of educational materials that can be used by users without payment in any form (Butcher, 2011). Open educational resources have received close attention in connection with the development of distance learning. Many researchers and international organizations note the significant positive impact of open educational resources on the methodology and content of education systems (UNESCO, 2010).

Margulis (2005) developed the open education resources for open courses withi the frame of MIT education activities. The model includes content, instruments and distribution resources.

Anderson (2009) noted that for the effective use of open educational resources, it is necessary to combine the various components of the pedagogical process, both the main ones and the auxiliary ones that provide them.

Conol and Alevisou (2010) described various resource models used in distance learning and the theoretical approaches underlying their application. Diallo, Wangeci, and Wright (2012) described an environment of the model that combines the possibilities of joint learning activities (creation, management, distribution) using open educational resources. This approach is based on the needs of competence development, regardless of the regional and cultural characteristics of the trainees.

The use of OER leads to a rethinking of the content of university education, to the technology and methodology for developing information materials for courses, their distribution and reuse (Conole & Culver, 2009; Lane & McAndrew 2010, McAndrew, 2010). Lane and McAndrew (2010) explore the systemic nature of open educational resources and how they can be applied to hands-on learning.

The paper (Khanna and Basak, 2013) proposes an OER architecture that can combine the main components of the practical use of distance learning methods in educational organizations of various types. The main element of the

proposed architecture is a web portal containing a set of databases, knowledge bases and other repositories, which together simplify the provision of information during the educational process as part of distance learning.

3.2 OER ARCHITECTURE FRAMEWORK

An analysis of well-known works devoted to the study of OER, as well as taking into account the target tasks of the portal, allows us to formulate the general architecture framework of the OED model.

The structure of architecture (Fig. 3.1) contains vertical components: academic, pedagogical, managerial, financial, technological and ethical (Khan, 2001). These vertical components are structured into horizontal levels: management support systems, IT infrastructure and services, online teaching and learning, open content development and maintenance, and student assessment and evaluation.

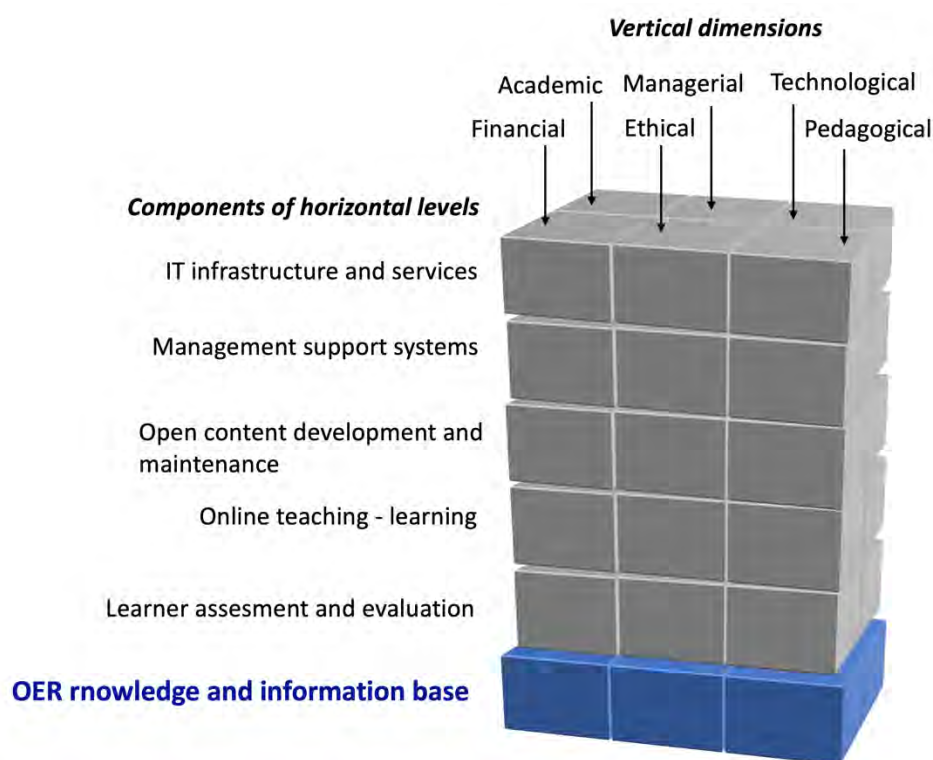


Figure 3.1: OER architecture framework

The indicated basic architecture of interconnected components of open resources can be used as the basis for the demo portal being developed.

The content of the main components of the architecture is shown in Table 3.1.

Table 3.1: The content of the main components of the framework

Dimensions of framework	Content of dimension
Pedagogical	The pedagogical dimension of the OER framework refers mainly to OER-based open initiatives such as open online courses, openly public teaching and learning, open study groups, and so on. This dimension pertains to issues concerning open initiatives such as content analysis, goal analysis, design approach, organisational methods, and ODL strategies.
Technological	The technological dimension of the OER framework refers to the technological infrastructure of distance learning environments, including issues such as infrastructure planning, designing hardware and software, and technical design for learning programmes, such as page and site design, content design, animation, multimedia, navigation, and usability testing.
Managerial	The managerial aspect of the OER framework is mainly concerned with the planning and management of administration and other educational activities required for fostering a distance learning environment. It also includes the execution of open management policies along with opening up access to the institution's managerial information and data. Further to this, management involves framing educational policies and decision making with regard to development and maintenance of a good learning environment, delivery of quality distance education, distribution of associated information, and so on.
Academic	The academic dimension of the framework involves the creation and use of online courses and programmes and teaching and learning materials and techniques, including open text books and SLMs (self-learning materials), for the benefit of students.
Financial	The economic/ financial dimension mainly refers to budgeting (i.e., management and availability of finances for the development, implementation, and maintenance of educational systems based on OER). It would also involve developing a sustainable and cost-effective business model, educational systems, and processes for the associated OER initiatives.
Ethical	The ethical considerations of open and distance learning relate to geographical diversity, learner diversity, legal issues (such as licensing), and information accessibility as related to the general institutional information.

The key component areas and support services of the framework is shown in Table 3.2.

Table 3.2: The key component areas and support services of the framework

Key component areas	Support services
IT infrastructure	An appropriate OER-based IT infrastructure is required for the proper operation and management of the concerned ODLI. The IT infrastructure should further help in proper dissemination, sharing, and utilization of OER so as to provide good quality course/programme content, e-content, instructional processes, web-based systems, and others. Using an appropriate IT infrastructure, several systems and solutions are to be developed and made available by the relevant services deployment and delivery departments. So, OER-supportive IT infrastructure services would involve open source education technologies that support the development of tools, techniques, and processes for the distance education system, including the creation of online systems to provide support for online learners. As such the IT infrastructure would involve open source software (OSS) and their applications, Internet, open web, online systems, learning management systems (LMS), and others.
Management support systems	The management support system (MSS) works mainly through the use of FOSS for coordinating and managing the various functions and activities pertaining to the DES in India. The implementation of an OER approach by the MSS leads to adopting the open pedagogies that would leverage open source educational technologies, online instructional design, and open practices of teaching and learning. Thus considering the above mentioned functions the MSS would plan to establish an OER-based organisational system in the DES of India which would streamline its functional activities in such a way that the OER materials would be developed, shared, and adapted efficiently and effectively.
Open content development and maintenance	One of the main tasks to be performed under this OER architecture is to make appropriate arrangements for the development of open content of reasonably good quality. Such open content would mainly be used for teaching-learning purposes and include materials related to courses and course components such as learning objects, teaching content, manuals related to practical labs, textbooks, and so on. It may also include production of audio, video, and animated educational programmes. In addition to this, it may develop new OER and also make use of existing OER (available locally or globally) to subsequently create complete online credit-based academic programmes. Each academic programme would be in the form of a suite of several interconnected applications, structured and presented in a wiki system (Srivathsan, 2009). The OERbased open content is to be developed in two groups, namely an open course guide (OCG) and an open program guide (OPG). The overall TEL components for a programme to be developed would broadly include the following: an open program guide, an open course guide, a wiki, a discussion forum, querying services, SMS, mlearning services, and so on. However, it is suggested that every academic programme of an ODLI may establish the OPG and associated OCG(s) for each course.
Open (online /public) teaching and learning	Using OERs, educators can undertake the work of teaching by using blogs and wikis. The course syllabus, modules, activities, and assignments would be publicly visible. Students can be officially enrolled in any course and their learning would be viewable by the public. An efficient OER-based TEL system of open e-learning is required to be established so as to provide support for all educational programs/courses offered by the concerned ODLI. Every ODLI would be required to set up various technology systems, such as Web sites, Internet, TV channels, EduSat, IP-TV, mobile, and community radio, so as to have an efficient DES which can help distance learning students. Such a TEL system is also required to support the delivery of various programs/courses offered.
Learner assessment and evaluation	Learner assessment is involved in certifying the academic level of performance achieved by the students in a particular course of study. A course evaluation system (CES) would be developed for learner assessment through the use of OCG/OPG. It would also support a testing and evaluation system along with the associated learner database management system. Arrangement for practical, design exercise, and term papers, and timely evaluation of answer scripts and term papers would be ensured.

3.3 KNOWLEDGE AND INFORMATION BASE

As a basic development strategy, the following stages of creating the components of the knowledge base and other open information can be proposed:

- A self-developing dynamic collection of national and international information resources on all components is created.
- A database of open information on pedagogy is being created.
- A database of open sources of research activity, correlated with academic topics, is being created.
- A database of experts from the business community is being created who are interested in disseminating their information and who can create open academic content.
- A repository of academic learning resources of learning objects is created.

All open information resources placed in the repository are accompanied by a lot of metadata containing additional information (copyrights, licenses, source location of the resource, etc.), which helps to increase the speed and efficiency of searching for the necessary information on the portal.

The open resources of the portal may include all types of information resources: text and audiovisual materials, specialized interactive learning tools, webinars, teleconferences, and others.

The main advantages of using OER within the framework of the developed model of the educational ecosystem are as follows:

1. Accessible educational content allows it to be used by all categories of users, while the pedagogical component can focus on developing the most appropriate teaching methods, and not on creating the actual content.
2. It is advisable to use an open license as a copyright for open educational resources developed as part of the development of the portal. This will expand the audience of users and democratize the possibility of a more active and wider use of the content of the portal.
3. The use of full information resources under an open license provides several advantages in ensuring an effective educational process:
 - Trainees save time searching for complete (rather than annotated) and most informative information resources, including multimedia. This does not require scanning and searching through the entire information space of the Internet but limits it to the already selected and most effective information materials of the portal itself.
 - Students can share and collaborate on the same resources, creating a better platform for cooperative forms of learning.
 - Learners can receive detailed ranked information with previous user ratings before choosing courses and using information resources.
 - In the process of use, students can form feedback with requests for additional information or an indication of the aging of the content of certain types of information, which ensures their self-renewing updating.

3.4 CONCLUSION

The proposed structure of open resources will ensure the standardization, efficiency and effectiveness of the educational process, creating a single information platform for a long-term self-renewing process of functioning of all the basic components of the educational ecosystem model being developed and the demo portal implementing it.

REFERENCES

1. Anderson, T. (2009). Are we ready for open educational resources? Presentation at the 23rd ICDE World Conference, Maastricht, The Netherlands.
2. Butcher, N., Kanwar, A., & Uvalic ´-Trumbic, S. (2011). A basic guide to open educational resources (OER). Vancouver, Canada: Commonwealth of Learning, and Paris, France: UNESCO.
3. Conole G., & Alevizou P. (2010). A literature review of the use of Web 2.0 tools in higher education. Open University, UK.
4. Conole G., & Culver J. (2009). Cloudworks: Social networking for learning design.
5. D'Antoni S. (2009): Open educational resources: Reviewing initiatives and issues. Open Learning: The Journal of Open, Distance and e-Learning, 24(1), 3-10.
6. Diallo, B., Wangeci, C., & Wright, C. R. (2012). Approaches to the production and use of OERs: The African Virtual University experience. In R. McGreal, W. Kinuthia, & S. Marshall (Eds), Open educational resources (working title). Athabasca, Canada: AU Press.
7. Hilton, J., III, Wiley, D., Stein, J., & Johnson, A. (2009). The four R's of openness and ALMS analysis: Frameworks for open educational resources
8. Hylén, J. (2006). Open educational resources – Opportunities and challenges Paris, France: OECD-CERI.
9. Khan, B. H. (2001). A framework for e-learning. Advanstar Communications.
10. Khanna P., & Basak P.C. (2011). An integrated web-based information system for open and distance learning institutions in India. The Journal of Information Technology Impact (JITI), 11(2), 153-168.
11. P. Khanna and P. C. Basak. (2013) An OER Architecture Framework : Needs and Design. The International Review of Research in Open and Distance Learning. Vol. 14, N 1, 65-83.
12. Lane A., & McAndrew, P. (2010). Are open educational resources systematic or systemic change agents for teaching practice? British Journal of Educational Technology, 41(6), 952-962.
13. Margulies, A. (2005). MIT Opencourseware – A new model for open sharing.

14. Plotkin, H. (2010). Free to learn: An open educational resources policy development guidebook for community college governance officials. San Francisco: Creative Commons.
15. UNESCO (2002). Forum on the impact of open courseware for higher education in developing countries: Final report.
16. UNESCO (2007). Education for all by 2015. Will we make it? Paris: UNESCO Publishing & Oxford University Press.
17. UNESCO (2010). Global trends in the development and use of open educational resources to reform educational practises (Policy Brief). Moscow: UNESCO Institute for Information Technologies in Education.
18. Wiley,D. (2009). Defining open. Open <http://opencontent.org/blog/archives/1123>
19. Wright, C. R., & Reju S. A. (2012). Developing and deploying OERs in sub-Saharan Africa: Building on the present. The International Review of Research in Open and Distance Learning, 13(2).

LIST OF AUTHORS

1. Boris Misnevs
2. Igor Kabashkin
3. Olga Zervina

4 . A3.4 Development of the model for job application support. (UMU)

This section presents a basic mechanism to support the accuracy of recruitment processes. A proposal to facilitate the match between the competences obtained by work candidates during their academic stage and the competences demanded in a job position is described. In addition, a validation of a competence model built on the basis of the harmonisation of competence standards along with skill and knowledge models is carried out.

4.1 CURRENT SCENARIO

The job search can be a complicated experience. On the one hand, recruiting companies have to deal with heavy workloads, sifting through CVs. On the other hand, the candidate has to go through the uncertainty of knowing whether the profile will be a good match for the job. The digitalisation of recruitment has brought new opportunities for both sides. In the first, companies can search for specific profiles and filter automatically candidates. In the second, workers can apply for jobs all over the world.

Interaction between employee and employer is at hand. However, there are still shortcomings when it comes to selecting the most suitable candidate. As an example, soft skills considered on a paper without actually being demonstrated in a real scenario can lead to inaccuracies in the recruitment process. Consequently, job application actions should be improved by having precise recruitment methods in place (Gürtzgen, Lochner, Pohlen, & van den Berg, 2021).

After an overall reduction in employment during the months of greatest impact of the pandemic, there has been a recovery in the number of digital job vacancies published by European countries since 2021. By the end of 2021 in Spain, the total number of digital job offers had grown by 21% compared to 2020, and this figure was 29% lower compared to 2019¹. This trend shows a recovery of the situation.

The European Union (EU) has promoted initiatives to strengthen the labour market. In particular, the European Employment Strategy focuses on four key domains: boosting labour demand, improving competences to reduce gaps in education systems, improving the functioning of labour markets and providing equal opportunities for all². It is on the objective 3 where the focus of this work lies.

¹ <https://www.jobmarketinsights.com/jmi-area/Landing>

² <https://ec.europa.eu/social/main.jsp?catId=101&intPageId=3427>

The aforementioned strategy is succeeding in reducing the rate of early school leavers. According to data collected by Eurostat, in 2021 this rate has decreased for both males and females. In particular, the overall rate fell to 13.3% in Spain. In addition, the decrease was more pronounced among women with 9.7% compared to 16.7% among men. In the EU this rate stood at 10.0% by 2020³. To make this happen, prevention, intervention and compensation measures have been followed. Prevention measures address structural issues that could lead to early abandonment. On the other hand, intervention measures tackle the issues students face by enhancing the quality of instruction and offering tailored support. Finally, compensation measures give opportunities to get a degree for those who prematurely quit their studies (Network & Training, 2014). These initiatives are important in order to make the labour market more effective (Gürtzgen, Diegmann, Pohlen, & van den Berg, 2021).

The EURES (EUROpean Employment Services) network is another example on improving the labour market by the EU. Its purpose is the cooperation with the State Public Employment Services of all the member countries. In addition, the network is responsible for providing citizens with information, counselling and recruitment services. The EURES job portal has approximately 4 million vacancies as of October 2022. Apart from that, there are nowadays a multitude of online platforms specialised in job search. Indeed and Monster are considered the best overall⁴.

Traditionally, study selection has been based on students' academic preferences (Subu et al., 2022). However, the future students are not really aware of what studying a degree entails in terms of competences (Sakamoto, 2022). The degrees courses offered by universities vary from one to another, but the training should respond to the employment curiosities of university graduates. It is therefore desirable that academic courses can be constructed according to the competences that students wish to acquire (Gleason et al., 2021). In this way, a double objective could be achieved. On the one hand, students may be more satisfied with the training they have received, especially when applying for a job that suits their tastes. On the other hand, there may be greater accuracy when companies carry out recruitment events, which could result in the satisfaction of employers at the same time by observing greater coverage of the skills demanded in the job offer (Thomas et al., 2022).

The work in this manuscript consists of the following sections. Section 5.2 describes the validation of the model with the steps carried out and the proposed refinements to the competence model. Section 5.3 sets out a basic taxonomy that could be used to categorise competences and make more effective use of the model. Section 5.4 details the proposed mechanism for finding out the percentage coverage of a job offer against a syllabus in terms of competences. Finally, Section 5.5 outlines some conclusions and future work.

4.2 MODEL VALIDATION

³ <https://umubox.um.es/index.php/s/iWXunnVm4VlzudM>

⁴ <https://www.thebalancemoney.com/top-best-job-websites-2064080>

As aforementioned, the starting point is a competence model based on the competence standards IEEE 1484.20.1 (IEEE, 2008) and CWA 16655-1 (European Committee for Standardization, 2013), the Skills Framework for the Information Age (SFIA) model (available at sfia-online.org) and the Technological Pedagogical Content Knowledge (TPACK) model (available at tpack.org). The model is divided into 4 dimensions based on those proposed in the European Digital Competences Framework (DigComp) (Carretero, Vuorikari, & Punie, 2017). The dimensions of the model are the following: dimension 1, context/areas of competence; dimension 2, competence descriptors and titles; dimension 3, competence levels; and finally, dimension 4, knowledge and skills. This model was built with the purpose of representing competences in e-Learning digital systems, enabling the easy distribution of competences by electronic means. **Ошибка! Источник ссылки не найден.** shows the attributes of each one of the dimensions in the model.

Table 4.1 Simplified version of the competence model

DIMENSION 1	Degree name		Scale
	▪ Certificate level		▪ Max scale
	▪ Know-Skill pair		▪ Min scale
	▪ Keywords		▪ Scale threshold
DIMENSION 2	▪ Competence identifier		▪ Years of experience
	▪ Title or name	DIMENSION 4	▪ Skill name
	▪ Description		▪ Code
	▪ Definitions		▪ Description
	▪ Extra identifier		▪ Autonomy
	▪ Abbreviations		▪ Business skills
	▪ Date of creation		▪ Influence
	▪ Date of modification		▪ Complexity
	▪ Validation start date		▪ Knowledge
	▪ Date of issue		▪ Skill level name
	▪ Author		▪ Skill level number
	▪ Topics		▪ Cognitive soft skill
	▪ Credits		▪ Affective soft skill
	▪ Level		▪ Psychomotor soft skill
	▪ Version		▪ Role-based hard skill
	▪ Explicit metadata		▪ Skill-based skill
DIMENSION 3	▪ Level		▪ Technology Knowledge
	▪ Max level		▪ Pedagogy Knowledge
	▪ Min level		▪ Content Knowledge
	▪ Level threshold		▪ Combination

After a general analysis of the main sources of competences, it was considered appropriate to carry out a validation of the model. The aim of this task was multiple. On the one hand, the attributes proposed in the model were checked to see whether they are sufficient to represent a competence. On the other hand, it showed whether

there are unnecessary or repeated attributes. Finally, it allowed to propose improvements that can lead to a more effective management of the competences in the model.

The validation of the competence model consisted of the following steps. Firstly, syllabus in the field of Project Management and Software Engineering were taken. In addition, information was collected from job offers. From both sources, competences were obtained for analysis (step 1). On the other hand, a basic competence management mechanism was proposed based on the handling of Knowledge-Skill pairs (attribute Know-Skill pair at **Ошибка! Источник ссылки не найден.**) along with the generation of labels (attribute Competence identifier at **Ошибка! Источник ссылки не найден.**) for the identification of competences (step 2). Finally, the textual descriptions of competences were transferred to the model, together with the Know-Skill pairs and the labels (step 3). Figure 4.1 depicts the iterativity of these steps.

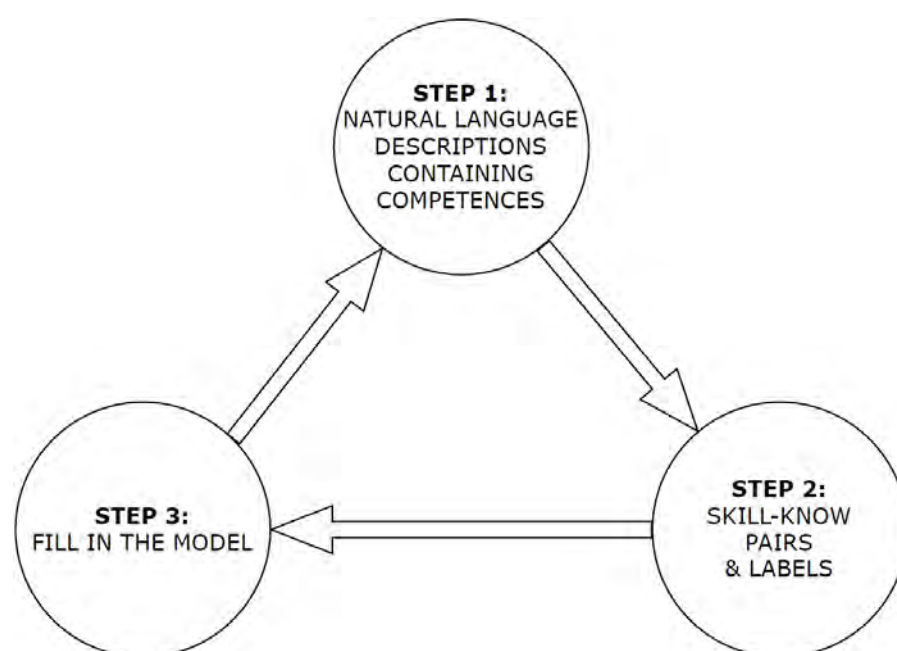


Figure 4.1. Steps for the validation of the compente model

4.2.1 VALIDATION STEPS

The steps mentioned in the previous section are explained in more detail below. These steps apply to existing competence sources such as syllabus and job offers. In addition, they can be applied to textual descriptions of competences made during the completion of the model.

STEP #1:

In the first step it was necessary to evaluate sentence by sentence whether the textual descriptions contained in the syllabi and in the job offers had academic competences. This task could be performed automatically or semi-

automatically with the assistance of a user. In particular, sentence separation is a basic task that does not present any difficulty. However, it is in the detection of competences in sentences that complications may arise.

In section 5.1 Job Skill Fundamentals of the manuscript 5.A2.5.2. Development of model for competence-based job skill building (UM) it was stated that competence is a sum of a definition of skill and knowledge (*Competence = Skill + Knowledge*). In fact, it is common to find in the definition of competences both components as a block. Some examples of competences defend this result: "Building functional and technical requirements", "Mentoring other engineers in Software Engineering", etc. In both cases the verbs, to build and to mentor, are naming skills, which are practical qualities that people acquire over time through practice, commitment and exposure to others (Frezza et al., 2018). On the other hand, the remaining words, "functional and technical requirements" and Software Engineering, define concepts of knowledge, which is related to the mastery of fundamental ideas and subjects in order to achieve cognitive or intellectual qualities (Frezza et al., 2018). This basic idea could be used as a guide to detect competences in sentences.

Job summary

We are a passionate team working to build a best-in-class healthcare product designed to make high-quality healthcare easy to access. We are looking for a Software Development Engineer who can build new engineering solutions with a motivated team that will delight our customers. The problem space is well-defined but the solution space is not and strong technical acumen and ability to spot problems at architecture stage would be important. This role requires deep technical expertise, and gives you the opportunity to engineer systems and build reliable and secure services for healthcare. You have an eye towards quality, and insist on the highest standards. You are motivated to tackle ambiguous situations with technologies to rapidly solve problems. You are a team player, contribute to the execution of team objectives, and leverage technical knowledge and engineering best practices to rapidly deliver solutions that have a broad business impact.

Key job responsibilities

We are looking for software developers with expertise and passion for building large scale distributed systems and services. In this role, you will have responsibility for:

- Building functional and technical requirements into detailed architecture and design
- Coding and testing complex system components
- Participating in code and design reviews to maintain our high development standards
- Overall system architecture, scalability, reliability, and performance
- Mentoring other engineers, defining our challenging technical culture, and helping to build a new fast-growing team

About the team

Amazon's mission is to make it dramatically easier for customers to access the healthcare products and services they need to get and stay healthy. Towards this mission, we (Health Storefront and Shared Tech) are building the technology, products and services, that help customers find, buy, and engage with the healthcare solutions they need.

BASIC QUALIFICATIONS

- 1+ years of experience contributing to the system design or architecture (architecture, design patterns, reliability and scaling) of new and current systems.
- 2+ years of non-internship professional software development experience
- Programming experience with at least one software programming language.

PREFERRED QUALIFICATIONS

- * 3+ years developing software solutions
- * Understanding software design principles and computer science fundamentals
- * Experience designing and implementing RESTful APIs at scale
- * Curiosity and drive to learn new technologies and business domains
- * Prior experience building services on a cloud platform (AWS or otherwise)
- * Prior experience working for a consumer-facing technology company
- * Strong knowledge of data structures, algorithms, distributed systems, and asynchronous architectures
- * Track record of mentoring other SDEs

Figure 2 Sentences and competences extraction from a job offer

Competences can be worded in very different ways depending on the author's style. It may be the case that the meanings of two competences are very similar, despite being worded differently. The use of synonyms generates

this kind of situation, "Teaching Software Engineering" vs "Explaining Software Engineering". The extraction of the concept of each competence is fundamental so that they can be distinguished with precision from each other. The use of Natural Language Processing (NLP) techniques could allow this step to be automatic or semi-automatic as well (see **Ошибка! Источник ссылки не найден.**). These techniques offer computationally efficient mechanisms that facilitate human-machine communication by means of natural language. The models being used in this technologies emphasise human cognitive processes for language comprehension. Moreover, these techniques include machine learning algorithms for language processing, which represents a significant advance in this field that can be exploited in this type of repetitive tasks. Some of the most common NLP techniques that could be used are showed in Figure 4.3 **Natural language processing techniques**

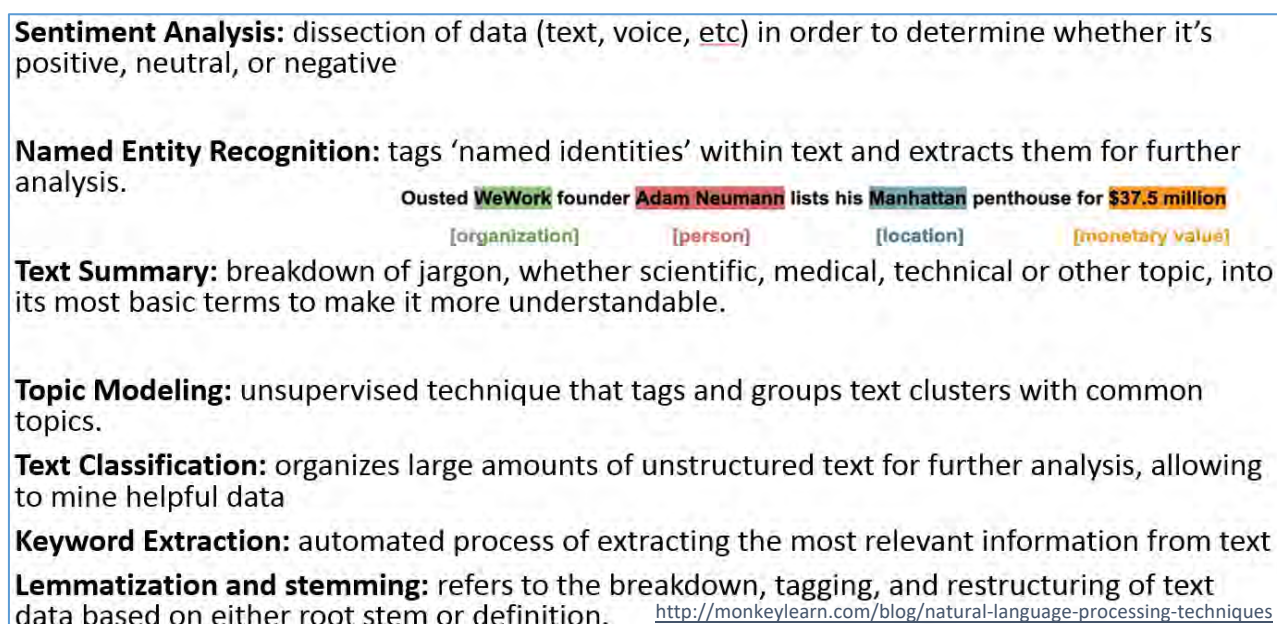


Figure 4.3 Natural language processing techniques

STEP #2:

The word pairs Knowledge-Skill were then constructed for management purposes. These word pairs have the quality of representing in natural language the concept of competence in a very summarised form. In this task the ideas aforementioned in *step 1* were taken into account. These are, the verb defines the skill and the noun the knowledge. When selecting the noun to represent the knowledge it is important to consider the word that has the most communicative power. If necessary, two words can be used instead of one to be more specific in defining the skill and the knowledge. In case a competence has no skill or knowledge, three dots are written in its place. A slash separates the two components.

Storing the concept of competence and associating it with Knowledge-Skill pairs is important for effective competence management. This technical requirement could be covered by the NLP techniques shown in Figure 2. Additionally, a competence labelling mechanism was proposed to strengthen competence management. The labelling is built with the first letter of the skill and the first letter of the knowledge. In addition, an incremental number is added, and its value depends on whether both letters have been used previously. When a competence has only skill or knowledge, a letter is used in the label along with the incremental number. Some examples of *step 2* can be seen in Figure 4.2.

STEP #3:

The information gathered in the previous steps was used to fill in the competence model (see **Ошибка! Источник ссылки не найден.**). This step made it possible to detect the strengths and weaknesses. In particular, in Dimension 1, related to the context/area of competence, the Know-Skill pair attribute was filled in. All other attributes remained empty in most cases. In Dimension 2, with attributes to define a competence, the following attributes were mainly filled in: competence identifier, with the label generated in the previous step; description, with the whole textual description of the competence; and in cases where sufficient information was provided in the syllabus or in the job offers: Extra identifier, Author, Topics, Credits, Level. The subject's code was added in the Extra identifier. In the Author attribute, the professor's or the company's name were filled in. In Topics, related subjects that appeared in the syllabus were entered. In the variable Credits the number of subjects' hours was collected, and finally, in level, whether the subject was a first, middle or last year subject.

Job summary
 Build a best-in-class healthcare product **Build-Product (bp1)**
 Make high-quality healthcare easy to access **MakeEasy-HealthcareAccess (mh1)**
 Software Development Engineer who can build new engineering solutions with a motivated team-Engineer (e1) **Build-Solution (bs1)**
 Delight our customers **Delight-Customer (dc1)**
 Strong technical acumen-Acumen (a1)
 Ability to spot problems at architecture stage **Spot-Problem (sp1)**
 Deep technical expertise-Expertise (e2)
 engineer (verbo) systems **Engineer-System (es1)**
 Build reliable and secure services for healthcare **Build-Service (bs2)**
 Have an eye towards quality **HaveEye-Quality (hq1)**
 Insist on the highest standards **Insist-Standard (is1)**
 You are motivated to tackle ambiguous situations with technologies to rapidly solve problems **Tackle-Situation (ts1)**
 You are a team player-Player (p1)
 contribute to the execution of team objectives **Contribute-Team (ct1)**
 leverage technical knowledge **Leverage-Knowledge (lk1)**
 Engineering (verbo) best practices **Engineer-Practice (ep1)**
 Rapidly deliver solutions **Deliver-Solution (ds1)**
 Have a broad business impact **HaveImpact-Business (hb1)**

Figure 4.2 Some examples of Knowledge-Skill pairs along with their labels

In Dimension 3, related to the level of competence, the attributes that were mainly filled in were the following, Level and Years. Finally, the attributes of Dimension 4 were the most frequently used. In Code, the labels of the

competences with skills were written. The variable Description stored all the text of the competence with skill. All competences with skill(s) were copied back into the attributes Autonomy, Business skills, Influence, Complexity, Cognitive, Affective, Psychomotor, Role-based, Skill-based depending on their meaning. The knowledge component and the competences with only this component were split between the attributes Technology, Pedagogy, Content and Combination.

4.2.2 MODEL REFINEMENT

After carrying out the validation process, some conclusions were obtained in order to refine the competence model. The variable Level is repeated in both Dimension 1 and 2. Since Dimension 2 is specific to the experience and level of the competence, it was proposed to remove this attribute from Dimension 1. On the other hand, it is common to find in job vacancies competences where years of experience are required. This information fits with the meaning of the variable Years in Dimension 3. However, sometimes the number of years is not implied but the importance of having a specific competence is emphasised. Some examples show this idea: "programming experience with at least one software programming language", "experience designing and implementing RESTful APIs at scale"... It was proposed to add the attribute General at the same level as Years so that this type of situations can be covered in the model.

The Dimension 4 variables Skill name, Skill code and Description were considered unnecessary as this information was stored in the variables of Dimension 2 Title or name, Competence identifier and Description respectively. Moreover, the variable Knowledge of Dimension 4 was also considered dispensable as the attributes Technology, Pedagogy, and Content from the TPACK model were proposed to store this information. Finally, the variables Skill level name and Skill level number were not considered necessary because they do not contribute anything new with respect to the attributes of Dimension 3.

Another change was the order of the components of the Knowledge-Skill pairs in Dimension 2. In order to make the meaning of competence more evident, it was decided to put skill first and then knowledge. This change is due to the fact that this the order in which the components of competence appear. Finally, *step 3* of completing the model would therefore be summarised in the following lines.

1. DIMENSION 1 > Skill-Know pair
2. DIMENSION 2 > Competence identifier < Labels
3. DIMENSION 2 > Description < Whole description of the competence
4. DIMENSION 3 > Experience level > Years < Competences in which a period of time is indicated
5. DIMENSION 3 > Experience level > General < Competences in which experience is required but no period of time is indicated
6. DIMENSION 4 > Autonomy
> Business skills

- > Influence
- > Complexity
- > Cognitive
- > Affective
- > Psychomotor
- > Role based
- > Skill based < Distribute competences with skills among these attributes according to their significance

7. DIMENSION 4 > Technology

- > Pedagogy
- > Content
- > Combination < Distribute the competences with Knowledge among these attributes according to their significance

Although not detailed in this section, many of the remaining attributes could be automatically populated with information from the system or from the user's session. Some examples could be the author, which would be the name of the user, the creation, modification and validation dates of a competence, which could be taken from the system depending on when the events to which these variables refer occur. Another example are the levels of competences which could be the same when entered into the system for a same subject. It is left to the discretion of the developers as to the most convenient way for users to complete the rest of the attributes. In any case, all of them are considered relevant to cover competences in any academic field. **Ошибка! Источник ссылки не найден.** shows the changes in the model.

Table 4.2 Competence model with the changes proposed (in grey to remove, in green to change)

DIMENSION 1	Degree name		Scale
	▪ Certificate level		▪ Max scale
	▪ Skill-Know pair		▪ Min scale
	▪ Keywords		▪ Scale threshold
DIMENSION 2	▪ Competence identifier		▪ Years/Generic
	▪ Title or name	DIMENSION 4	▪ Skill name
	▪ Description		▪ Code
	▪ Definitions		▪ Description
	▪ Extra identifier		▪ Autonomy
	▪ Abbreviations		▪ Business skills
	▪ Date of creation		▪ Influence
	▪ Date of modification		▪ Complexity
	▪ Validation start date		▪ Knowledge
	▪ Date of issue		▪ Skill level name
	▪ Author		▪ Skill level number
	▪ Topics		▪ Cognitive soft skill
	▪ Credits		▪ Affective soft skill

	▪ Level		▪ Psychomotor soft skill
	▪ Version		▪ Role-based hard skill
	▪ Explicit metadata		▪ Skill-based skill
DIMENSION 3	▪ Level		▪ Technology Knowledge
	▪ Max level		▪ Pedagogy Knowledge
	▪ Min level		▪ Content Knowledge
	▪ Level threshold		▪ Combination

4.3 COMPETENCE CATEGORISATION

The management of academic competences can become complicated if there is a large number to consider. Thanks to the automatic mechanisms provided by the information systems, this task can be simplified. In order to facilitate the management of the competences in the model, a simple classification technique based on an existing taxonomy was proposed (Paquette, 2007). This taxonomy considers 3 generic classes of skills. In the first one, competences are separated according to whether they are Receive, which refers to learning a skill; Reproduce, which is based on applying the skill in examples and exercises; Produce/Create, which considers applying the skill to generate work results; and finally, Self-management, which is based on influencing with the acquired skill. The remaining subdivisions can be seen in Figure 4.3. The competences in the job offer of **Ошибка! Источник ссылки не найден.** were categorised concerning this model in **Ошибка! Источник ссылки не найден.**

Generic Skills Classes			Active meta- knowledge (Pitrat)	Generic problems (KADS)	Cognitive objectives (Bloom)	Skills cycle (Romiszowski)
1	2	3				
Receive	1. Acknowledge					Attention
	2. Integrate	2.1 Identify 2.2 Memorize			Memorize	Perceptual acuteness and discrimination
Reproduce	3. Instantiate/ Specify	3.1 Illustrate 3.2 Discriminate 3.3 Explain	Knowledge Search and Storage		Understand	Interpretation
	4. Transpose/ Translate					Procedure Recall Schema Recall
	5. Apply	5.1 Use 5.2 Simulate	Knowledge Use, Expression		Apply	
Produce/Create	6. Analyze	6.1 Deduce 6.2 Classify 6.3 Predict 6.4 Diagnose	Knowledge Discovery	Prediction, Supervision, Classification, Diagnosis	Analyze	Analysis
	7. Repair			Repair		Synthesis
	8. Synthesize	8.1 Induce 8.2 Plan 8.3 Model/ Construct		Planning, Design, Modeling	Synthesize	
Self-manage	9. Evaluate		Knowledge Acquisition		Evaluate	Evaluation
	10. Self- control	10.1 Initiate/ Influence 10.2 Adapt/ control				Initiation, Continuation, Control

Figure 4.3 Paquette's taxonomy (Paquette, 2007)

Table 3 Skill competences in Figure 2 categorised with Paquette's taxonomy class 1 (Paquette, 2007)

Receive > 2 items Reproduce > 6 items	Learn-Business (lb1)	Produce/Create > 26 items (cont.)	Design-Restful (dr1)
	Learn-Technology (lt1)		Develop-Software (ds2)
	Leverage-Knowledge (lk1)		Engineer-Practice (ep1)
Produce/Create > 26 items	Maintain-Standard (ms1)	Self-manage > 16 items	Engineer-System (es1)
	Spot-Problem (sp1)		Implement-Restful (ir1)
	Test-Component (tc1)		Participate-CodeReview (pc1)
	Understand-Design (ud1)		Participate-DesignReview (pd1)
	Understand-Fundamentals (uf1)		Tackle-Situation (ts1)
	Build-Architecture (ba1)		Work-Consumer (wc1)
	Build-CloudService (bc1)		Define-Culture(dc2)
	Build-Design (bd1)		Get-Healthy (gh1)
	Build-Product (bp1)		Insist-Standard (is1)
	Build-Requirements (br1)		Delight-Customer (dc1)
	Build-Service (bs2)		HaveEye-Quality (hq1)
	Build-Service (bs2)		HaveImpact-Business (hb1)
	Build-Solution (bs1)		HelpBuy-Solution (hs2)
	Build-Team (bt1)		Help-Costumer (hc1)
	Build-Technology (bt2)		HelpEngage-Solution (hs3)
	Coding-Component (cc1)		HelpFind-Solution (hs1)
	Contribute-Architecture (ca1)		MakeEasy-HealthcareAccess (mh1)
	Contribute-Design (cd1)		MakeEasy-HealthcareProduct (mh2)
	Contribute-Reliability (cr1)		MakeEasy-HealthcareService (mh3)
	Contribute-Scaling (cs1)		Mentor-Engineer (me1)
	Contribute-Team (ct1)		Mentor-SDE (ms2)
	Deliver-Solution (ds1)		Stay-Healthy (sh1)

The categorisation of competences can be a very powerful tool for the search, as it focuses the target on related competences. Another help provided by the categorisation of competences is when selecting a present competence in the model to be assigned to a learning object. These tasks can be done more quickly if there is a previous competence categorisation. The knowledge component was not considered for categorisation as existing knowledge classifications such as the UNESCO Nomenclature for Fields of Science and Technology (available in Spanish at <http://skos.um.es/unesco6>) can be used for this purpose in the model.

4.4 COVERAGE PERCENTAGE

In order to find out how well an academic subject can meet the needs of a job offer in terms of competences, a simple comparative method was proposed. This method consisted of analysing the competences in **Ошибка!**

Источник ссылки не найден., which is a job offer, with the academic competences in a total of three subjects of the Transport and Telecommunication Institute in Latvia, namely: Software Engineering (B-019-04), Project Management (M-080-04), Modern Software Engineering (M-047-04). Figure 5.4 shows the competences extracted from the syllabus M-047-04.

SYLLABUS 3 (Modern Software Engineering): 40 Items

Modern processes of software engineering ...SoftwareEngineer (s1)

XP development process ...XP (x1)

Agile modeling ...AgileModeling (a1)

Refactoring ...Refactoring (r1)

Test first design under extreme development ...Test (t1) Test-Design (td1)

Management in agile software engineering ...ManagementAgile (m1)

Scrum ... agile software development process ...Scrum (s2)

Basic concepts of design patterns ...DesignPatterns (d1)

Patterns in software development ...SoftwarePatterns (s3)

Aspect-oriented software engineering ...AspectSoftware (a2)

Modern methods and processes for managing software projects ...ModernMethods (m2) ...ModernProcess (m3)

Software project time management ...TimeManagement (t2)

Resource planning in the schedule development Plan-Resource (pr1) Develop-Schedule (ds1)

Managing of the project duration ...ProjectDuration (p1)

Cost management: monitoring of project progress ...CostManagement (c1) Monitor-ProjectProgress (mp1)

Advanced concepts and techniques used throughout the software life cycle ...AdvancedSoftware (a6) ...TechniquesSoftware (t3)

Effective production and management of large, complex, and long-lived software systems ...EffectiveProduction (e1) ...ManagementSystems (m4)

Holistic perspective of technical and non-technical factors involved in developing useful and safe software systems in complex social and organisational contexts ...HolisticPerspective (h1) ...UsefulFactors (u1) ...SafeSoftware (s4) ...ComplexContext (c2)

Modern processes of software engineering. Introduction. ...ModernProcess (m3)

XP development process ...XP (x1)

Agile modeling ...AgileModeling (a1)

Refactoring ...Refactoring (r1)

Test-driven development ...TestDevelopment (t4)

Agile project management ...ManagementAgile (m1)

Scrum ... agile software development process ...Scrum (s2)

Basic concepts of design patterns ...DesignPatterns (d1)

Patterns in software development ...SoftwarePatterns (s3)

Aspect-oriented software engineering ...AspectSoftware (a2)

Modern methods and processes for managing software projects ...ModernMethods (m2) ...ModernProcess (m3)

Plan Agile software project activities Plan-AgileActivities (pa1)

Participate effectively in Agile teamwork, lead an independent Agile development, and conduct the quality evaluation of a particular software product Participate-AgileTeamwork (pa2) Lead-AgileDevelopment (la1) Conduct-QualityEvaluation (cq1)

Support software maintenance Support-Maintenance (sm1)

Have a sufficient training in the field of Agile software engineering, which allows to contribute effectively in the Agile project work activities Train-Agile (ta1) Contribute-AgileProject (ca1)

Follow up with new developments in the area of project management and quickly adapt to new scientific and practical achievements in Agile software engineering FollowUp-Development (fd1) Adapt-ScientificAchievements (as1) Adapt-PracticalAchievements (ap1)

Understand limitations of Agile software engineering processes Understand-AgileLimitations (ua1)

Understand the most popular Agile software development methodologies Understand-AgileMethodologies (um1)

Be master of methods and facilities supporting teamwork, planning and effective organization of Agile software development process Master-Methods (mm1) Master-Facilities (mf1) Support-Teamwork (st1) Plan-Organization (po1) ...AgileProcess (a3)

Be able to comprehend and illustrate the gained knowledge at scientific and technical and professional level Comprehend-Knowledge (ck1) Illustrate-Knowledge (ik1) ...ScientificKnowledge (s4) ...ProfessionalKnowledge (p2)

Know the advanced concepts and techniques used throughout the software life cycle with special attention to Agile requirements modeling methods Know-Concept (kc1) Know-Technique (kt1) ...AgileRequirement (a4) ...KnowMethods (k1)

Agile methods of analysis, refactoring, pattern-based design and Agile methods of project management ...AgileAnalysis(a5) ...Refactoring (r1) ...DesignPatterns (d1)

Figure 5.4 Competences extracted from the syllabus Modern Software Engineering (M-047-04)

#	S1	S2	S3	JOB
1	0.5	0.0	0.5	Build a best-in-class healthcare product Build-Product (bp1)
2	0.5	0.0	0.0	Make high-quality healthcare easy to access MakeEasy-HealthcareAccess (mh1)
3	1.0	0.5	0.5	Software Development Engineer who can build new engineering solutions with a motivated team ...Engineer (e1) Build-Solution (bs1)
4	0.5	0.0	0.0	Delight our customers Delight-Customer (dc1)
5	0.5	0.0	1.0	Strong technical acumen ...Acumen (a1)
6	1.0	0.0	1.0	Ability to spot problems at architecture stage Spot-Problem (sp1)
7	0.5	0.0	1.0	Deep technical expertise ...Expertise (e2)
8	0.5	0.5	1.0	engineer (verbo) systems Engineer-System (es1)
9	0.0	0.0	0.0	Build reliable and secure services for healthcare Build-Service (bs2)
10	0.5	0.5	1.0	Have an eye towards quality HaveEye-Quality (hq1)
11	0.5	0.0	1.0	Insist on the highest standards Insist-Standard (is1)
12	0.5	0.5	1.0	You are motivated to tackle ambiguous situations with technologies to rapidly solve problems Tackle-Situation (ts1)
13	1.0	1.0	0.5	You are a team player ...Player (p1)
14	1.0	1.0	1.0	contribute to the execution of team objectives Contribute-Team (ct1)
15	0.5	0.5	1.0	leverage technical knowledge Leverage-Knowledge (lk1)
16	0.5	1.0	1.0	Engineering (verbo) best practices Engineer-Practice (ep1)
17	1.0	0.5	1.0	Rapidly deliver solutions Deliver-Solution (ds1)
18	0.0	0.5	0.0	Have a broad business impact HaveImpact-Business (hb1)
19	0.5	0.5	0.5	Software developers with expertise and passion for building large scale distributed systems and services ...Developer (d1)
20	1.0	0.5	1.0	Building functional and technical requirements into detailed architecture and design Build-Requirements (br1) Build-Architecture (ba1) Build-Design (bd1)
21	0.0	0.0	0.0	Coding and testing complex system components Coding-Component (cc1) Test-Component (tc1)
22	0.5	0.5	0.5	Participating in code and design reviews to maintain our high development standards Participate-CodeReview (pc1) Participate-DesignReview (pd1) Maintain-Standard (ms1)
23	0.5	0.0	0.5	Overall system architecture, scalability, reliability, and performance ...Architecture (a2) ...Scalability (s1) ...Reliability (r1) ...Performance (p2)
24	0.5	0.5	0.5	Mentoring other engineers, defining our challenging technical culture, and helping to build a new fast-growing team Mentor-Engineer (me1) Define-Culture (dc2) Build-Team (bt1)
25	0.0	0.0	0.0	Make it dramatically easier for customers to access the healthcare products and services MakeEasy-HealthcareProduct (mh2) MakeEasy-HealthcareService (ms3)
26	0.0	0.0	0.0	Need to get and stay healthy Get-Healthy (gh1) Stay-Healthy (sh1)
27	0.0	0.0	0.0	Building the technology, products and services, that help customers find, buy, and engage with the healthcare solutions Build-Technology (bt2) Build-Product (bp1)crep> Build-Service (bs2) Help-Find-Solution (hs1) Help-Buy-Solution (hs2) HelpEngage-Solution (hs3)
28	0.0	0.0	0.5	14+ years of experience contributing to the system design or architecture (architecture, design patterns, reliability and scaling) of new and current systems Contribute-Design (cd1) Contribute-Architecture (ca1) Contribute-Reliability (cr1) Contribute-Scaling (cs1)
29	0.0	0.0	0.0	2+ years of non-internship professional software development experience ...ProfessionalSoftware (ps)
30	0.0	0.5	0.5	Programming experience with at least one software programming language ...ExperienceSoftware (es)
31	0.0	0.0	0.0	3+ years developing software solutions Develop-Software (ds2)
32	1.0	0.5	1.0	Understanding software design principles and computer science fundamentals Understand-Design (ud1) Understand-Fundamentals (uf1)
33	0.0	0.0	0.0	Experience designing and implementing RESTful APIs at scale Design-Restful (dr1) Implement-Restful (ir1)
34	0.0	0.5	0.5	Curiosity and drive to learn new technologies and business domains Learn-Technology (lt1) Learn-Business (lb1)
35	0.0	0.0	0.0	Prior experience building services on a cloud platform (AWS or otherwise) Build-CloudService (bc1)
36	0.0	0.0	0.0	Prior experience working for a consumer-facing technology company Work-Consumer (wc1)
37	1.0	0.0	1.0	Strong knowledge of data structures, algorithms, distributed systems, and asynchronous architectures ...DataStructure (d2) ...Algorithm (a3) ...DistributedSystems(ds3) ...AsynchronousArchitecture (a4)

Figure 7 Outcome of the competences assessment

Each competence extracted from the job offer was carefully read (see **Ошибка! Источник ссылки не найден.**). Subsequently, the competences in each syllabus were checked to assess the degree of coverage. In case the

competences of the syllabus covered the competence of the job offer a score of 1 was given. If the competence was partially covered a score of 0.5 was given. If the competences in the syllabus were not related at all, a score of 0.0 was given. These scores were named Matching Coefficients. The coefficients for each syllabus were summed. The total sum was divided by the total number of competencies in the job offer. As a result, a percentage coverage was obtained. This percentage was highest for the subject Modern Software Engineering (M-047-04) with a value of 51%. **Ошибка! Источник ссылки не найден.** shows the result of these assessments.

4.5 CONCLUSIONS

The classification of competences in order to relate them to the requirements of a job is an idea that has been explored before. The multilingual classification of European Skills, Competences, and Occupations (ESCO, available at <http://esco.ec.europa.eu/es/about-esco/what-esco>) is an example. ESCO is a classification of skills and occupations, which is intended to serve as a dictionary of occupational demands in the European labour market. This resource can be used by various types of electronic systems to match jobseekers with jobs based on their competences, advise them to follow training courses to retrain or upgrade their skills, etc. The existence of the ESCO system shows the need for further work on digital management approaches of competences.

The development of the system proposed in this work should rely on additional intelligence systems such as expert systems or machine learning techniques to be able to extract the competences of the sentences, analyse the meaning, and evaluate the degree of coverage with the requirements of a job. The use of structured languages could be of great help in the process of extracting competences from natural language descriptions. There are currently a multitude of proposals such as the definition of User Stories in the well-known agile software development methodologies that could be adapted. On the other hand, the use of pre-established syntactic structures or the proposal of recommendations for the writing of competences could be another great help for the subsequent detection of competences, as already occurs in the writing of technical project requirements.

As future work, it is proposed to analyse the databases mentioned in the manuscript such as UNESCO Nomenclature for Fields of Science and Technology and ESCO in order to carry out a more detailed validation of the model. In addition, it is considered to study comparative competence techniques that can generate more accurate results. In fact, such techniques have already been proposed in the literature with semantic models based on skill ontologies together with matching algorithms, which calculate the weights between nodes of a skill tree based on meaning (Lv & Zhu, 2006), or look for automatic matching of skill graphs extracted from CVs and job offers (Phan et al., 2021). The use of scales with a higher number of degrees such as Likert scales in obtaining the Matching Coefficients will also be tested.

REFERENCES

- Carretero, S., Vuorikari, R., & Punie, Y. (2017). DigComp 2.1: the digital Competence Framework for Citizens with eight proficiency levels and examples of use [Marco de Competencias Digitales para la Ciudadanía]. Retrieved from <https://epale.ec.europa.eu/es/content/marco-europeo-de-competencias-digitales-digcomp>
- European Committee for Standardization. (2013). InLOC - Part 1: Information Model for Learning Outcomes and Competences (CWA 16655-1), 1–48. Retrieved from <http://www.inloc.org/inloc/Home>
- Frezza, S., Daniels, M., Pears, A., Cajander, Å., Kann, V., Kapoor, A., ... Wallace, C. (2018). Modelling competencies for computing education beyond 2020: A research based approach to defining competencies in the computing disciplines. In *Annual Conference on Innovation and Technology in Computer Science Education, ITiCSE* (pp. 148–174). Retrieved from <https://doi.org/10.1145/3293881.3295782>
- Gleason, K. T., Commodore-Mensah, Y., Wu, A. W., Kearns, R., Pronovost, P., Aboumatar, H., & Himmelfarb, C. R. D. (2021). Massive open online course (MOOC) learning builds capacity and improves competence for patient safety among global learners: A prospective cohort study. *Nurse Education Today*, 104, 104984.
- Gürtzgen, N., Diegmann, A., Pohlen, L., & van den Berg, G. J. (2021). Do digital information technologies help unemployed job seekers find a job? Evidence from the broadband internet expansion in Germany. *European Economic Review*, 132, 103657.
- Gürtzgen, N., Lochner, B., Pohlen, L., & van den Berg, G. J. (2021). Does online search improve the match quality of new hires? *Labour Economics*, 70, 101981.
- IEEE. (2008). IEEE 1484.20.1-2007, IEEE Standard for Learning Technology—Data Model for Reusable Competency Definitions. *Training*. Retrieved from <https://doi.org/10.1109/IEEESTD.2008.4445693>
- Lv, H., & Zhu, B. (2006). Skill ontology-based semantic model and its matching algorithm. In *2006 7th International Conference on Computer-Aided Industrial Design and Conceptual Design* (pp. 1–4). IEEE.
- Network, E. E. I., & Training, E. C. for the D. of V. (2014). *Tackling early leaving from education and training in Europe: Strategies, policies and measures*. Education, Audiovisual and Culture Executive Agency, Brussels, Belgium.
- Paquette, G. (2007). An ontology and a software framework for competency modeling and management. *Educational Technology and Society*, 10(3), 1–21.
- Phan, T. T., Pham, V. Q., Nguyen, H. D., Huynh, A. T., Tran, D. A., & Pham, V. T. (2021). Ontology-based resume searching system for job applicants in information technology. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 261–273). Springer.
- Sakamoto, F. (2022). Global competence in Japan: What do students really need? *International Journal of Intercultural Relations*, 91, 216–228.
- Subu, M. A., Al Yateem, N., Dias, J. M., Rahman, S. A., Ahmed, F. R., Abraham, M. S., ... Alnaqbi, A. R. M. (2022).

Listening to the minority: A qualitative study exploring male students' perceptions of the nursing profession and reasons for choosing nursing as a career. *Nurse Education Today*, 116, 105442.

Thomas, M., Wallace, C., Jones, G., O'Kane, J., Wilson, L., Dale, F., & Pontin, D. (2022). Using an integrated competence model to evaluate a health visitor cascade training programme for the Family Resilience Assessment Instrument and Tool (FRAIT). *Nurse Education in Practice*, 62, 103336.

LIST OF AUTHORS

1. José A. García Berná
2. Begoña Moros Valle
3. Joaquín Nicolás Ros
4. Juan M. Carrillo de Gea

5 . A3.5 Design of the Search Engine (UoI)

5.1 INTRODUCTION

Search engines are tools that discover and sort content on the internet. In other words, they make our lives easy by aggregating the results they think will interest users more than others.

They achieve this through three functions. First, they crawl. This means that they use bots (or crawlers) that locate and follow links that exist on websites, in order to find other content that they add to the machine's index. Indexing is the second task that search engines do, as they categorize web page content using keywords.

It is of great importance to successfully do SEO (Search Engine Optimization) for a website. Precisely because search engines use keywords to understand content and then rank websites based on that, depending on how relevant it is to the user's search. Rankings depend on various factors, but regardless it is important to make the necessary improvements so that the crawlers can index the pages.

Our goal is to build a search engine with which users can effectively search for courses/competences on the LMS platform. This work will be done for pedagogical purposes and the main goal is to enhance learners' desire for knowledge and to facilitate access to the various courses using more contemporary media. In order to achieve that, machine learning and data mining techniques will be used.

Machine learning is a field of computer science that deals with the creation of algorithms that aim to "learn" from data, i.e., to acquire additional knowledge through interaction with the environment in which they operate and the ability to improve through repetition the way they perform that action. Various technologies of machine learning will be used in order to make the search engine more efficient and more accessible to learners.

Data mining transforms raw data into useful information. Imported data is stored in various formats, either in a central repository or distributed across regions. Various data types are used for these data and for this reason, in the next stage they are pre-processed to obtain a suitable form for analysis and knowledge extraction. In this case, data mining will be used in order to improve the search engine and provide a more personalized search result by using the training information from the same user.

Using these techniques, the main function of the search engine will be:

- Searching courses/competences through the platform according to keywords in users' query. This feature will be based on machine learning algorithms and their applications in search engines.
- Providing results which conclude a set of courses/competences that are linked with the metadata of the LMS platform. The platform's database will conclude courses/competences and their metadata in order for the search engine to successfully provide the right results, after the completion of the searching in the database.
- Adapting features efficiency and constantly improving the function of the search engine in order to enhance the pedagogical purposes of the platform, provide a more user-friendly environment and help learners easily find what they are looking for.

In addition, users will have the chance to give their feedback in order for the engine to improve and deliver a more reliable and personalized outcome of competences/courses.

5.2 SEARCH ENGINES

WHAT IS A SEARCH ENGINE

A search engine is a software system that is designed to extract from the World Wide Web (WWW) the most relevant information that corresponds to specific searches of their users. These searches are called keywords or search queries.

Search results are presented in the form of a list known as organic results or search engine results pages (SERPs). The information corresponding to the organic results can have various forms such as hyperlinks that lead to specific pages, images, videos, infographics, articles, scientific studies or can even be direct answers to user questions.

Unlike web directories that are managed solely by humans, search engines rely primarily on software to operate. Search engines can draw information from a small part of the Internet called the Surface Web, which accounts for about 4% of the Internet's total information. The remaining 96%, which corresponds mainly to the Deep Web and to a lesser extent to the Dark Web, is inaccessible to search engines (Economides, 2021).



Figure 5.1: Search Engines (Warren R.,2020)

WHAT ARE SEARCH ENGINES FOR?

The main function of search engines is to provide timely answers to queries submitted by users. The search engine is always ready to find the various web pages that contain information related to the user's request. The real job of what a search engine does is to deliver the best ranked pages to the top. However, this does not mean that they will be the ones with the best quality in their content (Seguidores.online n.d.).

THE HISTORY OF SEARCH ENGINES

The multitude of discoveries surrounding the Internet have contributed significantly to its enormous growth and to it being nothing like what it was when it was first created. In addition to the technological improvements and discoveries that were applied to it, a multitude of websites were created and the internet was no longer a simple source of information, but connected to important larger business activities. In its first stage of operation, it consisted of a number of Ftp (File transfer protocol) sites where users could download or upload files. Nevertheless, the search and finding of these files was distinguished by significant difficulty and this resulted from the need to know the exact address where they were located. This search was distinguished as a rather difficult, time-consuming process and required a lot of patience.

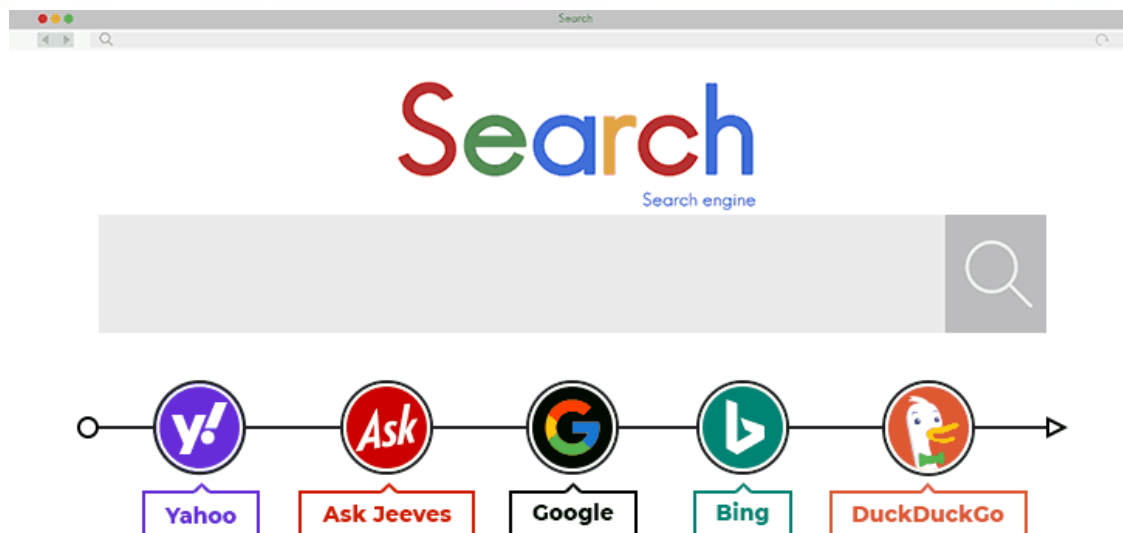


Figure 5.2: The History of Search Engines (Compton, 2019)

The first really important step to organize the contents of the Internet was the development of "Gophers" (Randolph Hoch, 2001), computer-based collections of Internet addresses that were registered with the help of a menu (The term "Gopher" comes from Minnesota's university mascot, where the first Internet "Gopher" was created there). "Gophers" were not HTML-based and their indexing was based on file titles or very brief descriptions, but if someone knew how to get to a "Gopher" it would allow them to "download" selected files.

The discovery of Gopher created the need for programs that can locate information through Gopher directories. This data helped build the program Veronica (Very Easy Rodent-Oriented Net-wide Index to Computerized Archives) and Jughead (Jonzy's Universal Gopher Hierarchy Excavation and Display) which search for files and texts stored in the Gopher system. The operation of these two programs was distinguished by the same procedures, allowing users to search the catalogs of information using words or phrases. In November 1993, the second Allweb search engine appeared, but it did not use a web robot, but was informed by the administrators of the websites about the existence of the latter.

"Gopher" prevailed for a few years before being overshadowed by the rapid development of the World Wide Web, which allowed the use of hyperlinks, full-text searching, graphical browsers, and also the development of search engines.

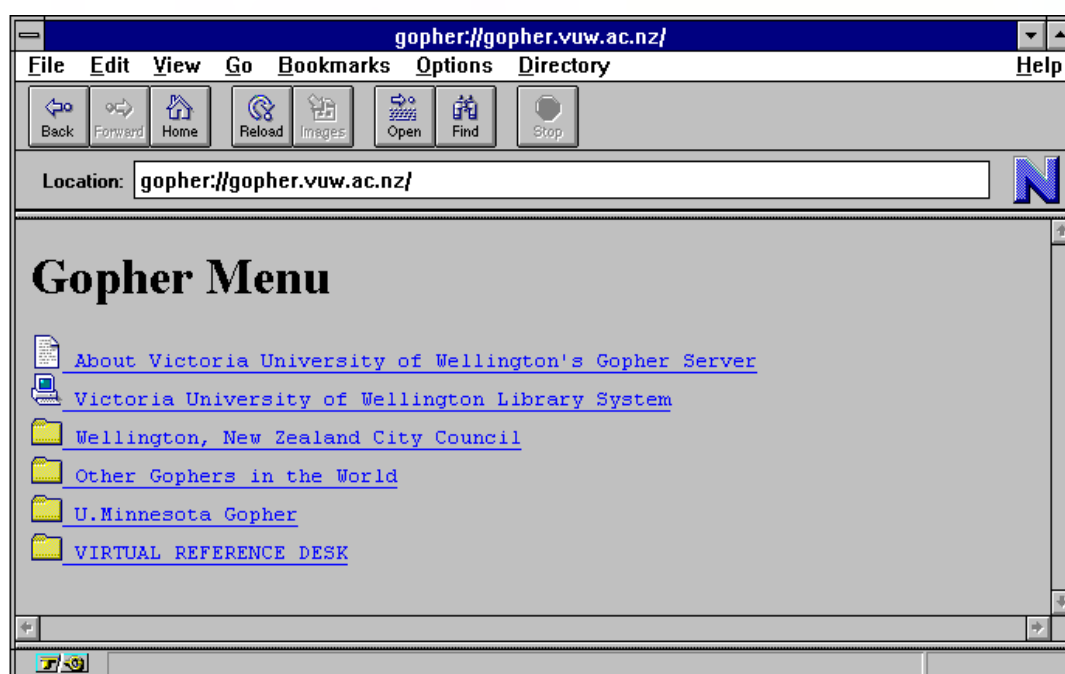


Figure 5.3: Goprher (<https://evert.meulie.net/tag/gopher/>)

The first search engine developed was WebCrawler, which came from the University of Washington and started to be used on April 1994. Within a year, three competitors emerged: the Lycos search engine, the Infoseek search engine, and the OpenText search engine. In late 1995 AltaVista and Excite made their appearance. It is interesting to emphasize that for a large part the search was done in a similar way to that in today's search engines, such as using logical operators (Boolean), using truncation, etc. Unfortunately, none of these search engines took advantage of Searching Technology. Furthermore, neither the search engines, nor the thematic directories, took advantage of the extensive Subject Classification Theory and practice of the last hundred years. These points relate in a very practical way as the professional browser must recognize the fact that most search engines were and are being developed for the everyday browser and not for those who aim to take advantage of more sophisticated approaches and techniques.

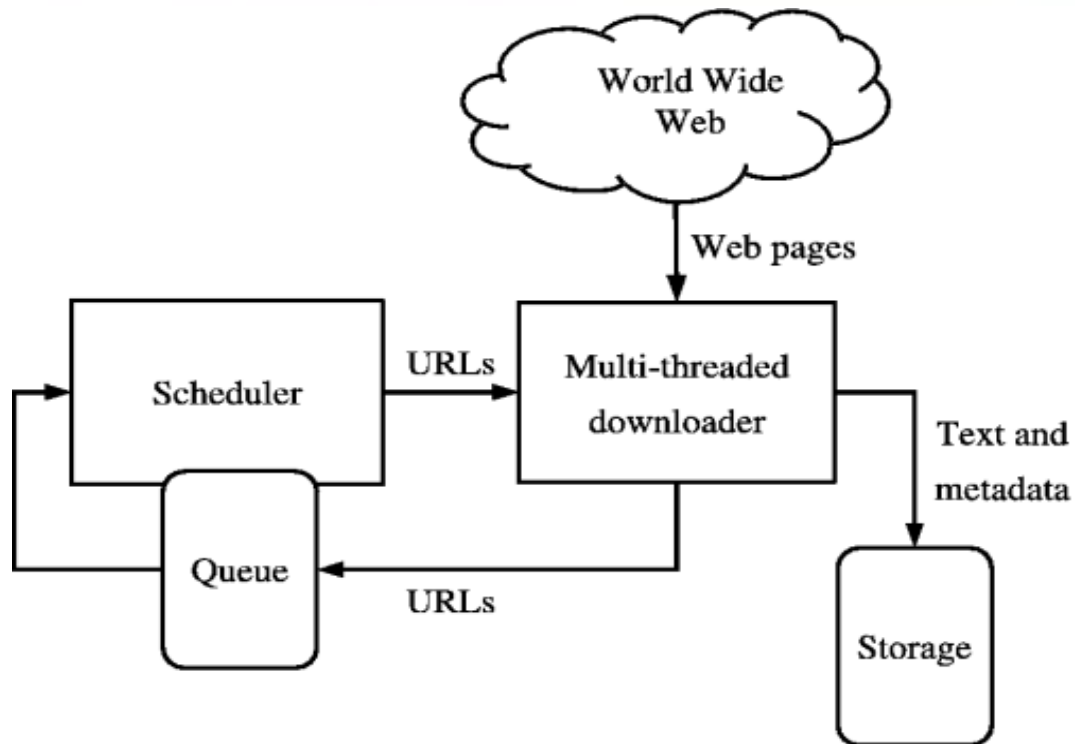


Figure 5.4: Web Crawler (https://en.wikipedia.org/wiki/Web_crawler)

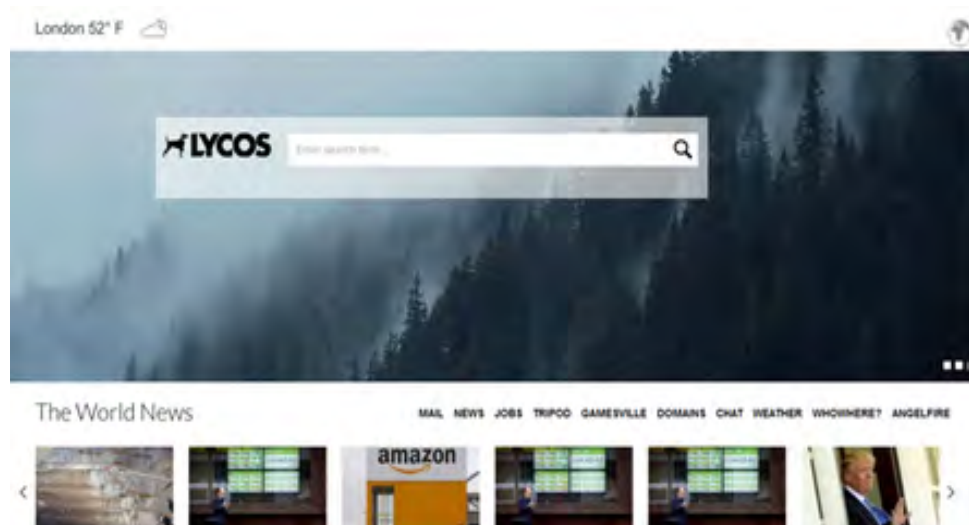


Figure 5.5: Lycos (<https://en.wikipedia.org/wiki/Lycos>)



Figure 5.6: Infoseek (<https://wiki.preterhuman.net/Infoseek>)

The HotBot search engine appeared in 1996 and Northern Light in 1997 (Randolph Hoch, 2001). HotBot featured a more complex, but at the same time an easy-to-use interface and was connected to a very large database (until the end of 1997, its database was considered the largest available). Northern Light brought an integration to searching the web and the information it contains. The Google search engine appeared in 1998, and its file classification along with an extremely simple interface combined effectively to produce an engine that quickly managed to gain popularity among both casual and long-term researchers. Meanwhile, the race for the largest search engine by number of pages had narrowed somewhat, until the appearance in 1999 of the search engine Fast Search, which consisted of a database of over 200 million records. This incentive, along with some other features, meant that the race for size was back on, with four machines reaching 200 million records by January 2000.



Figure 5.7: HotBot (<https://www.reddit.com/r/90sdesign/comments/6tg32d/hotbot/>)

Among the "first" search engines, Open Text took the first big step. At the beginning of 1998 it was no longer available, it had been abolished. Like the rest of the business world, search engine companies are extremely sensitive to the trends of the time and as a result are affected by them. In 1996 and 1997, the trend was to make sure that the search engine one was using or building was an "advanced" version of the others, regardless of whether the advanced version actually had more features than the others. Greater in importance in terms of benefits, 1998 brought the "personalization" of the portal. The idea of "portal personalization" or "Web Gateway" was evident in the identified and user-selected categories of news that appear on the home page, such as local weather and local TV programs, personal monitoring of portfolios, personal diaries, etc. The desire of search engine developers to follow this example and the realization that this approach could generate advertising revenue led these two very closely related standards to quickly become the universal business standard for the major search engines. Although many users did not realize it yet, this approach to personalization was an important step forward in terms of bringing the web to the level of a family and generally being always accessible, simple, used more often, and, in addition, to provide the specific and obvious benefits.

In the years 1999 and 2000 there was a great effort in the concept of the portal. During the first year, the added tools (such as directories, etc.) were mainly designed on the home page in the hope that users would use them. In 1999 there was a significant shift towards automatically incorporating page content into the final results - at the same time that the search engine's web database is indexed, searches the subject directory, business directory, etc., and presents these results along with the normal search results. This existence (completion) of tools (resources) has significantly improved the quality of the search results since it provides the researcher with particularly relevant results, without having to perform the search separately in different tools.

The next step has to do with both the users and the creators of the search engines. Tools that receive user attention will be maintained, enhanced, replicated, and evaluated. The problem, initially with web search engines, is that the person likely to use them is not the typical (everyday) user of search engines. The typical user would be less interested in the more complex and research-oriented features. The extent to which this is true is very apparent if someone look at the searches of a typical user. The Lycos search engine provides an interesting, if sometimes very compressed, list of favorite search addresses. In one week, the top 50 searches include 46 searches that refer to the entertainment, sports, or gaming categories. The relevance of this is not a matter of information snobbery, but the need to face reality and the primary position it advocates, which is that most search engines do not make money from the searcher who uses the web for business purposes. The only positive is that the total number of users is increasing, as well as the number of people using search engines for professional reasons, for investment, as well as for their basic education in science subjects such as humanity, business, and medicine, etc. There are many more reasons for search engine developers to pay more attention to the extreme

searcher. But the serious researcher must also use the salient features of a machine so that those features remain and are enhanced (Mapes, 2008), (Economides, 2021).

OPERATION OF SEARCH ENGINES

The order in which a search engine works is as follows:

- Web crawling
- Indexing
- Searching

Web Crawling

Web crawlers, also known as special software, are programs that visit web pages to:

1. identify the new pages-addresses that are to be added to the search engine
2. identify page-addresses that have already been explored and changed.

This special software gathers information about the content of the pages from the addresses it visits and provides this information to the search engine database. Much could be said about how this is done, but for the searcher it doesn't matter, although it is understandable, because some search engines find certain pages that other search engines do not, even if the page is in the second search engine's database. In many search engines, the most popular pages (such as those that are visited very often by users or those that have many links) are explored more thoroughly and more frequently by crawlers than less-popular pages.

Web crawlers can be programmed to explore web pages in depth or breadth, or both. What is programmed for deep exploration not only identifies the main site pages, but also identifies the linked pages within them. This software, which is programmed for the scope of site exploration, is typically interested in finding more of the main pages, but not necessarily in identifying all the linked pages of a main page. As search engines in today's era have evolved a lot and become even more competitive, there has been a tendency to merge depth and breadth.

Indexing

In terms of which pages will actually be retrieved by a user's query, indexing may play a more important role than the crawling process. The indexing program examines the information stored in the database and creates the appropriate index entries. When a query is submitted, this is used to identify the files that are relevant to the user's query.

Most search engines claim to index by looking at all words from each page. What matters most is what search engines choose to consider a 'word'. Some have a list of 'stop words' (small, common words that are considered insignificant enough to be ignored) that are not included in the index. Some search engines do not index such words as articles and links. Others omit widely used but potentially valuable words such as 'web' and 'internet'. Sometimes numbers are also omitted.

All major search engines compile their index by looking at the title and URL (Uniform Resource Locator). Metatags are usually, but not always, indexed (Metatags are words, phrases, or sentences placed in a special section of HTML (Hypertext Markup Language) code describing the content of the page). The Metatags cannot be seen when a page is visited, however there is the ability to ask the browser to show the 'page source'. It is understandable how useful the contents of metatags prove to be for retrieving information. However, some search engines deliberately do not index metatags because they are a part of the page that can be very easily abused and altered. This caution results in missing valuable indexing information.

HTML connoisseurs know that frames are used in millions of web pages. (Frames are an HTML device that treat different parts of a page as independent windows). Some search engines do not index by exploring the boxes, thus potentially missing areas relevant to the user's query. This disadvantage is compensated by the fact that the website builder creates a version of the page without frames as well as the version with frames. In addition, with the evolution of web construction, frames are used much less than in the past.

By understanding these different ways of indexing policy, it becomes clear why the relevant pages, even if they are registered in the engine's database, are not retrieved after a few searches. It also explains why a page can be retrieved by one machine and not another, even when the same page is on both machines.

Searching

When a query is entered into a search engine (usually using keywords), the engine examines its content and provides a list of the best options, i.e., websites that match the query posed, usually presenting a short summary that contains the title of the document and sometimes parts of the text. Content is created from the information and data stored. Most search engines support the use of the Boolean operators AND, OR and NOT to further specify the search query. Also, some search engines provide an advanced feature called proximity search, which allows users to set the distance between keywords. There is also concept-based searching where research involves using statistical analysis on pages containing the words or phrases being searched for. Additionally, the user has the ability to type a question in the same form as they would to a human with natural language queries that enable this functionality. The effectiveness of a search engine depends on the relevance of the set of results it returns. There are millions of web pages that include a specific word or phrase included in the query, but there are some pages that may be more relevant, popular or authoritative than others. Most search engines use

algorithms and methods to rank results to provide the "best" results first. As the Internet evolves, the specific methods and algorithms also evolve and change. Most web search engines are financially supported by advertising revenue. Thus, they use the practice of allowing advertisers to pay money to rank their listings higher in search results. Search engines that do not use this tactic are financially supported by serving search-related ads. Money is earned every time someone clicks on one of these ads.

The Recovery Engine

This is the program that takes a user's query and then searches the index to identify and deliver the files that match their query. In reality, two important events occur during this process:

1. the recovery engine identifies the files referred to in the query using a "recovery algorithm", and
2. the search engine then sorts the recovered files into a specific order and displays them to the user.

These can happen more or less simultaneously, or they can be quite distinct processes.

Retrieval Algorithms are programs used to apply criteria to determine which files contain particular words, phrases, or combinations of them. They can also match them against other user-defined criteria, such as whether a particular page contains audio or video files.

The part of the search engine that calculates the relevancy of files can be built into the retrieval algorithm, or it can be a separate process. Even when it is a separate process, the difference may not be obvious to the user, and usually shouldn't be.

Text Search Methods

- Keyword search: Most search engines use keywords to display the most relevant content.
- Concept Searching (Clustering): Concept-based search systems try to determine every time what the user might mean, not just what they type. Excite is the most relevant example using concept-based search.
- Meta-Search Engine: A meta-search engine sends user queries to many other search engines and/or databases and aggregates the data and results obtained (Seymour, Frantsvog, Kumar, 2011).

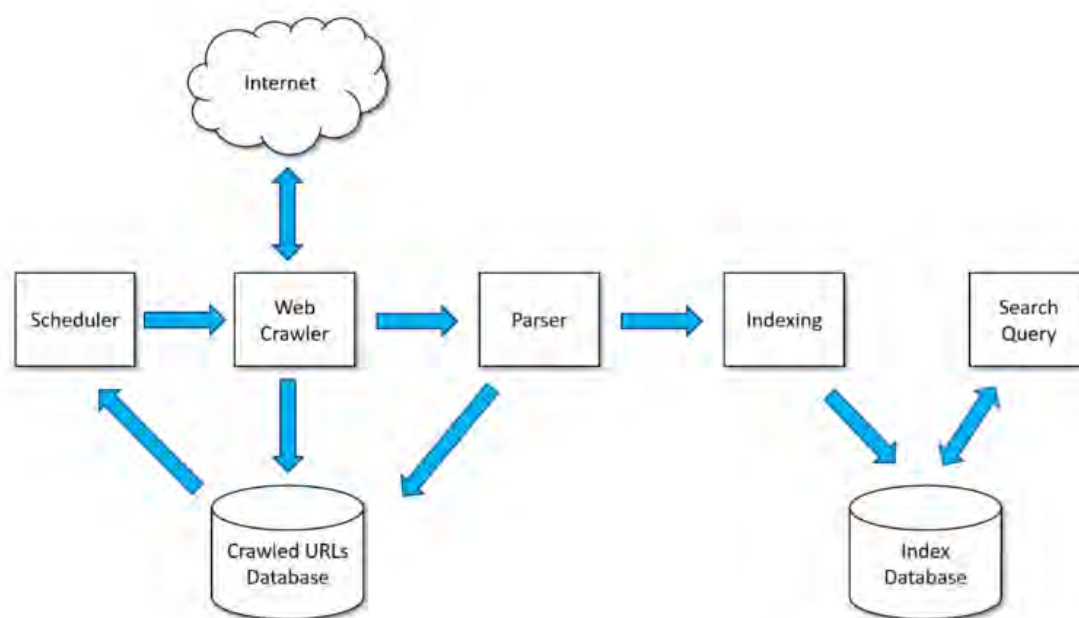


Figure 5.8: How Search Engines Work (Warren R., 2020)

In addition, some other essential parts of a search engine that are concluded in the operating process are the engine database and the HTML GUI.

The Engine Database

The total collection of information stored for each of the web pages constitutes the search engine's database. The collection consists of pages identified by crawlers, but increasingly also includes pages identified by other sources or techniques. A very large number of pages that are added to search engines come from the direct registration request of the creators of the website. If a user looks at the home page of any search engine, there will probably be a link that allows each user to submit a page to that particular search engine. As long as the page is not a case of 'spamming', the submitted pages will probably be added to the search engine's database. All or most search engine designers review submitted pages for spam (programs used by developers trying to cheat to illegally increase the chances of a page being retrieved). A service may also apply other criteria, but with the exception of spam, chances are very good that the submitted page will end up in the search engine's database.

Other sources may also feed the search engine database. The database may, for example, contain pages with subordinate titles from a subject index such as Open Directory or Yahoo. It is sometimes easy to forget that when you use a search engine, it is not searching the web directly, but searching a database that contains its files, which describe a portion of those pages that exist on the web. Knowing this can help avoid unrealistic expectations of what a search engine can actually achieve.

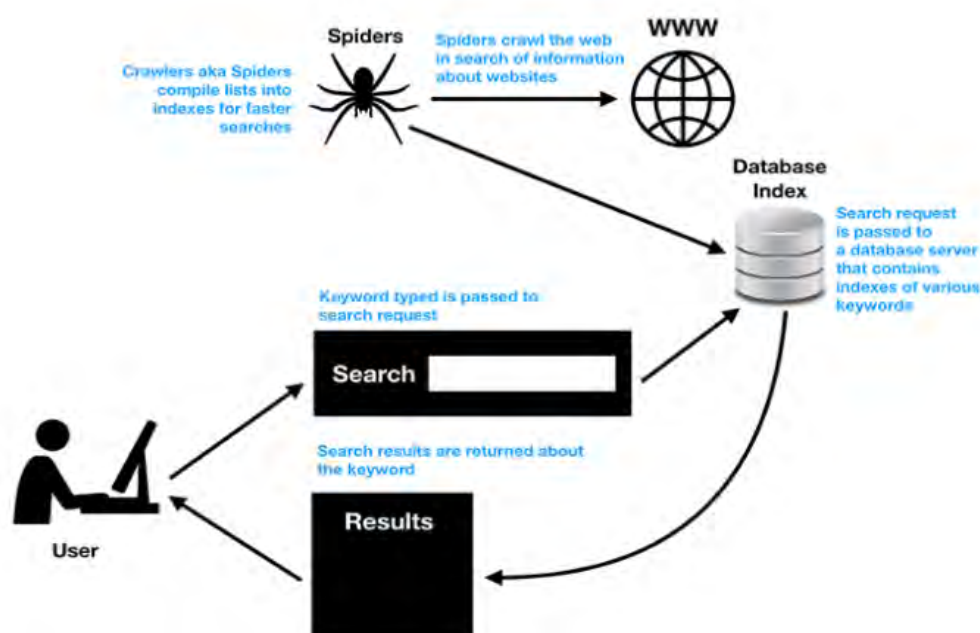


Figure 5.9: Search request passed to database (Tabora, 2019)

The HTML GUI

What users see whenever they connect to a search engine is the HTML-based interface. This interface gathers query data from the user, and sends that data to the search engine to retrieve the pages. Its most obvious function, of course, is to provide the user with a way to submit their question. However, the interface also provides other functions to the user, such as providing a space for advertisements, providing access to various features, and providing links to help pages and other information about the service in general.

The way the results are presented to the user tends to be standardized, since most search engines now provide a short summary along with the reference to the specific information as well as a percentage of relevance in relation to the requested term, as it was put by the user.

The spider returns to sites regularly (e.g., weekly) to check for any changes and update the database, ensuring network coverage is up-to-date and extensive. This results in a huge number of results for almost any search.

Furthermore, the automatic creation of the search engine database means that there is no segregation in terms of the quality of the information retrieved, which is necessary, given that anyone can publish information over the internet. In general, the lack of quality control of web resources means that the vast amounts of information retrieved can range from high-quality and search-relevant material to information of extremely dubious value.

Although search engines aim to perform the same function, they each approach it in a different way, sometimes leading to strikingly different results. Factors affecting results include database size, update frequency, and search capabilities. Also, search engines differ in their speed, the design of the search environment, the way results are displayed, and the amount of help they provide (Hock, 2001).

TYPES OF SEARCH ENGINES

Search engines have a lot of web pages stored on their database. However, the main goal is to provide accurate information according to requested keywords. Search engines are classified as follows:

- Hierarchical search engines
- Directory-based search
- Meta search engines
- Hybrid search engines
- Speciality or subject specific search engines

There are other types of search engines too, each one according to its operation (natural language, gateways and virtual libraries, intelligent agents).

Hierarchical search engines

Hierarchical search engines are related to web crawlers, where the search engine goes through web pages gathering information. When a user makes a specific query in a search engine, data files and results similar to that search are displayed. At different periods of time, search engines make revisions to different web pages, to update their database of possible changes that have occurred in their content. Every hierarchical search engine consists of three parts:

- The software that scans the network
- Programs that create databases
- Everything the user uses to explore the database

Directory Based Search Engines

Directory-type search engines offer users relevant information about listed web pages found in a link index. They don't contain any content information and they don't explore web pages either. They are only responsible for recording the page details, title and all descriptive details requested during registration. The information they deliver to the user is related to the body and topics covered by the website.

Meta-search engines

This search window uses multiple engines and with this action it is possible to get wider results. This is possible thanks to the fact that the answers to the various questions of the users use several search engines, thereby speeding up and feeding the information provided.

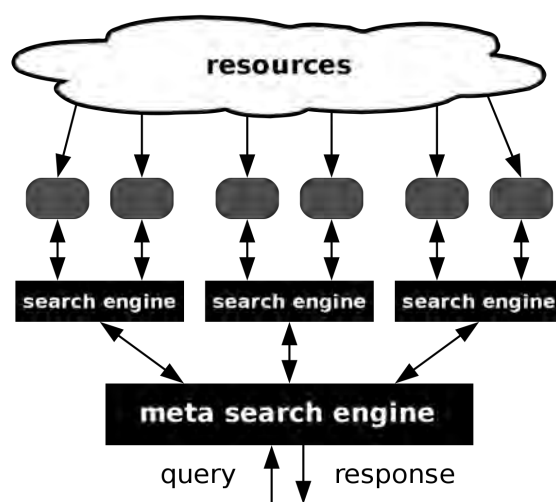


Figure 5.10: Meta-search Engine (https://en.wikipedia.org/wiki/Metasearch_engine)

Hybrid Search Engines

Both crawler-based results and human-powered listings are presented by hybrid search engines. These days, there is a combination of both results. Crawlers serve as the primary mechanism while human-powered directories serve as the secondary mechanism in the majority of crawler-based search engines, like Google. Google might use descriptions of webpages from human-powered directories in order to display them in search results. Hybrid forms of search engines are shifting more and more toward becoming crawler-based as human powered directories disappear. Even yet, copied and spammy websites are still removed manually from search results. In case a site is reported for spammy behavior, the website owner should take corrective action and resubmit the site to search engines. Before re-including the submitted site in the search results, the experts manually review it. Although the operations are controlled in this way by the crawlers, manual control is used to track and display the search results in a natural way (Ann Sunny, 2012).

Subject Specific Search Engines

These search engines do not attempt to index the entire web. Instead, they focus on searching for websites or pages within a pre-defined subject area, geographic area, or resource type. Because these specialized search engines aim for extensive coverage within a single domain and at coverage across subjects, they can often index

documents that are not included in even the largest search engine databases. For this reason, it is often a useful starting point for specific searches.

Natural Language Search Engines

These search engines solve the problem of finding information using a different approach. They do not use Boolean operators, but questions are formulated by the user using natural language. Examples of such search engines are AskJeeves.

Gateways and Virtual Libraries

As the internet has grown, the need to find information on it has become more imperative. Of course, this leads to two main problems - finding the information and evaluating it when found. The virtual library is the answer to these two questions. They are designed to offer quick and easy ways to find quality information that will help the user in his work. In particular, portals and virtual libraries are collections of high-quality information sources of a specific subject area, concerning a defined audience (Seguidores.online n.d.).

Intelligent Agents

It is commonly accepted that over time the effective use of the internet becomes more and more difficult. It is therefore obvious that locating and accessing information, retrieving, filtering and evaluating it is a particularly difficult process to be successfully carried out by a human user. Search engines in their traditional form prove to be ineffective under these conditions and the need to use modern technologies becomes demanding. Intelligent agents try to adapt to the profile and needs of their users and successfully carry out the above process by doing exactly what their users would do if they had the necessary time.

EXAMPLES OF SEARCH ENGINES

It is quite possible that some users think that there is only one search engine and that is the mighty Google. Although it is the most popular in the world, there are other search engines that also have their special benefits. Although there is no established classification for the engines, there are engines for queries on general topics and others specialize in image searches. In addition, there are engines for exclusively educational topics and also for children (Seguidores.online n.d.).

Google

Google started with the idea that websites are simply part of a huge popularity contest, and the more popular a site is, the more people link to it to recommend it to others. Therefore, the more links a website has, the better it should be for users, and the better it should rank in their new search engine.

In a 1998 university paper, it was determined that Google essentially translates a link from page a to page b as a vote from page a to page b. Google evaluates how important a page is by the votes it receives. This formed the most important part of Google's algorithm, known as PageRank. Over the years, Google has tweaked its algorithm to give less value to some things and more to others (Adamantios, 2021).

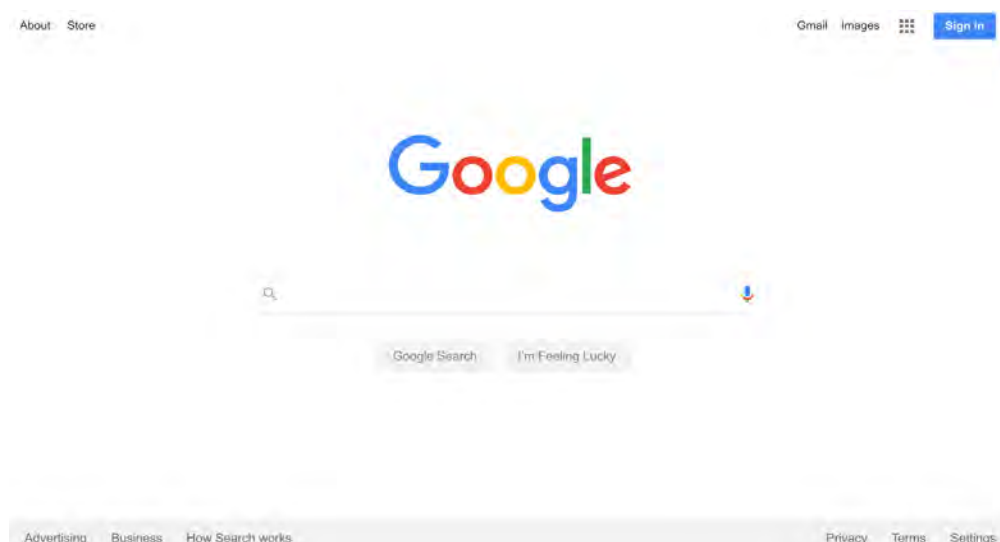


Figure 5.11: Google (https://en.wikipedia.org/wiki/Google_Search#/media/File:Google_Homepage.svg)

Bing

This Microsoft search engine is quite similar to the way Google offers answers to user queries. The interface is also very similar to Google but with some variations such as the ability to play video thumbnails while other video queries are made. The Bing search engine also offers a translation program, just by clicking on an extra tab (Seguidores.online n.d.).



Figure 5.12: Bing (Warren, 2020)

Yahoo!

Although the first impression users have when using this search engine is that they are entering a web page, it has millions of followers worldwide. Yahoo! offers the ability to have an email, enjoy online shopping, have a discount on airline tickets and tour packages, play games and endless other services (Seguidores.online n.d.).



Figure 5.13: Yahoo! (J Law, 2022)

Ask

It is a pretty accurate search engine; you just need to enter the word you want to search and the search result will be exclusively for the topic you entered (Seguidores.online n.d.).



Figure 5.14: Ask (https://play.google.com/store/apps/details?id=com.ask.browser&hl=en_US&gl=US)

DuckDuckGo

This search engine does not associate users with their Internet Protocol or IP data. It provides search results by categories. It also returns recipe results if the search term includes food names (Seguidores.online n.d.).

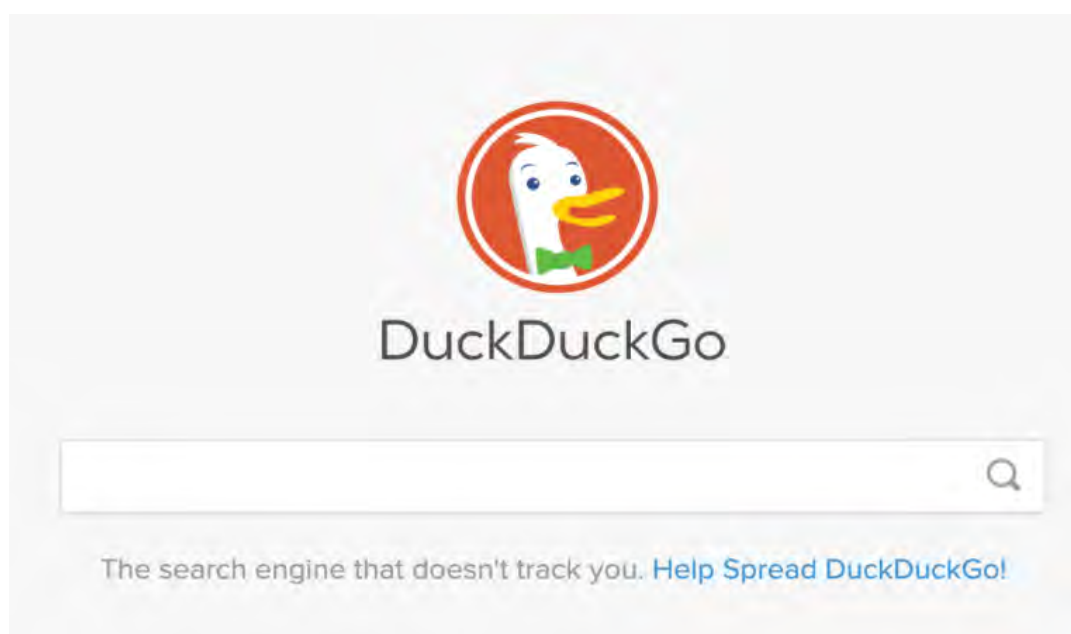


Figure 5.15: DuckDuckGo (Lomas, 2019)

Image search engine

In addition to queries on various topics, search engines offer the possibility to consult the various images stored on the network. Although it is not possible to determine which is the best image search engine. All offer good picture results and child protection filters, like the one below:

- Google Image Search: It has a collection of millions of images, which are available to all users. It offers different filters, which help users to search for the best result. The same image search option is offered by search engines Bing and Yahoo! (Seguidores.online n.d.).

Educational search engines

Each of the search engines has its areas of expertise, that's why the ones oriented towards educational topics and for the lovers of scientific topics are born. Query results are related to scientific research publications and specialized journals in a specific field of science, among others. The most commonly used engines for these purposes are:

- MetaGer: It is very useful in providing its services, since it offers high security to users' IP addresses, because it does not keep any traces during and after navigation. That is, it gives the user the possibility of history without leaving a fingerprint.



Figure 5.16: Metager (<https://metager.org/>)

- Base: Through it you can consult the various scientific articles, from recognized libraries to the available digital archives.



Figure 5.17: Base (https://en.wikipedia.org/wiki/BASE_%28search_engine%29)

- Google Scholar: With this search engine, you can consult the various scientific essays available. It also offers the possibility of viewing books or excerpts of them that are related to the query made (Seguidores.online n.d.).

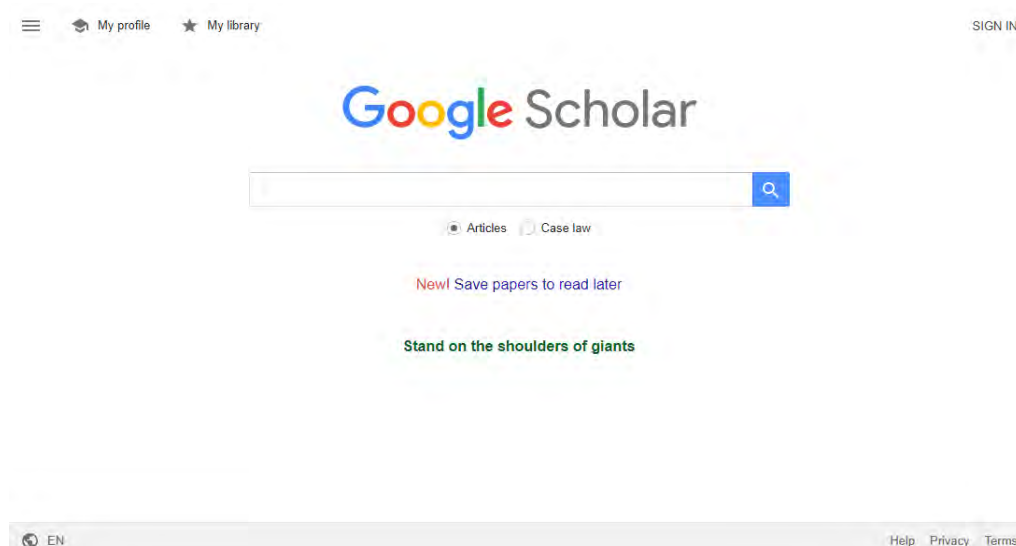


Figure 5.18: Google Scholar
(https://en.wikipedia.org/wiki/Google_Scholar#/media/File:Google_Scholar_home_page.png)

Search engine for kids

The options available for information seeking are not limited to just one sector of the population. They also include children, as part of a counseling education and everything under an approach that ensures appropriate content for them. Among the most used are fragFinn and BlindeKuh (Seguidores.online n.d.).

PROS AND CONS

Search engines offer a wealth of information and facilitate users in searching for any query. Knowledge has entered a new era thanks to search engines. However, although the advantages are very important and much more than their disadvantages, there are some elements that reduce their credibility and effectiveness, such as:

- They do not remember the previous search.
- They do not personalize the answers (e.g., they return the same URLs to University professors and primary school learners).
- They do not update periodically on new developments.
- They do not filter out useless information.
- They do not process selected information.

THE FUTURE OF SEARCH ENGINES

Today search engines are not just any index of web pages. Thanks to artificial intelligence, they are able to better understand the intent of their users' searches and provide personalized results.

Voice Search

Specific searches have a significant percentage of total searches. Google has reported that about 20% of its searches are done by voice. Voice searches use multi-word keywords (long-tail keywords) phrased in the form of a conversation. A good practice to take advantage of this new reality is to structure part of the website content in the form of questions and answers.

Visual Search

Users today are able to take a photo of an object (e.g., a product), upload the photo to search engines and thus start the customer journey. Today, 24% of website traffic originates from image searches.

This trend is expected to strengthen in the near future. Good practices for digital marketers regarding the exploitation of this trend is the use of quality and, if possible, original images. It is also necessary to use the keywords related to each image and specifically to its file name and ALT (Alternative Text).

It is also worth mentioning that Visual Recognition technology using Artificial Intelligence is developing rapidly, which helps search engines to understand the content of photos, without the need for accompanying descriptive texts (Economides, 2021).

Amazon Product Searches

Amazon is the largest e-commerce platform in the world. When it comes to product searches, at least in the US, it's the benchmark. The creation of content is of crucial importance for achieving a high ranking of the products in the specific engine but also in general.

In this context, it is becoming more and more important to use artificial intelligence for the automated creation of content on a large scale for e-shops and other websites (Economides, 2021).

Multitask Unified Model (MUM)

The MUM technology announced by Google in May 2021 is expected to significantly change the user experience of its machine.

It is an artificial intelligence model that uses transformer technology, which understands the meaning of words in relation to the context in which they are included and no longer perceives them as individual elements.

This development may enable Google to act as a virtual research assistant on search topics that will emphasize direct answers to user searches (Economides, 2021).

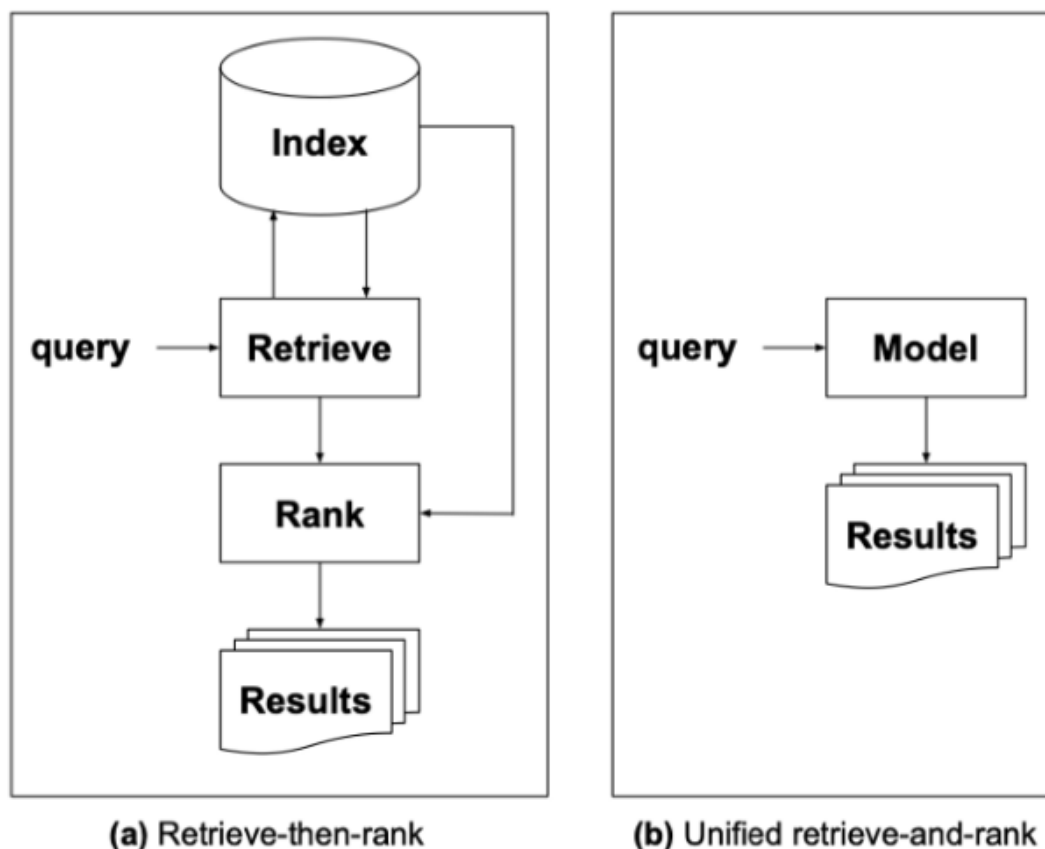


Figure 5.19: Unified Model (Marius, 2022)

TOP 10 SEARCH ENGINES IN THE WORLD IN 2022

Which are the top search engines today? Although Google is the most popular, there are other search engines that may not be so well known but still serve millions of search queries per day.

There are a number of alternative search engines that want to take Google's throne but none of them is ready (yet) to even pose a threat. The best Google alternatives are presented below and they are the 10 best search engines in 2022, ranked by popularity (Chris, 2022).

1. Google
2. Microsoft Bing
3. Yahoo
4. Baidu
5. Yandex
6. DuckDuckGo

7. Ask.com
8. Ecosia
9. Aol.com
10. Internet Archive

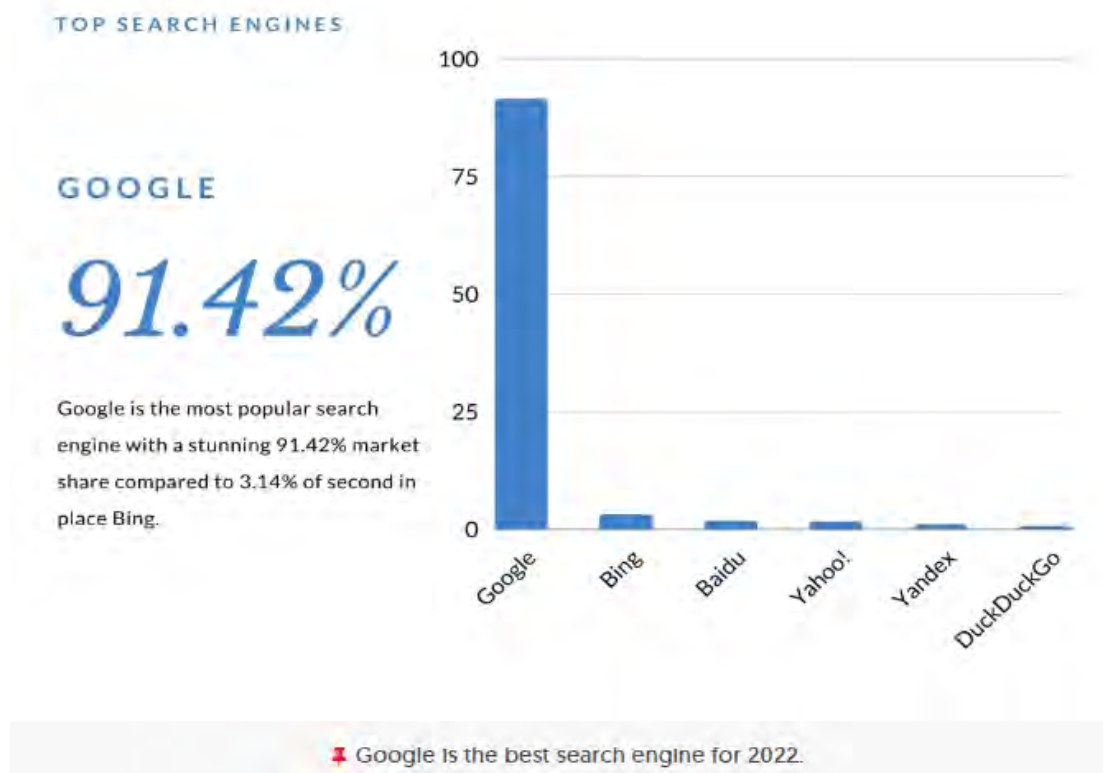


Figure 5.20: Top Search Engines for 2022 (Chris, 2022)

5.3 MACHINE LEARNING

Machine Learning is an important part of computer science and is related to the field of Artificial Intelligence. Machine Learning involves algorithms, which extract information from data by interacting with the environment they are in, and then aim to improve that action through iterations. According to Arthur Samuel (Samuel, 1959), machine learning "enables computers to learn without being explicitly programmed".

The most formal definition was given by Tom M. Mitchell (Mitchell, 1997), according to which: "A computer program is said to learn from an experience E, relative to a series of projects T and performance measured by P, if the performance on projects T, measured by P, improves with experience E".

Machine learning is found in many applications that do not involve explicit design and programming, such as search engines, visual character recognition, spam filters, etc. Machine learning prediction or categorization techniques are used by data mining. A system, through machine learning, has the ability to predict, which with feedback, leads to learning. In addition to feedback, learning is also based on examples and stored knowledge. In case there is a similar case in the future, the same or completely different prediction can be made using the feedback. In machine learning programs the science of statistics is particularly important, as the results obtained from the predictions must be statistically significant.

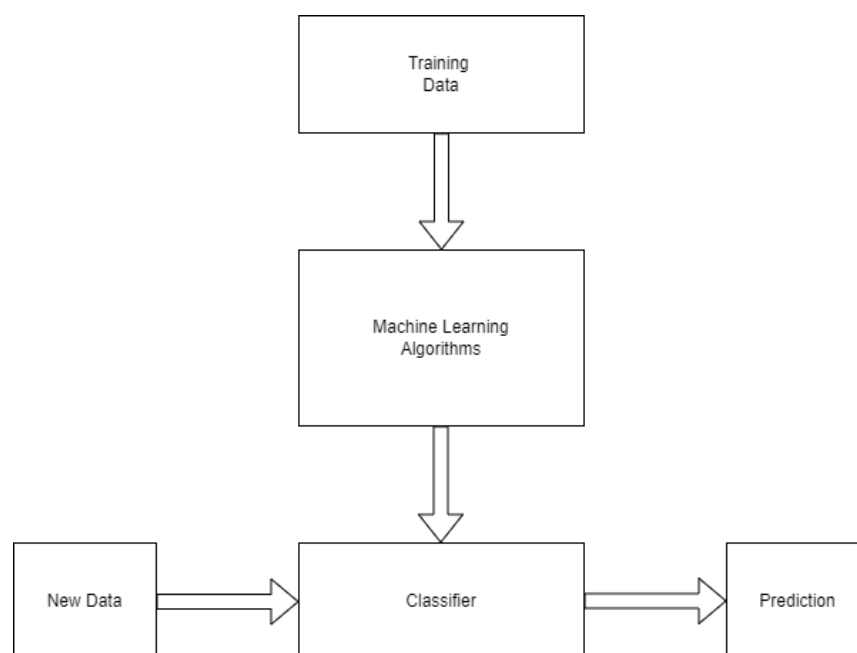


Figure 5.21: Machine Learning

TYPES OF MACHINE LEARNING

Machine learning, depending on the result and the way the learning system works and is fed back, is divided into four types, which are:

- Supervised learning
- Unsupervised learning
- Semi-supervised learning
- Reinforcement learning

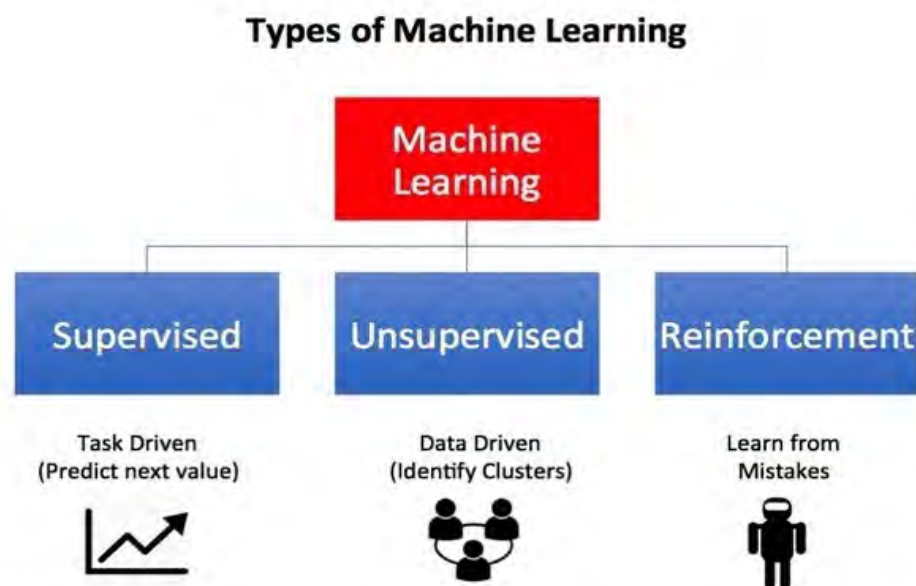


Figure 5.22: Types of Machine Learning (Heidenreich, 2018)

Supervised Learning

The term supervised learning refers to the process of feeding an algorithm with records in which an output variable of interest is known, and the algorithm tries to "learn" how to predict the value with new records where the outcome is unknown. That is, they model a response variable based on one or more explanatory variables (input variable). The data set is a collection of input and output pairs of the form (x_i, y_i) , $i = 1, 2, \dots, n$. x_i is called a feature vector, where each value of the vector describes the i -th element of the data set (record). y_i is the "label" and is usually an element from a set of categories or a real number, but can also be a more complex structure (vector, graph, etc.). x_i is the set of independent variables and y_i is the dependent variable. Practically, the model with a set of training data finds relationships between the values of the feature vector, since it knows what the desired output is. After the training process and after having satisfactory results, the model should be able to predict the category (or label) of an unknown record, with a high probability, correctly. Algorithms based on this learning are classification and regression algorithms (Burkov, 2019; Heidenreich, 2018).

Known supervised learning algorithms/techniques are:

- Decision Trees
- Support Vector Machines (SVM)
- Naïve Bayes
- K-Nearest Neighbors (KNN) etc.

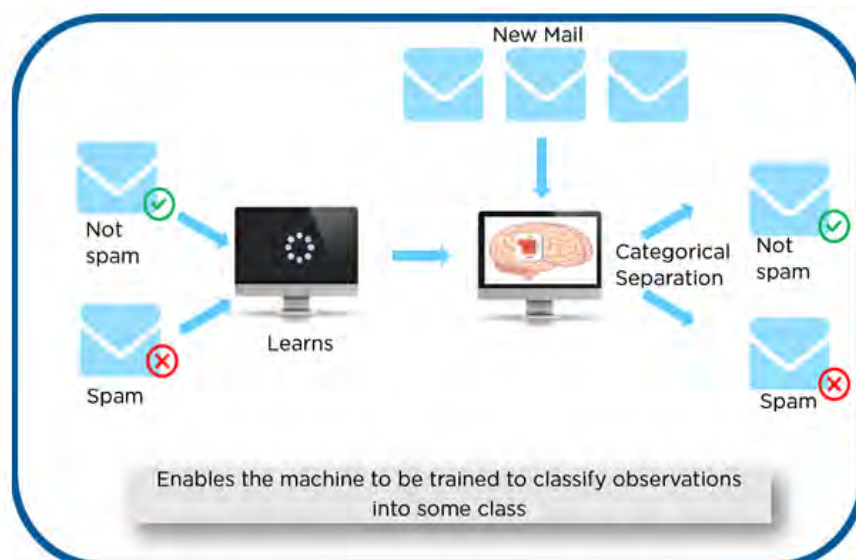


Figure 5.23: Example of Supervised Learning. Email categorization (Heidnreich, 2018)

Unsupervised Learning

Unsupervised learning is used when there is no response variable to predict or classify. More specifically, the term unsupervised learning refers to the analysis that someone attempts to learn something else about the data beyond predicting the value of a variable of interest, such as, for example, whether it belongs to a cluster. Unsupervised learning techniques are used when there is no domain to predict but rather, the relationships between the data are explored to discover their general structure.

In other words, it is automated generation of new knowledge, where the desired outputs for the training set are not known and the model is generally built for some set of inputs with the aim of finding the structure of that set. Indicative algorithms of this class of learning are:

- The Apriori algorithm
- K-means
- Density-based Spatial Clustering of Applications with Noise (DBSCAN)
- Autoencoders
- Local Outlier Factor etc.

Semi-supervised Learning

In semi-supervised learning, the data contains both labeled and unlabeled data. The number of labeled data is usually much higher. The purpose of semi-supervised learning is the same as supervised learning, with the

difference that using a lot of unlabeled data leads to the creation of a better model, for some cases (Burkov, 2019; Heidenreich, 2018).

Reinforcement Learning

Reinforcement learning is a technique where the model is influenced by the environment. The model perceives the state of the environment as a vector of features. The model is able to act. Each action of the model is rewarded or punished, and can change the state of the environment. The purpose of the model is to perform the best action, in each state of the environment. Reinforcement learning is used in computer games, robotics, simulations, and resource management systems. Indicative reinforcement learning algorithms are:

- Monte Carlo
- Q-Learning
- SARs
- DQN

APPLICATIONS OF MACHINE LEARNING IN SEARCH ENGINES

Machine Learning can be applied in various areas related to search engines. Some applications of machine learning in search engines are presented below and the more interested ones are then explained:

- Pattern detection
- Identifying new signals
- Custom signals based on specific query
- Image search to understand photos
- Identifying similarities between words in a search query
- Improve ad quality
- Query understanding
- URL/Document understanding
- Search features
- Crawling
- User classification

- Search ranking
- Synonyms identification/Query expansion
- Intent disambiguation

Pattern Detection

Search engines use machine learning to find patterns that can be used to identify spam or duplicate content. Although there are still human quality raters, machine learning has made it possible for Google to automatically scan web pages and pick out low quality ones without having to examine them first. Machine learning gets more accurate as it analyzes more pages.

Identifying the meaning of words based on their usage

Search engines might not be able to provide an exact definition of a word or phrase if it is extremely new and has not yet gained widespread usage. The machine learning algorithms of a search engine will gather data and attempt to determine the meaning of a word or phrase when more and more people start using it and it is printed in numerous places on the internet. Then, the search engine can eventually comprehend it perfectly.

Eliminating spam and low-quality contents from search results

Search engines use machine learning to detect duplicate, spam and low-quality content. Common characteristics of such content are several outbound links that actually lead to unrelated pages, usage of stop words and synonyms in abundance, etc. In order to give users more relevant content and improve user experience, search engines work to remove such contents from their search results. The amount of human work needed to spot low-quality information has been significantly decreased thanks to machine learning. Although there are still human quality raters, overall human involvement has drastically decreased.

Image search to understand photos

An example of this application is Google Image Search where users can upload photos and receive details about the image or similar looking images, etc. On the internet, there is a ton of data in the form of photos. As there are so many photos, search engines can employ machine learning in order to power their image searching functionality.

Improve ad quality

The advertisements that search engines display on their websites are a major source of income for them. Users are more likely to actually buy the offered product or service when appropriate adverts are suggested to them, which benefits both the company providing the advertised product or service and the search engine that

placed the advertisement. Both require payment. Search engines utilize machine learning to pinpoint the correct target audience for displaying various advertisements. Relevant adverts are displayed to the user in response to the search engine queries that he or she enters.

Query understanding

Machine learning is used to decipher the user-typed search queries. One issue that machine learning helps to overcome is query classification. Different classifiers are used to the search query by search engines; hence, search queries are divided to:

- Navigational search queries
- Informational search queries
- Transactional search queries

URL/Document understanding

This feature is about every action that can be done in order to understand a URL (Uniform Resource Locator). For example, spam detection, page classification, etc.

Search Features

Machine Learning is used in order to generate search features like site links, related searches, knowledge graph data, etc.

User classification

User classification is a feature that concerns users and the process of trying to figure out what kind of a user a person is. This feature is mainly useful for personalized search.

Understanding user queries

Every time users enter a question into a search engine, like Google, Bing, etc., it becomes crucial for the search engine to comprehend what users attempting to ask. A search engine which cannot understand what users are trying to ask or cannot understand a query clearly, will be useless, as it will not be able to provide users with the right responses. Using machine learning is significant in this situation. While typing their requests into search engines, users are possible to make spelling mistakes. It is impossible to assume that a user will spell every word correctly. A search engine will display the proper spelling of a word in case users enter it incorrectly. Even if users make a spelling error, the search engine is intelligent enough to recognize the term they are typing. So, search engines use machine learning for spelling correction.

Synonyms Identification

The best engines can correctly respond to users' searches in case they use synonyms, even the least common ones. Search engines can figure out what users are looking for. In this case, machine learning is also applied. Users occasionally ask questions that are a little vague. An effective search engine should be able to recognize the uncertainty and find a solution. Machine learning is used in this case too. A search engine can also categorize a query into one of the other categories, such as whether it is informational, transactional, navigational or belongs to any other category. The classification of the queries is done using machine learning. The search engine may provide the appropriate supplementary information depending on the category of the queries. For instance, if users type "Westminster Abbey", they will find information about it and also see its location on Google Maps.

Intent disambiguation

For better understanding this feature, an example is necessary. If there is a user who searches for eagles, the search engine should be able to figure out if the query concerns the band eagles, or Philadelphia eagles or the bird eagle or all of them together. Machine Learning is required in these types of scenarios.

Search Ranking

The term "ranking" describes how a website or page is displayed in search engine results. A position in a search engine is a typical term for a webpage's ranking there. The position in which a specific website appears after conducting a search is known as search engine ranking. Ten websites are normally listed on each page of the search results, though occasionally videos, photos, and local listings are also included. In fact, a higher ranking in the search results is correlated with a lower number, whereas a lower ranking is correlated with a greater number. Since websites that are ranked higher often obtain a larger percentage of click-throughs and more visitors are attracted by them compared to lower-ranked websites, many website owners run SEO efforts to raise their search engine rating and move their website closer to the top of the results. The ranking of a website in a search engine is affected by a wide range of variables, including its age, the caliber of its link profile, the relevance of the page, social signals, and amount of competition among others. Instead of ranking an entire website, search engines rate specific pages. This implies that a deep internal page might be listed on the third page, while the homepage might rank #1 for specific keywords (Ann Sunny, 2012).

5.4 DATA MINING

Today, a huge amount of data is generated from various sources, natural and technological, as well as human. Most of this data is in raw form (raw data). The recording of this data can come from measurement collections, mobile devices, social networks, sensor data, tracking data, querying, web browsing, etc. The low cost

of storage has allowed the preservation of this data in a primary form, in order to carry out knowledge through their analysis.

Data Mining is used to address the challenges of managing this data and uncovering hidden patterns (pattern recognition) and encoding the information collected. Data is extremely useful, provided its collective intelligence is harnessed effectively. Essentially the raw data is in a form that the human mind, with its limited analytical ability, cannot extract useful information from. Data mining is a solution to this problem. This is achieved with the help of techniques and algorithms applied by data mining, in combination with other sciences (statistics, machine learning, pattern recognition, etc.), to analyze the data.

DATA MINING AND KNOWLEDGE DISCOVERY

Data mining is an integral part of knowledge discovery from data (KDD), which is the process of transforming raw data into useful information. This process consists of a series of transformations, from pre-processing the data to post-processing the data mining results.

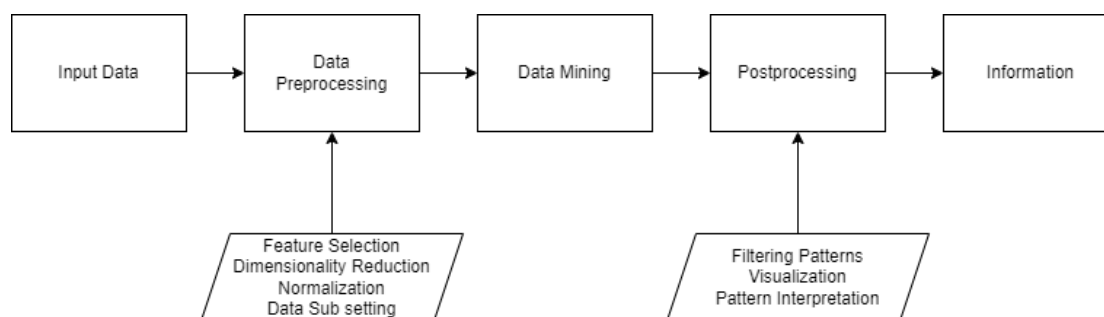


Figure 5.24: Process of discovering knowledge from Data

First, it should be noted that the data is available in a variety of formats and may come from different categories. Also, depending on the format of the analysis and the way it is designed, it is possible that they should be used in more than one data category. For example, if one considers the example of data mining from a relational database, then only the data that exists recorded in the database is used, which in the vast majority comes from common categories. Whereas in the case where mining is to be done from systems designed to operate on the internet, things are completely different, as there is a huge amount of data with a different category each.

Therefore, the purpose of pre-processing (or preparation) is to transform all existing data into a suitable form to assist the subsequent operation of the mining algorithm. This goal, in some cases, is easily achievable, while in some, more modern, approaches it is particularly complex, due to the variety and unevenness of the data. Some basic aims are to eliminate the noise contained in the data, to fit all elements in a common system in a certain way, and others.

Then, after running the mining algorithm, there is further analysis. This analysis mainly aims to make the data as friendly and understandable as possible to the user, who will use it and ultimately receive the information. More specifically, it states that the information created is intended to feed another system, such as a recommendation system or a decision support system, and generally to help analysts easily interpret patterns and correlations between data. In order for such a thing to be possible, there must be a presentation of the correlations in a simple and comprehensible way for humans. Therefore, at this stage, statistical techniques, visualization diagrams and other tools are used, which are generally proven to have a better effect on user understanding.

METADATA

Metadata is data that describes or explains other data, making it easier to retrieve, process and manage information. They are stored in databases and usually aim to extract other more important and exploitable data. They facilitate information extraction and can even help organize content.

There is metadata associated with search engine results. SEO (Search Engine Optimization) is a service whose demand is increasing rapidly. More and more owners of websites or online stores request that their website or their online store appear higher in the search engine results.

Metatitles, metatags, metadescriptions and many other elements are extremely important to be declared correctly in order to achieve better SEO. In other words, these are elements, which may not be immediately apparent to the user, but are nevertheless the critical information based on which Google classifies the web pages in the results of a search. Search engines use this type of metadata, trying to "serve" the user results that are very close to what the user is really looking for.

PRACTICAL DIFFICULTIES

Data analysis techniques often face some difficulties, which can be handled by data mining. These difficulties can come from the huge amount of data and its great diversity and are presented below:

- **Scaling:** Data mining algorithms are capable of handling the size of data sets (gigabytes, terabytes or even petabytes).
- **High dimensionality:** The increase in the size of the data also results from the multitude of characteristics of the data that can be measured in hundreds or thousands (e.g., temporal or spatial components) or from the high frequency of data collection (e.g., temperature measurements).

- Heterogeneous and complex data: Data sets usually contain different types of features, either continuous or categorical. The nature of the objects (e.g., collections of web pages containing semi-structured text and hyperlinks, climate data consisting of time series of measurements, etc.) as well as the need to record the relationships between them (e.g., temporal / spatial autocorrelation, relationships parent-child) accounts for increasing data complexity. It is important that data mining can handle different types of features.
- Data ownership and distribution: In data analysis there is often the issue of ownership or transfer which is a problem as the data is either stored in one location or not owned by an organization or is geographically distributed between resources owned by multiple entities. Some issues that need to be addressed are:
 - reducing the amount of communication required to perform distributed computations,
 - the efficient consolidation of data mining obtained from multiple sources, and
 - dealing with data security.
- Non-traditional analysis: Traditionally, the analysis process is based on the hypothesis-and-test statistical approach. This process is extremely energy-intensive and today requires automated hypothesis generation and evaluation processes and may involve the analysis of temporal data.

CATEGORIES OF DATA MINING SYSTEMS

The field of data mining has expanded enormously in recent years, since more and more sciences are making use of its techniques, with the aim of improving the decisions made and generally exploiting the possibilities provided by the abundance of existing data. It is therefore natural that the great appeal has led to the creation of different requirements, which in turn lead to different design and operation of new algorithms. So, a point has been reached where there is a plethora of algorithms available. These algorithms are quite different from each other and each of them suits different classes of data. However, a primary division into two basic types of models is possible, whose individual goals are satisfied by different classes of algorithms.

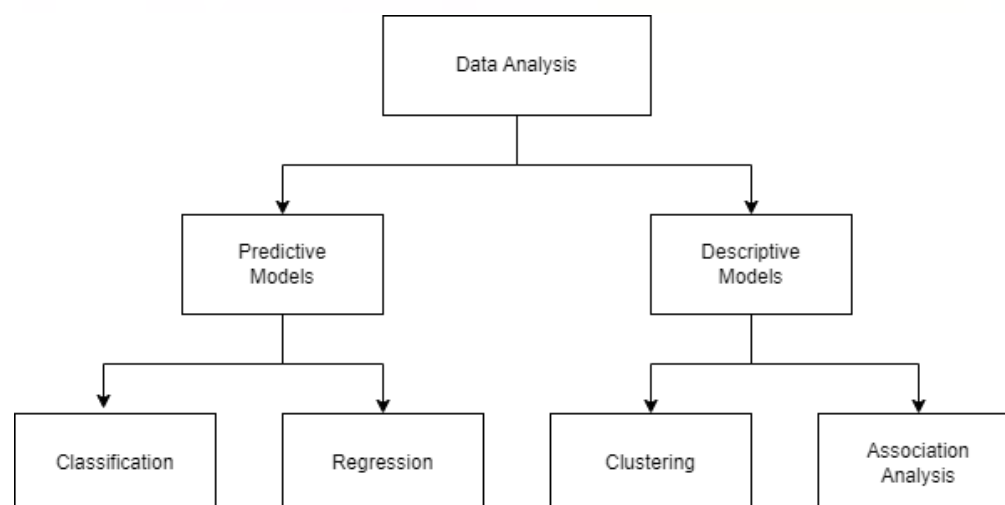


Figure 5.25: Data Mining and Analysis Techniques

Regarding the division in terms of the type of models, there are two main types, the first is made up of models that predict (predictive models) and models that describe (descriptive models - Tan, Steinbach, Karpatne and Kumar, 2006). These two types differ mainly in terms of their operating objective. As the respective names suggest, the first models try to predict values and requirements of certain basic characteristics under consideration, while the second ones work with the aim of combining data in order to create patterns and relationships between the data and its properties.

The above separation falls at a very initial level of analysis, while on a more technical level it can be noted that there are certain categories of algorithms, which perform specific separations and operations between the data and their properties. These categories satisfy both the goals of descriptive and the goals of predictive models. Additionally, at this point it is stated that there are too many categorizations and correspondingly too many algorithms that fall into each of them.

Briefly it is stated that these categories are classification, regression, clustering and association analysis.

DATA MINING METRICS

Data mining metrics are a set of measurements that help determine the effectiveness of a data mining technique or algorithm. They are used to find the best possible technique. Each type of data mining is different and has its own metrics. To evaluate an outcome, in many cases, a single metric may not be enough. In these cases, to achieve the accuracy of the assessment, multiple metrics are used. Choosing the right metrics for any data mining application is a critical part.

Data mining metrics are divided into the following categories:

- Accuracy

- Reliability
- Utility

Accuracy

Accuracy is a measure of how effectively a model can relate a result to the characteristics of the data. Accuracy metrics are different and depend on the data used. In reality, however, there are cases of missing values, or their change after multiple processing. In the process in which the model is researched and developed, there is an acceptance of a margin of error, especially in the case of data with uniform characteristics. There is the case of combining accuracy metrics with reliability metrics.

Reliability

Regarding reliability, this is a metric in which a data mining model is evaluated on different datasets. The reliability of a data mining model is achieved when it produces the same type of predictions, or generally the same kinds of patterns, regardless of the test data provided.

Utility

Utility is a category of metrics that includes various metrics that inform whether useful information is provided by the model. For example, some models may not be useful because they lack the ability to explain the phenomenon or generalize to other data sets, even if the models are accurate and reliable.

Measuring effectiveness or usefulness is sometimes a difficult process. In fact, different techniques could use different metrics and also based on what and how useful a model's output is. For example, when it comes to entrepreneurship, it is the return on investment (ROI) that is useful. The ROI metric looks at the difference between the cost of the model and the benefits of using the model. Of course, quantifying the return is a difficult process. One way to measure it is by watching for increased sales, or decreased advertising spend, or both.

DATA

Data mining requires the "preparation" of the data. Data in the real world is usually noisy, huge in volume and can come from different sources. In data mining the first and most important step is to pre-process the data, which requires knowledge about it. In the data analysis process, it is important to clarify the attributes that make up the data, the values that the attributes take, which attributes are discrete and which take real values, how the values are distributed, if there are ways to display the data, if there are outliers, if there is a way to measure similarity between some records.

Entries and Attribute Types

The records that make up datasets are entities, which are described by their attributes. Other names that may be encountered as records are samples, examples, snapshots, data points, or objects. In some types of files (e.g., csv files), or in a database, records are represented by rows, while their attributes are represented by columns.

Depending on the field of science and literature, the attribute is also referred to as a dimension or variable. In Machine Learning, as well as in data mining, the term attribute is preferred, in statistics the term variable, while in Data Warehouses the term dimension. Records are described by sets of features, called feature vectors. For example, a customer may have an identifier, name, and address, which are feature vectors.

The type of each attribute depends on the values that the attribute takes and can be:

- Nominal - Categorical
- Binary
- Ordinal
- Numeric

Categorical attribute

A categorical attribute contains values, which can be symbols or names. The values are not sorted and each value represents some kind of category, code or status. In computer science, categorical attribute values are also known as enumerations (Han, Kamber and Pei, 2012).

In addition to symbols and names, values may also be represented by numbers. However, at these values mathematical operations do not make sense, so the numbers are not used quantitatively.

In categorical values it is meaningless to find the mean or median for an attribute, given a set of data, since, as mentioned, the classification is meaningless. However, it is interesting to calculate the most common value of the feature, as it is important for measures of central tendency (Han, Kamber and Pei, 2012).

Binary attribute

A binary attribute is a categorical attribute that has only two states: 0 or 1. 0 represents the absence of an attribute and 1, on the contrary, represents the presence. Binary attributes are also referred to as Boolean in case the two states they contain correspond to true and false.

In case the two states of a binary attribute are equally valuable, i.e., the result must not necessarily be coded to 0 or 1, then the attribute is called symmetric. An example of such a case is gender. Conversely, there is also the asymmetric binary feature in which the states are not equal, such as the positive and negative result of a medical test for the coronavirus. Conventionally, the most significant outcome is coded as 1, which is usually the rarest, and the other as 0 (Han, Kamber and Pei, 2012).

Hierarchical attribute

An attribute in which the values have a rank to each other, but the magnitude between successive values is not known (e.g., small – medium – large), is a hierarchical attribute. Hierarchical features are used to register ratings that cannot be measured objectively. Therefore, the use of the specific features is mainly in questionnaires for evaluations.

Categorical, binary, and hierarchical values are qualitative measurements, that is, they describe a characteristic of an object without giving an actual size or quantity (Han, Kamber and Pei, 2012).

Numerical attribute

An attribute, which is quantitative, is a numerical attribute. This means that it is a measurable quantity that can be represented by integer or real values and allows arithmetic operations to be performed between two values. Numeric features can be expressed as interval-scaled or ratio-scaled.

Equally spaced features express a measurable quantity, which is measured by equally sized intervals. In this type of scale measurements, the values can be negative, zero or positive. So, in addition to being able to rank, it is also possible to quantify the difference between two features. Examples of interval values are temperature and IQ.

The analog measurement scale is a numerical characteristic that has an inherent zero point, which means that it can only take non-negative values. In this measurement, one value can be characterized as a multiple of another (ratio). In addition, the prices can be sorted and there is the possibility of calculating the difference between prices, average price, prevailing price, etc. Typical examples are weight, age, years of experience, height, coordinates, etc.

Data Quality

Data mining applications are often applied to data that was collected for another purpose or for future but unspecified applications. For this reason, data mining cannot usually take advantage of the significant benefits of "addressing quality issues at the source". Rather, the design of experiments or surveys, in which a predetermined level of data quality is achieved, is the preoccupation of a large part of statisticians. Data quality problems are

usually unavoidable. For this reason, data mining relies on detecting and correcting data quality problems, as well as using algorithms that tolerate poor data quality. The first step, detection and correction, is cleaning the data.

Measurement and Data Collection Issues

It is impossible for data to always be in perfect condition. Human errors, limitations of measurement devices, or flaws in the data collection process are some of the problems that may exist. Also, values or entire data objects may be missing, while in other cases, the objects may be fakes or duplicates, i.e., multiple data objects that all correspond to a single "real" object. For example, for a person who has recently lived at two different addresses there may be two different records. Even if all the data is there and clear, inconsistencies are still possible.

Measurement and Data Collection Errors

Measurement error is any problem that may arise from the measurement process. It is possible that the value recorded is different from the actual value, which is a common problem. For continuous properties, the numerical difference between the measured and the true value is called an error. In data collection, an error can be the omission of data objects or attribute values, or the inappropriate inclusion of a data object. Measurement errors and data collection errors are divided into systematic and random.

Noise and Artifacts

In a measurement error noise is a random component which may include distortion of a value or addition of spurious objects. The term noise is often used in relation to data that has a spatial or temporal component. To reduce the noise in such cases, techniques from signal or image processing are used, which help to discover patterns (signals) which may be "lost in the noise". Eliminating noise is a difficult process. For this reason, data mining focuses on devising robust algorithms that produce results that, even in the presence of noise, are acceptable. A streak at the same point in a set of photos is a data error that is the result of a more deterministic phenomenon. Such deterministic distortions of the data are often referred to as artifacts.

Validity, Bias and Accuracy

In statistical and experimental science, the quality of the measurement process and the resulting data is measured by validity and bias. Validity is the closeness of repeated measurements (of the same quantity) to each other. Bias is a systematic quantity that is measured. Typically, validity is measured as the standard deviation of a set of values, while bias is the difference between the mean of the set of values and the known value of the quantity being measured. Determination of bias is limited to objects with a measured quantity known by means other than the current state. For the degree of measurement error in the data, the term "accuracy" is commonly used.

Accuracy refers to the closeness of measurements to the true value of the measured quantity. It is characterized by its dependence on validity and bias, but as it is a general concept, with respect to these two quantities, there is no specific formula. A special feature of precision is the fact that it uses significant figures. The basic purpose is to use only as many digits to represent the result of a measurement or calculation as are justified by the validity of the data.

In data mining, significant figures, validity, bias, and precision are often ignored, but that doesn't mean they aren't important to statistics and science. A frequent phenomenon is that data sets are not accompanied by information about the accuracy of the data, and furthermore, the programs used for the analysis return results without any such information. However, it is possible for an analyst to make significant errors in analyzing the data if they do not understand the accuracy of the data and results.

Extreme Values

Outliers are either data objects whose characteristics differ from most of the other data objects in the dataset, or values of a characteristic that, relative to the typical values for that characteristic, are unusual. Definitions of outliers, proposed by the data mining and statistics communities, are many and varied. Also, the concept of noise should not be confused with that of extreme values. Outliers may be legitimate objects or data values, so they are of some interest relative to the noise. For example, fraud and network intrusion detection efforts focus on finding unusual objects or events among a large number of normal ones.

Absent Values

It is possible for an object that one or more attribute values are missing. This can happen in cases where information was not collected, such as when some people did not want to give their age or weight during a study. Also, in other cases, certain attributes may not apply to all objects. For example, often when filling out a form there are parts that only need to be filled in if the person answered a certain way to a previous question. However, all fields are stored as normal for simplicity. Regardless, missing values should also be taken into account when analyzing data.

There are several strategies (and variations of these strategies) to deal with missing data, each of which may be appropriate in specific situations. Some of them are described below.

Delete data objects or attributes

One of the strategies to deal with missing data is to eliminate data objects or attributes that contain missing values. However, data objects contain specific information, and in case there are many objects with missing values, then the analysis will not be as reliable. Conversely, if the objects with missing values are not that many, then it might be better to skip them. In addition to eliminating objects, there is also the elimination of

attributes that have missing values. However, there is the case that omitted features play an important role in the analysis, so elimination should be done with great care.

Estimate missing values

There are times when a reliable estimate of missing values can be made. For example, in a data set that contains many similar data sets, the feature values of points that are close to the point with the missing value are usually used to estimate that value. In case the feature is continuous, the average feature value of the nearest neighbors is used. If the attribute is categorical, then the most frequently occurring attribute value is used.

Ignore missing values during analysis

In data mining many approaches can be modified so that missing values are ignored. For example, suppose objects are being grouped and a calculation of similarity between pairs of data objects should be performed. In case in a pair one or both objects have missing values for some features, then the similarity calculation can be done using only the features that do not have missing values. Of course, the similarity will have been approximated, but unless the total number of features is small or the number of missing values is large, this uncertain estimate may be insignificant. Similarly, many sort schemes can be modified to work with missing values.

Inconsistent Values

It is possible that the data contains some inconsistent values. For example, in an address field that contains the zip code and city, the specified zip code area may not exist in that city. This may be due to, for example, human error where a person has entered the information incorrectly. Regardless of the causes, inconsistent values should be identified and problems that have occurred corrected as much as possible.

Some types of inconsistent values can be easily detected. For example, one can easily understand that a person's height should not have a negative value. It may also, in another case, require an outside source of information to intervene.

When an inconsistency is detected, there is the case for timely correction of the problem in the data. For example, a product code may have "check" digits, or a product code may be double-checked against a list of known product codes, then corrected if the code is incorrect but close to a known one code. When an inconsistency is corrected, additional or redundant information is required.

Duplicate Data

A data set may contain data objects that are duplicates or near duplicates of each other. There are cases where duplicate submissions are received from people who have almost the same name and this is contained in a

database. In order to identify and eliminate such duplicates, two critical issues must be addressed. First, if a single object is actually represented by two objects, then the values of the corresponding attributes may differ, and there should be a resolution of these inconsistent values. The second issue concerns data that is similar but not duplicated, such as when two different people have exactly the same names. In this case, these data objects should not be mixed randomly. A term often used when dealing with these issues is the term “deduplication”.

There are cases where two or more objects are identical with respect to the attributes measured by the database, but still represent different objects than them. In this case, duplicates are legal, but there are algorithms that can run into problems if the fact that identical objects can exist is not taken into account during their design process.

Application-Related Issues

Data quality can also be considered during the implementation process, as expressed by the statement "data are of high quality if they are suitable for their intended use". In business and industry, this particular approach to data quality has proven quite useful. This approach is also prevalent in statistical and experimental sciences, with an emphasis on the careful design of experiments to collect data relevant to a particular hypothesis. As with quality issues at the level of measurement and data collection, there are many issues that are specific to specific applications and fields.

Timeliness

Data, from its collection onwards, begins to "age". This is especially the case if the data concerns phenomena or processes that are constantly evolving or changing, such as customer purchasing behavior or web browsing patterns. In fact, this data represents a snapshot only for a specific and limited period of time. The result, therefore, of the existence of old data is the existence of old models and patterns based on that data.

Relevance

An application should contain the necessary information from the available data. For example, suppose a model is built that has the ability to predict accident rates for drivers. If information related to the driver's gender and age is not included, then the model may be inaccurate, unless such information can be made available indirectly through other characteristics.

There is also the difficulty of ensuring that in a data set, objects are related. Sampling bias is a common problem. This occurs when a sample does not contain different types of objects in proportion to their actual occurrence in the population. For example, a survey contains data that describes only those individuals who participate and answer the survey questions. An incorrect analysis can usually result from sampling bias, as the data present may only be reflected by the results of the data analysis.

Knowledge of Data

Ideally, data sets should be documented, containing descriptions of their different aspects. Of course, the existence of documentation may not help or facilitate the analysis that follows. For example, if several features are identified by the documentation as closely related, then there may be a lot of redundant information contained in those features, and so it may be decided to keep only one. In case a poor documentation has been done, then the analysis of the data may be incorrect. Also, important characteristics are the precision of the data, the type of characteristics (nominal, ordinal, interval, ratio), the scale of measurement (e.g., meters or feet for length), and the origin of the data.

Measures of similarity and dissimilarity

There are applications in data mining, such as clustering or outlier analysis, in which a way to evaluate the similarity of two objects is necessary. An example is a store, which may wish to group its customers who have similar characteristics (e.g., income, age, residence) for marketing purposes. In the case of outlier analysis, the aim is to find objects that do not look like other objects, so that possible frauds can be identified.

There are several approaches to measuring similarity between two objects. One of them is to consider the objects as vectors in n -dimensional space and to calculate their similarity as the cosine of the angle they form:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

where $x \cdot y$ the inner product of the vectors and $\|x\|$ the Euclidean norm of the vector x . This similarity is known as the cosine similarity.

Another approach to similarity between objects is their association. One correlation method is the Pearson correlation. Given the covariance Σ of the data points x and y and their standard deviation σ , calculate the Pearson correlation as:

$$\text{Pearson}(x, y) = \frac{\Sigma(x, y)}{\sigma_x \sigma_y}$$

Other examples of approximations are Minkowski distance, Euclidean distance and Manhattan distance.

DATA MINING PROCESS

As for how data mining systems work, it is very specific and clearly defined. More specifically, there are recorded algorithms that are applied depending on the desired final result, the format and nature of the data and other factors, while before and after the operation of each algorithm, there are the pre-processing and post-processing procedures of the data, until they reach such a form that they are capable of providing knowledge either to the system or to a user.

Others see Data Mining as a step in the knowledge discovery process, which consists of the following steps:

1. **Data Cleaning**, to remove noise, incorrect data and information that is unnecessary.
2. **Data Integration**, where multiple data sources can be combined.
3. **Data Collection and Selection**, where the data related to the analysis process is retrieved from the database.
4. **Data Transformation**, where data is transformed or aggregated into forms suitable for mining, performing summarization or aggregation operations for example.
5. **Data Mining**, the process of using intelligent methods to discover patterns.
6. **Interpretation / Evaluation of Patterns**, to recognize the really interesting patterns that lead to the creation of knowledge. The evaluation is based on some indicators that show the effectiveness, and they are reliable and supportive.
7. **Presentation of Knowledge**, where presentation techniques are used, data is visualized and results are most often presented in the form of diagrams.

There is the case of repetitions of some of the stages of the discovery process, and in fact, several times. Selection, preprocessing, and transformation are steps that are often referred to as data preparation steps. At each process stage, the effort required changes.

Data Cleaning

In the data cleaning stage, the process starts from the source data, which may be stored in various sources, such as transaction tracking systems, independent databases, independent archives, external sources, etc. In addition, the source data may contain various defects, such as errors, contradictions, missing values, etc. The initial data from the various sources should be collected, homogenized and cleaned. Problematic data containing incorrect, outliers, or missing values can mislead mining algorithms and result in invalid and incorrect patterns being extracted. Some data mining algorithms have the ability to deal with data problems intrinsically. Of course, this is not the case in general and the way to deal with it may not always be the most effective. For this reason, it is preferable for the analyst to undertake data cleaning as an independent task and in a manner controlled by him. Usually, when the data problems are cleared, then it is stored in a Data Warehouse.

Data Collection

When selecting the data, the dataset in which to search for patterns is also determined. The process of selecting the data is quite important, since, through it, the results that will be obtained are indirectly determined.

Many times, the way the data is stored is not compatible with discovery algorithms required to extract the data and organize it into structures that will be accessible to them.

Data Transformation

Before the algorithms are applied, the data preprocessing process is an important and necessary stage of data mining. The collection of data is usually done by processes that can be difficult to extract knowledge from, resulting in noise and quality being negatively affected. Procedures include data collection from questionnaires, measuring devices such as sensors or medical instruments, interviews, observations or the internet. Human errors in entering values or problems arising from devices can cause noise and inconsistent values.

The data is what guides the data mining methods (data driven). This means that every result is drawn directly from the data. For this reason, it is important to choose the appropriate data. Proper selection of data means that the appropriate features or characteristics should be selected first. The feature selection process is a process that directly depends on the task the analyst is performing. There are traits that may be useful for one job and traits useful for another. The analyst is the one who is initially responsible for selecting the features that, in his opinion, contain essential information relevant to his analysis.

The initial subjective selection of features is not enough. In the initial stage, features, which are obviously not relevant to the analysis, are excluded by the analyst. But then the choice may not be so obvious, as the same size may be recorded in different ways.

During the feature selection process, the data transformation process also takes place. For example, numeric values can be reduced to other numeric values, or numeric values can be converted to nominal values. Such tasks are usually performed to adapt the data to the requirements of the analysis methods. For example, in some classification methods there may be a strong influence from fields with large values and a weak influence from fields with small values. In such cases, the orders of magnitude of the values in the different fields should be comparable. From this process, what emerges at the end is a data set that will be used in pattern extraction.

Training Process

During the training process, the technique to be followed is chosen, i.e., the algorithm to be applied. The type of knowledge to be sought is what determines it. Information patterns and prediction patterns are the two categories of knowledge that can emerge. Essential knowledge discovery and data mining are often substitutes for each other. According to the above process, the data mining stage is a single step in the whole process of knowledge mining.

Interpretation of results

After the creation of the model is completed, the results should be evaluated and their significance interpreted. For example, accuracy can be a measure by which to determine which model to choose. In categorization problems, the "confusion matrix" is a good tool to understand the results. In order to better understand and interpret the results, various graphical representations should be created. Therefore, statistical techniques, visualization diagrams and other tools are used in this stage, which generally affect the better understanding of the users.

Presentation of knowledge

Data visualization is the display of information in graph or table form. Successful visualization is done by converting data (information) into a visual form so that data characteristics and relationships between data elements or properties can be analyzed. The goal of visualization is the interpretation of the visualized information by a person and the formation of a mental model of the information.

In everyday life, the preferred approach to explaining the weather, the economy, and the results of political elections are visual techniques such as graphs and tables. Similarly, data mining includes visual techniques that can play a key role in data analysis, even though algorithmic or mathematical approaches are often emphasized in most technical disciplines. Sometimes the use of visualization techniques in data mining is referred to as visual data mining.

DATA MINING AND COMBINATION WITH OTHER FIELDS

Data mining draws methodologies from many scientific disciplines, some examples being Statistics, Artificial Intelligence, Machine Learning, Databases, Search Engines, etc. These disciplines alone make it impossible to extract knowledge from large volumes of data, or the process is very slow and, in some cases, inaccurate. For example, Statistics, although it offers data analysis solutions, does not take into account the large volume of data, similarly, Machine Learning and Pattern Recognition (Kyrkos, 2015). On the other hand, the Database industry can store large volumes of data, but cannot offer data analysis techniques. Data mining combines these disciplines, providing a solution to the problems of each discipline that they alone cannot address.

Statistics

The science of Statistics helps in the analysis of data as it can identify patterns hidden in the relationships between variables in the data. Statistics can provide a mathematical interpretation of data, but this cannot be generalized for very large datasets. Statistics is used in data mining to dig deep into data and gain insights from it.

Statistical techniques are applied to the algorithms used in data mining, which allows the acquisition of new knowledge from the data.

Database Systems and Data Warehouses

Database systems research focuses on the creation, maintenance and use of databases for organizations and end users. In DB systems, well-recognized principles have been established in data modeling, query languages (e.g., SQL), query processing and optimization methods, data storage and, finally, data indexing and access methods, resulting in DBs are known for their high scalability in processing very large volumes of data.

Many data mining tasks have to handle large data sets or even streaming data. Therefore, data mining can make good use of DB systems to produce high-performance and scalable results on large datasets. In addition, data mining can be used to extend the capabilities of DBs and meet the needs of demanding users with exceptional data analysis requirements.

Modern DB systems incorporate data analysis capabilities with the help of data warehouses and data mining. A data warehouse consists of data, which comes from multiple sources and different time frames. It unifies the data in multidimensional space, to form data cubes. Cube data models not only facilitate online analytical processing (OLAP), but also promote data mining on multidimensional data (Han, Kamber and Pei, 2012).

World Wide Web and Search Engines

In recent years with the rapid growth of mobile devices, the World Wide Web has become the most popular way to communicate and disseminate information. New websites, online stores, research articles, educational content and software are created daily. To understand the size and importance of the contribution of the development of the Internet, it should be realized that the amount of information that exists so far on the Internet is impossible to measure precisely. Google's search engine performs 5.6 billion searches per day, with each query in the search engine not exceeding two seconds in time. This is a huge success of data mining, firstly because such a large volume of data is searched in a very short time and secondly because the first results in any query are usually the most useful. Thus, the user quickly and easily receives only the essential information he wants.

Web data mining is considered an extension of the operation performed with traditional keyword-based search engines. Web crawlers play a key role in mining. Web data mining can be distinguished into three categories:

- Web content mining
- Mining from usage data (Web usage mining)

- Mining from the structure data (Web structure mining)

Web content mining examines the content of web pages. Content includes both text and graphical data. Content mining is similar to the process performed by basic information retrieval techniques but goes further than simply using keywords to search.

Usage data mining includes techniques and tools for tracking user requests.

Finally, structure data mining uses graph theory to analyze a web page, the structure of how web pages are connected to each other, and tree structure to analyze and describe the HTML and XML source code of web pages (Hofmann, 2013).

DATA MINING TECHNIQUES AND ALGORITHMS

Data mining software packages today are considered to be automated. However, there is a requirement for users to give some guidelines. The expected data mining algorithm method is one of these requirements. Therefore, for users to use data mining tools, they should have a basic knowledge of these methods. The different types of data mining methods can be classified interchangeably. They are generally divided into six broad categories which are:

- Description of items
- Dependence-correlation analysis
- Classification and prediction
- Clustering – grouping
- Analysis of extreme values (outliers)
- Development analysis

A categorization of the different types of data mining is:

Prediction

- Categorization
- Price Prediction

Compartment

- Clustering - Grouping

Connection Analysis

- Discovering correlations

Discovery of sequence patterns

- Discovery of similar time series

Discrepancy Discovery

- Statistics
- Illustration

Main Categories of Algorithms

Classification

Classification is one of the most widespread categories of algorithms in data analysis. This category mostly satisfies the goals of forecasting models. More specifically, these algorithms try to create the conditions so that any kind of object to be evaluated can be classified into already existing data categories. This classification is based on the characteristics of the specific object and based on how well they fit with other categories. An important issue is how much similarity is considered sufficient to classify an object into a broader group. Also important is the issue that arises in cases where the available data are such that no distinct groups can be formed.

The above issues are resolved, to the extent possible, by each algorithm with different approaches, which will be presented below during the analysis of the algorithms. As can be seen from the above, two separate databases are absolutely necessary for the execution of the algorithms of this category. The first database is used as an auxiliary, with the aim of feeding the model with data and by extension creating the groups (training data set). The second one, is the main base that contains the data, the classification of which is sought.

Additionally, it is worth noting that such algorithms can be used either in databases containing discrete or finite data (classification) or in databases with a continuous type of data (regression) with small variations. However, both algorithms are the same, with minor variations. In conclusion, it is stated that the dominant categories of algorithms, used in the context of classification, are decision trees and neural networks.

Regression

The next major category noted is regression. It is a branch of statistics that examines the relationship between an output value (dependent value) and an input value (simple regression) or more values (multiple regression) (Gorunescu, 2011). Mathematically, a regression model is a function by which the value of the independent variable can be predicted through the independents.

The first form of regression used to find the equation of a straight line is the method of least squares, used by Legendre (Legendre, 1805) and Gauss (Gauss, 1809) to determine the orbit of objects around the sun. However, the term regression was coined by Galton in 1886 (Galton, 1886), where he described the phenomenon that the height of offspring who had tall ancestors tends to fall towards the mean. For Galton, the term regression had only a biological meaning. Later, Yule (Yule, 1897) and Pearson (Pearson, 1903) generalized the term Regression and combined it with statistics. In their publications, the joint distribution of the dependent and independent variables is assumed to be normal (Gaussian). Fisher (Fisher, 1922-1925), in contrast, considered that the hypothetical distribution is normal, but the joint distribution need not necessarily be normal.

Common regression algorithms are:

- Linear regression
- Non-linear regression
- Generalized linear models
- Decision trees
- Neural networks

During the regression process, the model to be estimated is first selected and then the method for estimating the model variables is selected. The models consist of:

During the regression process, the model to be estimated is first selected and then the method for estimating the model variables is selected. The models consist of:

- Unknown parameters, which are usually represented by a vector b .
- Independent variables, which are observed in the data and represented by a vector X_i , where i , the order of the data.
- Dependent variable, which is also observed in the data and is denoted as Y_i .
- Error terms, which are not immediately apparent in the data and are usually represented as e_i .

Most models consider Y_i to be a function of X_i , β , with e_i denoting an additive error term that is either an unmodeled component of Y_i or random statistical noise:

$$Y_i = f(X_i, \beta) + e_i$$

Figure 26 shows a simple example of linear regression.

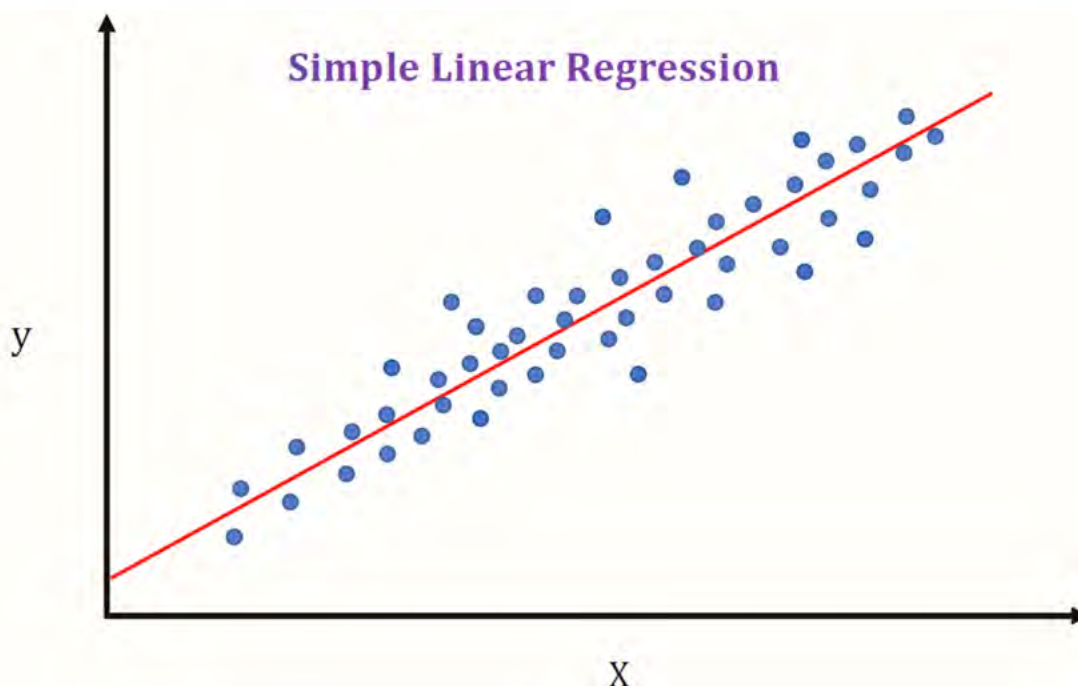


Figure 5.26: Linear Regression (Source: https://miro.medium.com/max/2584/1*Nf2tTTkALYq6RTMQmhjo1A.png)

Clustering

The next basic category is clustering. This particular category contains algorithms, whose operation tries to satisfy the objectives mandated by the descriptive data analysis model. More specifically, the main objective of the techniques of this category is to divide a data set, with inhomogeneities between them, into specific clusters with compact and related data content. In simpler words, from a general set it is sought to create special groups of data. Clustering is a completely different process from classification, as it does not base its operation on the classification of new data into already constructed groups of data, but tries to create these groups of data or clusters, as indicated in the context of data analysis. Under this light, it could be considered that clustering is a necessary process for the proper conduct of the classification, subsequently. In the end, these are two different processes, which, however, must coexist.

There are many clustering algorithms and they have as common the set of data. This fact results in the use of various clustering models, which are divided into the following categories:

- Connectivity models: such as hierarchical clustering that creates models based on distance.
- Centric models: such as the k-means algorithm which depicts each cluster with a simple mean vector.
- Distribution models: in which clusters are modeled using statistical distributions.
- Density models: which consider densely covered regions of the data space as clusters.

- Subspace models: to model the clusters, both cluster members and related features are used.
- Group models: where there is not have a clear model for the results, but information about their grouping.
- Graph-based models: which use cliques to create clusters.
- Signed graph models: each path has a sign resulting from the product of the edge signs. Thus, using the axiom that no circle has exactly one negative edge, there are more than two clusters or subgraphs with only positive edges.
- Neural models: they are usually models that resemble some of the previous ones, with the most well-known being the self-organizing map.

There is also separation based on the number of clusters an object can belong to. If each object belongs to a cluster or not then this is the case of hard clustering, while if each object belongs to each cluster to a certain degree, this is the case of soft/fuzzy clustering.

In general, there are the following types of clustering:

- Partitioning Clustering
- Hierarchical Clustering
- Fuzzy Clustering
- Crisp Clustering
- Kohonen Net Clustering
- Density-based Clustering
- Grid-based Clustering
- Subspace Clustering

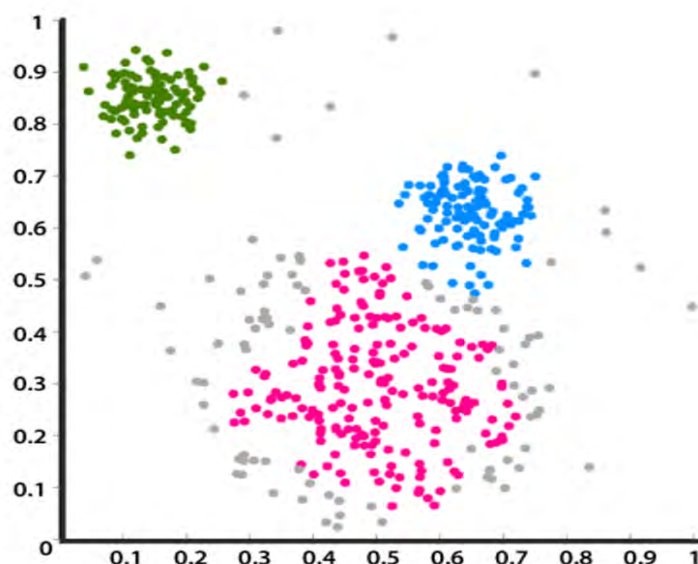


Figure 5.27: Clustering with density model (DBSCAN). (Source: <https://cdn1.byjus.com/wp-content/uploads/2019/10/Cluster-2-2.png>)

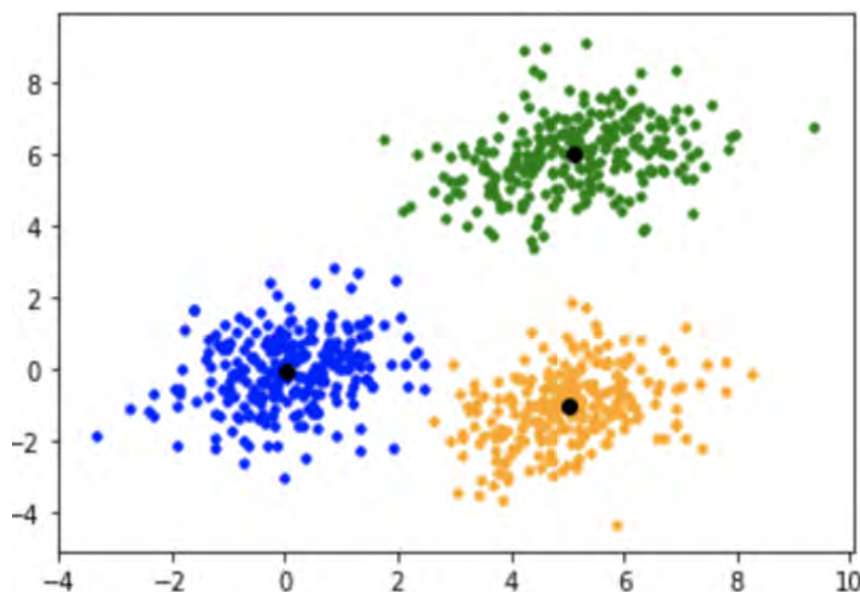


Figure 5.28: Clustering with centroid model (k-means). (Source: <https://media.geeksforgeeks.org/wp-content/uploads/20190812011831/Screenshot-2019-08-12-at-1.09.42-AM.png>)

Figure 5.27 and Figure 5.28, show the clustering results using a density model and using a centroid model.

Evaluating the results is just as complex a process as the clustering itself. Some popular approaches are internal evaluation, in which the clustering is summarized into a quality score, external, in which the results are compared to an existing classification result, manual, which is done by humans, and indirect, in which the clustering is evaluated based on its usefulness in the particular application.

Association Analysis

Association analysis is the next class of algorithms to be presented. This category is also used in the context of achieving the goals of the description models. In addition, from a multitude of works, it is considered quite a modern and useful category and because of this it has gathered numerous applications.

Techniques in this category focus on finding patterns and relationships that you think exist between different categories of data, but are not easy to spot in the first place. It is also quite common to have strong relationships between data that are completely dissimilar to each other, and in fact these relationships are sought to be identified. While, on the other side, the relationships between related, but not identical, groups of data are often found. In contrast, with the non-obvious cases, the patterns between related categories are not so interesting as they do not provide any particular insight into the patterns of association, only the very basic one.

One of the most well-known applications of this category, which was to a large extent the starting point for its establishment, is the analysis of the supermarket basket. According to the specific analysis, the rationalization of the choices made when purchasing products, by various buyers, is sought, while in the second stage an effort is made to create patterns and to generalize them. This tactic was considered to be effective and for this reason, several super markets, mainly in America, have adopted it and are trying to create the layout of both their space and their shelves, following association analysis models.

Data Mining Algorithms

Linear Regression

Linear regression is one of the most popular prediction techniques in statistics and machine learning. It is used to find relationships between one or more independent values and a dependent value. There are two kinds of linear regression:

- Simple linear regression
- Multiple linear regression

Let, a vector $x^t = (x_1, x_2, \dots, x_p)$, where p is the number of variables and x^t is the inverse vector. The prediction of the dependent value y can be calculated using the following formula:

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^p x_j \hat{\beta}_j,$$

In the case that an ace is added to the vector x and the coefficient β_0 is added to the vector of coefficients, the formula can be written as a product of 2 vectors as follows:

$$\hat{y} = x^t \hat{\beta}$$

The most widespread technique for building the linear model on a training set, describing a phenomenon, is the method of least squares (MET-RSS). In this method, each coefficient is selected and the objective is to minimize the relationship. The type is:

$$RSS(\beta) = \sum_{i=1}^n (y_i - x^t \beta)^2,$$

where n is the number of observations in the data set. The function 3 is quadratic, so there is always a minimum value. The equation can be written:

$$RSS(\beta) = (y - X\beta)^t (y - X\beta),$$

where X is an $n \times p$ matrix, with each row of the matrix being a training set vector and y being an n -sized vector containing the output values. Finally, if the product $X^t X$ is an invertible matrix, the equation can be written:

$$\hat{\beta} = (X^t X)^{-1} X^t y.$$

Other well-known linear models are Ridge and Lasso regression algorithms. These algorithms normalize the parameters, and, in the case of Lasso, can eliminate some coefficients altogether. Ridge's normalization is also known as L2, while Lasso's normalization is known as L1 (Hastie, Tibshirani and Friedman, 2009).

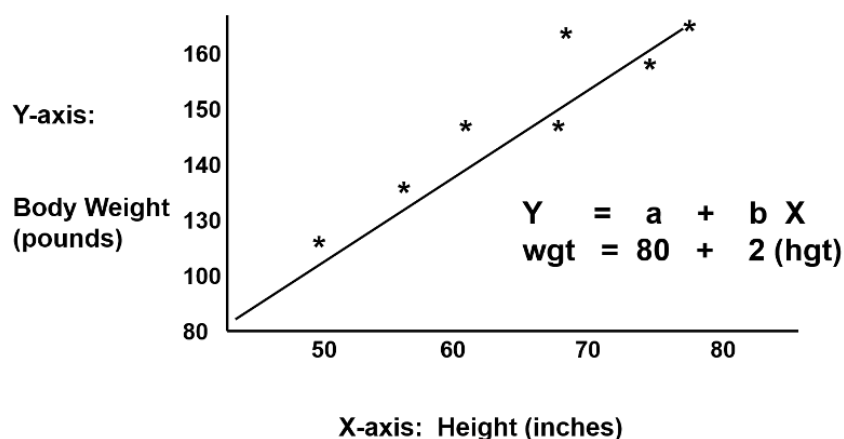


Figure 5.29: Simple Linear Regression Example (https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704-EP713_MultivariableMethods/)

Nearest Neighbors

The K-Nearest Neighbor (KNN) algorithm is a simple supervised learning algorithm that can be used for both classification and regression. With KNN, the assumption is made that similar records have a short distance between them. To use the appropriate distance measure for the given problem, there is a popular approach that is often used and that is the Euclidean distance.

With KNN, the distances between the record are found, and the goal is to also find the label of all the other records in the data set. A parameter that states the number of neighbors to consider to find the label is that of the hyperparameter k and is a feature of this algorithm. In order to find the appropriate value of the hyperparameter k , the algorithm should be run several times, each time with a different value of k , and that value should be chosen which minimizes the number of errors and at the same time preserves the ability of the algorithm to predict accurately.

In the case of a classification problem, it takes the category of the majority of the k nearest neighbors, while in the case of regression the label value is equal to the average of the k nearest neighbors.

The algorithm is as follows:

Step 1. Initialize the hyperparameter k .

Step 2. For each record in the dataset:

1. Calculate the distance between the new record and the current record.
2. Add the distance and index of the record to a collection.

Step 3. Sort the collection containing the distances in ascending order.

Step 4. Select the first k distances, along with the indices of the entries.

Step 5. Find the tags from the nearest records.

Step 6. If this is a regression problem, return the mean of the nearest entries.

Step 7. If it is a classification problem, return the majority of the category from the closest records.

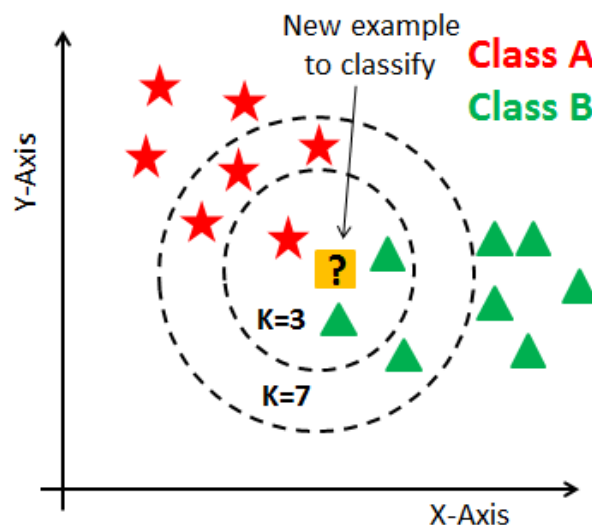


Figure 5.30: Example of the KNN algorithm. (Source: https://medium.com/@sudhanshugupta_66164/k-nearest-neighbor-knn-algorithm-for-machine-learning-1b506eb2c4a4)

Decision Trees and Random Forests

Decision trees in both regression and classification use tree-shaped models. A Decision Tree consists of:

- Root: It is the beginning of the tree that represents all the data.
- Decision nodes: They are the nodes that split into two or more nodes.
- Leaf/Terminal node: Nodes, which are no longer separated. Usually, these nodes are the end result.

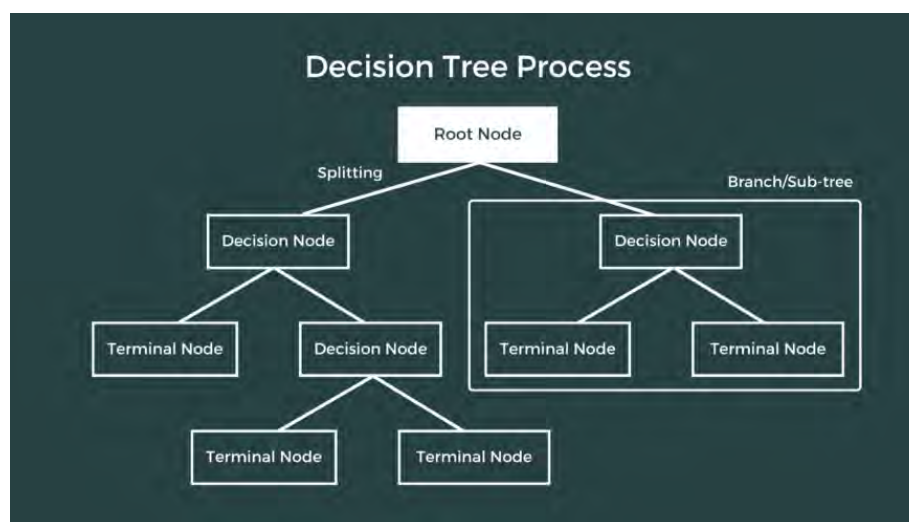


Figure 5.31: Decision Tree Structure (Source: <https://www.theclickreader.com/decision-tree-regression/>)

The node splitting process starts at the root and is followed by a subtree leading to a terminal node containing the final prediction. The construction of the decision tree starts from the top down, choosing at each node a variable that optimally partitions the data set.

At each node the accuracy of the model is calculated with a cost function. Usually, the RMSE (ROOT MEAN SQUARE ERROR) formula is used and is:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}}$$

Random Forests are a set of decision trees, where each decision tree randomly selects some features. This approach helps limit error due to bias and variance, making Random Forests a more robust technique than decision trees (Liberman, 2017).

Artificial Neural Networks

In artificial neural network (ANN) models, the function that simulates the relationship between data is $y = f^*(x)$, where $f(x)$ is the composition of several functions and is of the form $f(x) = (f^{(n)} \circ f^{(n-1)} \circ \dots \circ f^{(1)})(x)$, where $f^{(i)}$ is the i -th level of the network, n is the depth of the network and $f^{(n)}$ is the output of the network. The remaining layers are called hidden layers. The representation of an ANN can be done with a directed acyclic graph, which describes the composition between the functions. The hidden layers are responsible for recognizing the various patterns that may be present in the data. The output layer is driven by the training data so that the results it produces are as close as possible to the actual output. The training algorithm has the responsibility to exploit the hidden layers in such a way that acceptable solutions are obtained.

Among the simplest forms of a neural network is the Perceptron, which was invented by Frank Rosenblatt in 1958. The Perceptron consists of n inputs, one neuron, and one output, where n is the number of features in a data set. There are two phases in an ANN: forward propagation and backpropagation. The process by which data is passed into the ANN is feedforward, while feedback updates the weights.

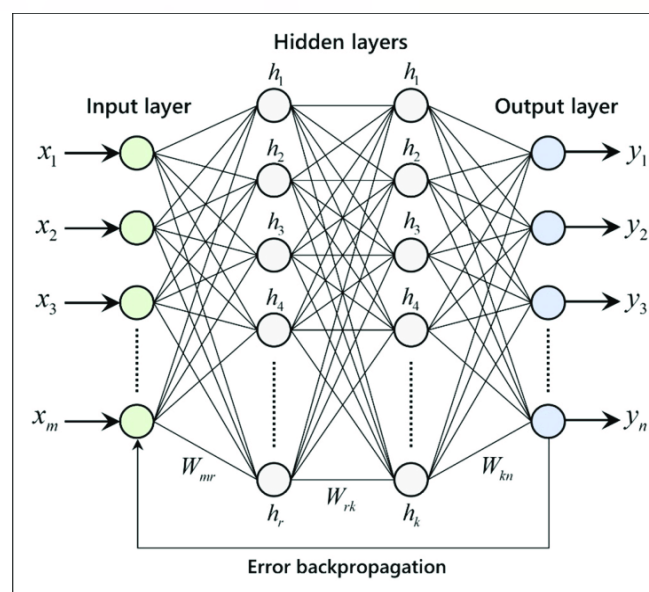


Figure 5.32: A Multi-Layer Artificial Neural Network (Source: https://www.researchgate.net/figure/Architecture-of-multilayer-artificial-neural-network-with-error-backpropagation_fig3_329216193)

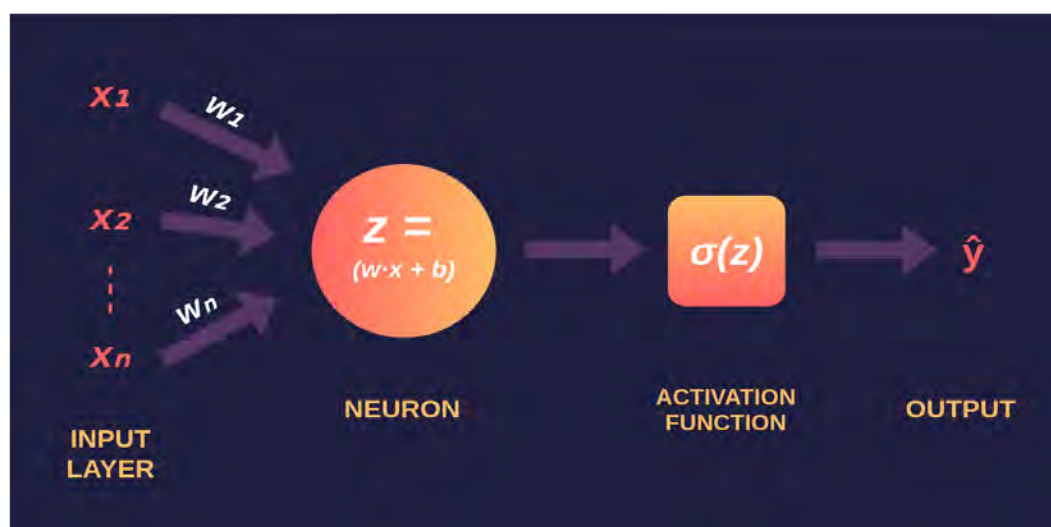


Figure 5.33: A simple artificial neural network. (Source: <https://towardsdatascience.com/introduction-to-math-behind-neural-networks-e8b60dbbdeba>)

Each input x_i is multiplied by the corresponding weight of w_i and the products are summed. Therefore, the relation is $\Sigma = (x_1 w_1) + (x_2 w_2) + \dots + (x_n w_n)$. If x and w are two vectors, where $x = (x_1, x_2, \dots, x_n)$ and $w = (w_1, w_2, \dots, w_n)$, then the sum of the products can be written $\Sigma = x \cdot w$. The strength between neurons is represented by the weights, which already decide how much each feature affects the output of the neuron. That is, the output of the neuron is more affected by a feature with a higher weight value than by one with a lower weight value. Also, a bias value is added to the final sum, from which the activation function is compensated and shifted right or left to produce the desired output values. The relationship created is as follows: $z = x \cdot w + b$, where b is the polarization

value. In the Perceptron case the value of z is passed through an activation function, while when there are more hidden layers, more are passed (one activation function for each hidden layer). The training time of the ANN is contributed by the activation functions. The function most commonly used as an activation function is the logistic (Figure 5.34), also referred to as the sigmoid. Using accounting as the activation function, the forecast ANN is calculated by the formula:

$$\hat{y} = \sigma(z) = \frac{1}{1 + e^z}$$

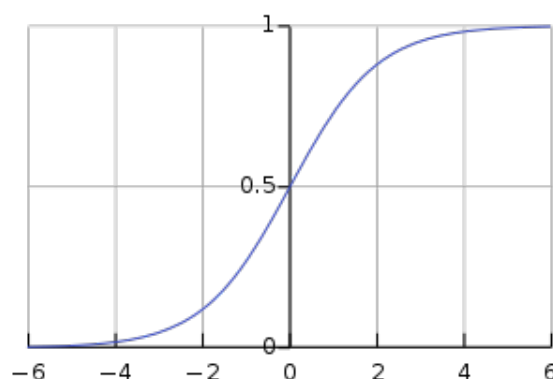


Figure 5.34: Logistic function (Source: https://en.wikipedia.org/wiki/Logistic_function)

After the prediction value has been calculated by the ANN, the neural network feedback process begins, starting with finding the error. One of the most widespread methods of finding errors is that of MET and is a procedure similar to the one mentioned in linear regression. To optimize the training of artificial neural networks, the gradient descent algorithm is usually used. The feedback and derivative descent algorithm are processes that iterate until they converge and the weights are updated as follows:

$$w_i = w_i - (a \cdot \frac{\partial C}{\partial w_i})$$

$$b = b - (a \cdot \frac{\partial C}{\partial b}),$$

where $C = \text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$, and

$$\frac{\partial C}{\partial w_i} = \frac{2}{n} \cdot \text{sum}(y - \hat{y}) \cdot \sigma(z) \cdot (1 - \sigma(z)) \cdot x_i$$

$$\frac{\partial C}{\partial b} = \frac{2}{n} \cdot \text{sum}(y - \hat{y}) \cdot \sigma(z) \cdot (1 - \sigma(z)).$$

Support Vector Machines

Support Vector Machines (SVMs) is a technique that belongs to the group of learning machines and aims to process data. It is used in classification problems and in approximating the form of the function in regression problems. The logic of a learning machine is to give the value y_i of a function (unknown to us) corresponding to a given point x_i . This is done as follows. For given set I of points $x_i \in R^n$ and having the corresponding values $y_i \in R$ that the unknown function takes, the learning machine is trained to learn the relation connecting x_i to y_i . That is, the machine learns the mapping $x_i \rightarrow y_i$ and so for a point x_m , different from those in the learning set I , it will give the value y_m that the unknown function would take.

In the case of SVM classification, the set of points I consists of two subsets k and n . Thus, the result of the function will be $+1$ or -1 ($y_i = +1$ or $y_i = -1$) depending on which subset the given point x_i belongs to. These two subsets are called classes and the value $+1$ (-1) is the "label" of the class. That is, in this case the SVMs learn to correctly classify the points x_i into the two classes. The points x_i and their corresponding values, y_i , constitute the training set. The points x_i are called training patterns, while the values y_i corresponding to them, training targets. To decide which class a new point belongs to, it should be found where the boundary of each class is, i.e., find a line (two-dimensional space) that separates the two classes. SVMs, to achieve class separation, use straight lines. Thus, from the relative position of the point to be classified and the dividing line, it will be possible to draw the conclusion to which class the point belongs.

The reasoning behind support vector machines is that if it has been chosen the one that maximizes the margin it is less likely to misclassify an unknown object in the future.

Bayesian classifiers

In Naive Bayes classification probabilities are used to solve the classification problem. These are models that assign class "labels" to the elements of a problem, where the classes come from a finite set. They are based on the principle that for a given attribute, its value is independent of the value of any other given attribute of the class. The main benefits of Naive Bayes classification are that it remains unaffected by isolated noise points and irrelevant features, and can handle missing values by ignoring them in the likelihood estimation. However, the assumption of data independence may not hold for some properties, which could be correlated.

Bayes' theorem calculates the conditional probability $P(H|X)$, i.e., the probability that hypothesis H is verified given that event X is true. According to Bayes' theorem, the probability $P(H|X)$ is given by the equation:

$$P(H|X) = \frac{P(H) * P\left(\frac{H}{X}\right)}{P(X)}$$

where $P(H)$ is the prior probability that hypothesis H is true, $P(X)$ is the prior probability that event X will occur, and $P(X|H)$ is the probability that event X will occur given that case H applies.

K-Means

K-means is an unsupervised learning algorithm in which the data set is divided into a fixed number of K non-overlapping subgroups, where each record belongs to a single group. The algorithm aims to make the records in a group as similar as possible to the center of the cluster (centroid), based on their characteristics. Similarity is calculated through distance measures, with the classic example being the Euclidean distance.

The steps of the K-Means algorithm are as follows:

1. Group the data into K groups.
2. Randomly select K points as the centers of the groups.
3. Match the objects to the nearest cluster center according to the Euclidean distance function.
4. Calculate the centroid in each group (average of all objects in each group).
5. Repeat steps 2, 3 and 4 until the same points are assigned to each group and there is no change in the centroids.

The algorithm is evaluated based on the number of groups. A popular approach to this assessment is the Elbow method. With this method the best number for the parameter K (number of clusters) can be found. It is based on the sum of the squared distance (Sum Squared Error – SSE) between the data and the centroids. The number K is chosen at the point where the SSE begins to straighten and forms an elbow. Sometimes it is difficult to understand a good number of clusters to use, as the curve decreases monotonically and may show no bend (Dabbura, 2018).

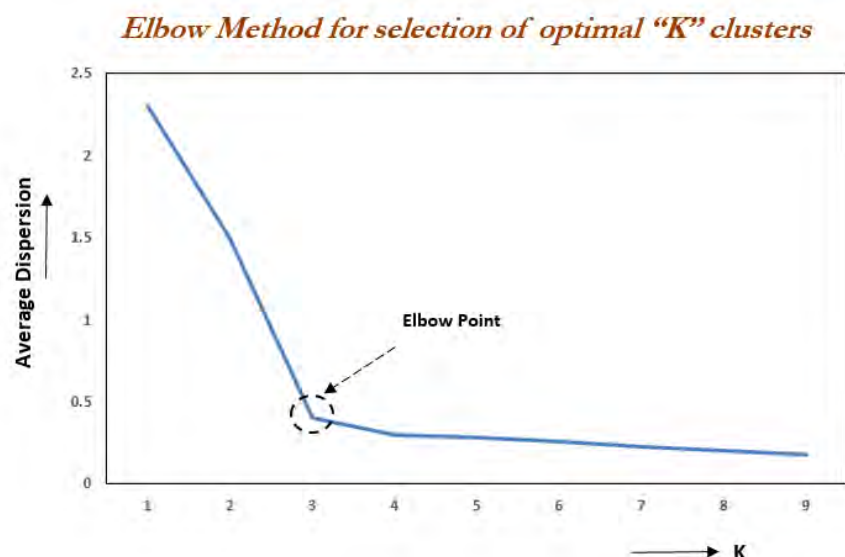


Figure 5.35: Representation of SSE based on the number of Ks (Source:

<https://www.oreilly.com/library/view/statistics-for-machine/9781788295758/c71ea970-0f3c-4973-8d3a-b09a7a6553c1.xhtml>)

Density-based clustering

These algorithms are based on the density around some point that is considered a cluster. A representative algorithm is DBScan. Similar to connection-based clustering, it is based on connection points within certain distance limits. However, it only connects points that satisfy a density criterion and in the original variant is defined as a minimum number of objects inside this radius. A cluster consists of all connected objects (which can form a cluster of an arbitrary shape, unlike many other methods) plus all objects that are within range of those objects. However, the method expects some density reduction to detect cluster boundaries and cannot detect the intrinsic cluster structures that are prevalent in the majority of real-life data.

Hierarchical Clustering

Link-based clustering, also known as hierarchical clustering, has as its central idea that objects are more related to objects that are close to them than to those that are further away. These algorithms connect "objects" to form "clusters", based on their distance. A cluster can largely be described by the maximum distance required to connect parts of the cluster. At different distances, different groups will form, which can be represented using a dendrogram, which also explains the common name "hierarchical clustering". These algorithms do not provide a single segmentation of the data set, but instead provide an extensive hierarchy of clusters that merge together at certain distances. However, a unique partition of the data set is not produced, but a hierarchy from which the user will again have to select the appropriate clusters and furthermore they are not very robust against outliers, which will either appear as additional poles or cause other clusters to merge.

TF-IDF

Each text is represented as a vector in a multidimensional space, where each dimension represents a unique term of a collection of texts. Also, every dimension corresponds to a real number which depends on the frequency of occurrence of each term each time in the text. The TF-IDF method aims to weight all the terms of a collection of texts. In short, its goal is to assign the corresponding weight to each term and by extension to each dimension of this multidimensional space. This happens because the simple numbering of a term in a text is not enough to inform about the importance of this term and the weight of the information it contains. This method consists of the quantities TF and IDF. The quantity TF (Term Frequency) indicates how many times a term appears in a text. On the other hand, the IDF quantity indicates how widespread a term is in a text and also in the entire collection of texts. The goal of this method, through the TF-IDF weight, is to select those terms that best capture the content of a text. To determine the weight of a term, both TF and IDF quantities are equally important. It is assumed that N is the total number of items that can be recommended to users and that the keyword k_i appears in n_i of them. It is further assumed that $f_{i,j}$ is the number of times that keyword k_i appears in document d_i . Then $TF_{i,j}$ is the term frequency of keyword k_i in document d_i and is defined as:

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}}$$

where the maximum is calculated over the frequencies $f_{z,j}$ of all keywords k_z that appear in document d_j . However, there are keywords that appear in documents but are not useful in separating relevant from non-relevant documents. For this reason, the inverse document frequency of the keyword k_i is used as

$$IDF_i = \log \frac{N}{n_i}$$

Then the TF-IDF weight of keyword k_i in document d_j is determined as $w_{i,j} = TF_{i,j} * IDF_i$.

Association Rules

Association rules are a rule-based method used by Machine Learning that aims to find relationships between variables in large databases. Metrics of interest are used to find these relationships. Association rules are widely used in market analysis, but also in Bioinformatics, attack detection, etc.

The problem of association rules is defined as follows (Agrawal, Imielinski and Swami, 1993):

Let $I = \{i_1, i_2, \dots, i_n\}$ be a set of n binary attributes called objects and let $D = \{t_1, t_2, \dots, t_m\}$ be a set of transactions called a database. Each transaction D has a unique transaction number and contains a subset of objects from I . A relation of the form $X \Rightarrow Y$, $X, Y \subseteq I$ is defined as a rule. Each rule consists of two sets of objects, X , Y , with X called the antecedent or left-hand side and Y called the consequent or right-hand side. A very simple

example of an association rule is the following: {eggs, flour} \Rightarrow {milk}, meaning that if eggs and flour have been purchased, milk will also be purchased.

To select the correct rules from all possible combinations of rules that may exist, various metrics are used. The most widespread are support and confidence, in which prices should pass certain minimum thresholds.

Let X, Y be sets of objects, $X \Rightarrow Y$ the association rule, and T a set of transactions of a given database.

The support represents how often the set of objects appears in the database, and as the support of X with respect to T , is defined as the percentage of transactions t in the database that contain the set of objects X :

$$supp(X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|}$$

Confidence represents how often the rule has been proven correct. More specifically, the confidence of a rule $X \Rightarrow Y$, with respect to a set of transactions T , is the percentage of transactions that contain both X and Y :

$$conf(X \Rightarrow Y) = \frac{supp(X \cup Y)}{supp(X)}$$

Confidence can be thought of as an estimate of the conditional probability $P(E_X)$, where E_X, E_Y , the events that a transaction involves the sets of objects X, Y respectively. It essentially represents the probability of finding the right side of the rule in transactions, given that the left side of the rule is also contained in those transactions.

5.5 eMEDIATOR SEARCH ENGINE

A CUSTOMIZED SEARCH ENGINE

Most e-learning platforms use general search engines (see Section 5.2.5) including several varied disciplines to find competencies and learning outcomes related to their studies, but it seems less useful because of the time and effort that learners must spend to find items related to their learning requirements and favorites. Universal search cannot and probably should not meet the precise needs of disciplines. The need for enhancement is an important challenge in this case.

Here we will design a personalized search as shown in Figure 36, that will be included within the developed LMS platform and will include some of the machine learning and data mining technologies described in Sections 5.3 and 5.45 for overcoming the boundaries of the general search.

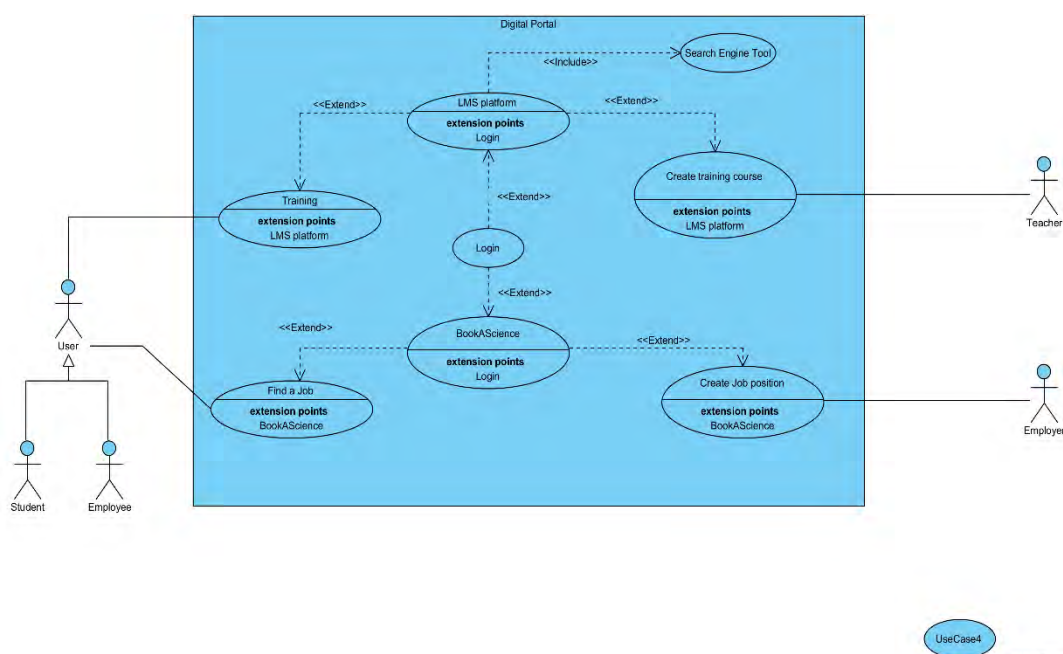


Figure 5.36: Search Engine

The search engine will be a specific tool that can be a component of the proposed e-platform. The proposed search tool will allow learners according to search and indexing mechanisms to direct their own competence/course. Through the process of discovery, or guided discovery thereby the learner learns the facts, concepts, and procedures of the searched competence/course. Its function will be based on providing courses/competences and their information according to keywords during the searching process of a user. The type of the search engine will be hybrid, as it will offer a combination of crawler-based results and human-powered directories. This method automatically combines different types of search results, and complements content-based search with ontology-based search and vice versa.

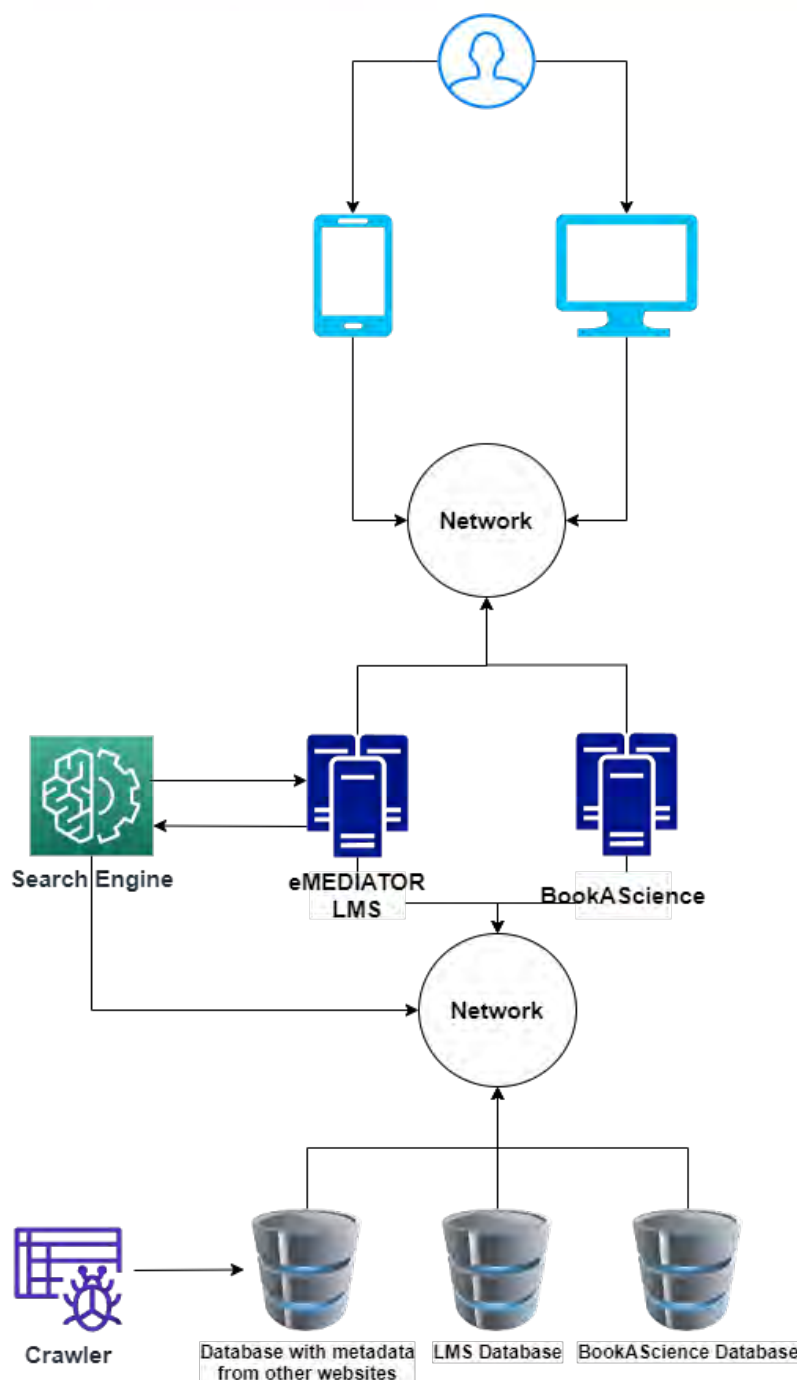


Figure 5.37: System's Architecture

Users can use the search engine to find both Competences and Courses based on their needs. As shown in the following picture, users can conduct a keyword search and the system automatically will search from both eMEDIATOR database as well as popular online databases, retrieve the results, classify them and rate them based on user profile and return it to the user.

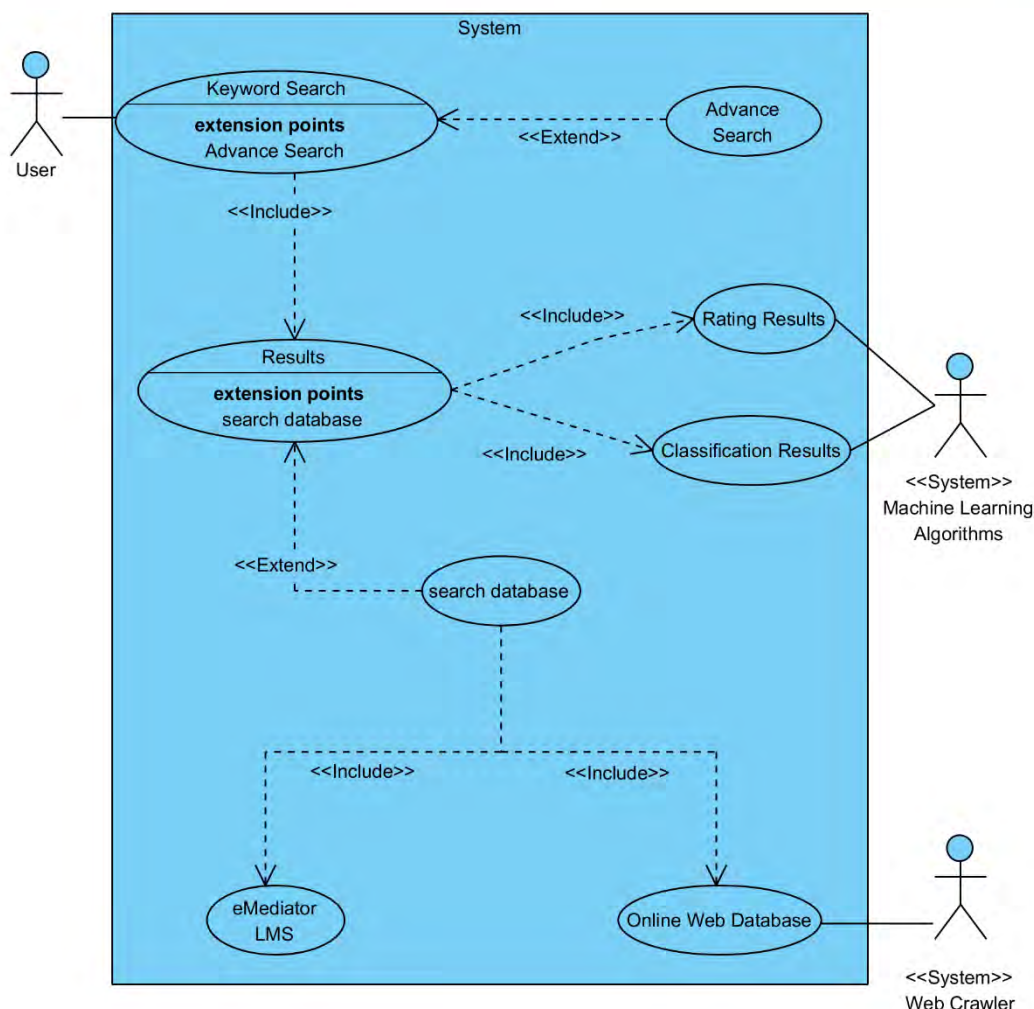


Figure 5.38: Use Case of Search Engine

In order to make the search engine as effective as possible, machine learning algorithms will be used along with some features that will improve user experience. The approach steps, which are explained in more details below, are:

1. Creating datasets with statistical indicators.
2. The statistical indicators most related to courses/competences will be obtained by implementing techniques and algorithms like decision tree algorithms. The statistical indicators that will be used by a greater number of algorithms, will be the ones most related to courses/competences.
3. Once the statistical indicators most related to courses/competences will be discovered, they will be used to construct the metric.
4. Relating this metric to the courses/competences.

Keyword Extraction

Users typically care about keywords while looking for courses/competences since they enable them to rapidly understand information relevant to a specific course/competence. In light of this, having a keyword extractor will help the search engine significantly, especially since some courses lack relevant keywords on their main page for crawling.

Users' search terms can be precisely matched with the keywords extractor's and it will also support the recommendation of courses/competences in relevant fields based on courses/competences with similar keywords. Based on a set of related key phrases, it is possible to determine the purpose of the query. Additionally, terms in a certain domain can be precisely matched and quickly recognized.

A basic property here is the generation of relevant metadata from learning outcome and competences for a specific course. Automatic generation of pertinent metadata for competences is still a very difficult problem. The fact that the related educational information depends in large extension on the context and the information generated automatically fills only the simple fields which have low value (section 5.4.2).

Web scrapers will be used, to retrieve data from the online databases including courses' details and metadata from webpages like Udemy. When users complete a search, the results will include all the relevant courses or competences gathered to the database of the system, along with their metadata. Furthermore, using the user profile, the results will be sorted and ranked based on their preferences and based on the preferences of other users with similar profile.

Grammar and Spell Corrector

In order to improve user experience in searching, it was decided to implement grammar and spelling correction tools, which are now frequently found in open-source search engines. Achieving the highest accuracy possible may be the biggest obstacle in the process of improving the search engine. With the use of a powerful abnormal detector for characters or words, it is possible to generate a list of potential search terms in the fields in order to make them more resistant to unclear or insufficient searches.

Clustering courses/competences

Another approach to improve the search engine is clustering, in order to classify courses/competences according to keywords.

The main objective here is to create and group learning objects that can be used and reused in different learning contexts. There are several algorithms that are used for listing research results, like linear regression and

nearest neighbors (Section 5.4.8.2.2). A clustering tool will provide the ability to launch an activity/competence by direct request.

A clustering algorithm is used in data mining and follows the next steps:

1. Groups instances according to their similarity by applying the Euclidean metric.
2. When the instances are grouped, it is important to study each group to find similar features in the instances of the same group and discover different ones in different groups, in other words, finding each group's significance in relation to the instance features.

The analysis with clustering techniques is meant to group courses/competences according to keywords. The different groups that will be obtained reflect users' preferences.

There are many clustering algorithms and their selection depends on the problem to be solved. In this case, various algorithms will be reviewed and the ones with the best performance will be selected.

APPLYING MACHINE LEARNING TECHNIQUES FOR CLASSIFYING COMPETENCES/COURSES

Machine learning techniques can be applied regularly and frequently, supporting an automated analysis process. This feature makes them suitable for classifying courses and competences.

Classification algorithm

It is important to note that users tend to have difficulty to choosing the proper keywords. The use of classification algorithms for improving users' decision can help users for choosing the most relevant competence. Various algorithms can detect the favorites and requirements of learners by applying feature-based knowledge segmentation and data mining (Section 5.4.8.2).

In this case, the goal is to obtain a method that identifies the key factors to classify courses and obtain the classification results frequently and regularly. There are two different perspectives on categorization: the classical view and the prototype theory. The Classical view indicates that a category contains an object, or individual, as long as all and each of certain characteristics or attributes are obeyed. The Prototype theory indicates that a category contains an object, or individual, if it is similar to another that is a prototype of this category. Once the course/competence metric value will be obtained for each user, a comparative study has to be done.

PROVIDE COMPETENCES RECOMMENDATION BASED ON REPUTATION SYSTEMS RATING

A reputation system functions by making it easier to gather, aggregate and disseminate information about an entity, which may then be used to describe and forecast that entity's future behavior. The major goal of

reputation systems is to facilitate the development of trust between unfamiliar users. The rating aggregation process is a main part of reputation system to produce global opinion about quality of the outcoming competences.

As mentioned above, users will be able to give feedback in our platform in order to deliver a more reliable competence outcome from a specific course. This is the feature in which our reputation system will be based on. In addition, machine learning algorithms will be used during the rating aggregation process in order to provide a more reliable competences recommendation.

From the rating dataset, various user-related variables are extracted, including:

- User tendency, which measures user's behavior in providing ratings, a feature that will be expressed by three variables (number of positive ratings, number of neutral ratings and number of negative ratings given by a user).
- User fluctuation, which measures the variance of user ratings from the ratings provided by the community.
- User experience, which is the ratio of number of ratings provided by each user to the total number of ratings in the system.
- User reliability, which measures the average of errors for all ratings provided by a user. This variable shows the reliability of a user in providing ratings, which measures the closeness of user ratings to the average ratings of a course.

A description of user's ratings will be represented by the extracted dataset and each row will represent a user data whereas the columns will represent the extracted variables. User reliability will be predicted as a form of weight by the extracted dataset entered to the machine learning algorithm. Reliability is considered an output variable while tendency, fluctuation and experience variables are considered input variables. Multiple machine learning algorithms can be used including, Linear Regression (LR), Support Vector Regression (SVR), K-Nearest Neighbor (KNN) and Regression Tree (RT). The algorithm or the algorithms with the best performance will be chosen after an extended review and testing.

There are some basic features that will help with improvement of the reputation system, as well as its functionality.

- History: A user's history is the collection of data, which has been saved, that documents their previous interactions and the results. This feature is essential to the concept of reputation since it is frequently used to predict the expected outcome of ongoing or upcoming searches for courses/competences.

- Collection: The behavioral data of users should be recorded for a reputation system to build trust. A reputation system can facilitate the collection of data on interactions between users using a variety of ways.
 - Direct: information that is directly derived from personal contact or through the observation of other people's interactions.
 - Indirect: data can be collected from other entities (individuals or groups) based on interactions that the entity doing the query was not aware of.
 - Derived: Data are collected from sources that were not intended to be used as reputation sources in the given situation.
- Representation: This feature represents the structure that is used to define, exchange and interpret reputation data.
- Aggregation: The reputation score of a user is calculated with a way that it is described by aggregation. The sum of all reviews – both negative and positive – for a user is the most basic method of reputation aggregation. Every positive rating raises the total by one, while every negative rating lowers it by one. All the entities in a system can be ranked using the final score. Using the average of all the ratings to generate a single rating for each entity is a marginally improved method.

REFERENCES

- Stefanos Economides, 2021, *Search Engines: Learn All About Them! - SEO In Greece* [Online]. Available at: <https://www.seoingreece.org/%CE%BC%CE%B7%CF%87%CE%B1%CE%BD%CE%AD%CF%82-%CE%B1%CE%BD%CE%B1%CE%B6%CE%AE%CF%84%CE%B7%CF%83%CE%B7%CF%82/>
- Warren R, 2020, *How Search Engines Work* [Online]. Available at: <https://aeroadmin.com/articles/en/2020/how-search-engines-work/>
- Neenu Ann Sunny, 2020, *Machine Learning in Search Engines*, Volume 8, Issue 2.
- Seguidores.online, *Search Engine: What is it? And what is it for?* [Online]. Available at: <https://seguidores.online/que-es-un-motor-de-busqueda/>
- Patrick Compton, 2019, *The History of Search Engines and Their Algorithms: How Do They Impact SEO* [Online]. Available at: <https://www.crazydomains.com/learn/search-engine-algorithms/>
- Beverly Mapes, 2008, *A History of Search Engines | Top of the List* [Online]. Available at: <https://topofthelist.net/a-history-of-search-engines/>
- Tom Seymour, Dean Frantsvog, Satheesh Kumar, 2011, *History of Search Engines* [Online]. Available at: https://www.researchgate.net/publication/265104813_History_Of_Search_Engines
- Vincent Tabora, 2019, *How Search Engines Retrieve Information — Search and Rank Algorithms* [Online]. Available at: <https://medium.com/swlh/how-search-engines-retrieve-information-search-and-rank-algorithms-53703f194eac>
- Randolph Hock, 2001, *The Extreme Searcher's Guide to Web Search Engines*, Edition.2, CyberAgeBooks
- Adamantios, 2021, *Search Progress* [Online]. Available at: <https://seo363.com/2021/08/22/%CE%B7-%CF%80%CF%81%CF%8C%CE%BF%CE%B4%CE%BF%CF%82-%CF%84%CE%B7%CF%82-%CE%B1%CE%BD%CE%B1%CE%B6%CE%AE%CF%84%CE%B7%CF%83%CE%B7%CF%82/>
- Tom Warren, 2020, *Bing is now Microsoft Bing as the search engine gets a rebrand* [Online]. Available at: <https://www.theverge.com/2020/10/5/21502315/microsoft-bing-rebrand-search-engine-logo>
- Thomas J Law, 2022, *Meet the Top 10 Search Engines in the World in 2022* [Online]. Available at: <https://www.oberlo.com/blog/top-search-engines-world>
- Natasha Lomas, 2019, *Google has quietly added DuckDuckGo as a search engine option for Chrome users in ~60 markets* [Online]. Available at: <https://techcrunch.com/2019/03/13/google-has-quietly-added-duckduckgo->

as-a-search-engine-option-for-chrome-users-in-60-markets/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlMmNvbS8&guce_referrer_sig=AQAAAAAnKU-E6bxGKgfTzysOgwNcpFR4TEp

Hucker Marius, 2022, *RIP BERT: Google's MUM is coming* [Online]. Available at: <https://towardsdatascience.com/rip-bert-googles-mum-is-coming-cb3becd9670f>

Alex Chris, 2022, *Top 10 Search Engines in the World (2022 Update)* [Online]. Available at: <https://www.reliablesoft.net/top-10-search-engines-in-the-world/>

Samuel, A., 1959. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*.

Mitchell, T., 1997, *Data mining in agriculture*. Machine Learning. MacGraw-Hill Companies. Inc., Boston. Mucherino, A., Papajorgji, P. and Pardalos, P., 2009, Dordrecht: Springer, pp.19-20.

Burkov, A., 2019. The Hundred-Page Machine Learning Book by Andriy Burkov. Education, I., 2022. *What are Neural Networks?* [online] Ibm.com. Available at: <<https://www.ibm.com/cloud/learn/neural-networks>> [Accessed 17 April 2022].

Hunter Heidenreich, 2018, *What are the types of machine learning?* [Online]. Available at: <https://towardsdatascience.com/what-are-the-types-of-machine-learning-e2b9e5d1756f>

Kyrkos E., 2015, *Mining Knowledge from Data* [Online]. Available at: https://repository.kallipos.gr/bitstream/11419/1233/2/Kef_6.pdf [Accessed 13 April 2022].

Hofmann, M., & Klinkenberg, R. (Eds.), 2013, *RapidMiner: Data mining use cases and business analytics applications*.

Han, J., Kamber, M. and Pei, J., 2012, *DATA MINING*. 3rd: MORGAN KAUFMANN, p.26, 41-65.

Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009, *The Elements of Statistical Learning*, Data Mining, Inference, and Prediction, Second Edition, Springer.

Tan, P., Steinbach, M., Karpatne, A. and Kumar, V., 2006, *Introduction to data mining*. pp.2,3.

Gorunescu, F., 2011, *Data mining*. Berlin: Springer.

Galton, F. 1886, *Regression towards mediocrity in hereditary stature*. The Journal of the Anthropological Institute of Great Britain and Ireland, pp. 246–263.

Yule G U, 1897a, *On the theory of correlation*. Journal of the Royal Statistical Society, pp. 812 -854.

Pearson, K. and Lee, A. 1903, *On the Laws of Inheritance in Man: Inheritance of Physical Characters*. Biometrika, pp. 357-462.

Fisher RA, 1925, *Theory of statistical estimation*. Proc. Cambridge Philos. Soc., pp. 700–725.

Imad Dabbura, 2018, *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks* [Online]. Available at: <https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a>

Liberman, N., 2017, *Decision Trees and Random Forests*.

Agrawal, R., Imieliński, T. and Swami, A., 1993, June. *Mining association rules between sets of items in large databases*. In Proceedings of the 1993 ACM SIGMOD international conference on Management of data (pp. 207-216).

LIST OF AUTHORS

1. Katerina Florou
2. Salmas Dimitrios
3. Vasiliki Liagkou

6 . A.3.6. Development of the mock-up testing procedure and test case requirements (AU)

This section describes the work of Aalen University in the 3rd period of the eMEDIATOR project.

6.1 eMEDIATOR Mock-Up FUNCTIONALITIES

The development of the Mock-Up has been successful. Currently, it is available under: <http://92.50.75.83:8080>

The Mock-Up combines already different functionalities such as:

1. Login and User Management Functionality: It is needed for the further work with the portal according to the planned achievements and integrations. There will be different users working with the portal. Currently, you can already manage and work with the personal information and the displayed user's data. You can also upload profile images as well as other information:

Account Settings

Information

Organizations

Memberships

Roles

Password

Apps

Information

USER DISPLAY DATA

Screen Name *


nicolas

Email Address *

nicolas.dolle@p-a-systems.com

User ID

40224



Change
Delete

PERSONAL INFORMATION

Language

English (United States)

Job Title

Prefix

Mr

Birthday

08/09/1995

First Name *

Nicolas

Middle Name

Figure 6.5: Illustration of the user management

2. Blog-Functionality: It is needed for news and important information for the users:

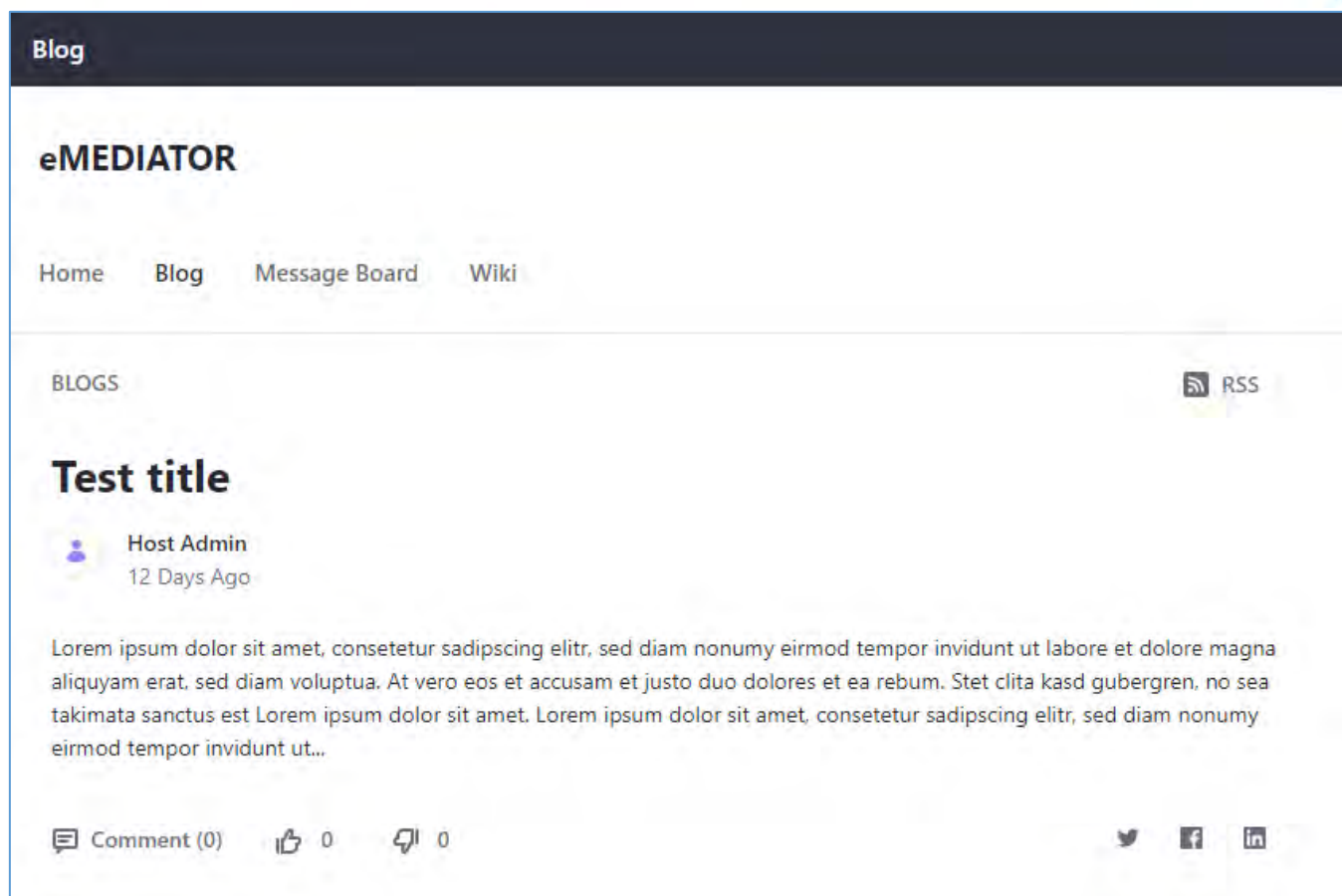


Figure 6.6: Blog functionality illustration

3. Message board: The message board has currently in the Mock-Up the functionality of the communication tool in the portal. It is working like a common forum. Users can open threads and discuss topics. Usually, there is also an internal Chat Tool planned, which will make it even easier to communicate between the users of the portal:

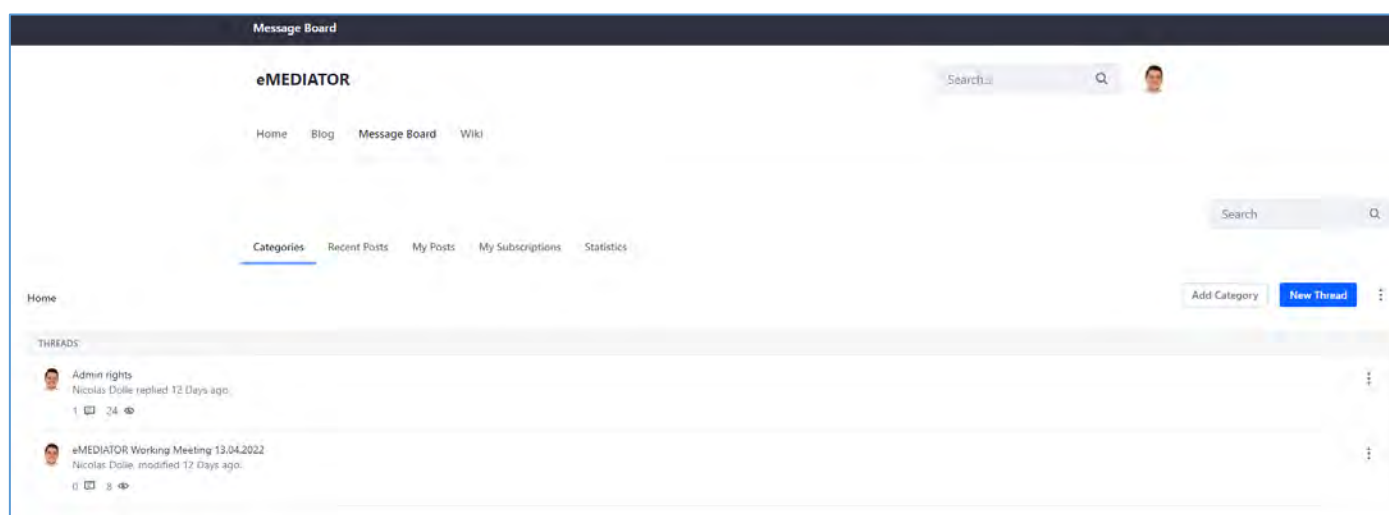


Figure 6.7: Message board illustration

4. Wiki-System: For the eMEDIATOR Portal it is planned to implement a Wiki-System, which will support the information management of the solution. Ideally, user will be able to work with the system using the Wiki-System for onboarding new users and roles on the portal. Additionally, all processes and workflows should be documented later in this Wiki-System.

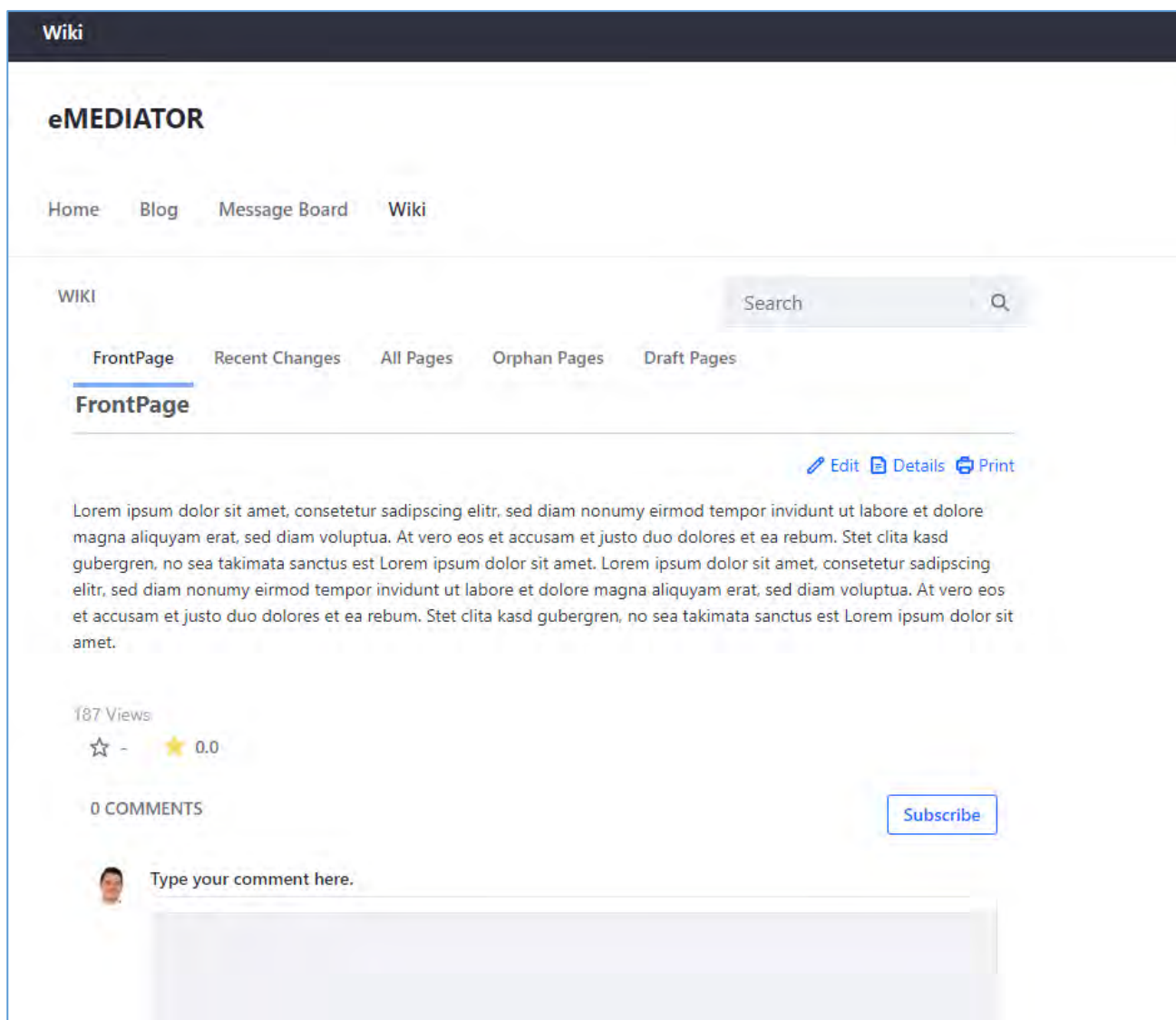


Figure 6.8: Wiki-System illustration

5. Workflow Management System: The chosen technology, Liferay, provides intense integration functionalities. That is why it is the perfect basis for the eMEDIATOR portal. Currently, the AU team works and experiments with a workflow management system, which is very important for adding new processes as well as workflows to the eMEDIATOR portal. A potential engine could be „Koleo” that is working great with Liferay. The AU team tested it by a classical workflow of paper approval for conferences:

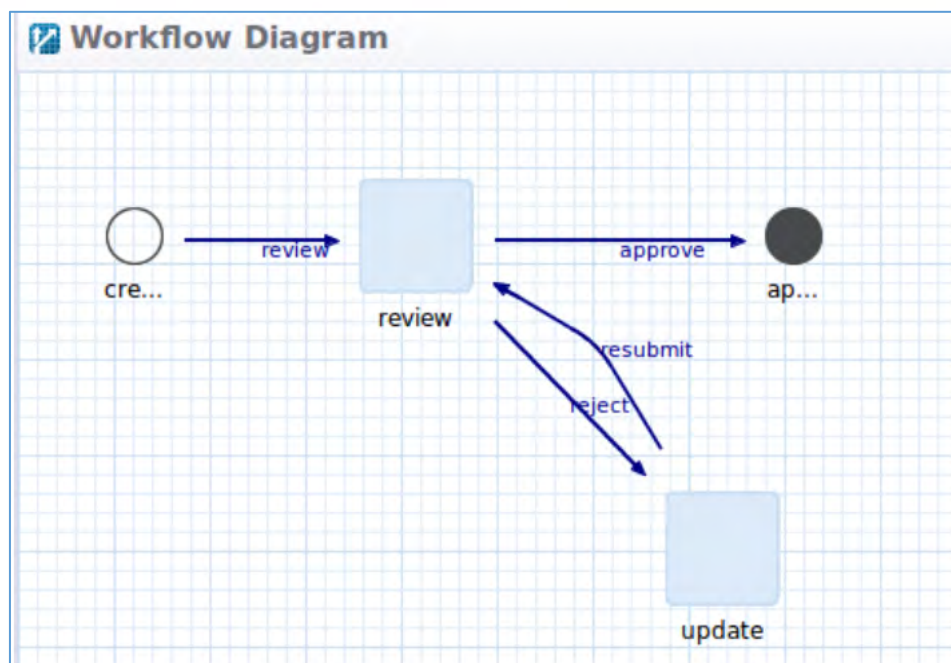


Figure 6.9: Workflow Management System Paper Approval Test

All these processes and workflows have been implemented already in the Mock-Up according to the requirements gathered in the first two iterations of requirements engineering.

6.2 TEST CASE REQUIREMENTS

The requirements for the test cases have been developed during the discussion of the partners of eMEDIATOR moderated by Aalen University. The requirements engineering workshop has been the basis of this work. It was focused on the generation of a common understanding of the most important test requirements. The procedure followed a red line that has been practically done during commercial system development workshops.

1. **Objectives:** What are the objectives of the new system / solution?
2. **Context / Scope:** What is the framework of the new system / solution?
3. **Stakeholder:** Who is responsible for what? Who needs to be involved? Who is user?
4. **Functional Requirements:** What functions should the system have?
5. **Non-Functional Requirements:** What other requirements are there that are not functional?

6.3 TESTING PROCEDURE / ROADMAP

Testing of the eMEDIATOR portal is currently done by the partners as well as by the technical team of Aalen University. The developed sites and functionalities are tested continuously. These tests can be considered as Alpha-Tests of the portal. It is planned to execute as many Alpha-Tests as possible in the internal eMEDIATOR

team. As the Alpha Demo Portal is plan for the next period, the partners have agreed to have the first pilot testers included in the platform solution. These are: Students, Companies and Academic Staff. After integration of further functionalities, the testers can start moving around in the portal and think about possible optimizations. The roadmap for the next periods is displayed in the following figure:

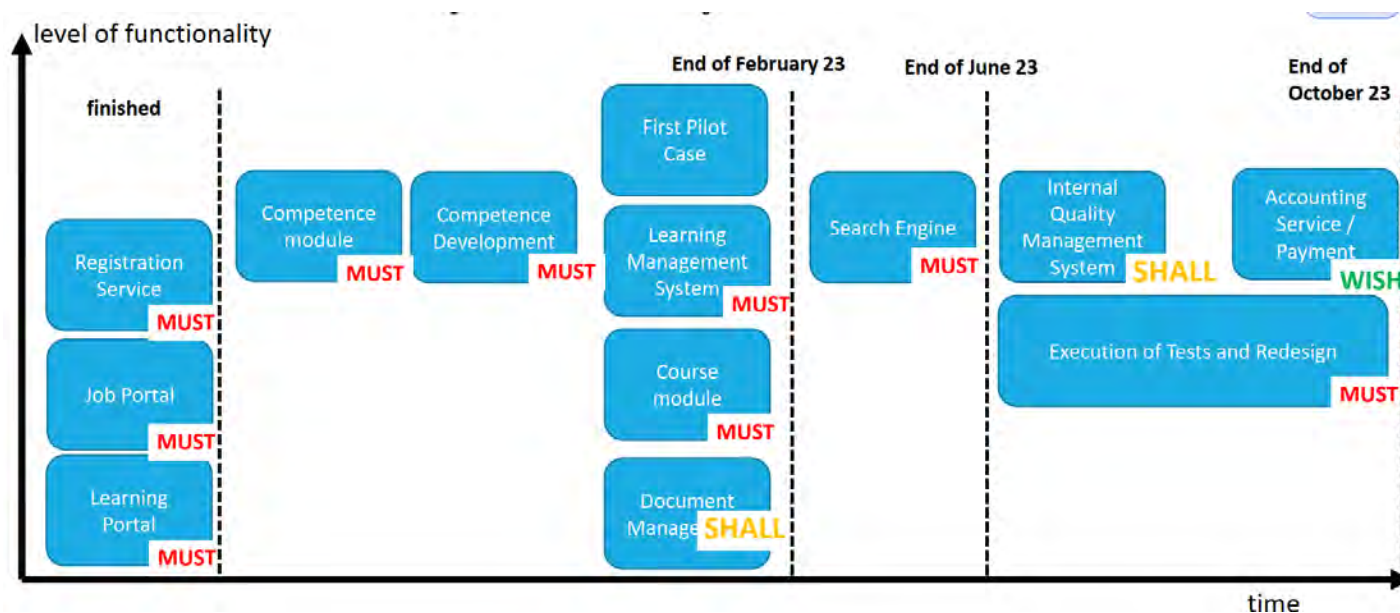


Figure 6.10: Roadmap Development

The whole documentation of the Workshop can be found in the Appendix 5.

LIST OF AUTHORS

1. Nicolas Dolle – Project Manager and Researcher Aalen University

APPENDIX 1. Examples of Functional Architecture
Requirements testing procedure and test case
requirements (TTI)

APPENDIX 2. Examples of Learning Delivery Model
Requirements testing procedure and test case
requirements (UL)

APPENDIX 3. Examples of Education Data Structuring and
Storage Experience testing procedure and test case
requirements (UM)

APPENDIX 4. Examples of Technologies and API used for
Digital Education System testing procedure and test case
requirements (UoI)

APPENDIX 5. Examples of Digital Platforms for Education Service Delivery testing procedure and test case requirements (AU)

APPENDIX_5_eMEDIATOR_IoU_AU_Workshop_Documentation.pdf

APPENDIX 6. Examples Miscellaneous
