UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

MEJORAMIENTO DE UNA METODOLOGÍA PARA LA IDENTIFICACIÓN DE WEBSITE KEYOBJECT MEDIANTE LA APLICACIÓN DE TECNOLOGÍAS EYE TRACKING, ANÁLISIS DE DILATACIÓN PUPILAR Y ALGORITMOS DE WEB MINING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INDUSTRIAL

GUSTAVO ADOLFO MARTÍNEZ AZÓCAR

PROFESOR GUÍA:
SR. JUAN DOMINGO VELÁSQUEZ SILVA

MIEMBROS DE LA COMISIÓN:
SR. PEDRO MALDONADO ARBOGAST
SR. ALBERTO CABEZAS BULLEMORE

SANTIAGO DE CHILE
DICIEMBRE 2013

Financiado por el proyecto FONDEF-CONICYT CA12I10061 - AKORI

Resumen Ejecutivo

RESUMEN DE LA MEMORIA
PARA OPTAR AL TITULO DE
INGENIERO CIVIL INDUSTRIAL
POR : GUSTAVO MARTÍNEZ AZÓCAR
FECHA: 11/12/2013
PROF. GUIA: SR. JUAN VELÁSQUEZ

El crecimiento acelerado de internet ha creado un aumento sostenido de los sitios web para todo tipo de empresas, organizaciones y particulares, provocando un nivel de oferta inmensamente alto. Estos sitios comienzan cada vez más a ser un importante canal tanto de comunicación directa con el cliente como de ventas, por lo que se hace necesario tratar de generar estrategias que permitan atraer a más usuarios al sitio y además hacer que los actuales usuarios continúen utilizándolo. Esto lleva a preguntarse qué tipo de información resulta de utilidad para el usuario final y como poder identificar esa información.

Anteriormente se ha tratado de abordar este problema mediante técnica de *web mining* a las áreas de contenido, estructuras y usabilidad de un sitio web, de modo de poder encontrar patrones que permitan generar información y conocimiento sobre estos datos. Estos a su vez permitirían tomar mejores decisiones respecto de la estructura y contenido de los sitios web.

Sin embargo este tipo de técnicas incluía la conjunción de datos objetivos (*web logs*) con datos subjetivos (encuestas y focus group principalmente), los cuales poseen una alta variabilidad tanto personal como interpersonal. Esto provoca que el análisis posterior de los datos pueda contener errores, lo que redunda en peores decisiones.

Para resolver en cierta manera eso, este proyecto de memoria desarrolló algoritmos de *web mining* que incluyen análisis de exploración visual y neurodatos. Al ser ambas fuentes de datos objetivas, se elimina en cierta parte la variabilidad de los resultados posteriores, con la consecuente mejora en las decisiones a tomar.

El resultado principal de este proyecto son algoritmos de *web mining* y modelos de comportamiento del usuario que incluyen información de análisis de exploración visual y datos obtenidos a través de ténicas de neurociencia. Se incluyen también una lsita de website keyobjects encontrados en la página de prueba para este proyecto.

Se incluyen además una revisión general acerca de los principales temas sobre los que el proyecto se basa: la web e internet, el proceso KDD, *Web Mining*, sistemas de *eye tracking* y website keyobjects. Por otra parte se especificaron los alcances del proyecto de memoria, tanto técnicos como de investigación.

Se concluye que el resultado del trabajo fue exitoso, incluso siendo el resultado de los algoritmos similares a la metodología previa. Sin embargo se abre un nuevo camino en cuanto al análisis de sitio dadas las relaciones encontradas entre el comportamiento pupilar y el análisis del sitio. Son incluidas ciertas consideraciones y recomendaciones para continuar y mejorar este trabajo.

Abstract

ABSTRACT OF THE THESIS
TO OBTAIN THE TITLE OF
CIVIL INDUSTRIAL ENGINEER
BY: GUSTAVO MARTÍNEZ AZÓCAR
DATE: 11/12/2013
GUIDE PROF.: MR. JUAN VELÁSQUEZ

The rapid growth of the Internet has created a sustained increase in web sites for all types of businesses, organizations and individuals, causing an immensely high level of supply. These sites begin to be, increasingly, an important communication channel for customers and also a sales channel, so it is necessary to generate strategies for attracting more users to the site and also make existing users continue using it. This raises the question of what kind of information is useful to the end user and how to identify this information.

Previously this problem has been addressed by web mining techniques in the areas of content, structure and usability of a website. This techniques allow to find patterns that will generate information and knowledge on these data. In the same way this data allows to take better decisions about the structure and content of the websites.

However, this type of techniques includes the combination of objective data (web logs) with subjective data (surveys and focus groups mainly), which have a high variability both personal and interpersonal. This causes the subsequent analysis of the data may contain errors, resulting in worse decisions.

For solving this in some way, this project developed web mining algorithms that included visual exploration and pupil dilation analysis. Being both objective data sources be eliminated to a certain part of the variability of the subsequent results, with consequent improvement in decision making.

The results obtained were web mining algorithms and user behavior models that incorporate visual exploration and analysis of data obtained through techniques of neuroscience, applied and validated. Also a list of website keyobjects found through this techniques is presented.

It also includes an overview on the main topics on which the project is based: the web and the internet, the KDD process, web mining, eye tracking systems and website keyobject. In the same way were specified the thesis technical and research scope.

Is concluded that the results obtained in this project thesis were successful, even being the results similar to the previous methodology. However a new way of exploring website using this techniques is opened due to the relations founded between the pupil dilation and the analysis of the site. Some considerations and recommendations are included for continuing this work or improving this work.

En cualquier lugar que estuvieran, recordaran siempre que el pasado era mentira, que la memoria no tenía caminos de regreso, que toda primavera antigua era irrecuperable, y que el amor más desatinado y tenaz era de todos modos una verdad efímera.
- Gabriel García Márquez, Cien años de soledad

Acknowledgements

Después de un largo y costoso proceso finalmente este proyecto de memoria ha terminado. Sin embargo su realización no hubiese sido posible sin la ayuda de muchas personas que de alguna forma u otra colaboraron en su desarrollo, a quienes me gustaría agradecerles.

A todos los integrantes de "la salita": Yerko Covacevich, Claudio Aracena, Manuel Castro, Cristian Rodriguez, Edison Marrese y Alfonso Abadía. Gracias a todos por su ayuda en el desarrollo de las soluciones, los análisis, en los problemas con el informe y su preocupación y compañia.

A todo el equipo del laboratorio de Neurosistemas que siempre tuvieron la mejor disposición para ayudar y enseñar a manejar los equipos o explicar el funcionamiento de los programas. Gracias en especial a Christ Devia, Enzo Brunetti y Caro Astudillo que fueron con los que trabajé y que estuvieron preocupados constantemente.

A mi amigo personal Felipe Bravo quien me colaboró fuertemente con su amplio conocimiento de mining, algoritmos, análisis, computación, etc. sin lo cual no habría podido avanzar o llegar a buen puerto.

A Kristopher Muñoz por dedicar gran parte de su tiempo y trabajo en realizar algoritmos y análisis para el desarrollo de mi trabajo, sin los cuales este proyecto hubiese tomado mucho más tiempo.

A todos los que se dieron el tiempo de venir a realizar el experimento a esta facultad. Javier Ojeda, Camila Paniagua, Pedro Abarca, Belen Castro y todos aquellos que no conocí personalmente pero que participaron voluntariamente en el.

A mi profesor guía Juan Velásquez que me dio la libertad y confianza para desarrollar este proyecto y apoyarme en todo momento con el. De igual manera al profesor Pedro Maldonado por confiar en mi y en el trabajo que realicé.

Y finalmente a todos los que se preocuparon constantemente de mi trabajo, pregúntandome sobre el, dándome ánimo y apoyándome constantemente, incluso desde los más remotos lugares.

A todos ustedes, muchas gracias.

Gustavo Adolfo Martínez Azócar

Index

Index of Tables

Index of Figures

Chapter 1

Introduction

1.1 Work context

Internet use has increasingly massified over time [1], with a resulting increase in the number of existing web sites. This increased supply of information available on the Web creates the challenge of establishing differentiation from competition and for this reason it is necessary to generate useful and interesting information for users or customers, in order to keep them visiting, stay longer and also attract other users to the site. This leads to the need to develop the question of what type or kind of information is useful to the end user of the site and how to define it.

Previously, to address this, web mining techniques have been used in the areas of content, structure and on web site users to retrieve such information and knowledge. These techniques are based primarily on data mining analysis of the data from these primary sources of information.

With these techniques it has been possible to find the website keywords, which were defined as "a word or possibly a set of words that are used by users in their search process and characterize the content of a page or website because" [2], whose function is to determine the most important words for users of a site based on the content and navigation through it.

However getting website keywords neglected one of the most important resources on a site: the objects. An object or website object was defined by Dujovne as "any structured group of words or a media resource that is present on a web page that has meta-data describing the contents" [3]. And together with this, a website keyobject was defined as "a website object

that captures the attention of users and characterizing the content of a website." [3]

Dujovne and Velasquez developed a methodology that allowed the identification of website keyobjects on a site by the amount of time a user spent on a website object. To do this they used algorithms that allowed the reconstruction of the sequence of pages a user looked at during a visit to the site and the implementation of a survey on the same control group that visited the site.

Later Gonzalez in [4] improved this methodology, replacing the survey for an eye-tracking system, which allowed him to more reliably determine the object permanence on each website. This is why the natural step in the line of investigation is to quantify the user interest in each website object using the information obtained from pupillary dilation and the correlation between the level of dilation and interest in what is observed.

1.1.1 Project description

The project to develop as a thesis project is mainly an improvement to the existing methodology for identifying website keyobjects (based on eye-tracking systems) through the introduction of information about pupil dilation produced by the individual and neural data (or extract data through techniques of neuroscience) as new sources.

First, a state-of-the-art analysis of the three main components of the project will be developed: eye-tracking systems, information models obtained through the measurement of pupil dilation, neural data and models of web user behavior.

Then the data generated by the pupil dilation examination along with the information obtained through neuroscience techniques (neural data) will be properly incorporated into the development of the web mining algorithms.

After that the algorithms and analysis techniques to detect patterns in neural data will be developed, and this in conjunction with the data provided by the eye-tracking system will allow the type of and interest level on a website object to be established, plus the time spent on it. In the same way a change in the methodology designed by Gonzales will be made to incorporate these new techniques.

Finally both methodologies (that developed by Gonzalez in [4] and the one developed in this project) will be applied to a test site, and the results of both compared in order to quantify the possible existence of improvement between the new techniques.

1.1.2 Project description

The methodology developed by Velasquez and Dujovne had limitations in the data analysis, mixing objective data sources (web logs) with subjective data sources (surveys), thus skewing the information obtained. This is mainly due to the subjective data showing high intrapersonal variability, mainly a product of seasonality and individual emotions, together with interpersonal variability, for example, which was added in a focus group.

Gonzalez' work continued along the line of detecting the level of interest in web objects, the main difference being the replacement of the surveys in the method proposed by Velásquez and Dujovne by an eye-tracking system, which would avoid the use of subjective data and replace it with objective data. While this was useful, improving on the previous method by a certain percentage, the system lacked information about the type of interest (positive or negative) that a website object generated.

This project aims to improve the previous methodology, by adding the information obtained through neural data and eye-tracking systems (which have a high correlation with user preferences) for quantifying the interest level shown by the user, along with the introduction of more objective data.

Finally, the hypothesis presented in this thesis project is: "It is possible to improve the effectiveness of web mining algorithms using web intelligence techniques, adding visual exploration analysis, pupil dilation information and analysis of patterns obtained from neural data."

1.2 Objectives

This work pursues the following objectives:

1.2.1 General objective

The main objective of the thesis project is:

Develop, apply and validate web mining algorithms and user behavior models by adding visual exploration data and pupil dilation analysis.

1.2.2 Specific objectives

The specific objectives for this thesis project are:

1. Analyze the state of the art of web user behavior
2. Analyze the state of the art of eye-tracking systems
3. Analyze the state of the art of pupil dilation analysis
4. Characterize the data produced by the eye-tracking system and neuroscientists
5. Develop techniques that allow detection of patterns in the data produced by the eye-tracking system
6. Develop a functional prototype for applying the techniques on a test web site
7. Analyze and conclude based on the results, and propose future developments

1.3 Methodology

A first step in the methodology to be followed is a literature review of the three main components of the project: a model of web user behavior, eye-tracking systems and pupil dilation.

After this step the rest of the project will proceed using the CRISP-DM methodology, adapting it so that it can be applied to developing web mining techniques.

The CRISP-DM methodology consists of six interrelated and cyclical steps and a constant feedback, enabling the development of standard data mining analysis [5]. These steps are:

1. Understanding the business: the phase in which it is necessary to understand the business, by understanding their objectives, goals and current situation. It is also where the project's development objectives are defined.
2. Understanding the data: this stage is where the data is analyzed in order to arrive at a full understanding. It is also where the necessary information requirements are determined to develop the project.
3. Data Preparation: Once the available data have been identified and understood, the next step is to proceed to perform the tasks of selecting, processing and cleaning.
4. Modeling: Here is where modeling is done, using the proper algorithms or software for doing it. As a process that needs feedback, new algorithms can be added according to the level of knowledge about the data.

5. Evaluation: This is the stage where the results of the models obtained above are evaluated, within the business context.

6. Implementation: Finally at this stage is where the hypotheses are tested and where new knowledge is also discovered.

In the case of this thesis project, the resulting adjustment would be:

First seek to understand the objectives of the project and what the expected results are. Also define the environment where the thesis will be developed, i.e. technologies to be used, the test site for developing the algorithms and the necessary control groups.

Then the data capture needed for developing the project will proceed. Among this data was the chosen site, along with its respective objects, web server log files and data on the interests of users, gathered through eye-tracking systems and neuroscience techniques.

The following step will be transformation of the captured data needed for the project. This includes the process of sessionalization and characterization of user interest.

The next step is to develop algorithms for detecting patterns of information in the data collected. Different strategies will be used according to the level of development achieved.

Finally, results will be evaluated, comparing them with previous methods in order to verify the possible existence of improvements.

1.4 Expected results

The expected results of the thesis project are:

1. A report documenting the results obtained through the research of the state of the art of the eye-tracking systems.

2. A report documenting the results obtained through the research of the state of the art of information about pupil dilation.

3. A data model including the definitions and descriptions of the data vector obtained and its utility for the project.

4. A methodology proposal, along with the respective algorithms, for identifying website keyobjects.

5. A report with the results obtained after implementing the methodology.

6. A report detailing the analysis made and the results obtained.

1.5 Scope

The scope of the project is the same context where it will be applied, i.e. the web area and its content and structure, and not therefore to develop models and generic algorithms. However, it would be considered possible to use in other areas or industries adapting the required components.

The research and development scope will focus on the web intelligence area, taking the issues related to the eye-tracking technology and neuroscience as input or data sources. However a state-of-the-art analysis of both fields will be included.

Finally, within the technical scope, it is noteworthy that the data and information related to eye-tracking technology and neuroscience analysis will be obtained using the equipment provided by the Universidad de Chile. External equipment needs will depend on the resources available for the project. For this reason, the memory project aims to develop models and algorithms using the latest technologies, but also those which are available for student use.

1.6 Content Structure

The rest of this document is structured as follows. In the first place, in Chapter 2 the conceptual framework of this project will be explained. It will give all the necessary definitions and explanations used during the development of the thesis and which will be used later. Among these concepts some basic information is included about the Web and Internet, Web mining and its derivatives, the human eye and its components, the pupillary response and its relationship with human behavior and the state of the art about eye-tracking systems.

Later in Chapter 3 the design and requirements of the experiment will be explained, and all its different components defined. Concepts like the experimental group, the devices needed, the web sites for testing, how the data will be captured, etc. will be fully detailed here. Also some of the necessary data transformation requirements are explained thoroughly.

In Chapter 4 the whole experiment's implementation is explained, detailing which devices will be used, the characteristics of the experimental group selected, the web sites chosen, etc.

It will also detail the practical work and how it was developed.

Chapter 5 shows the results obtained by the experiment, detailing every component and including a comparative analysis with the previous methodologies.

Finally in Chapter 6 the conclusions of the project, based on the results of the previous chapter and the research done, are presented. Some guidelines and basic baselines are also given for further development.

Chapter 2

Conceptual framework

This conceptual framework seeks to give to the reader a general and clear explanation of the essential and basic concepts from which will be developed the thesis project. This will involve a brief review of the main concepts, definitions and performance.

2.1 Web and Internet

These two terms are often used indistinctly, but each has its own meaning. Internet, in technical terms, is a decentralized collection of interconnected networks, which enables the connection between geographically separated teams through a particular protocol, which makes it work as a single global network. The Web, on the other hand, is the name given by Tim Berners-Lee to the system of pages and related objects linked together through hyperlinks.

For the proper working of the Web the existence of a document transfer protocol is necessary. In this case the protocol is called HTTP (Hypertext Transfer Protocol) [6], which is maintained by the World Wide Web Consortium, an international organization that is responsible for maintaining standards for the World Wide Web (abbreviated WWW).

Web architecture is formed by three main components: Hypertext Transfer Protocol (HTTP) [6], Hypertext Markup Language (HTML) [7] and Uniform Resource Locator (URL) [8] which are detailed below:

- **Hypertext Transfer Protocol (HTTP)** is a protocol for transferring documents between two devices in a network. The specifications of this protocol, as mentioned

before, are maintained by the World Wide Web Consortium.

- **Hypertext Markup Language (HTML)**: Hypertext markup language is used mostly for creating web pages. It describes the structure and content of the document as a text, and also allows the generation of links or connections between documents and the insertion of objects on pages such as images, videos, etc. An image showing some code from the site *http://www.mbauchile.cl* can be seen in figure 2.1.

Figure 2.1: HTML code of site http://www.mbauchile.cl

- **Uniform Resource Locator (URL)**: The string that associates each web element with an address, which allows it to be located on the Web. It generally consists of the transfer protocol, the domain being accessed and the particular resource. For example in http://www.docode.cl/index.php, http corresponds to the protocol, www.docode.cl corresponds to the domain and index.php corresponds to the particular file.

2.1.1 Information produced by the web

Web operation produces three basic data types: structure, content and navigation data generated by users. These are detailed below:

- **Structure**: the structure of a web page is defined by the HTML language, which also defines the content of the page. It also allows the generation of connections to other web

sites through links (or hyperlinks) that create common information communities [9]. In general the main structure contains the beginning tags, the header, metadata, styles, functions and the body of the document. Similarly, the structure of a web site is established by the existence of hyperlinks linking the pages to one another and which allow users to navigate between documents.

- **Content**: The content refers to the different components (or objects) present within a web page such as text, images, videos, etc. The text on a page can be easily described, but the media (sounds, videos or pictures) is not so easy, so the existence of metadata that refers to content in this type of objects is necessary.

- **Navigation data**: Each time a user navigates through a web site, the web server that runs the site leaves a record of each request made by the user in a file called a web log [10]. Web log files store the following fields:

  - Host IP: The IP address from where the request was made.
  - User and User ID: If the site has user authentication mechanisms, it stores the username and his numeric identifier.
  - Timestamp: Date and time when the request was made.
  - Application Method: Form in which the request was made.
  - URI: Name and location of the requested file.
  - Protocol: HTTP protocol version of the software that makes the request
  - Status: Status of the petition outcome. This result is defined by the status code that is delivered to the browser.
  - Bytes: Document size in bytes sent.
  - Reference: Refers to the page from which the user logged in to the document requested. It is empty if he/she has entered directly through the browser.
  - Agent: Identification code of the agent that made the request.

  The analysis of these records allows reconstruction of the session [11] generated by each user on a web site, which would enable understanding the process of navigation on a site.

2.2 KDD process

The large amount of data generated and stored by different industries, such as retail, mining, finance, etc. makes it almost impossible to do a manual analysis of it. Given this

scenario, it is necessary to generate an automated or semi-automated technique that allows the efficient extraction of information from these large volumes of data.

This process of extracting information from large data sources is known as Knowledge Discovery in Databases (or the KDD process), and was defined by Fayyad et al. [12] as "the non-trivial process of identifying valid patterns, original, useful and understandable about data." Data is defined as a set of facts stored in a data source, and pattern is defined as an expression in a language that describes a subset of the data or a model relevant to that subset.

Figure 2.2 shows the components of the KDD process, which consists of five main steps. It is important to notice that this process is iterative and interactive, and for this reason it is possible to return to a previous step. Weighting the results with an expert in the relevant business is also recommended. These steps are detailed here:

1. **Data Selection**: This is the stage where the data to study is chosen, being generally a subset of the universe of data available. To make the choice it is necessary to analyze the requirements and objectives of the project from the viewpoint of the business expert, since a bad choice can lead to erroneous test results.

2. **Preprocessing data**: A processing is done by cleaning programs on the chosen data to effect noise reduction, handle missing data, etc. After this processing a clean data set is obtained.

3. **Transformation**: From the data set obtained in the previous step certain fields will be chosen for analysis, considering in the process only useful data variables. This helps to further reduce the dimensionality of the data set. The result of this process is a refined database.

4. **Data Mining**: This is the core of the KDD process and involves the discovery of patterns in data. After defining the objectives of the process, different algorithms can be applied to find information and patterns related to the objectives. The selection of these algorithms depends directly on the objectives, some being more appropriate than others. These algorithms permit different functions to be applied to the data such as data grouping (cluster generation), classification, regression, predictions, etc. Tests are generally performed with more than one algorithm to compare and verify the results.

5. **Interpretation and Analysis**: This is where the results of the previous process are obtained, mainly patterns that allow the generation of knowledge. In this stage the

business expert plays a key role because he is the most suitable to analyze these patterns, and to verify the usefulness and validity of the knowledge generated.

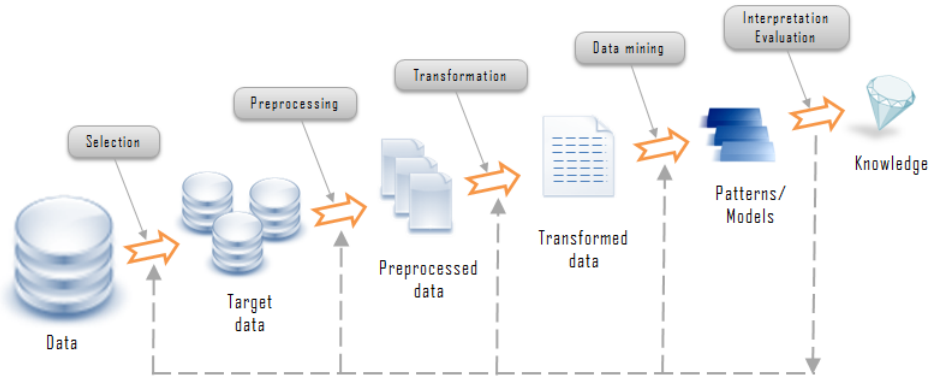

Figure 2.2: Steps of KDD process

2.3 Web Mining

Given the large amount of data existing on the Web, and their characteristics such as heterogeneity, time-variance, high dimensionality, etc., they are good candidates for treatment with data mining techniques. The use of these techniques on such data has been called web mining [13] and research in this area is growing given the increasing use of the Web [14]. This field is the product of the union of several research areas such as databases, information retrieval, machine learning, language processing [15], etc. The main purpose of these techniques is to find patterns, trends and user behavior on the sites, on the basis of which it is possible to make improvements in the structure, content and usability of a web site.

Web mining techniques are separated into three main areas according to the source from which data is extracted. These areas are:

2.3.1 Web Content Mining

Web content mining (WCM) focuses on finding information from web documents, among which include content, data and files. This type of mining involves the analysis of text, images and videos, where the analysis of the last two is a sub area called multimedia data mining. In order to analyze the text of a document a preprocessing is first required to be done. This transformation of web pages is called vector space modeling and consists of transforming the document into a characteristics vector.

2.3.2 Web Structure Mining

This area of web mining is responsible for analyzing the hyperlinks between pages and the documents linked with them. In this way a web site can be represented as a directed graph, where nodes represent pages and links represent the hyperlinks between pages. Each link goes from the node where the hyperlink is included and ends in the node representing the page where the link goes. This definition allows the assignment of values according to the importance of each node, and with this information the pages or sites which are more important within a network can be seen. These techniques are used by search engines like Google [16] to order search results.

2.3.3 Web Usage Mining

These techniques are designed to find information patterns in user behavior when they navigate through the web pages on a site [17]. To perform this analysis it is necessary to apply mining techniques to the user's browsing session on the web site. The first step is rebuilding the entire session on the site.

Those techniques are neither independent nor exclusive. The data originated on the Web allows analysis to be done through different kinds of web mining and even combining them to get better results.

In this project the techniques used are mainly Web Usage mining for the analysis of navigation patterns and user behavior on the site through the logs stored in the web log files. Web content mining is also used for the analysis of contents (mainly objects) inside the web pages. With these analyses it will possible to get information about the behavior of the users on the site and the most important objects on every page.

2.4 Human eye

The human eye is an organ which reacts to and detects light, and as a sense organ, allows vision in mammals. Different components of the eye allow light perception, color differentiation and depth perception, and permit us to distinguish about 10 million colors [18]. Specialized cells in the retina receive the light signals which affect the adjustment of the size of the pupil and the regulation of different hormones. [19]

Some of the components of the eye are:

- **Cornea**: Is located in the front part of the eye and covers the iris, pupil and anterior chamber. This structure is transparent and refracts the light passing through it. This characteristic is used by some eye-tracker systems.

- **Sclera**: Is a membrane located in the external part of the eyeball. It is white, thick and strong and its function is to give shape and protect the internal components. Usually the visible part of the sclera is called the *white part of the eye* [20].

- **Choroid**: Is a membrane located between the retina and sclera and is the vascular layer of the eye, containing multiple connective tissues. Its functions are to provide nutrients to the components of the eyeball and keep the temperature constant.

- **Pupil**: Is a hole located in the center of the iris, which allows light entrance into the eyeball [21]. It is an expandable and contractile structure, and its function is to regulate the amount of light received by the retina. It appears black because light rays are mostly absorbed by the tissue inside the eyeball.

- **Retina**: Is a light-sensitive tissue located on the rear inner surface of the eye. The light impinging on the retina creates a sequence of chemical and electric phenomena which produce nerve impulses. These impulses are sent to the brain through the optic nerve. This is part of the first stage of visual perception [22].

- **Iris**: Is the colored and rounded membrane which separates the anterior and posterior eyeball's chambers, and its function is to control the size and diameter of the pupil [23]. The iris color is often called *eye color*.

- **Macula**: Is a place near the center of the retina shaped as an oval with a yellow pigmentation. It contains a greater concentration of cone cells, allowing high-resolution vision. This structure works when something is observed or focused on with interest [22].

The location of these components are described in figure 2.3.

2.4.1 Eye movement

The eye can perform different kinds of movements for fixating, acquiring and tracking visual stimuli, and in this way the person can develop a cerebral image of the scenario [24].

These movements can be classified into three main categories:

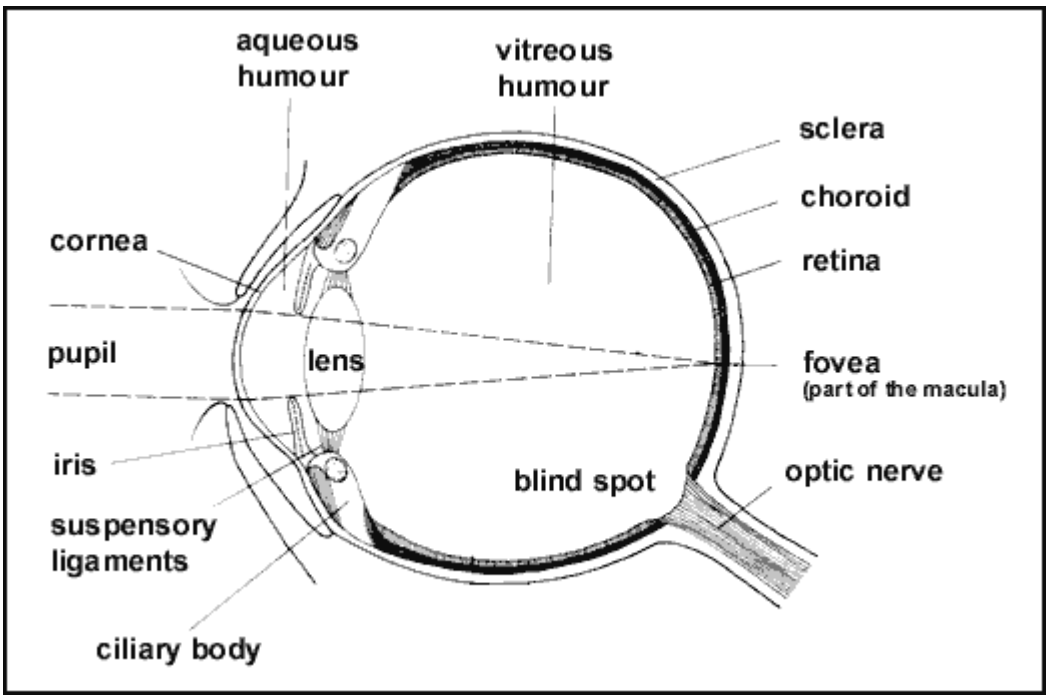

Figure 2.3: Parts of the human eye

- **Pursuit Movements**: Are the movements performed by the eye to keep track of a moving object.
- **Vergence Movements**: Are the movements of both eyes to the same object with the purpose of maintaining binocular vision.
- **Saccades**: These movements occur when the subject is scanning a visual scene, trying to find the most interesting parts of a scene in order to focus on them. The speed of the saccade cannot be consciously controlled so the eyes move as fast as they can [25].

There is another eye mechanism not considered a movement– fixation. Visual fixation occurs when the subject maintains focus on an object, in order to get a high-resolution image of the important details of the scene. Usually when a subject gets visual stimulation he/she starts tracking the scene, fixating on details and performing saccades between the different fixation points [24].

2.4.2 Visual attention

One of the most important questions is: why it is necessary to track the eye movements of the subject? Answered in a simple way, humans move the eyes to bring a particular portion of the visible field of view into high resolution to see it in fine detail. This is often done to drive the focus of concentration to that particular point (or region) of interest, even for

short lapses of time. Therefore the main conjecture which can be made is, if it is possible to track the eye movements of a subject, it is possible to register the path of attention followed by the participant. This can thus provide a good source of information about what looks interesting, what draws attention, how the world is perceived, which components of a scene are more noticeable, etc.

Notwithstanding the previous reasons for tracking eye movement, it is necessary to establish a definition for what attention is, and whether eye movement gives some information about the process known as visual attention.

Regardless of having been a topic of study during the last century, there is not a definitive complete definition of visual attention. One definition was given by the psychologist William James [26]:

> Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others...
>
> When the things are apprehended by the senses, the number of them that can be attended to at once is small, *'Pluribus intentus, minor est ad singula sensus.'*

The last part of the phrase, in Latin, can be translated into something like *'Many filtered into few for perception'*. It is supposed that attention is the filter in this case.

The human mind cannot attend to all the stimuli which it receives at the same time, due to its finite capacity. Generally attention is the way the mind focuses its capacities onto a selection of the whole range of stimuli, in order to process it in an adequate way. Due to the limited capacity of the brain, it is necessary to concentrate on specific components of the whole "sensory realm" to examine in a detailed way the most interesting smells, sounds, sights, etc. rather than the peripheral stimuli. This is particularly noticeable in vision, where a scene is inspected using parts of the whole scene, which are highly detailed due to the attention process, and are later mixed with other detailed parts, to construct a coherent representation of the whole scene.

In simple terms, the human vision system has two parts working together; a central zone with a very high resolution called foveal vision and the rest of the visual range with a low resolution called *peripheral vision*. Thus, when the gaze is directed at a particular place, the

foveal vision gives a high resolution image of that particular range of the scene, and the rest of the scene is watched in low resolution through peripheral vision.

Visual attention and eye movements

If visual attention is considered in terms of "what" and "where", i.e. "where" the noticeable things are on a scene (through parafoveal vision) and "what" these noticeable things are (using foveal vision), it is normal to consider that eye movements work in a way that support this dual-attentive hypothesis. This means that vision behaves in a cyclical process composed of these steps:

1. Given a scene as stimulus the whole frame is seen through peripheral vision and thus mostly at low resolution. In this step the interesting components of the scene may "pop out" in the field of view and engage the attention to their location for further detailed inspection.

2. Foveal location disengages attention and the eyes are moved toward the region that attracted attention.

3. After the eyes are pointed to the area of interest, the fovea is directed at the region of interest and attention is engaged to that point to perceive it in high resolution.

This type of visual attention represents a bottom-up approach. This means that the interesting areas of the visual stimulus, or in this case scene, attract enough attention to prevent looking at the rest of the scene. On the other hand, a top-down approach is driven by other cognitive factors such as knowledge, expectations, goals, etc. Under this model, the subjects are more likely to look around using peripheral vision. For example a persons who drives regularly will notice gas stations more than another person who does not drive often [27].

2.4.3 Eye tracking techniques

Eye-tracking systems are devices that allow, using different techniques, measurement of the eye movements of a person. In this way it is possible to know where the subject is directing their gaze during its use, and with these data it is possible to determine the sequence of movement of the eyes along with the time they stopped to observe an object.

In this section the main techniques for tracking eye movement will be presented, belonging to two main categories, the ones which measure eye position relative to head position and the ones which measure the orientation of the eyes in space, which is called point of regard [28].

Electro-Oculography (EOG)

This technique was widely used during the 1950s (and is still used in some places today). It consists of electrodes positioned around the eyes to measure the differences in the skin's electrical potential. This technique measures the eye movement relative to head position, so it is not very useful for measuring point gaze unless it is used with a head tracker to measure the movements of the head. An example of this is the electrodes on a subject as shown in figure 2.4

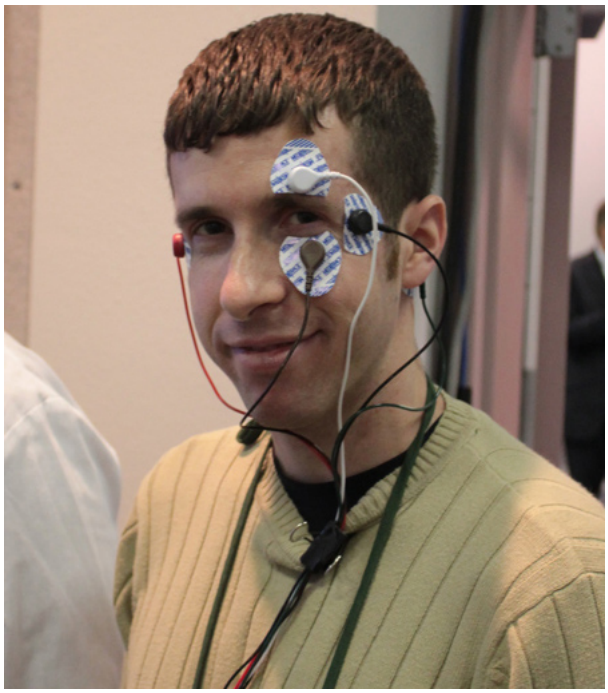

Figure 2.4: Subject using an electro-oculography technology

Scleral Contact Lens / Search Coil

A scleral contact lens is one of the most precise methods for measuring eye movements. This device consists of a mechanical or optical reference object mounted on a contact lens, which is worn directly on the eye. The contact lens for this device is necessarily large, extending over the cornea and sclera, to avoid any sliding movement over the eye.

Different kinds of mechanical and optical devices have been used on the stalk attached to the contact lens but the most popular is the wire coil. This kind of device can be easily

measured while moving through an electromagnetic field. A figure showing the size and the position on the eye is shown in figure 2.5

Despite being one of the most precise techniques, is also the most intrusive method, causing discomfort to the user during its use. Also it measures the movements relative to head position, and it is not generally suitable for point-of-regard measurement.

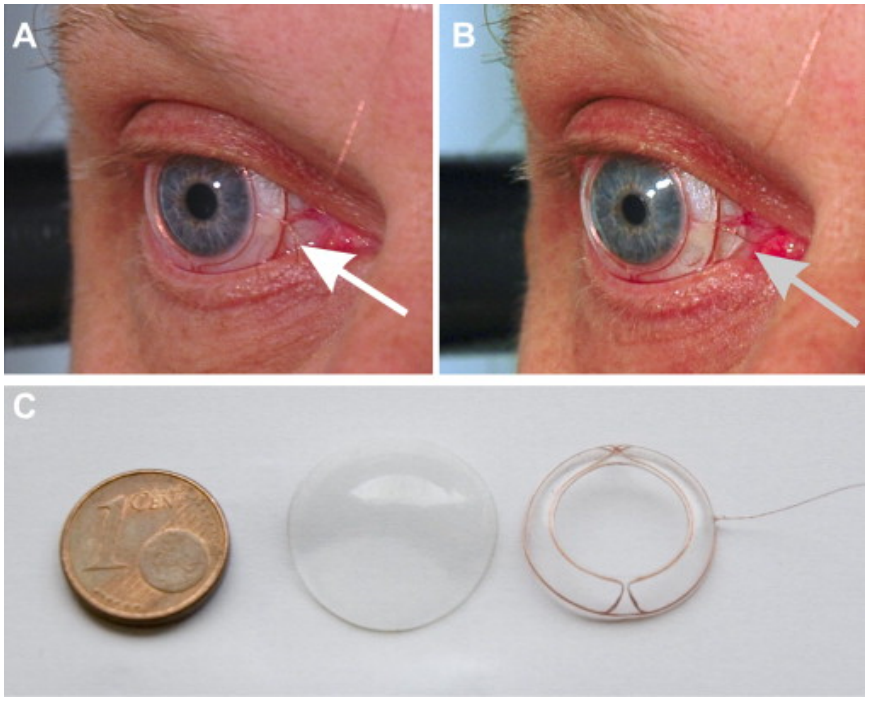

Figure 2.5: Scleral Contact lens with a search coil.
A subject wearing a scleral search coil in conventional manner (A), the wire exiting directly (white arrow). When wearing a bandage lens on top of the coil, the wire is applied to the sclera and exits the eye near the medial canthus (B, gray arrow). The bandage lens (C, center) is 5 mm. larger in diameter than a scleral search coil (C, right). [29]

PhotoOculography (POG) or Video-Oculography (VOG)

This category of eye-tracking techniques groups a wide variety of techniques for measuring eye movements which involve the measurement of different features of the eye during rotation/translation movements. Some of these features are the shape of the pupil, the position of the limbus or the reflection of the cornea under light exposure (usually infra-red light).

These techniques work differently, but are grouped together because they often do not provide point-of-regard measurement. Further, results may be gotten automatically, but in some cases may involve having to visually inspect the recorded eye movements. This kind of analysis can be extremely tedious and prone to error, and it is limited to the sample rate of the device used.

Video Based Combined Pupil / Corneal Reflection

The main innovation of this technique is the provision of point-of-regard measurement. Point-of-regard measurement requires having the subject's head fixed so that the eye's position relative to the head and point of regard coincide. Otherwise it is necessary to measure other features of the eyes to disambiguate head movements from eye rotation.

Two useful features for measurement are the corneal reflection and the pupil center. Video-based trackers incorporating relatively inexpensive cameras and image processing hardware are used to compute the point of regard in real time. The apparatus and set up can be table-mounted (as shown in figure 2.6) or worn on the head (as shown in figure 2.7). The techniques are the same in both systems, and the only difference is the size of the device. Often the head mount is more precise, because it moves with the head, but it is also more invasive.

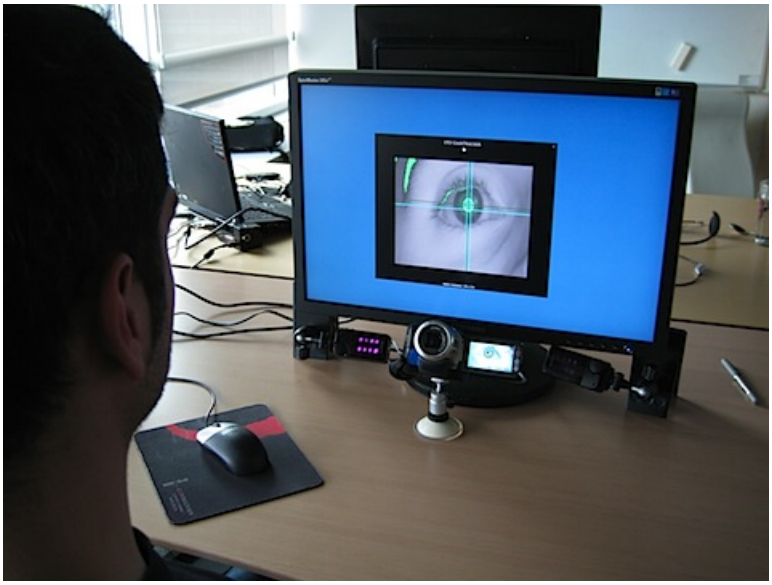

Figure 2.6: Example of table-mounted video-based eye tracker.

Figure 2.7: Example of head-mounted video-based eye tracker.

The corneal reflection of the light source (usually an infra-red light) is measured relative to the location of the pupil center, where a clearly identifiable bright disc (the pupil), and a small one (the corneal reflection) can be detected. Using a calibration system of 9 points and some basic trigonometric calculations it is possible to estimate the position of the cornea and the pupil, therefore the point of regard of the subject.

Data generated by eye-tracker devices

Modern eye-tracking systems provide a wide variety of data. Among the most important are:

- time in milliseconds when the data was collected
- horizontal and vertical position, on the monitor, that observes each eye separately
- horizontal and vertical location of the pupil for each eye separately
- distance in millimeters from the eye tracker to each eye separately
- diameter of the pupil of each eye in millimeters

2.5 Pupillary Response

Pupillary response is a physiological response that varies the size of the pupil, producing a dilation or constriction. This is due primarily because of the light reflex and the accommodation reflex. The first one is produced by the exposure of the eye to a light source, generating a dilation when it is exposed to dim light and a constriction in intense light. The second reflex refers to changes in pupil diameter produced by trying to focus on an object.

Other kinds of smaller changes in pupil diameter are reflections of cognitive processes or behavioral activities. Among these, researchers have found correlations between pupil dilation and positive stimuli, pupil constriction and unpleasant stimuli and pupil diameter and the learning process, information processing, short-term memory and nonverbal communication. It can also indicate sexual stimulation [30].

The diameter of the pupil is determined by the activity of two iris muscles, the sphincter pupillae and the dilator pupillae, and the major function is to change the pupillary diameter according to the light environment in order to adjust the amount of light allowed to enter the eye. The pupil in humans can dilate to about 8 to 9 mm., constrict to 1.5 mm. and react to stimuli in 0.2 s with peaks in 0.5 to 1.0 s. [31], [32]

Pupil movements (constriction and dilation) are produced by the iris's muscles controlled through the ANS (autonomic nervous system). In particular, pupillary constriction is produced by the innervation of the circular fiber of the iris, a process caused by the neurons of the PNS (peripheral nervous system). Pupil dilation is produced by the excitation of radial fibers of the iris, produced by the SNS (sympathetic nervous system) neurons.

2.5.1 Pupil Dilation Measurement

Pupil measurement has been a topic in eye research throughout the last century but only during the last 40 years have accurate measurements been possible. This has been facilitated by the development of more precise and reliable kinds of instrumentation, even under diverse experimental conditions and under various types of psychological and physiological states.

One of the basic problems of pupil measurement is the constant fluctuation of pupil size in both eyes during waking hours, which can be about 1 mm. in amplitude. This is due to the brain mechanism that regulates pupil size according to light stimulation. Under normal light conditions it is important to consider this behavior when taking measurements.

In the beginning, the measurement of pupil size was done using infrared photography

which allowed photographing pupil diameter even under bad light conditions. Some photoelectric methods were also used, devices which measured the light reflected on the iris. Another device, described by Hess [33], used a 16 mm. camera to film the pupil. This device, consisting of a movie camera, a projector, a screen and reflecting mirrors, allowed filming of the subject's pupil regardless of their eye color through an infrared sensor. To measure the dilation 20 individual frames were averaged and used as pupil size. This kind of technique was the most affordable but the frame-by-frame measurement involved a lot of errors, both human and mechanical, resulting in questionable reliability. Hess highlighted the importance of such variables as stimulus brightness and brightness contrast, and how they have to be maintained constant during the whole trial.

In modern laboratories and research centers a video-based pupillometer is often used. This device consists of a closed-circuit TV for recording the eye and a signal processor for measuring and displaying the pupil diameter. In these kinds of devices infrared light illuminates the eye and a silicon matrix tube camera is used to record the size of the pupil. Pupil size and variations are presented as numerical readouts or on a chart, and enable measurement over a 0 to 10 mm. range. A diagram explaining the system is presented in figure 2.8.

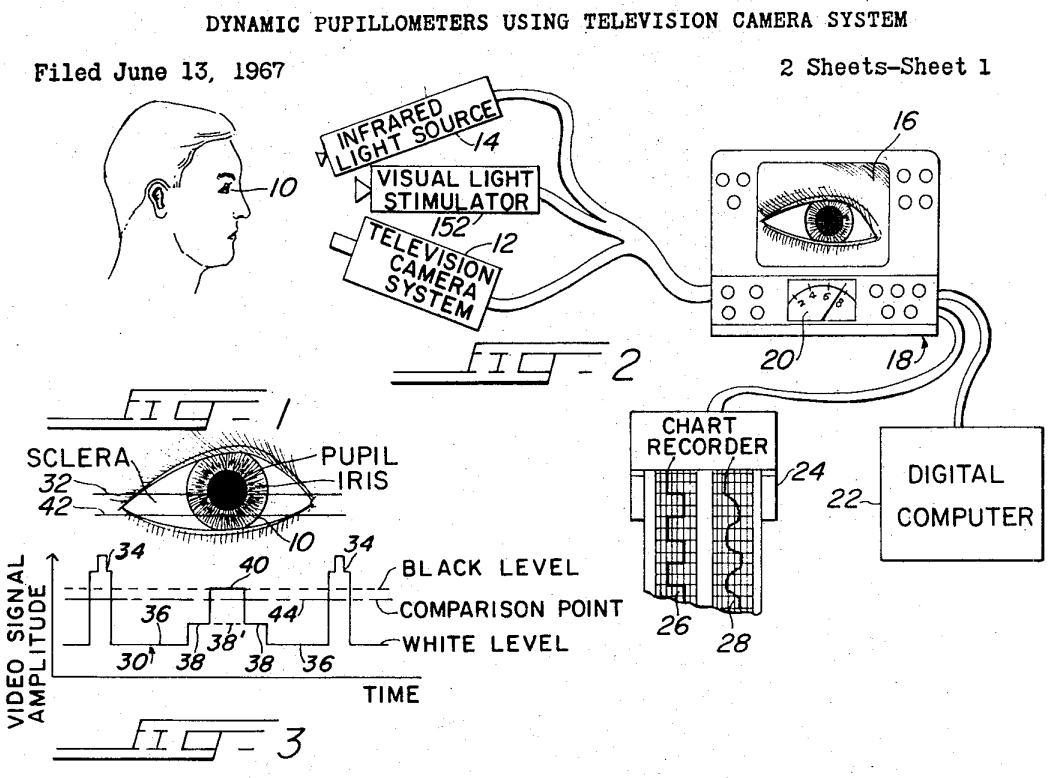

Figure 2.8: Diagram of Whittaker Pupillometer
Source: Diagram 1 in [34]

Recent research uses momentary estimates of pupil diameter which enables analysis using computational techniques. This approach is useful for task-evoked pupillary response (TEPR), where it is necessary to measure the pupil size at specific times during a task [35]. With the availability of this kind of data it is possible to specify the events during the trials in the experiment, and the background variations can be canceled out. According to Stern and Dunham in [36], the TEPR is an appropriate index of pupillary response because the TEPR is measured in fractions of millimeters and usually the background pupillary response is around one millimeter or more. Averaging the results enables the response to stand out from the "background noise", but this procedure requires establishing a baseline for measuring changes in pupil size evoked by tasks.

2.5.2 Pupil Dilation and behavioral correlation

Modern pupillometry techniques allow researching in diverse fields, human behavior being one of the most complex and interesting. The most influential researcher in this field was Eckhard Hess who suggested the correlation between mental activities, perception, interests, information processing, attitudes and pupillary response. He was not the first to suggest this, but his work led to an increase in research in this field. In 1975 he launched a book called "The Tell-Tale Eye: How Your Eyes Reveal Hidden Thoughts and Emotions."

Over the years a lot of research has been done in this particular field. Some of the most important conclusions and results obtained are presented in this section.

Pupil size and mental activity

The correlation between mental activity and pupillary response was first shown by Hess and Polt [37] through a mental multiplication test. They asked the participants to mentally calculate the result of a multiplication of two factors, increasing the difficulty from two factors with just one digit to two factors with two digits each, with the results showing the relation between difficulty level and pupil size. The range of size increase was between 4% to 30% during the period between pre-question to pre-answer, decreasing immediately after the answer was given. Thus, this result shows that pupil activity reflects the information-processing load produced by cognitive tasks. Later, Polt in [38] used electric shocks as a threat for incorrect answers on the same test. This resulted in increased amounts of effort to solve the problems, and as a consequence, in bigger pupil dilation.

Later Beatty, in [39], concluded that the amplitude of the task-evoked pupillary response (TEPR) is an index of a common factor related to the cognitive demands of memory, language processing, reasoning and perception. In fact, pupillary responses have proven to be very

important indicators of mental efforts involved in task resolution.

Attitudes

According to research done by Hess in [33], the pupil could be a better index of attitude towards persons or things than other traditional systems such as questionnaires or interviews. In research done by Barlow in [40], he showed images of three political leaders of different political factions (Lyndon Johnson, George Wallace and Martin Luther King Jr.) and an unknown person to subjects classified as either "liberal" or "conservative". While the subjects classified as liberal showed pupillary dilation when Lyndon Johnson and Martin Luther King Jr., were shown and pupillary constriction when Wallace was shown, the ones classified as conservative showed an opposite pattern. This implies that the pupil dilates when they see someone with whom they agreed and constricts in the ones who disagreed. However, this result was questioned by Ertas in [41], due to the results obtained in a similar experiment. They showed pictures of two presidential candidates (Richard Nixon and George McGovern) and a picture of an unknown person to supporters of either of those candidates. In another stage of the experiment, the candidates' last names and the name Smith were also spoken. The three pictures were associated with pupillary constriction, and the names were associated with pupillary dilation. These results showed a conflict with the previous results. Thus, the relation between pupillary response and attitudes requires additional research and study to be accepted.

Perception

Perception has been related to pupil size through various different approaches. In [42] Kahneman and Beatty tried to relate the difficulty encountered by subjects in a pitch discrimination test with pupillary dilation. The participants had to judge whether the tone presented was higher or lower than the standard tone. In the cases where the difficulty in distinguishing between the two became greater, participants showed higher dilation. In another experiment, Hakerem and Sutton in [43] measured pupillary response to a barely perceptible threshold visual stimuli. When the light flashes were not detected or when the subject was not asked to detect, pupillary dilation did not occur. When they were required to detect whether a flash was present or absent, or when they correctly discriminated a flash, pupillary dilation occurred. Similar results were obtained by Beatty in [44] when participants were asked to detect a weak tone only present in half of the trials. In this case pupil dilation occurred only when the subjects detected the presented signal. Beatty's conclusion was that pupillary dilations reflect changes in nervous system activation due to perceptual processing and the difficulty of the task.

Affective value of stimuli and pupil size

One of the results obtained in the original study made by Hess in [30] was the relationship between pupil size and affect or "feeling tone" generated by different pictures. Hess and Polt reported in this research that when female subjects viewed pictures of male nudes or babies they experienced greater pupil dilations than men. On the other hands male subjects reacted in the same way with nude females. They concluded that gender differences indicated greater interest in nudes of the opposite sex. Later in [45] Hess, Seltzer and Shlien found that homosexual men had greater pupil dilation when they were seeing pictures of male nudes compared to female nudes, whereas results obtained with heterosexual men showed opposite pupillary response.

Nonverbal communication and pupil size

In 1975 Hess, in [46], concluded that pupil size is an important variable during nonverbal communication. This conclusion was obtained after an experiment where Hess showed pictures of a woman to male participants, where the pupil size in the images was manipulated previously. Subjects described the woman with large pupils as soft, feminine or pretty. On the other hand they described the woman with small pupils as hard, selfish or cold. These results led Hess to conclude that pupil size seems to act as an unlearned mechanism in the facilitation of certain social behavior such as sexual interest.

In other studies conducted by Hess, large pupils were associated with happy faces and small pupils with angry faces. He concluded that larger pupils are associated with attractiveness, sexual interest and happiness whereas small ones are related to the opposite characteristics. This conclusion was not completely accepted by other investigators such as Janisse in [47], where he pointed out that the validity of this data was questionable because the data was collected with visual stimuli where the light reflex could not be ruled out as a possible cause of pupillary change. Hess developed another approach in which participants were allowed to draw pupils on line drawings of faces. The results showed that participants consistently drew large pupils on happy faces and small pupils on sad faces. These same results were obtained with both adults and children as subjects.

Hicks in [48] tested Hess' results with 223 college students. They found that subjects drew large pupils on happy faces and smaller ones on sad faces, the same conclusion as the previous research.

Information Processing, learning and pupil size

Poock in [49] concluded that pupil diameter was related to information-processing speed. First he determined a maximum processing capacity by making the subjects press buttons corresponding to displayed numerals as fast as possible. Then he made the participants process numerals at different percentages (between 50% to 125%) of maximum capacity. When the subjects were required to process information at 75% and 100% significant increases in pupil diameter appeared. However when the requirement was raised to over 100% to 125% of capacity, pupillary constriction occurred.

In another study Peavler in [50] required 14 college students to reproduce a string of digits after hearing them, varying the size of the string between 5, 9 and 13 characters. A trend toward increased pupil size with each successive digit in the five and nine digit condition was noticed. The experiment also showed the dilation leveled off immediately after presentation of the 10th digit in the 13-digit string. These results suggested to Peavler that the information processing effort was momentarily suspended at that instant. In this experiment, nine digits was approximately the short-term memory capacity of the subject, and correspond to the point of no further dilation. Even when Peavler and Poock got similar results for the first parts of their experiments (more processing load corresponded to a bigger pupil size) they disagreed about what happens afterward, when there is an overload.

2.6 Website Keyobject

The content of a web site is composed mainly of text formatted using HTML standard, but also includes other types of contents, called multimedia content, which include images, sound, animations, videos, etc. Nowadays these kinds of contents are some of the most popular on the World Wide Web, and sites like Youtube, Facebook, Instagram, etc. are among the most popular sites in the world.

One of the problems presented by this kind of content is the difficulty of analyzing it. Because the format of these data does not allow the direct use of web mining techniques due to ignorance about their content, the option of applying automatic techniques is disabled. Web mining techniques implemented directly on a web site will not consider these kinds of data, because they give inaccurate results.

This section will try to show a complete presentation of *website keyobjects*. First the definition of *website objects* and *website keyobjects* is given, then an explanation of how to implement them on a web site. A comparison measure for the objects will then be established

and finally a methodology for finding them will be detailed.

2.6.1 Definition

Website Keyobjects are based on the definition of an object on the Web, but this definition can cover a broad range of meanings, because the Web is defined as a set of pages or resources with links between them. For this project, object will be considered as any group of words having some kind of structure, and in the same way, multimedia files which are shown on the web site pages, including all kinds of pictures, images, sound, animations, etc. Objects based on word structure need to be inside delimitators such as paragraphs, tables or other kinds of tag separation.

An example of the previous definition can be seen in figure 2.9, which corresponds to a well-known news web site in Chile. It is possible to watch some of the objects on the page (not all of them).

Figure 2.9: Example of web site objects on the site www.emol.com

In the figure three objects are defined: a banner, an image and the title of a story, and each one has different formats: an animation, an image and a text with an image. But the separation between them is subjective, because a human can understand the difference between them, but an automated system cannot do that analysis. It is necessary to define metadata for the objects in order to define their content.

Therefore a web site object was defined by Dujovne and Velasquez as "a structured

group of words or media, which are present on a web page, having metadata describing their content" [3]. This definition clearly states the need for metadata, which are necessary to build the content representation vector of the page, and will allow the comparison of different objects between them (a video with a picture for example).

The previous definition allows any object on a web site to be described, and by doing so it is possible to process them using different algorithms and analyzing technologies. This definition is also the baseline for the definition of a *Website Keyobject*.

A web site keyobject was defined as "one or a group of website object that attract the user attention and characterize the content of a page or website" [3]. This definition states which web objects get more attention and are more interesting to the user and therefore, identify which object types would help to improve the presentation, usability and content of the web site.

2.6.2 Representation

HTML standard does not have any kind of structure for defining components of the web objects which incorporate metadata. For this reason it is necessary to create another way to do it.

One solution presented in [3] was inspired by the integration of metadata using MPEG-7, where every multimedia file is associated with an XLM[1] file which stores the metadata of the multimedia file. Using this method it is easy to give a basic definition for every object on the page without adding other elements to the source code, and in the same way it will be easier to get the defining concepts of each object.

In an XML file the necessary data for describing the objects is added, and it will keep a link between the objects and the page where they belong. The data in this file will be:

- Page Id
- Object
  - Object Id
  - Object format
  - Concepts associated to the object

[1]XML is an acronym for Extensible Markup Language, a markup language that defines rules for encoding documents in a format that is readable by humans and also by machines.

With this data it is possible to get the relationship between a page and its objects. The format of the object can also be obtained along with the concepts which define it.

A basic definition of object concept could be a set of words which describe the concept in a partial or total way. Using XML syntax it would be something like:

<concept>Analysis tool</concept>

In the example, the concept *Analysis tool* defines that the object is a tool for doing some kind of analysis, but does not say anything about the relations between the concepts.

To handle the previous problem each concept will be classified into various categories which will allow create relationships between them. Every concept can belong to one or more categories, depending on the context of the object. An example of this could be:

<concept category="marketing">Analysis tool</concept>

<concept category="computation">Analysis tool</concept>

In the previous example, the object *analysis tool* will be a tool related to marketing on certain pages, but on others could be a tool designed for computation analysis. This definition will also allow easily getting all the concepts related to certain categories.

For the concepts definition three substantives will be used. According to [51], this will not give a complete definition of the object, but a sufficient one. However this could be used for comparing two concepts.

2.6.3 Objects comparison

Web mining algorithms usually need a way to make comparisons between pages. For example in [2] the similarity measure was made using a relation between the permanence time on the pages.

Page comparison is not sufficient to compare web site objects, because a page can contain more than one object, hence the use of the previous methodology is not direct.

A similarity measure unit, similar to the one used to find web site keywords in [2], considers two objects O_1 and O_2 where:

$$|O_1| = N \text{ and } |O_2| = M \; \exists N, M > 0 \wedge N \leq M$$

The i-esim concept of the object O is also defined as:

$$C_i(O) \; i = \{1...M\}$$

The main idea is to pair up similar concepts for every object, and based on this result compare the objects. The pseudo-code for the algorithm would be:

- $\forall \; C_i(O_1) \; i = \{1...N\}$:
  - $\forall \; C_j(O_2) \; j = \{1...M\}$:
    - Compare the 3 words which define $C_i(O_1)$ with the 3 words which define $C_j(O_2)$ using a thesaurus[2]
    - If the words are the same 1 is added. In case of being synonyms 0.5 will be added.
    - The pair of concepts with their sum closer to 3 will be paired.

2.6.4 Transforming categories in strings and comparison

It was previously defined that every concept could belong to one or more categories, and in this way it is possible to define an object using the categories of their concepts. This would be expressed in this way:

$$O = Category1, Category2, ..., CategoryN$$

Category	**Associated char**
Category 1	A
Category 2	B
Category 3	C
Category 4	D
Category 5	E

Table 2.1: Association between categories and chars

And supposing that each category is associated with a character, according to the table 2.1, every object could be represented through the characters of each category. For example:

[2]A thesaurus is a reference work that lists words grouped together according to similarity of meaning, used for representing concepts.

$$O_1 = ABB$$

$$O_1 = BBC$$

If every category is codified and the objects are represented through those characters, it is possible to define a string for each object. This string will represent the categories which define every object and, after being paired, it will be possible to compare those strings to discover if they are similar.

To compare two strings the *Levenshtein distance* is used, also known as *Edit Distance*, which is a string metric for measuring the difference between two string sequences. This metric is defined as the minimum number of edits (insertion, deletion or substitution) required for transforming one string into the other.

The application of this technique to the string results in a conceptual comparison, because a difference between both represents a difference between the concepts. If two objects have a Levenshtein distance equal to 0 or near to 0 it will mean that they are conceptually similar. On the other hand, if the distance is far from 0 it will mean that the objects are different.

Finally the similarity between objects will be calculated through the relation between the distance equation defined previously and the maximal length between the two strings. The equation can be seen in the equation 2.1.

$$do(O_1, O_2) = 1 - \frac{L(O_1, O_2)}{max\{|O_1|, |O_2|\}} \tag{2.1}$$

The value of the equation will be 1 when both objects are equal and 0 when they are totally different.

2.6.5 Methodology for finding web site keyobjects

The methodology proposed by Dujovne and Velasquez in [3], and improved by González in [4], for identifying web site keyobjects is defined by two main processes: Data transformation and clustering.

Data transformation

The data to transform in this stage comes from different sources so it is necessary to perform separate transformation processes to each data set.

- **Sessionalization**: The process of rebuilding the sessions of different web site users. In this case, a session is the sequence of web pages that were visited while browsing the site. This information can also give the user's time spent on each page.
- **Adding metadata**: In general the objects on a page suffer a lack of metadata, for which reason it is necessary to incorporate it in order to do further analysis. First it is necessary to identify the different objects from a web site page, and then, in conjunction with the web master of the site, define the concepts that characterize each object.
- **Time spent on the objects**: Dujovne and Velasquez suggested conducting a survey on a control group to determine the time spent on each object. This was replaced by Gonzalez introducing an eye-tracking system which allowed him to incorporate objective data of the time spent on each object to the sessions data.
- **Vector of user behavior**: For each session identified above objects were chosen that caught the attention of the user. With this it was possible to define the vector called Important Object Vector (IOV) defined by the equation:

$$v = [(o_1, t_1), ..., (o_n, t_n)] \tag{2.2}$$

  Where o_i represents the object and t_i represents the time spent on each page.

Clustering algorithms

In order to apply clustering algorithms to the identified sessions it is necessary to define a distance measure between vectors. A distance measure between Important Object Vectors was defined for this.

- Similarity measures for sessions: In [3] the similarity between two IOV was defined through the equation:

$$st(\alpha, \beta) = \frac{1}{i} * (\sum_{k=1}^{i} min(\frac{\tau_k^\alpha}{\tau_k^\beta}, \frac{\tau_k^\beta}{\tau_k^\alpha}) * do(o_k^\alpha, o_k^\beta)) \tag{2.3}$$

Where α and β correspond to the user sessions identifier to be compared. τ_k^α corresponds to the time spent by the user α on the object o_k and $do(o_k^\alpha, o_k^\beta)$ is the similarity between objects using the Levenshtein distance.

The last equation was defined between 0 and 1, where 0 represents no similarity at all and 1 represents total similarity.

Later this measure was used as an input parameter for these clustering algorithms: Association Rules, K-Means and Self Organizing Feature Maps.

2.6.6 Clustering Algorithms

In the previous methodologies for finding web site keyobjects, created by Dujovne and Velásquez in [3], the main objective was grouping the user behavior vectors using three different techniques: SOFM (Self Organizing Feature Maps), K-means and Association Rules. The results of these techniques gave different sets of vectors, where their inner components were similar among themselves but different comparing them with the other sets. The criteria for selecting the *Website Keyobjects* was to select the objects which appeared more times in the set of vectors after applying the techniques. It is important to notice that the similitude measure between two IOVs using the equation 2.3 will be used to compare to sessions.

Algorithms

First the algorithms used for this process will be presented, along with how they work and how they were applied to that project.

The first algorithm presented will be SOFM, because the number of clusters obtained after its application will be used as an entry for the *K-means* algorithm. Then the *K-means* clustering algorithm will be shown and finally the algorithms to get the *Association Rules*.

SOFM The Self Organizing Feature Maps (also known as Kohonen Maps) are a type of artificial neural network (ANN) that uses an unsupervised learning algorithm which allows mapping high-dimensional elements into a grid of a few (usually two) variables, called a map. These kinds of artificial neural network are different than the classic ones in the sense that they use neighborhood function to preserve the topological properties of input space.

Every neuron in the network is associated with a weight vector which has the same dimensionality as the training vector, a relative position inside the grid and also a set of neurons called neighborhood. This set of neurons will define the type of topology of the

network, the rectangular and hexagonal ones being the most usual. There is a special case of networks with a toroidal topology where the first neuron on the map will have as neighbors the one to the right, the one below it, the last of its column and the last of its row (see figure 2.10). The neighbors of the rest of the neurons are defined in this same way.

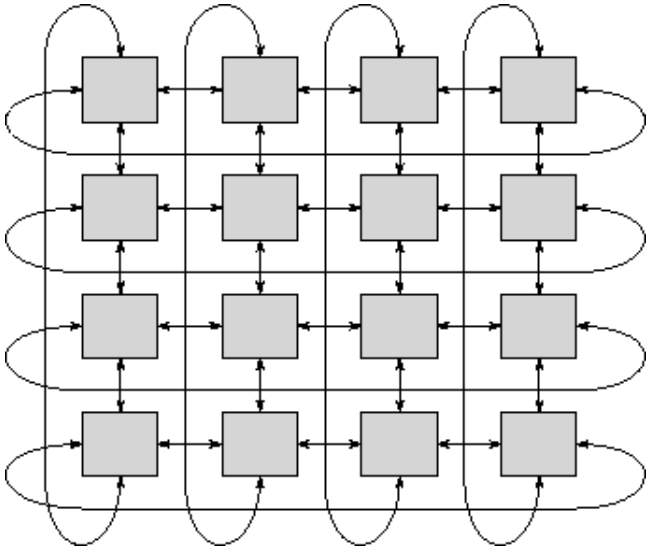

Figure 2.10: Example of a toroidal network

As was explained before, the neighborhoods in the algorithm define a topology and its influence radius declines through the iterations of the algorithm. This means that in the beginning of the algorithm process the pattern generates a lot of changes in the winner neuron and its neighborhoods, but in the following iterations these modifications will apply only to the nearest neighbors and at the end of the algorithm will affect only the winner neurons.

The training algorithm has a competitive nature because all the neurons "battle" between them to define which one is the most similar to the example. When the winner is found, arrangements to the network are made in such a way that the winner neurons become similar to the examples.

It is not possible to use the SOFM algorithm in a direct way for the clustering session process due to the nature of the data to be grouped. To do this every neuron will be defined as an IOV, and to update them during the training the similarity measure between the vectors will be used. For every IOV it will be necessary to find the neuron most similar to it and then update the net weights according to the distances calculated. This process is iterative and will end when the net weights are lower than ϵ.

K-Means The K-means algorithm allows the grouping of data sets into K clusters where the objects in a cluster are more similar to other objects in the same group (homogeneous) than when compared with the objects in the other clusters (heterogeneous).

The learning method in this algorithm is supervised because it requires the number of the K clusters to be created. This algorithm works using a *Top-Down* approach because it starts with a specified number of groups and later assigns the patterns to each one of them. In this algorithm there are no overlaps between groups because each object is assigned only to one cluster. It is also a *Two-Phase* algorithm because each iteration starts with an assignation of objects to the clusters and then the *centroids* are calculated.

The main steps of the algorithm are:

- **Clusters Initialization**: Using all the objects as sources, K are randomly selected, assigned to a cluster and set as centroids. Then the rest of the objects are compared with these centroids and assigned to the cluster which has the centroid with the minimal distance to it.
- **Calculating the cluster's centroid**: For each cluster the distance between the centroid and the objects in the cluster is compared. The one with the minimal distance is designed as the new centroid for that cluster.
- **Reassigning elements to the clusters**: For each object in each cluster a comparison is made with the centroids of the cluster and the object is moved to the cluster where the distance between the object and the centroid is minimal.

These steps can be seen in figure 2.11

This process is repeated until there are no changes between clusters or until the error between iterations stays stable (i.e. does not change).

Association Rules This is a "clustering" method which uses unsupervised learning to group objects, while attempting to understand the links or associations between the different variables or attributes of the group. The main rules of this method follow the pattern:

IF variable1=a **AND** variable2=b **THEN** variable3=c

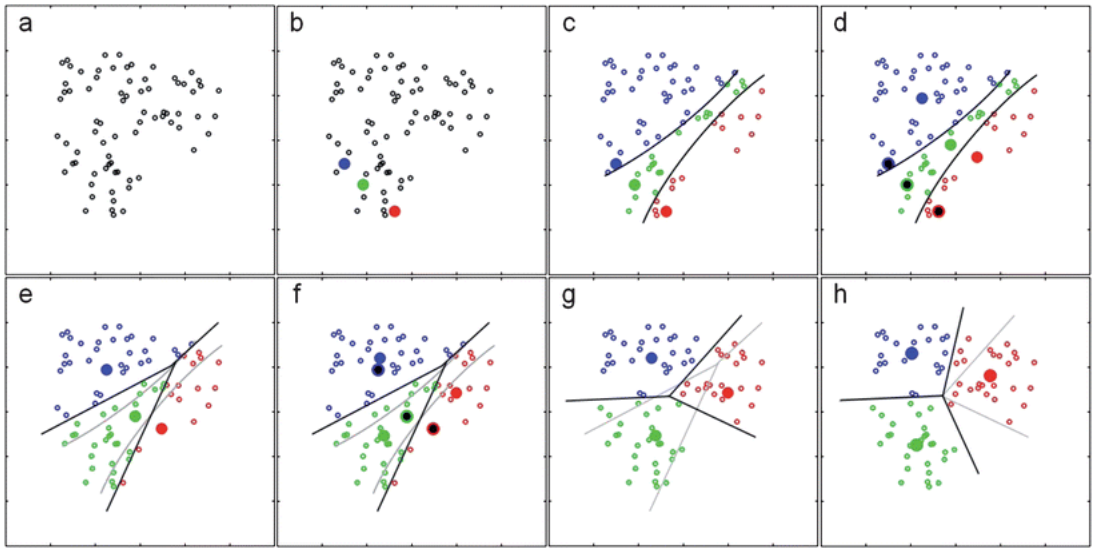

Figure 2.11: Steps of k-means algorithm

A good example of this method can be seen in [52]. Here a set contains different objects such as are shown in figure 2.12.

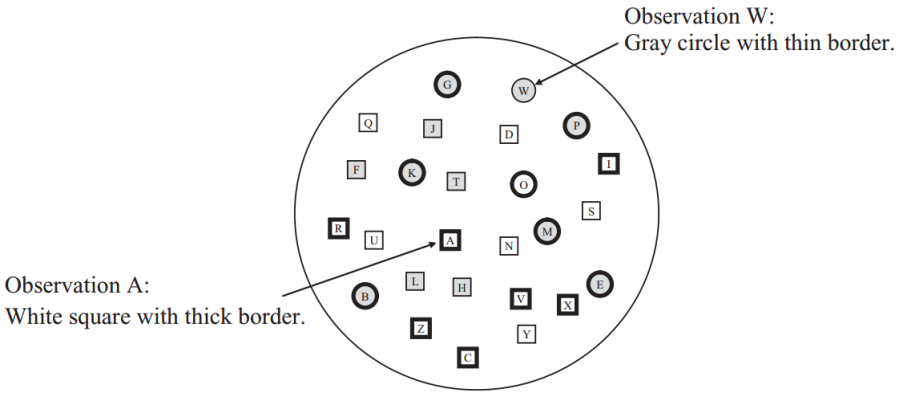

Figure 2.12: Example of Association Rules
Source: Figure 6.26. in [52]

In the figure 2.12 the objects have different kinds of shapes, colors and borders. One possible rule for this set could be:

IF Color = Gray **AND** Shape = Circle **THEN** Border = thick

The set has six objects that meet the rule (i.e. there are six objects which are gray, are circles and their border is thick). But this observation can be made by means of other rules using the same variables with different values. For example:

IF Border = thick **AND** Color = gray **THEN** Shape = circle

IF Border = thick **AND** Shape = Circle **THEN** Color = gray

To measure how good a rule is and to compare it with others there are three indicators:

- **Support**: Support is a way of measuring the total number of observations that the rules map. It is usually defined in terms of proportion or percentage. In the previous example, using the first rule, there are six gray circles with a thick border out of a total of 26 objects. Therefore the support for the rule is 6/26 or 23%

- **Confidence**: The rules are divided into two main parts, the *If-part* or *antecedent* and the *THEN-part* or *consequence*. The first part refers to the statements linked with the and in the first part of the rule, while the second part refers to the statement after the THEN. The confidence indicator is a measure for how predictable a rule is, and its formula is:

$$Confidence = groupsupport/IF - partsupport \tag{2.4}$$

  In the example the support value for the group was 0.23 and the support value for the IF-part of the rule (number of gray circles in the data set) is 7/26 or 0.27. Dividing these two values yields:

$$Confidence = 0.23/0.27 = 0.85$$

  This means that rule number 1 has a confidence of 85%.

- **Lift**: Lift is defined as the importance of the rule because it describes the association between the IF-part of the rule and the THEN-part of the rule. The calculation of the indicator is made by dividing the confidence value by the support value across all observations of the THEN-part

$$Lift = Confidence/THEN - partsupport \tag{2.5}$$

  For the example, using rule number 1, the confidence is 0.85. The THEN-part support of the rule is 13 out of 26 (there are 13 objects with thick borders out of a total of 26 objects), i.e. 0.5. Dividing these two values yields:

$$Lift = 0.85/0.5 = 1.7$$

When a lift value is bigger than 1 it indicates a positive association. On the other hand when it is less than 1 it indicates a negative association.

To find rules inside a data set there is an algorithm called *Apriori* defined in [53]. This algorithm tries to find subsets of items which meet a certain level of confidence C defined by the user. To do this, the algorithm uses a *Bottom-Up* approach where one element is added to the subsets for each iteration, to test if they comply with the restriction. To store the data the algorithm uses a tree structure, where it keeps the candidates for rules. This kind of data structure allows the application of *pruning* techniques to increase the efficiency of the algorithm.

Chapter 3

Experiment Design

This section will explain how the hypothesis will be tested. To do this it will be necessary to develop an experiment to check the improvements of the new methodology versus the previous one. With this information it will be possible to compare the results of both methods and verify if there is an improvement using the pupillary response as a new data source. This experiment, briefly explained before, will be thoroughly detailed in this section.

3.1 Requirements

To complete the experiment some requirements needed to be accomplished. These requirements are mainly related directly to the data capturing process because this is one of the most important steps in the experiment. The main idea is to get a clean environment and data, without any kind of bias or other kinds of distorting variables. First the characteristics required for the experimental group will be explained, then the type of web site required and its characteristics and finally the devices needed to get the eye movement data and for the measurement of pupillary response.

3.1.1 Experimental group

The experimental group is one of the most important components of an experiment, because the data obtained with this group will be used for extrapolating the results to the rest of the universe. For this reason the number of participants and their profiles will be the two main variables for choosing the group. A bad decision could lead to invalid results or inconsistency with reality.

If possible, it is highly recommended to have a sample size of 33 participants as a minimum because according to the central limit theorem, with this number the mean and the variance of the sample will be similar to the rest of the population. In this case, the main variable of the study is the permanence time of the users on the web objects.

If the last point is impossible to reach, it will be necessary to have a representative sample of the web users and potential customers of the site. To get this information it will be necessary to have the assistance of the site's web master or the site's business expert. Any other kinds of users will also be considered, because it is completely possible for anyone on the internet to access the site.

The main profile of the participants will be:

- Age range between 16 to 50 years, according to the range of world Wide Web users.
- Having a high school degree at least.
- Having used or knowing how to use a web browser
- Having passed the selection survey (see appendix A)
- Information about expertise level will be collected, but will not be an exclusion variable.

The sample selected represents the main characteristics of the internet users and, in particular, the users of the web site chosen. Does not include participants who could access to site, like children or elder people, or people who does not know how to use a web browser or computers, because the results gotten through those participants could add noise to the data for analysis, resulting in less accurate results. This attributes makes this sample a valid sample of the main internet user.

It is important to notice that each participant will have to sign an informed consent document allowing the usage of the data collected through experiments with them (see appendix B).

3.1.2 The web site

There are three main characteristics needed for the web site in this experiment. The first being the number of pages on the site, the second is the number of objects inside every page on the site, and the third is having a large number of visits.

The quantity of pages on a site is important because it will determine the complexity of finding the web site keyobjects. With an excessive number of pages the problem will be too complex for testing the hypothesis, adding more problems to the project. On the other hand, a site with too few pages will not benefit from the experiment approach. Nowadays on the Internet it is possible to find an enormous number of different web sites, containing different quantities of pages, from just one to hundreds and even thousands of pages. For this experiment it will be necessary to have a site with a number of pages between 50 and 100.

Secondly the number of objects per web page will determine the complexity of finding which ones are the important ones and this number is also closely related to the usability level of the site. In the case of a page having just one object it will not be necessary to check it, because the amount of time spent on the page will be the amount of time focused on the object. In the case of having too many objects per page it will be recommendable to redesign the experiment instead of testing it with that page. For this experiment it will be necessary to have an average number of 20 objects per page.

The third point is vital for doing the mining analysis. This kind of processing requires a large amount of data for finding patterns and then discovering knowledge, so it will be completely necessary to have at least a sufficient number of visits as well as access to the corresponding web log files.

Finally another requirement could be useful for this experiment, and that is the possibility of changing the pages on the site. This could help in further development, because if it is necessary, the improvements could be applied directly on the page to test the results and check if there are any kinds of changes in user behavior. In any case, the experiment is completely feasible without this feature.

3.1.3 Eye-tracking system and devices

Nowadays most of the eye-tracking systems available on the market have enough precision for this kind of experiment, but the main problem is their cost. For this project an eye-tracking system will be necessary which can allow the identification of the objects watched by the participant, while retrieving all the necessary data for the experiment. It is also necessary for the system to allow the researcher to measure the pupillary response of the participant, retrieving the data related to it.

3.2 Capturing data

This section describes the process of capturing data from the different sources including the experimental group, the web site, the users, etc. All these sources have different kinds of data, with different formats and specifications, and for this reason it is necessary to treat with them in specific ways for each one.

3.2.1 Experimental group

To capture the user interest in the objects on a page one approach will be used, an eye-tracking system which allows the measurement of pupillary dilation. A survey, as was conducted in [3], will not be done because the previous methodology showed that with the eye-tracking system it is possible to get better results compared to the use of a survey.

Eye-tracking system

To get a quantitative measurement of the time spent on a web object, and the pupillary response produced by it, an eye-tracking system will be used. This device will obtain the ocular movements of the user, and with this information it will be possible to determine the time spent on the objects. In the same way, the data about pupil dilation will be obtained, and combining this information with the ocular movements will allow the response to the web objects to be determined.

To get this information two approaches will be used, assigning a task to the user on the site and presenting the pages as a stimulus.

- **Task assignments:** It is usual in web usability experiments to assign a task to the users, and during this assignment to record the ocular movement of the participants. These assignments or tasks have to be directly related to the normal usage of the site. For example asking the participant to buy a ticket on an airline site or buy a product on an online shopping site. To choose a task assignment it is necessary to have a good level of knowledge of the site process, because the ocular movements are directly related to the kind of task requested [54]. If a task does not correspond to the normal usage of the site the experiment will not generate proper results. For this reason it is necessary to validate the tasks or assignments with the site's administrator because he is the one who has the best knowledge about the main process of the site.
- **Pages as stimulus:** This approach tries to retrieve data in a different way, mainly because this experiment does not want to get information about usability but about

web site keyobjects and the time spent on them. Instead of asking for a specific task or assignment, the pages of the site will be shown to the participants as visual stimuli, in a sequential order. The participant may change the page that is being visualized when he considers there is nothing else interesting on it. In any case each page may be watched a maximum amount of time before it changes automatically to the next one. This time will be obtained from the web log files.

3.2.2 The web site

All the components of the chosen site will be separated for analysis. To do this a web crawler will be used which will retrieve all the pages on the site and the links between them.

The pages of the site captured by the web crawler will be transformed to the *png* format. With this transformation it will be easier to map the objects on the page and, in the same way, to match data between the eye-tracker data source and the pages on the site.

Objects

The process of identifying and delimiting objects will be done, if possible, with the web master or site administrator, as the person who best understands the site components, or if this is not possible, at least validate the delimitation done. The steps developed by Dujovne and Velásquez in [3] will be followed to describe the content of the objects on the site. A list of word descriptors and their formats will be saved after this process.

3.2.3 Web logs

The web log file will be retrieved by the site's administrator, directly from the server machine. This file could be located in different places depending on the site's web server and the setup.

3.3 Data transformation

To analyze the data retrieved from the sources it is necessary to transform them. This will permit them to be processed with data mining algorithms, and then the information to be retrieved, this being step one of the critical steps of the KDD process. The transformation process is different for every data source, and this section will explain the detailed steps.

3.3.1 Web pages

The site's web pages will be transformed to a *png* file (image format), including data such as URL and size.

Objects

The objects on the pages will be detailed, and the information such as format, description, size, etc. will be completed and this information will then be stored in a relational database. A list of concepts describing the objects and their categories will also be stored, along with the relation between objects and concepts.

To know which objects are on which pages, a relational table will be created, to easily map the objects onto the pages. This relation will also store data about the left superior coordinate, and using the size data, it will be possible to map any pixel on the page onto an object. This thus enables the researcher to know which objects are being looked at by the participant during the experiment.

After the mapping step and storing process, the conceptual distance between objects will be calculated, according to the equation 2.1. This information will also be stored in a relational table, to be used by the data mining algorithms. This will avoid the calculation process during the KDD process, and improve the efficiency of the whole process.

The tables, keys, and fields of the database are shown in figure 3.1.

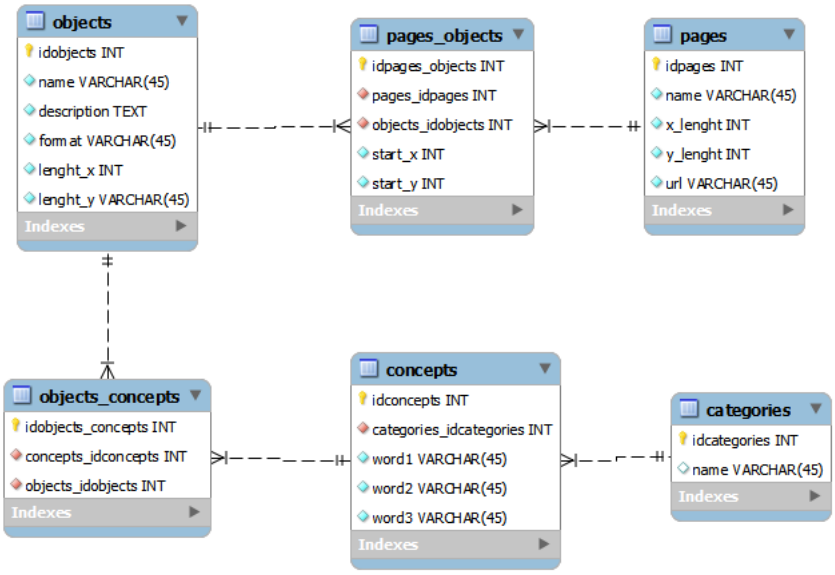

Figure 3.1: EER Model of the site and it's objects

3.3.2 Web logs

After retrieving the web log files from the server, they will be processed to extract the data as detailed in section 2.1.1). These data will be stored in a relational table on the database, allowing them to be worked with directly. This will be necessary for the sessionalization process, to obtain the information about the time spent on every page (see section 2.6.5). The results of this process will also be stored in another table in the database, allowing the researcher to know the page visiting sequence and the time spent on each one of them. The figure 3.2 shows the tables, keys and fields corresponding to the data that will be stored.

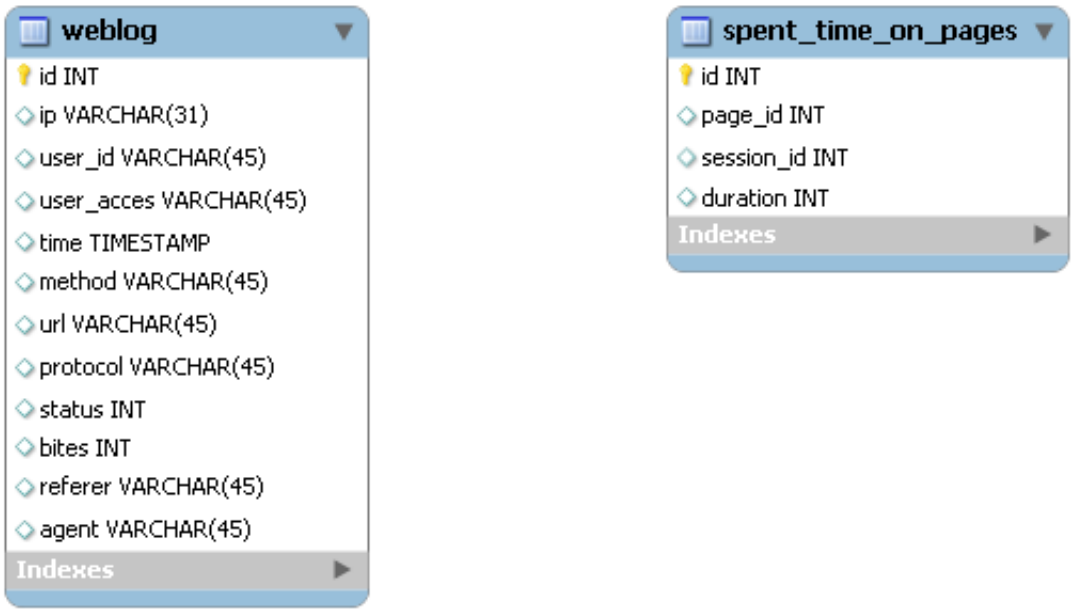

Figure 3.2: EER Model web log tables

3.3.3 User interest

Next the process of data transformation of the main data source of user interest in this experiment is explained, the eye-tracking system.

Eye-tracking system

The transformation process of the eye-tracking data is mainly for knowing the time spent on every object and the pupillary response produced by them. The object mapping of every page is needed (see section 2.6.1) for correlating that data with the eye movement data on the page, mixing the coordinates watched and time spent with the objects. With this information it will be possible to calculate the time spent on every object, as well as the

pupillary response produced by them. After that this data will be stored in a table in the database, according to figure 3.3.

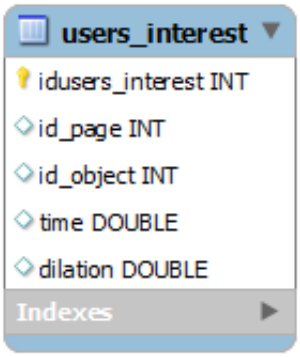

Figure 3.3: EER Model User interest table

The time spent on the objects will be transformed to a percentage of the total permanence on the page, and after that an average will be calculated with the participant's data. This will be done in an analogous way with the pupillary response of every object.

Then, using the data stored in the database after the sessionalization process and after the processes of transforming the eye-tracking data, the time spent on every page will be weighted with the permanence percentages on the objects obtained in the last two transformations.

3.3.4 Experimental group

Some data will be asked of the participants in the experiment, information such as age, gender, profession, level of internet knowledge, knowledge about the research and knowledge about the site. This will be stored in a database table along with the other tables.

The transformations here will be done to try to find some kind of correlation between the type of user and the interest in the objects, the knowledge of the research and the relation between the integrity of the answers (the relation between the answers on the survey and the results of the eye tracker), etc.

3.4 Changes in the methodology

The main change between this methodology and the methodology proposed by Gonzalez in [4] is the addition of the pupillary response as a data source using the eye-tracking system. In the same way the main change between the methodology proposed by Dujovne and the one proposed by Gonzalez is the way of getting the interest data. The latter includes the time spent on every object using the eye-tracker machine as source, but for this methodology the

data about pupillary dilation is also going to be used as a significant variable. The experiment will use this variable on the web mining algorithms for identifying if the pupillary response can be related to the identification of web site keyobjects.

3.5 Results comparisons

The results of both methodologies (the one presented in this research and the one developed by Gonzalez in [4]) are a set of web site keyobjects. To compare the results the assistance of the site business expert will be necessary, who will decide if an object categorized as a keyobject was matched correctly. After this step both methodologies will be compared using the precision comparator which is useful for evaluating a classification algorithm precision.

Chapter 4

Implementation

This chapter explains the entire implementation of the experiment designed in Chapter 3, in order to achieve the proposed result, detailing its most important components.

First the details of the selected components for realizing the test are given. These include the web site, the participants and the devices.

Next the process of testing and capturing data are detailed, explaining the whole process of recruiting participants, the trials and the process of getting the data from the devices.

Finally the KDD process developed for the data is explained. Details about selection, preprocessing and transformation of data are given.

4.1 Requirements

This section describes the main requirements chosen for developing the experiment and the tests. Specifically it shows the chosen experimental group and their data, the web site selected for testing, the machines and devices for capturing data and the software and code used for processing the captured data and for getting the results.

4.1.1 Experimental group chosen

For a better process of choosing the participants in the experiment, first the main focus of the web site was established and the type of customer they are looking for. In this way it was possible to choose a representative sample of the main clients of the web site. This

information was retrieved through the web administrator, who is the business expert.

With this data a number of 15 were chosen for testing the site, of whom seven were male and eight were female. The average age of the experimental group was 26.1 years, with a variance of 2.2 years.

The group involved was mainly composed of university students and professionals, of different knowledge areas. Among them were such occupations as engineers, biologist, kinesiologists, students, etc.

According to their own experience (no test was developed for measuring this), one considered him/herself an expert on web navigation, 12 considered their knowledge as average and the other two think they are basic or amateur users.

In table 4.1 the data and attributes about every participant of the experiment can be viewed.

Number	**Sex**	**Age**	**Profession or Activity**	**User level**	**Knowledge about project?**
1	Male	29	Engineer	Medium	Yes
2	Female	25	Student	Low	Yes
3	Male	25	Student	Low	Yes
4	Female	24	Student	Medium	Yes
5	Female	27	Kinesiologist	Medium	Yes
6	Male	24	Student	High	Yes
7	Female	28	Psychologist	Medium	No
8	Female	28	Student	Medium	No
9	Male	29	Psychologist	Medium	No
10	Male	24	Student	Medium	No
11	Male	26	Student	Medium	No
12	Female	26	Student	Medium	No
13	Female	24	Student	Medium	No
14	Male	23	Student	Medium	No
15	Female	30	Student	Medium	No

Table 4.1: Experimental group characteristics

It is important to note that, in contrast with the previous work developed by González in [4], in this project it was impossible to differentiate between the people who had knowledge about the project and the ones who did not. This was mainly because the devices used for capturing data are different, and work in different ways for the same purposes. In the previous study it was possible to track the eye movements without the participant knowing about it, but with the new device this is not possible due to its physical characteristics. In any case, according to the previous study, this variable did not have any important consequence on

the results.

4.1.2 Web site chosen

The experiment was tested on the site *MBA Ingeniería Industrial - Universidad de Chile* developed by the Master in Business and Administration program of the Industrial Engineering Department of Universidad de Chile. This site has been running since January 2011 and provides information about the program offered by the University. Among the pages on the site it is possible to find information about the courses, the methodology of the plan, teachers, student profiles, etc., as well as some pages for applying to the program.

At the beginning of the experiment the site was composed of 32 pages and 359 objects, and they appear 1014 times in total. This means that an object can appear on more than one page, which behavior is quite usual with such common objects as banners, menus, footers, headers, etc. The average number of objects per page is 31.9 and their average size is 418.6 pixels wide and 100.1 pixels high.

The site's statistics, for the first seven months of this year (January to August), show that the average monthly number of visits is 5320.5, the average number of distinct sessions are 2960.5 and an average of 26512.5 pages are seen. In the table 4.2 it is possible to view the statistics for the last seven months.

Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth (in MB)
January	4889	9439	36592	219120	8.23
February	2181	5274	26850	107951	4.13
March	2597	6141	31772	145755	5.30
April	2779	6257	26119	135582	4.23
May	2083	5330	21711	101204	3.87
June	2542	6483	19937	93108	3.50
July	2505	6930	30611	105958	3.24
August	4108	7744	42673	138970	4.14
Total	**23684**	**42564**	**236265**	**1047648**	**36.64**
Average	**2960.5**	**5320.5**	**29533.1**	**130956**	**4.58**

Table 4.2: Site's statistics

4.1.3 Machines and devices

To capture eye movements and to measure pupillary dilation a fourth category eye-tracker system was used. This category, according to section 2.4.3, corresponds to an eye-tracking system which uses video-based combined pupil and corneal reflection. These kinds of devices are the most advanced system for measuring this kind of movements according to the degree of accuracy obtained with them.

In particular the device corresponds to eye-tracker model *Eyelink 1000*, developed by *SR Research* company. This device is composed of a main screen with a high speed camera and an infrared emitter, connected to a host computer which does the data processing. There is one more computer which is also connected to the host. This one allows the researcher to develop the experiments and get the data in an easy way. It also shows in real time the same stimulus that the participants see. The whole setup for a trial takes around 2 to 5 minutes.

The eye-tracker specifications are shown in the table 4.3

Sampling Rate	2000 Hz Monocular / 1000 Hz Binocular
Accuracy	0.25° - 0.5° average accuracy
Resolution	0.01° RMS, micro-saccade resolution of 0.05°
Real-time Data Access	1.4 msec (SD $<$ 0.4 msec) @ 2000 Hz

Table 4.3: SR Research Eyelink 1000 specifications

This system belongs to the Medical Faculty of the *Universidad de Chile* and they provided all the necessary information and components for working with it.

4.1.4 Software and code

To implement the algorithms and process the data it was necessary to choose an adequate set of tools. These were chosen for their relative simplicity, for their easy access and also for the general information about them. It is always necessary to choose tools which allow the researcher to achieve the goals of the project and fulfill the requirements.

The main tools used in this project were software for processing data and programming languages for capturing data. They are detailed below.

- **Ubuntu Linux Operating System:** Ubuntu is a computer operating system based on a Linux distribution called Debian. It is distributed as free and open source software, and is developed by Canonical Company. According to online surveys, in 2012 it was the most popular Linux distribution on desktop/laptop personal computer, which is also the main market of the company. This OS was chosen because it is a well-known system, stable and reliable for programming ambiances, as well as for the availability of fully documented applications and its user friendly interface called Unity.

- **Python Programming Language:** Python is a high-level programming language, used for general purposes. It runs on such different operating systems as Windows, Linux, Mac OS, etc., being also included in such Linux distributions as Ubuntu. Python supports such multiple programming paradigms as functional programming,

object-oriented and imperative. This programming language was chosen for these characteristics:

  - Easy learning language allowing the development of complex algorithms in a short time
  - Free for use even in commercial products
  - Clear and easy to understand syntax
  - Fully documented and extensive standard library
  - High number of third-party libraries
  - Intuitive object orientation and full modularity
  - High level dynamic data types such as lists, arrays and other data structures.

- **XML (eXtended Markup Language):** XML is a markup language for encoding documents in a format that is both machine-readable and human-readable, according to a set of rules which defines it. Their specification is mainly defined by W3C. This language is based on a textual data format emphasizing simplicity, generality and usability on the Internet. It is widely used for representing such arbitrary data structures as, for example, web services.

- **MySQL:** It is an open-source relational database management system (RDMS) which runs as server providing access to multiple data bases for multiple users. It is available for many platforms including Windows, Linux and Mac OS and it is widely used by different web applications and also for large-scale websites as Wikipedia, Facebook, Google and Youtube. The source code is available under GNU General Public License. This system was chosen mainly for being free for non-commercial systems, the availability of documentation and the support of different database engines.

- **Matlab:** It is a language programming language and a numerical computing environment developed by MathWorks Company. This system allows working with numerical data in an easy way, enabling plotting of functions and data, implementations of algorithms, matrix manipulations, creation of user interfaces. It also interacts with such other languages as C, Java and Python.

4.2 Capturing data

The data used in this kind of research comes from a lot of different sources, and the variety of data types is likewise large. This section will try to describe the process of capturing this

data and how the storing process was managed. The section is ordered according to the source of the data.

4.2.1 Data about the experimental group

The data about the experimental group is mainly the data obtained from the eye-tracking device. To collect this it was necessary to explain the experiment to the participant in every trial. They were located in front of the display and the eye-tracking device, according to the figure 4.1, and for every trial the first 20 participants were required to do a task related to finding information on a web page for making a decision about it. The exact instruction was: "You are interested in taking an MBA program, but you are still unsure about the decision. During the information searching process you found the site www.mbauchile.cl. This site belongs to the MBA program of the Industrial Engineering Department, and has the important information about that program. Starting at the home page, browse freely on the site until you have the necessary information for making a decision." This instruction tried to emulate the normal behavior of the user during their stay on the web site.

Figure 4.1: Eyelink 1000 layout with a subject

The remaining subjects were required to watch the site without browsing on it, but just watching random pages of the site. The subjects could change the page they were seeing to the next one at any moment, but the system only allowed a maximum stay of one minute on every page. If the user exceeded the maximum stay time the next page was presented automatically. The number of pages presented to the subjects was 32.

For every trial the eye-tracker device retrieves a log file with the data obtained with it. This file has multiple fields indicating the values of the variables during the whole trial. The typical sample line looks like this:

```
<time> <xpl> <ypl> <psl> <xpr> <ypr> <psr> <xvl> <yvl> <xvr> <yvr> <xr> <yr>
```

These variables are:

- <**time**>: timestamp in milliseconds.
- <**xpl**>: left-eye x position data.
- <**ypl**>: left-eye y position data.
- <**psl**>: left pupil size (area or diameter).
- <**xpr**>: right-eye x position.
- <**ypr**>: right-eye y position.
- <**psr**>: right pupil size (area or diameter).
- <**xvl**>: left-eye instantaneous velocity (degrees/sec).
- <**yvl**>: left-eye instantaneous velocity (degrees/sec).
- <**xvr**>: right-eye instantaneous velocity (degrees/sec).
- <**yvr**>: right-eye instantaneous velocity (degrees/sec).
- <**xr**>: x resolution (position units/degree).
- <**yr**>: y resolution (position units/degree).

4.2.2 Data from the web sites

To retrieve the pages on the site *http://www.mbauchile.cl* the *wget* program was used. Wget is a computer program that retrieves files from web servers, supporting downloads via HTTP, HTTPS and FTP protocols. This program is included with the Ubuntu distribution explained before.

Every page on the site was converted to a PNG[1] file for mapping the objects later. The data about the image was stored in the database shown in figure 4.2.

[1]PNG is the acronym for Portable Network Graphics, a popular image format on the World Wide Web

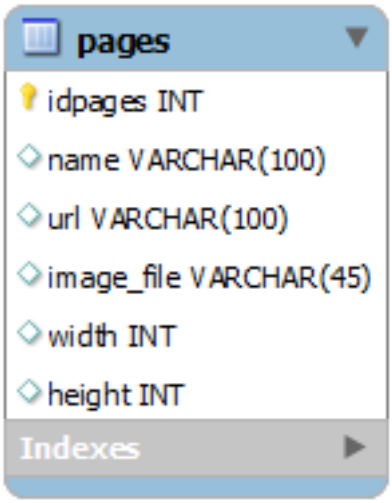

Figure 4.2: EER model of pages table

4.2.3 Web log data

The web log files were obtained directly from the web server where the site is running, with the help of the web site's administrator. It was not possible to get all the previous data, because a repository for these kinds of files does not exist. However, it was possible to recover the files from the month of August, 2013. During this time 4108 different persons visited the site in 7744 sessions. The number of pages seen was 42,673, the number of requests to the server was 138,970 and the traffic was 4.14 GB.

4.2.4 Objects on the site

After the site was completely retrieved from the server the objects definition process began. To identify the objects on every page the main variables considered were the concepts associated with each object and the spatial distance between them, in the same way as the previous work in [4]. After the process the result gave 359 objects and most of them appeared on several pages on the site.

To get the coordinates of each object the image of the file of each page was used. Using the data about width and height of each object, and knowing the top-left corner, the calculation of the four corners was easy.

4.2.5 Objects concepts

After the objects were defined, a list of concepts for defining them was made. This was done in the same way as the two previous methodologies in [3] and [4], through a manual

analysis of every object. This method was chosen instead of some automatic method, such as *Latent Dirichlet Allocation* or *Pachinko allocation*, because the number of objects was low and these kinds of techniques would add more variables for the analysis, adding more bias to the final result and giving more complexity the algorithms for finding patterns in the data.

The data about the concepts was stored in the same database shown in figure 3.1.

4.3 KDD Process

In this section the KDD process performed on the data obtained in the previous section is described. It is separated into three main topics: data selection, preprocessing and transformation and data mining. This whole process is the core of the experiment and yields the knowledge. More information about the KDD process is presented in section 2.2.

4.3.1 Data selection, preprocessing and transformation

This section describes the process done after getting all the data in the previous section. Each component of this process was separated according to the source of the data due to the differences among them.

Web sites

From the total of pages obtained in the previous step, two were excluded. The first one was the web syndication page (in http://www.mbauchile.cl/feed/rss/index.php). It was excluded because this page corresponds to an RSS[2] file for making the contents of the page available to other sites. This page is not made for people but for crawlers or other kinds of application such as feed readers.

The other page excluded was one that was completely the same as the other page, both as objects and the concepts representing them. They had different URLs, but their session data was included as a request to just one of them.

Objects and concepts

359 objects were found on the site and these results validated by the business expert. However some of the objects were joined with other ones, some were separated and some were deleted.

[2]RSS is an acronym for Rich Site Summary, a popular web feed format on the Internet

The concepts for the objects were established by the researcher, using the same techniques as the previous research. This process was then validated by the business expert of the site, who accepted the definition given to each one. The only modification made in this step was the reassignment of some concepts to the objects that were separated or joined in the previous step.

Then after defining each object's concepts, the algorithm used for calculating the conceptual similarity between objects, defined in [3], was implemented. The detailed description of this algorithm can be found in section 2.6. The results of this process was stored in a table called *concept_similitude* in the previous database (shown in figure 3.1).

Experimental group

The data about the experimental group was collected basically using the eye-tracker device, mixing it with other sources of data such as web logs and videos produced by the eye-tracker software. The survey used in the previous methodology for getting the data about interest was not used in this project.

For every trial of the experiment, the eye tracker generates a *EDF* file, which is a binary file containing the whole register made during the trial, highly compressed. To work with this data it is necessary to translate it to a more readable file. Using the software *EDF2ASC*, provided by the eye-tracker company (SR Research), it is possible to convert it to an ASCII file [3]. This file is easily readable for humans and, in the same way, for parsers that can extract the data contained in them.

The basic data registered by the eye tracker was presented in 2.6, but it also collected other data, possibly including the following:

- **Messages**: Message lines contain the data about different kind of events happening during the trial. These messages can be sent by some applications running on the display computer, and contain data for analysis. The most important are the ones related with mouse actions like moving, clicking, scrolling, etc. Every line includes the timestamp of the event.
- **Fixations**: The file includes data about fixations made by each eye. This line of data includes information about which eye data is being presented, the starting and end

[3] ASCII is an acronym for American Standard Code for Information Interchange, a common character-encoding scheme for writing files.

time of the fixation, the duration, the average position on the X and the Y axis and the average pupil size during the fixation. Every line is timestamped.

- **Saccades**: A timestamped line containing data about the movements made by each eye (a saccade) during the trial. This line includes data about which eye is being presented, the starting and end time of the saccade, the duration, the position of the eye at the beginning and at the end of the saccade on both axes, the visual angle covered by the saccade and the peak velocity during the saccade.

- **Blinks**: Blinks are the period of data when pupil information is missing. These events are registered on a timestamped line showing data about which eye blinked, the starting and end time of the blink and the duration of the event.

Of all the data collected by the eye tracker only some will be used. The most important are those related to fixations and events. For the fixations the data used will be the average position of each eye, the duration of the fixation and the average pupil size. The messages about mouse clicking, in combination with the log server information, will be used to know which links were clicked. The data about scrolling events will also be used for recalculating the relative position of the objects on the screen.

According to [22], the main range of fixation duration is between 150 and 600 milliseconds, so to understand something it is necessary to look at it for at least 150 milliseconds. In this case the eye tracker is configured at 500 hz resolution, which means the data is captured every 2 milliseconds, so there won't be a problem with distortions of the fixation data. In any case, only registries over 150 ms. will be considered.

To map this data into the web pages of the site it was necessary to develop an algorithm for processing the data made by the eye tracker mixed with the web log data. With this information it was possible to discover which pages the subject was looking at during the trial and which objects on those particular pages were looked at.

One of the biggest problems during this stage was the *scrolling* problem. Usually the pages on a website are larger than the screen resolution, and for reading the entire page it is necessary to scroll the page. When this happens, the relative positions of the objects on a page change, some of them appearing and others disappearing. To know the new positions of the objects during a scroll event, it was necessary to improve the previous algorithm by adding a checking feature to the video frame of the trial for every scrolling message. In this particular case it was necessary to add a small object to every page on the website called an *intensity bar* which is basically a long bar with different degrees of color, starting in the color

black and ending in the color white. This object did not affect the usability of the site nor the user behavior on it because the size was minimal, just enough to be noticed by an image analyzer, and was located on the left side of the pages, far from the main parts. Then, when the algorithm detected a scrolling message it checked the video frame corresponding to that timestamp, and with the help of an image analyzer algorithm, which checked the relative position of the intensity bar, it was possible to recalculate the new positions of the objects on the page.

Another problem presented during this stage was the floating menus. A floating menu is basically a navigation bar which reacts when the mouse hovers over it. When this event happens, the menu appears on the screen, increasing the size of the menu object, and covering the objects under it (figure 4.3). On the other hand, when the mouse is not over them, the page behaves in the normal way (figure 4.4). This is a problem because when this happens the eyes are looking at the menu, and not at the objects under it, so it was necessary to add this to the algorithm. To improve the previous algorithm an instruction was added which checked the mouse movement messages in the eye-tracker file, to check the position of the cursor. When the cursor position was over one of the options on the menu bar, the algorithm changed the position and size of the affected objects, making it possible to calculate the correct amount of interest on the objects. In the same way, when the mouse was not there, the algorithm corrected the position and size of the objects.

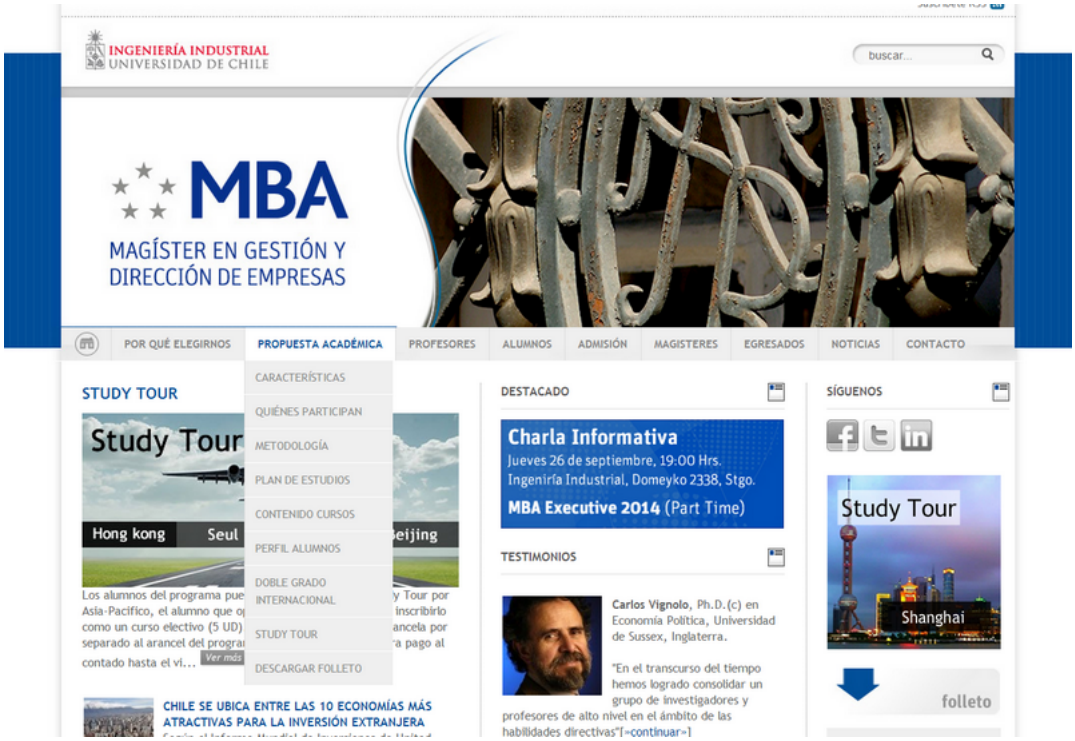

Figure 4.3: Navigation menu opened

After that process, the registries made by the algorithm were grouped by the objects,

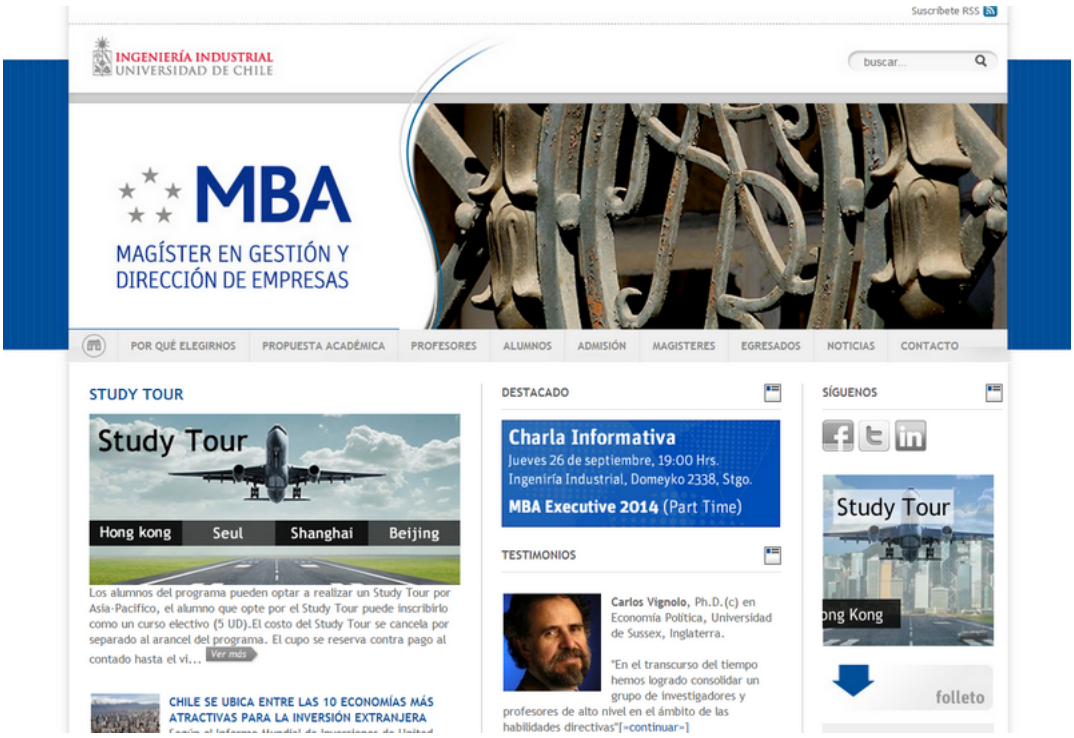

Figure 4.4: Navigation menu closed

and with this set the percentage of time that a user spent on each object and the average pupil size for them were calculated. When an object was not looked at, the percentage of time spent on it and the average pupil size were included as 0, for the correct calculations of average interest.

Finally the time spent on every object and the average pupil size were averaged, and the result was stored in a table called *averaged_interest_objects* (see figure 4.5) . The data stored here were the ID of the page, the ID of the object, the average time spent on in and the averaged pupil size.

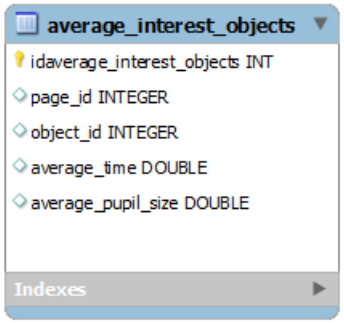

Figure 4.5: Table averaged_interest_objects

Web logs

The sessionalization process was realized following the same process as the previous methodologies [3] [4]. Those methodologies are based on the previous work by Fernández in [55], which is actually validated. The detailed steps of this process are not going to be explained here, being outside the main scope of this thesis project.

This process was performed directly over the web log files provided by the web master of the www.mbauchile.cl site. This web log file was separated by the web server into five different files containing the information specified in section 2.1.1.

The maximum session time was set at 30 minutes, and with this restriction a total of 6622 registries was obtained out of a total of 2219 sessions. This gave an average of 4.1 pages per session.

Then, in the same way as in the previous work, on each page the previously obtained results were transformed into the objects which composed it and then the permanence time was weighted with the percentage of permanence on the object. It is important to notice that some of the objects appear on several pages, and in those cases the subject could see the same object twice or more during the same session. In those cases the permanence times on both pages were added.

For the pupil information the contraction produced by the object during the fixation was used because this is related to the kind of object viewed according to the research. Figure 4.6 shows how the pupil reacts to different kinds of images (neutral, positive and negative).

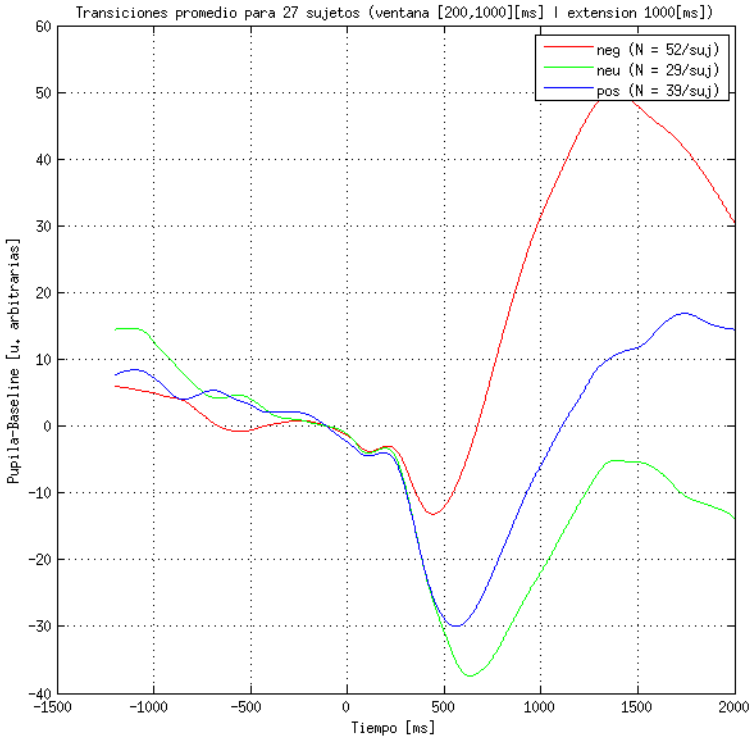

Figure 4.6: Pupil behavior for different kind of emotional valences

According to [56], the images with greater emotional valence are seen more frequently than the neutral ones. This can be seen in the previous figure, where the same results were obtained. On the other hand, in [57] it is shown that images with important content or meaning produce a bigger dilation than the ones without content.

The behaviors demonstrated in those studies are analogous in the way that both of them showed the same behavior for different kinds of stimulus (emotional valence and content information). For this reasons, a new indicator using the difference between the maximum dilation value and the lowest contraction point, for the time range between the start of the fixation until 500 ms. after seeing the object, could be useful for including the data about object importance (in content) and object emotional valence. This information will be calculated and stored in the same way as for contraction. This indicator will be called "delta" because it is the difference between two indicators.

Finally the n most important objects per session were selected, as was described before in section 2.6. There were three different criteria for selecting them. The first one was the same as the previous methodology, where the objects were selected based on the permanence time on them, to get a baseline for comparing results. The second criterion was similar to

the first one, but here the objects were ordered using two variables, first the permanence time and second the pupil dilation percentage. The third way was ordering the objects based on the permanence time, but later reorganizing them using the pupil dilation percentage. To calculate the n value the procedure was realized by first calculating the average (μ) and the standard deviation (σ) of the number of objects per session. Later, the n number was selected in the range $[\mu - 3\sigma, \mu + 3\sigma]$. All the data obtained after these processes was stored in the tables *permanence_object* and *conceptual_similarity* (see figure 4.7).

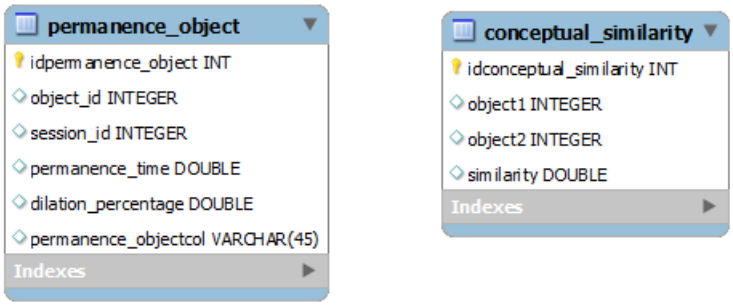

Figure 4.7: Tables object_permanence and conceptual_similarity

In the next steps these data will be used as sources for the data mining techniques. In the table *conceptual_similarity* the data about similitude between concepts are stored and in the table *permanence_object* the data about permanence time, dilation percentage, the object and the corresponding session are stored.

Chapter 5

Results

This chapter shows the algorithms applied to the data and their results, after the processing explained in the previous chapters and sections. These results are explained in a simple and ordered manner, including the transformations done to the data to apply the analysis algorithms.

5.1 Data Exploration

The amount of data retrieved during the experimentation process is quite large and includes many variables to consider. Among the objects only the ones which were seen during the tests were considered, which corresponds to 25% of the objects mapped on the site. This result is expected because many of the objects appear on parts of the pages that the users usually do not visit or which are not interesting.

The objects which were seen were defined on a matrix using five variables: height (in pixels), width (in pixels), area (the result of height per weight), the average time spent on it (in seconds), the delta indicator, the average contraction of the pupil during visualization (using an arbitrary measuring unit which allows comparison of results) and the average number of views.

Variable Correlation

First, Pearson tests were calculated to find out if there was any kind of correlation between the variables of the objects. The results of these tests are shown in table 5.1.

The previous table shows that there are no important correlations between the variables,

	Width	Height	Area	Time spent	Delta	Contraction	N of views
Width	1						
Height	0,2754	1					
Area	0,5754	0,8993	1				
Time spent	0,1419	0,2186	0,2078	1			
Delta	-0.0304	0,0902	0,0963	0.1028	1		
Contraction	-0,0710	0,1075	0,0812	-0,0566	0.4557	1	
N of views	0,04326	0,5129	0,5201	0,2299	0.0183	-0.1221	1

Table 5.1: Correlation between object variables

except for the expected one between height, width and area. The bigger correlation between variables is the one between area and number of views, which makes sense in the way that a bigger object, or an object which takes up more space on the page, will probably be seen more than the others.

Other relations between the variables were plotted to more easily notice whether there were important relations between them. Figure 5.1 shows the plot of the variables area and time, figure 5.2 shows the plot of the variables area and contraction and figure 5.3 shows variables contraction and time spent.

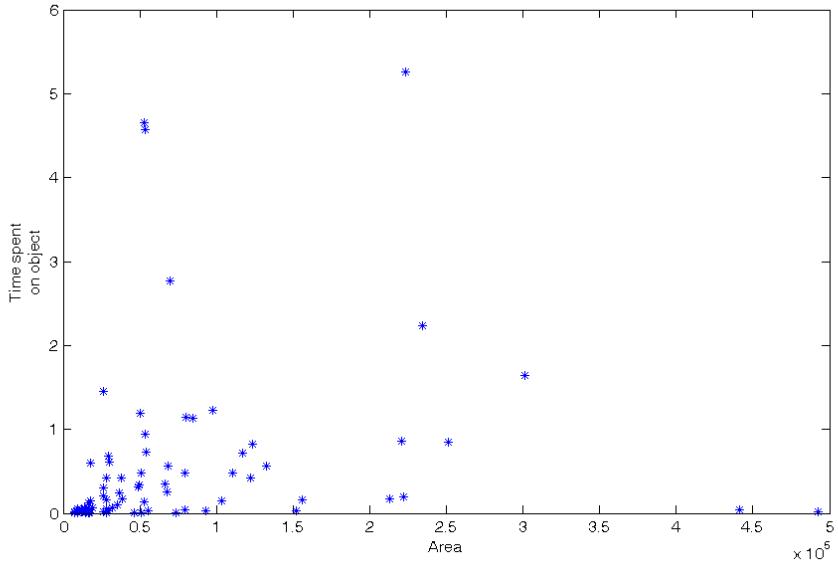

Figure 5.1: Area vs time spent on object

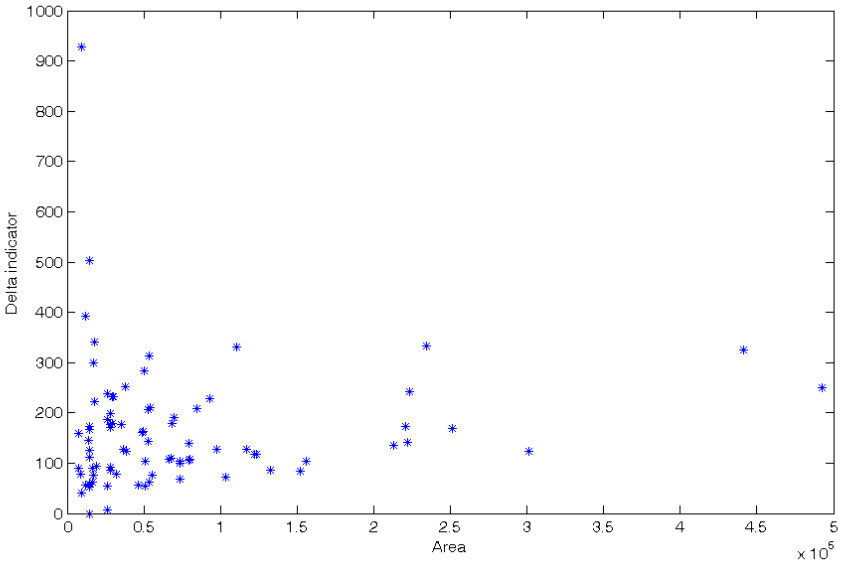

Figure 5.2: area vs delta indicator

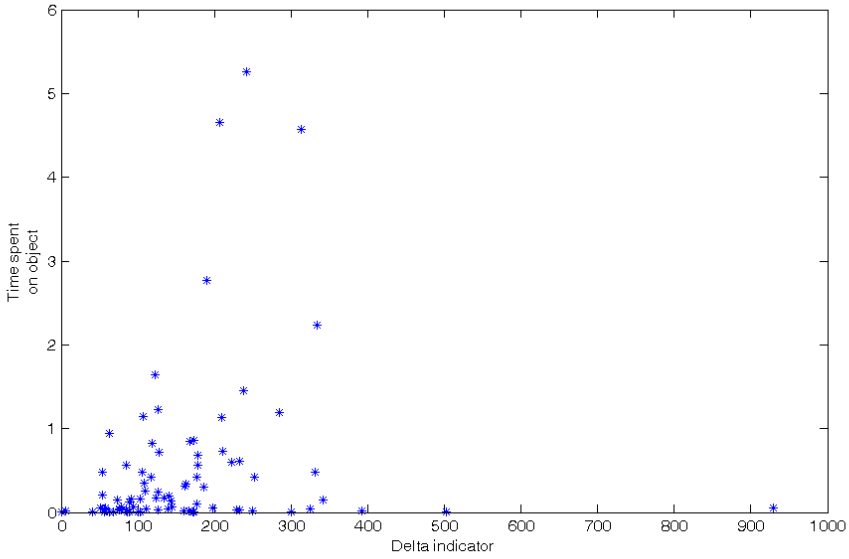

Figure 5.3: Delta vs time spent on object

These plots do not show any kind of relation in a simple way. This was an expected result based on the correlation table shown in table 5.1.

To know if there was some kind of strange behavior in the numbers of objects observed, a histogram of the number of objects per time spent on them was created. Figure 5.4 shows the result.

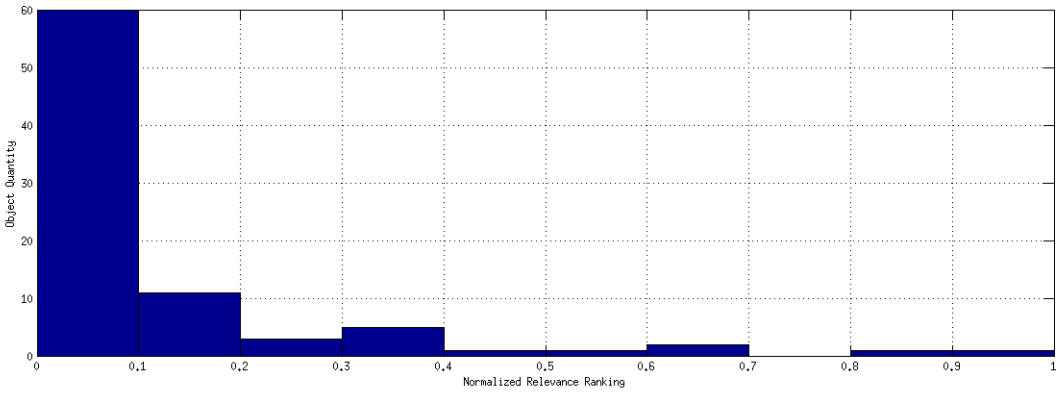

Figure 5.4: Histogram: number of objects per time spent on them (normalized)

The previous figure shows an expected result, where there are a large number of unimportant objects and a few with great importance.

These results just by themselves do not allow doing any kind of interpretation, other than analyzing if the tests were done correctly.

5.2 Association Rules

To apply association rules to the data in this project it was necessary to transform them. This step consisted of defining which objects produced any kind of pupil dilation for each session obtained after the sessionalization process. The algorithm *A Priori* was used to find the rules, using a minimum confidence of 0.9 and a number of 30 rules. The results of this process can be seen in table 5.2.

From the table 5.2 it is possible to determine that there is a cluster selecting the objects 1, 2, 3, 4, 5, 6, 7, 8, 9 (in that order) as some of the ones that produced dilation for the sessions. The objects found here will be used in combination with the other techniques.

Rule number	If part		Then part	Support	Confidence
Rule 1	Object 1	–>	Object 2	99,6999%	100%
Rule 2	Object 2	–>	Object 1	99,6999%	100%
Rule 3	Object 1	–>	Object 3	99,6999%	100%
Rule 4	Object 3	–>	Object 1	99,6999%	100%
Rule 5	Object 1	–>	Object 4	99,6999%	100%
Rule 6	Object 4	–>	Object 1	99,6999%	100%
Rule 7	Object 1	–>	Object 5	99,6999%	100%
Rule 8	Object 5	–>	Object 1	99,6999%	100%
Rule 9	Object 1	–>	Object 6	99,6999%	100%
Rule 10	Object 6	–>	Object 1	99,6999%	100%
Rule 11	Object 1	–>	Object 7	99,6999%	100%
Rule 12	Object 7	–>	Object 1	99,6999%	100%
Rule 13	Object 1	–>	Object 8	99,6999%	100%
Rule 14	Object 8	–>	Object 1	99,6999%	100%
Rule 15	Object 1	–>	Object 9	99,6999%	100%
Rule 16	Object 9	–>	Object 1	99,6999%	100%
Rule 17	Object 10	–>	Object 1	86,1345%	100%
Rule 18	Object 11	–>	Object 1	86,1345%	100%
Rule 19	Object 12	–>	Object 1	86,1345%	100%
Rule 20	Object 13	–>	Object 1	86,1345%	100%
Rule 21	Object 14	–>	Object 1	86,1345%	100%
Rule 22	Object 15	–>	Object 1	86,1345%	100%
Rule 23	Object 16	–>	Object 1	86,1345%	100%
Rule 24	Object 17	–>	Object 1	86,1345%	100%
Rule 25	Object 18	–>	Object 1	86,1345%	100%
Rule 26	Object 19	–>	Object 1	86,1345%	100%
Rule 27	Object 20	–>	Object 1	86,1345%	100%
Rule 28	Object 59	–>	Object 1	85,8343%	100%
Rule 29	Object 21	–>	Object 1	62,9652%	100%
Rule 30	Object 22	–>	Object 1	62,9652%	100%

Table 5.2: Rules created by association rules

5.3 K Means

For K Means clustering the delta indicator variables and the time spent on objects were used. What could be expected from this clustering is a methodology to parameterize the idea of relevance of an object. There were multiple variables such as area, height, width, etc. which could be used for this clustering process, but a qualitative analysis of this kind of objects (great height, great width or both) shows that there is no relation between this variable and the importance. For example the footer object is an object with a large area but is located on the lower part of every page on the site, and does not include any kind of important information, so it is not important for defining a web page. In this way, the variables selected much better represent the relationship with importance and relevance.

For the K Means process three clusters were chosen and the plot can be seen in figure 5.5.

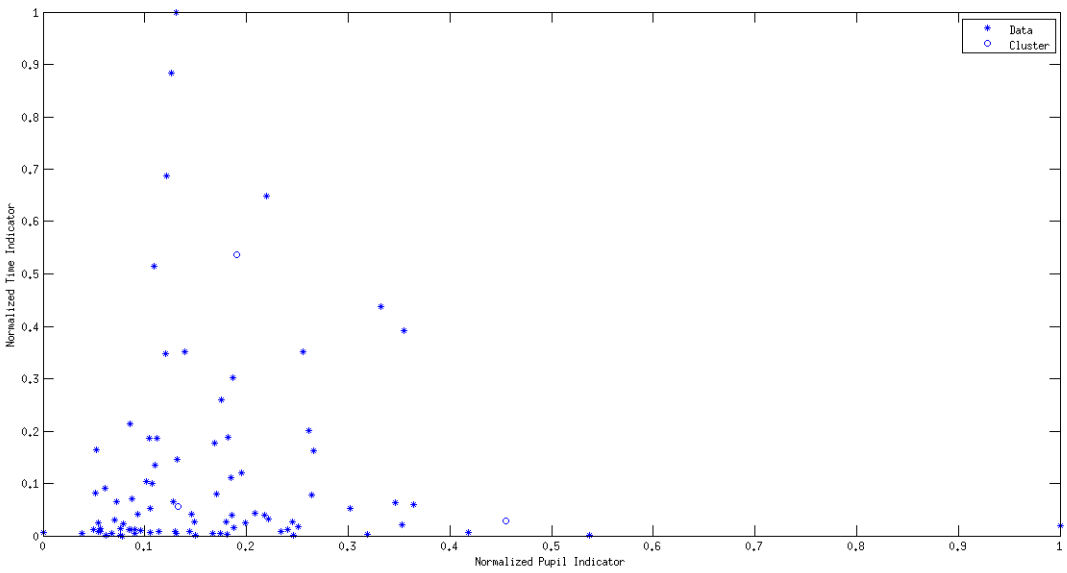

Figure 5.5: K Means clusters for K=3

The results of the clustering process gave three clusters. These were called "High time," "High delta" and "Low importance." The first one includes the objects in the middle-left part of the graph which are the ones with lower delta value, but with a high time spent on them. On the other hand, the "High delta" cluster includes the objects of the bottom-right part of the graph which are the ones with high delta value and low time spent on them. Finally the "Low importance" cluster are the ones with low delta value and low time spent on object value. These clusters are the ones with a higher number of objects. In table 5.3 and table 5.4 the objects belonging to the clusters "High time" and "High delta" respectively can be seen.

High time Cluster
Object 5
Object 195
Object 215
Object 243
Object 244
Object 245
Object 265
Object 267
Object 306
Object 356

Table 5.3: High time Cluster

High time Cluster
Object 130
Object 277

Table 5.4: High delta Cluster

5.4 Website Keyobjects

To combine the objects obtained through both techniques the number of times they appeared on them was counted. After that they were ordered using the indicator delta*time spent on each object, starting with the one with the highest value. The objects with that value being equal to 0 were not considered. The results of this procedure can be seen in table 5.5

Object Number	Website object	Name	Is keyobject?
1	Object 5	header picture	Yes
2	Object 356	picture detalle profe 1	Yes
3	Object 6	navigation menu	Yes
4	Object 195	paragraph detalle noticia 1	Yes
5	Object 4	logo mba	No
6	Object 267	paragraph quienes 1	Yes
7	Object 306	paragraph metodologia 1	Yes
8	Object 244	image plataforma 1	Yes
9	Object 1	ingenieria industrial logo	No
10	Object 265	paragraph corporacion 1	Yes
11	Object 307	paragraph metodologia 2	Yes
12	Object 215	paragraph propuesta 1	Yes
13	Object 243	paragraph plataforma 1	Yes
14	Object 277	title contenido 1	Yes
15	Object 245	paragraph perfil egresados 1	Yes
16	Object 130	title elegirnos 10	No
17	Object 3	search form	No
18	Object 59	titulo cuerpo	Yes
19	Object 23	main post text 1	Yes
20	Object 22	main post picture	Yes

Table 5.5: List of Website Keyobjects

5.5 Discussion

The results in table 5.5 show that there are 16 of 20 objects which were classified correctly. This corresponds to 80% precision which is the same amount of precision obtained with the previous methodology. However the results and the analysis performed on the data shows that, compared to the previous methodology, this technique includes more objects with text, as paragraphs or titles. This may be due to the fact that reading a long paragraph and understanding it requires more cognitive processing, which is directly related to pupil dilation.

In the same way, the indicator created by the multiplication of the time spent on an object with the delta, is a good measure of the importance of the objects on the page. This result could not be obtained before because the previous methodology did not include a way for measuring pupil dilation.

Finally the results obtained do not allow the objects to be related with any kind of emotional valence. This was an expected result because the object mapping process did not show any kind of special objects, the majority being completely neutral. These include the paragraph or text, because their content cannot be easily classified as positive, negative or neutral.

Chapter 6

Conclusions

This work proposes the utilization of pupil dilation as an improvement on the previous methodology for finding web site keyobjects. This methodology is the continuation of a series of studies related to finding these objects, previously begun by Dujovne and later by Gonzalez. The improvement includes the measurement of the pupil and eye movements with newer machines and devices, and includes new techniques and algorithms for processing eye movements and pupil dilation. These techniques are based on the assumption that there is a relation between object importance and pupil movements, in turn based on the previous work related to the cognitive process and its reflection on the pupil.

Thus, the pupil information adds a new dimension for measuring web user behavior, but this kind of information can be related to many more variables such as cognitive process, reading, emotional valences of the images or objects, etc. so the process of disambiguation of this variable is non-trivial.

The use of this variable in this project and the results obtained with it are considered a success because the results obtained were the expected, even in this case, where the results were similar to the previous methodology. In particular the use of this variable and the transformation done with it give a new tool for classifying objects and for finding which ones caused more cognitive processing by the subjects. This opens a new way of web site analysis and exploration, where new kinds of variables can be explored; for example the amount of mental processing needed by the subject, the emotions produced by it, stress level, etc., and not just the analysis of objects.

For the previous reason the main hypothesis of this project is validated, because the

relation between website keyobjects and attention on them and importance as web pages definers was showed through empiric demonstration. It is possible to conclude that pupil dilation it is an important factor to consider in any kind of user behavior on web pages research, and its analysis does not make more difficult these kind of research because the most of the new eye-tracking devices include pupil dilation analysis system.

In particular the results obtained here can also be considered for new improvements of design and layout of the site, when considering the kind of users and the kind of information that are being looked for. These modifications will directly affect the number of people who visit the site and the length of their sessions, which would result in more possible customer attention to the program.

Finally it is important to notice that this methodology and results can be an important improvement to the web analysis done by different companies in the world, which are currently based mostly on such subjective data as surveys and simple observation. To improve this it is necessary to automatize the different parts of the process like the object mapping, the retrieval of the web page, the log analysis, etc.

6.1 Further Development and Recommendations

Based on the research done on this project, there are certain considerations which can be considered for further development of a methodology for finding web site keyobjects, for improving the results and decreasing the bias. This can provide a baseline for developing techniques which include new types of information and variables in order to get better results.

Kind of web site

It is important to notice which kind of web site is used for testing. In this project an informational site was used that includes all the important things about the studies program offered. This kind of site is used primarily for getting a first impression about the program or to get additional information, for which reason users visit it just once on average.

It is necessary to consider this because the objects on the site have to be created with this kind of behavior in mind. On other kinds of sites, like blogs or news, the objects are probably created for repeated visits.

User tasks on the web site

The behavior of the user on the site is closely related to the previous point. The explo-

ration of the site differs based on the type of visits made to it. On a single-visit site all the objects will probably be explored, but on a daily visit site a lot of objects will probably be overridden because they were seen in the past. Accordingly, applying the sessionalization process to this site and supposing that every session is the same can introduce bias in the result.

In the same way, during the experimentation step, it is necessary to give different tasks to the participants for getting all the information about which objects were seen according to the task given. The behavior on the site could be completely different given one task or another.

Web log time

The web log time data has to be considered because it can introduce bias into the results. For example, on the site used for this project, the number of visits prior to the application process into the MBA program is greater than any other month, because people are looking for information for applying, and for the same reason will visit certain pages more than others. Thus, the web log can indicate different kinds of behaviors depending on the time of the year.

It is recommended to have a large amount of web log data, with a large number of requests, and of different time ranges to include all these variables in the analysis. It is also recommended to do an analysis of the user behavior of these sessions to detect whether there are differences according to the season of the year.

Object type

It is important to determine that the objects are of different types. This refers to the fact that some objects are there for "decoration" (as logos, headers and pictures), others are ads (for example banners, pictures or videos), others are the content of the site (paragraphs, text, titles) and others for navigation purposes (such as menus, links, buttons). The behavior on these objects is completely different based on for their functions. For example the menus are seen more than any other object because it is necessary to look at them to know where to go or how the site is arranged. The logos and pictures around the site are usually seen just once because they do not give much more information and are mostly useful for knowing about the owner of the site. The banners and ads are avoided by most experienced users who are "trained" to not look at them. Paragraph and text on the other hand are read to discover the content, and usually every page has different content, so much of time on a page is spent on those objects.

For these reasons it is important to classify them (prior to or after the experiment) and consider this as valuable information for analysis.

Web site page layout

The layout of the page can directly affect the exploration of the site. For example the logos and menus take up an important percentage of the page because they are usually located in the upper part of every page, and on the other hand the contact information is located on the lower part of the page, and it is necessary to scroll to look at them. It is logical to think that those objects that are shown first, on the upper part of page, will be looked at more than others, even if they do not add useful information about the site process.

Creating a classifier or giving a score to the objects, based on such data as length of the page, length and width of the page, area of the object, etc. can be a way of including a weight factor for them.

Cognitive process during the exploration

This project did not include information about the cognitive process during the exploration of the web site. For example it could be important to analyze if some of the objects require more cognitive processing to understand them (for example text too complicated to understand) or if the web site produces any kind of emotion in the subjects (frustration, anger, relaxation, etc.) to know which objects to include or which to exchange for others.

This kind of analysis can be performed using technologies such as EEG (electroencephalogram) for measuring the brain activity of the subject during the experiment, and then analyzing the combination of data obtained through the eye tracker and the EEG.

Web user behavior on the page

Usually a web site is scanned from the upper part to the bottom, and from left to right, but this behavior could be guided according to the objects on it. This is because there is a kind of relation between the objects (for example, after looking at the title of something the next step will be to look at the paragraph related to that title.)

It can be useful to consider these kinds of relations between the objects on a page, and how they are related to the behavior on a page. In other words, how decisions about what to look at and what to choose after each object are made. This kind of analysis could consider the time spent on the objects, the movement of the eyes on the page and the order in which

the objects were seen.

The results of this analysis could help to improve the layout of a site by guiding the behavior of the user on it.

Object mapping

One of the most important steps in this methodology was the object mapping on the site. If the site is big, with a large number of pages and objects on it, the process will be tedious and slow, taking a long time, and being prone to errors.

Therefore, the development of an automated or semi-automated object mapper could significantly decrease the time needed for doing this task, and could retrieve better results, with fewer errors. The development of this kind of system is not trivial because the page code and the image of the page need to be analyzed, and then that information combined to get the objects and their positions.

Appendixs

A Pre Selection Survey

This survey includes 8 simple questions for knowing if the subject is suitable for the experiment. The comments at the side of each question are for the person who is taking the survey.

At the beginning of the survey it is necessary to explain the instructions for the survey: *I'm going to ask some question about your eyes, because we will use devices which registers your eyes movement and the dilation of your pupil. Also we will register your behavior on the web site.*

1. Do you use glasses or contact lenses for reading on your computer or tablet?

   - [ ] **Yes** Continue with the next question.
   - [ ] **No** Go to question number 3

2. Are your glasses made for

   - [ ] **Reading?** Continue with the next question.
   - [ ] **Seeing far objects?** Continue with the next question.
   - [ ] **Both**

3. Can you read on a screen computer or a web page without using glasses or contact lenses?

   - [ ] **Yes** Continue with the next question.
   - [ ] **No** Finish survey.

4. Do you have cataracts?

   - [ ] **Yes** Continue with the next question.
   - [ ] **No** Continue with the next question.

5. Do you have eyes implants?

- [ ] **Yes** Finish survey.
- [ ] **No** Continue with the next question.

6. Do you have glaucoma?

   - [ ] **Yes** Finish survey.
   - [ ] **No** Continue with the next question.

7. Do you use a screen reader, screen magnifier or any other kind of technology for using computer or visiting web pages?

   - [ ] **Yes** Finish survey.
   - [ ] **No** Continue with the next question.

8. Are your pupils constantly dilated?

   - [ ] **Yes** Finish survey.
   - [ ] **No** Continue with the next question.

B Informed consent

Informed Consent

"IMPROVEMENT OF A METHODOLOGY FOR IDENTIFICATION OF WEBSITE KEYOBJECTS THROUGH THE APPLICATION OF EYE TRACKING TECHNOLOGIES, PUPIL DILATION ANALYSIS AND WEB MINING ALGORITHMS"

Main researcher name: Gustavo Martínez Azócar
Institution: Programa de Fisiología y Biofísica, ICBM, Facultad de Medicina, Universidad de Chile.
Telephone Number: 2978 6035

Invitation to take part: We are inviting to take part on the research project called "Improvement of a methodology for identification of website keyobjects through the application of eye tracking technologies, pupil dilation analysis and web mining algorithms", because are needed participants of general population.

Objectives: This research has as objective make research about pupil response as significant variable during the process of free navigation on a web site.

Procedure: If you accept the invitation for participant in the research, you will be tested during a single session for around 20 minutes. During this session you will be asked to watch a series of image sequences on computer's screen meanwhile a camera register the eyes movements and the changes on your pupils diameter. This techniques are harmless and noninvasive.

Risks: The registration of eyes movement and pupil diameter does not have other kind of effects on you. In any case, if you consider that there are some effects due to the procedures, you have to communicate with Gustavo Martínez on the telephone number 2978 6035.

Costs: The procedures specified here doesnt have any kind of cost for you. All necessary means for doing them will be provided by the Neurosystems Lab of Medicine Faculty of Universidad de Chile.

Benefits: This research implies a direct benefit to the knowledge progress, being a significant contribution to understanding of pupil response during cognitive processes.

Compensation: This research does not have any kind of economic compensation for your participation.

Confidentiality: All the information derived of your participation on this research will be stored with full confidentiality, including the access of researchers or research supervisory agencies. Any kind of publication or scientific communication of the rsults of this research will be completely anonymous.

Additional information: You will be informed if during the development of this research appears new knowledge of complications that may affect your willingness of continuing as a participant of this research.

Voluntarism: Your participation on this research is totally voluntary and you can reject it during any moment telling it to the researcher. In the same way the researcher could decide your retire if he considers that decision goes to your benefit.

Participant rights: If you need any other kind of information about your participation in this research you can call to: Researcher: Gustavo Martínez Azócar, phone number: 2978 6035 Institution director: Pedro Maldonado Arbogast, phone number: 2978 6035

Conclusion: After you have received and understood the information on this document, and being able to solve any kind of doubts, I give my consent to participate in this project.

Subject name: _____
Occupation: _____
Education level: _____
Age: _____
Sign: _____
Date: _____

Informant name: _____
Sign: _____
Date: _____

Researcher name: _____
Sign: _____
Date: _____

References

[1] Miniwatts Marketing Group, "Estadísticas de uso de internet y sitios web.." http://www.internetworldstats.com/stats.htm, 2011. 30/06/2012.

[2] V. J.D, R. S, B. A, Y. H., and A. T, "Towards the identification of keywords in the web site text context: A methodological approach," *Journal of web information systems*, pp. 11–15, 2005.

[3] J. D. Velásquez and L. E. Dujovne, "Identifying web site key objects: A methodological approach,"

[4] L. González and J. D. Velásquez, "Mejoramiento de una metodología para la identificación de website keyobjects mediante la aplicación de tecnologías eye tracking,"

[5] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining,"

[6] World Wide Web Consortium, "Http - hypertext transfer protocol." http://www.w3.org/Protocols/, 2012. 30/06/2012.

[7] World Wide Web Consortium, "Html 4.01 specification." http://www.w3.org/TR/1999/REC-html401-19991224/, 1999. 30/06/2012.

[8] World Wide Web Consortium, "Uniform resource locators." http://www.w3.org/Addressing/URL/url-spec.html, 2012. 30/06/2012.

[9] D. Gibson, J. Kleinberg, and P. Raghavan, "Proceedings of the ninth acm conference on hypertext and hypermedia: links, objects, time and space," pp. 225–234, ACM, 1998.

[10] World Wide Web Consortium, "Log files - apache http server." http://httpd.apache.org/docs/current/logs.html, 2012. 30/06/2012.

[11] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A framework for the evaluation of session reconstruction heuristics in web-usage analysis," *INFORMS journal on computing*, pp. 171–190, 2003.

[12] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, 1996.

[13] G. Chang, M. Healey, J. McHugh, and J. Wang, *Mining the World Wide Web*. Kluwer, 2001.

[14] R. K. Blockeel and H., "Web mining research: A survey," *Information Processing Society of Japan SIGNL Note*, pp. 1–15, 2000.

[15] O. Etzioni, "The world-wide web: quagmire or gold mine?," *Commun, ACM*, pp. 65–68, 1996.

[16] P. L. and S. Brin, "The anatomy of a large-scale hypertextual web search," *Computer networks and ISDN systems*, pp. 107–117, 1998.

[17] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: discovery and applications of usage patterns from web data," *SIGKDD Explor. Newsletter*, pp. 12–23, 2000.

[18] D. B. Judd and G. Wyszecki, *Color in business, science, and industry*. Wiley, 1975.

[19] C. Zimmer, "The brain: Our strange, important, subconscious light detectors," *Discover Magazine*, 2012.

[20] G. J. Tortora and B. Derrickson, *Principles of Anatomy and Physiology*. Editorial Medica Panamericana, 2006.

[21] B. Cassin and S. Solomon, *Dictionary of Eye Terminology*. Triad Publishing Company, 1990.

[22] A. T. Duchowski, *Eye tracking methodology: Theroy and practice*. Springer Verlag, 2003.

[23] Harvey Richard Schiffman, *Sensation and Perception: An Integrated Approach*. Wiley, 2000.

[24] J. Nielsen and K. Pernica, *Eyetracking web usability*. New Riders Pub, 2009.

[25] Neil R. Carlson and Donald Heth C, *Psychology: the science of behaviour*. Toronto: Pearson, 2010.

[26] W. James, F. Burkhardt, F. Bowers, and I. Skrupskelis, *The Principles of Psychology*. No. v. 1 in American science series–advanced course, Harvard University Press, 1981.

[27] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Percept.*, vol. 7, pp. 6:1–6:39, Jan. 2010.

[28] L. R. Young and D. Sheena, "Survey of eye movement recording methods," *Behavior Research Methods & Instrumentation*, vol. 7, no. 5, pp. 397–429, 1975.

[29] A. Sprenger, B. Neppert, S. Köster, S. Gais, D. Kömpf, C. Helmchen, and H. Kimmig, "Long-term eye movement recordings with a scleral search coil-eyelid protection device allows new applications," *Journal of Neuroscience Methods*, vol. 170, no. 2, pp. 305 – 309, 2008.

[30] E. H. Hess and J. M. Polt, "Pupil size as related to interest value of visual stimuli," *Science*, vol. 132, no. 3423, pp. 349–350, 1960.

[31] A. C. Guyton, *Basic human physiology: normal function and mechanisms of disease.* Saunders Philadelphia London Toronto, 1977.

[32] H. Davson, *The Eye. Vol. 3. Muscular Mechanisms.* Academic, 1970.

[33] E. H. Hess, "Pupillometrics: A method of studying mental, emotional and sensory processes," *Handbook of psychophysiology*, pp. 491–531, 1972.

[34] L. Stark and A. Troelstra, "Dynamic pupillometers using television camera system," October 1970.

[35] M. G. H. Coles, E. Donchin, and S. W. Porges, "The pupillary system," *Psychophysiology: Systems, Processes, and Applications*, pp. 43–50, 1986.

[36] J. T. Cacioppo and L. G. Tassinary, "The ocular system," *Principles of psychophysiology: Physical, social, and inferential elements.*, pp. 193–215, 1990.

[37] E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during simple problem-solving," *Science*, vol. 143, no. 3611, pp. 1190–1192, 1964.

[38] J. M. Polt, "Effect of threat of shock on pupillary response in a problem-solving situation," *Perceptual and Motor Skills*, vol. 31, no. 2, pp. 587–593, 1970.

[39] J. Beatty, "Task-evoked pupillary responses, processing load, and the structure of processing resources.," *Psychological bulletin*, vol. 91, no. 2, p. 276, 1982.

[40] J. D. Barlow, "Pupillary size as an index of preference in political candidates," *Perceptual and Motor Skills*, vol. 28, no. 2, pp. 587–590, 1969.

[41] W. Clark and M. Ertas, "A comparison of pupillary reactions to visual and auditory stimuli in a test of preferences for presidential candidates.," *JSAS Catalog of Selected Documents in Psychology*, vol. 1, pp. 20–25, 1975.

[42] D. Hahnemann and J. Beatty, "Pupillary responses in a pitch-discrimination task," *Perception & Psychophysics*, vol. 2, no. 3, pp. 101–105, 1967.

[43] G. Hakerem and S. Sutton, "Pupillary response at visual threshold.," *Nature*, 1966.

[44] J. Beatty, "Prediction of detection of weak acoustic signals from patterns of pupillary activity preceding behavioral response," Tech. Rep. 140, Los Angeles: University of California, Dept. of Psychology, 1975.

[45] E. H. Hess, A. L. Seltzer, and J. M. Shlien, "Pupil response of hetero-and homosexual males to pictures of men and women: A pilot study.," *Journal of Abnormal Psychology*, vol. 70, no. 3, p. 165, 1965.

[46] E. Hess, "The tell-tale eye," 1975.

[47] M. P. Janisse, *Pupillometry: The psychology of the pupillary response*. Hemisphere Publishing Corporation Washington, 1977.

[48] R. A. Hicks, S. L. Williams, and F. Ferrante, "Pupillary attributions of college students to happy and angry faces," *Perceptual and Motor Skills*, vol. 48, no. 2, pp. 401–402, 1979.

[49] G. K. Poock, "Information processing vs pupil diameter," *Perceptual and Motor Skills*, vol. 37, no. 3, pp. 1000–1002, 1973.

[50] W. S. Peavler, "Pupil size, information overload, and performance differences," *Psychophysiology*, vol. 11, no. 5, pp. 559–566, 1974.

[51] M. SECo, "Problemas formales de la definición lexicográfica," in *Estudios ofrecidos a Emilio Alarcos Llorach*, pp. 217–240, Universidad de Oviedo, 1978.

[52] G. Myatt, *Making Sense of Data: A Practical Guide to Exploratory Data Analysis and Data Mining*. Wiley, 2007.

[53] R. Agrawal, R. Srikant, *et al.*, "Fast algorithms for mining association rules," in *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, vol. 1215, pp. 487–499, 1994.

[54] A. L. Yarbus, B. Haigh, and L. A. Rigss, *Eye movements and vision*, vol. 2. Plenum press New York, 1967.

[55] J. I. Fernańdez, *Mejorando el Contenido Textual de un Sitio Web a Traveś de la Identificacioń de sus Web Site Keywords*. Departamento de Ingenieriá Industrial, Universidad de Chile, 2010.

[56] A. Rösler, C. Ulrich, J. Billino, P. Sterzer, S. Weidauer, T. Bernhardt, H. Steinmetz, L. Frölich, and A. Kleinschmidt, "Effects of arousing emotional scenes on the distribution of visuospatial attention: Changes with aging and early subcortical vascular dementia," *Journal of the Neurological Sciences*, vol. 229, pp. 109–116, 2005.

[57] C. M. Privitera, L. W. Renninger, T. Carney, S. Klein, and M. Aguilar, "Pupil dilation during visual target detection," *Journal of Vision*, vol. 10, no. 10, 2010.