

UNIVERSIDAD DE SANTIAGO DE CHILE
FACULTAD DE INGENIERÍA
Departamento de Ingeniería Informática



**MODELO DE PREDICCIÓN DEL COMPORTAMIENTO DE
BÚSQUEDA DE INFORMACIÓN EN LÍNEA EN ESTUDIANTES
DE EDUCACIÓN PRIMARIA**

Gonzalo Javier Martinez Ramirez

Profesor guía: Roberto Ignacio González Ibañez

Tesis de grado presentado en conformidad
a los requisitos para obtener el grado de
Magíster en Ingeniería Informática.

Santiago – Chile

2017

RESUMEN

Durante la última década, debido a los rápidos avances de las tecnologías de la información y comunicación ha aumentado la cantidad de recursos digitales en Internet, la diversidad de fuentes de información, y, además, se ha facilitado el acceso a estos. Asimismo, las búsquedas *web* han pasado a ser parte de las tareas comunes que realizan los estudiantes de los planteles educativos. Considerando la diversidad de fuentes de información y tipos de recursos en línea, resulta necesario desarrollar competencias informacionales durante el proceso de formación en los distintos niveles educativos (primaria, secundaria y universitaria).

En el marco del proyecto iFuCo (*Enhancing learning and teaching future competences of online inquiry in multiple domains*), formado por investigadores de Chile y Finlandia, el cual desea investigar y modelar los comportamientos y competencias de investigación en línea de estudiantes de enseñanza básica, se propone la construcción de un modelo de predicción del comportamiento de búsqueda de información en línea en estudiantes de educación básica el cual se vaya perfeccionando a través del registro de datos históricos y que de un feedback en tiempo real.

La investigación será guiada por la metodología KDD con el fin de descubrir patrones en los datos que permitan la creación de un modelo de predicción del comportamiento de búsqueda. Además, para apoyar el proceso de investigación, se desarrollará una plataforma que funcione como extensión de la plataforma NEURONE (*oNlinE inqUiry expeRimentatiON systEm*). La plataforma propuesta alimentará y perfeccionará el modelo de predicción y entregará predicciones en tiempo real. Esta plataforma se guiará bajo la metodología RAD (*Rapid Application Development*) la cual se orienta a un desarrollo iterativo e incremental para la rápida construcción de prototipos de *software*.

Palabras Claves:

ABSTRACT

Today

Keywords:

TABLA DE CONTENIDOS

Capítulo 1. Introducción	1
1.1 Antecedentes y motivación	1
1.2 Descripción del problema	2
1.3 Solución propuesta	3
1.3.1 Características de la solución	3
1.3.2 Propósito de la solución	4
1.4 Objetivos y alcances de la solución	5
1.4.1 Objetivo general	5
1.4.2 Objetivos específicos	5
1.4.3 Alcances	6
1.5 Metodología y herramientas utilizadas	7
1.5.1 Metodología a usar	7
1.5.2 Herramientas de desarrollo	8
1.6 Organización del documento	9
 Capítulo 2. Marco teórico	 10
2.1 Marco conceptual	10
2.1.1 Búsqueda de información	10
2.1.2 Alfabetización informacional	10
2.1.3 Competencias de investigación en línea	10
2.1.4 Técnicas de minería de datos	10
2.2 Estado del arte	10
2.3 Marco de investigación	11
2.4 Resumen	11
 Apéndice A. Capítulo Apéndice	 13
A.1 Sección del apéndice	13
A.1.1 Subsección del apéndice	13
 Apéndice B. Another Appendix Chapter	 14

ÍNDICE DE TABLAS

B.1. Ejemplo de una tabla.	14
------------------------------------	----

ÍNDICE DE FIGURAS

1.1. Ciclo de construcción y perfeccionamiento del modelo	4
1.2. Proceso de búsqueda de información de un estudiante	5
A.1. A scientific diagram using the <code>pgfplots</code> package by Christian Feuersaenger using the same colors which are also used for the layout	13

CAPÍTULO 1. INTRODUCCIÓN

1.1. ANTECEDENTES Y MOTIVACIÓN

La alfabetización informacional (conocida en inglés como *information literacy*) es definida como “el grupo de habilidades en las que se requiere reconocer cuándo la información es necesaria y tener la habilidad de encontrar, evaluar y usar efectivamente dicha información necesaria”¹ (Association *et al.*, 2000, p. 2). Durante la última década, debido a los rápidos avances de las tecnologías de la información y comunicación (TICs) ha aumentado la cantidad de recursos digitales en Internet y además se ha facilitado el acceso a ellos. Estos avances han provocado una brecha entre el ser humano y la habilidad de reconocer cuando la información es necesaria para satisfacer su necesidad de búsqueda, la cual se puede asociar principalmente a dos razones: En primer lugar, las competencias de alfabetización informacional no son enseñadas ni reforzadas a temprana edad. Segundo, las búsquedas *web* han pasado a ser parte de las tareas comunes que realizan los estudiantes, disminuyendo las visitas a bibliotecas y el uso de fuentes revisadas.

Considerando la diversidad de fuentes de información y tipos de recursos en línea, resulta necesario desarrollar competencias informacionales durante el proceso de formación en los distintos niveles educativos (básica, media y universitaria). La enseñanza de la alfabetización informacional se imparte principalmente por bibliotecas universitarias, y en menor medida en la etapa escolar obligatoria (Weiner, 2014). En Chile, la enseñanza de competencias informacionales es cubierta en bibliotecas universitarias y cursos introductorios de mallas universitarias (Marzal & Saurina, 2015). De acuerdo con Urrea y Castro (2016), los estudiantes universitarios de Chile presentan problemas con las competencias informacionales, ya que no aplican la búsqueda de información de forma crítica. Una de las posibles causas de por qué los estudiantes tienen dificultades con estas competencias es el hecho de que en los colegios y en el inicio de su educación se prioriza la reiteración de la información. Las consecuencias de no considerar cuándo y por qué se necesita la información, dónde encontrarla y cómo evaluarla, se ven reflejadas en la evaluación crítica de la información, y en el desempeño de los estudiantes (Urrea & Castro, 2016).

A través de encuestas a estudiantes universitarios, Head (2013, p. 475) establece que al momento de realizar investigaciones el 84 % de los estudiantes universitarios utiliza como fuente primaria de búsqueda Wikipedia² y un 87 % consulta a sus amigos, sin verificar la veracidad de la información que obtienen. Como consecuencia, los estudiantes al no ser instruidos en parafrasear, resumir o citar fuentes revisadas, caen al plagio de forma premeditada o no intencionada.

¹

²<https://es.wikipedia.org/>

A partir de los argumentos anteriormente expuestos, respecto a la enseñanza de competencias de alfabetización informacional, se puede ver que no ha sido completamente satisfecha y la brecha entre los usuarios e alfabetización informacional permanece abierta.

Esta propuesta de tesis se enmarca en el contexto del proyecto de investigación “*Enhancing Learning and Teaching Future Competences of Online Inquiry in Multiple Domains*”³ (iFuCo, desde ahora en adelante), el cual pretende abordar la temática de la alfabetización informacional en estudiantes de enseñanza básica con el objetivo de estudiar sus patrones de comportamiento y ofrecer modelos curriculares adecuados respecto al tema (Sormunen *et al.*, 2017).

1.2. DESCRIPCIÓN DEL PROBLEMA

En el contexto de la enseñanza de la alfabetización informacional, las evaluaciones de los cursos se centran principalmente en los resultados de los estudiantes sin tomar en cuenta el proceso formativo y factores asociados que podrían influir directa o indirectamente sobre los resultados finales y el desempeño de búsqueda de los alumnos.

En el contexto del proyecto de investigación iFuCo, el cual pretende realizar un análisis cuantitativo y cualitativo de la alfabetización informacional y las competencias de búsqueda en línea en estudiantes de enseñanza básica⁴ en los países de Chile y Finlandia, surgen las siguientes interrogantes (*research questions*, RQ desde ahora en adelante):

RQ 1 ¿De qué manera se puede estimar durante el proceso de aprendizaje de competencias informacionales la influencia de diversos factores en el desempeño de búsqueda de la información de los estudiantes?

RQ 2 ¿En qué medida es posible detectar situaciones anormales de conducta, y determinar las causas que llevan a un estudiante a fallar durante el proceso de búsqueda de información?

RQ 3 ¿De qué manera se puede implementar un módulo de clasificación y predicción del desempeño de los estudiantes en la búsqueda de información en herramientas de apoyo de la alfabetización informacional para proporcionar una retro evaluación oportuna a estudiantes y docentes?

³<https://www.researchgate.net/project/Enhancing-learning-and-teaching-for-future-competences-of-online-inquiry-in-multiple-domains-iFuCo>

⁴En otros países es conocido como enseñanza primaria

1.3. SOLUCIÓN PROPUESTA

1.3.1. Características de la solución

La solución consiste en incorporar un módulo en NEURONE (González-Ibáñez, Gacitua, Sormunen & Kiili, 2017) que clasifique y prediga de forma continua el desempeño de búsqueda de los estudiantes de enseñanza básica en un curso de alfabetización informacional, específicamente en el tema de investigaciones en línea⁵ (*online inquiry*).

Los datos son recopilados y almacenados por NEURONE, estos datos provienen de registros del proceso de búsqueda de información en línea en un sistema cerrado, los cuales son: historial de navegación, consultas realizadas, movimientos del *mouse*, escritura por teclado, número de *clicks* y tiempos de permanencia en páginas *web*. Además, se conoce con anticipación los documentos y párrafos ideales a seleccionar por parte de los estudiantes.

El módulo propuesto hará uso de Apache Spark⁶, el cual es un *framework* de código abierto para el procesamiento de datos masivos, el cual incluye librerías de minería de datos y aprendizaje de máquina. Este módulo se conectará con el sistema NEURONE, funcionando como una extensión del mismo, consultando su base de datos, alimentando y perfeccionando el modelo.

El ciclo de construcción, evaluación y optimización del modelo se ilustra en la Figura 1.1, donde a través de los datos históricos obtenidos de NEURONE construye el modelo, lo evalúa y lo optimiza en un proceso continuo, entregando como resultado la clasificación del desempeño de búsqueda y prediciendo de forma continua a partir del comportamiento actual de búsqueda de información del estudiante.

⁵

⁶<https://spark.apache.org/>



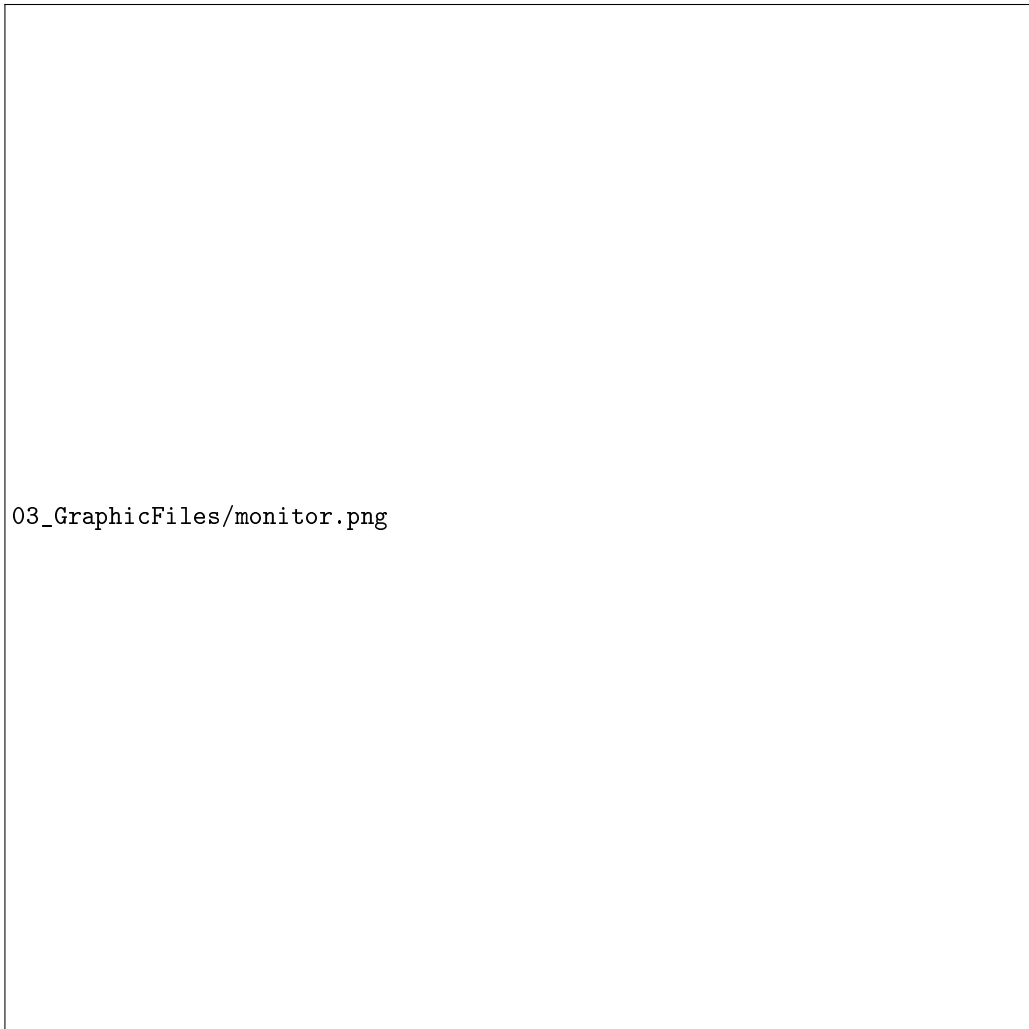
Figura 1.1: Ciclo de construcción y perfeccionamiento del modelo

Fuente: Elaboración propia, (2017).

1.3.2. Propósito de la solución

El propósito de la solución consiste en proveer evaluaciones de desempeño de búsqueda oportunas que permitan a los docentes aplicar acciones correctivas durante el proceso de formación y desarrollo de competencias informacionales en cursos de alfabetización informacional.

Con el módulo propuesto en este trabajo, el docente obtiene una estimación temprana del desempeño del estudiante en el proceso de búsqueda de información, de tal forma que él pueda guiar al estudiante en el proceso. Tal como ilustra la Figura 1.2, el estudiante interactúa con el sistema educacional, en este caso NEURONE y la plataforma propuesta a través de técnicas de minería de datos informa al docente de los patrones y predicciones del desempeño de búsqueda del estudiante con el objetivo de ayudar en la toma de decisiones al docente correspondiente para diseñar y planificar de mejor forma la entrega de contenidos hacia el estudiante.



03_GraphicFiles/monitor.png

Figura 1.2: Proceso de búsqueda de información de un estudiante

Fuente: Elaboración propia, (2017).

1.4. OBJETIVOS Y ALCANCES DE LA SOLUCIÓN

1.4.1. Objetivo general

Diseñar y evaluar un modelo predictivo del desempeño de búsqueda de información en línea de estudiantes de enseñanza básica.

1.4.2. Objetivos específicos

1. Realizar una revisión bibliográfica sobre trabajos recientes relacionados con minería de datos en el contexto educacional.

2. Realizar una exploración, limpieza, pre-procesamiento y transformación de los datos recopilados por la plataforma NEURONE (acrónimo de *oNlinE inqUiry expeRimentatiON systEm*).
3. Definir las características de comportamiento de búsqueda de los estudiantes para la construcción de modelos predicción.
4. Comparar, seleccionar e implementar algoritmos de minería de datos, para la construcción de modelos de predicción.
5. Implementar los modelos de predicción del comportamiento de búsqueda en línea de estudiantes de básica.
6. Implementar y evaluar una plataforma de enseñanza de competencias informacionales, la cual en base a los datos provistos por NEURONE prediga el desempeño de búsqueda de un estudiante.

1.4.3. Alcances

Los modelos se construyen a partir de un conjunto de datos específicos, estos datos tienen su propio contexto y origen que limitan la generalización de los modelos a construir. A continuación, se describen las principales limitaciones y alcances de la solución.

1. El curso de alfabetización informacional y sus respectivos registros de datos, pertenecen al proyecto iFuCo, el cual es un trabajo colaborativo entre universidades de Finlandia (University of Tampere, University of Jyväskylä y University of Turku) y de Chile (Universidad de Santiago de Chile y Pontificia Universidad Católica de Chile).
2. Los registros de datos provienen de un estudio enmarcado en un curso de alfabetización en información, aplicado al área de Ciencia y Ciencias Sociales, en ambos países.
3. Los datos son recolectados y almacenados por un sistema externo llamado NEURONE (*oNlinE inqUiry expeRimentatiON systEm*), trabajo de memoria de un estudiante de la carrera de Ingeniería de Ejecución en Computación e Informática de la Universidad de Santiago de Chile.
4. La solución funciona como un sistema predictor del desempeño del estudiante en la búsqueda de información, sin ofrecer acciones correctivas en caso de bajo desempeño.

1.5. METODOLOGÍA Y HERRAMIENTAS UTILIZADAS

1.5.1. Metodología a usar

El presente proyecto presenta una componente de investigación y desarrollo de *software* (I+D), esto debido a la relación que existe entre ambas componentes, la investigación necesita una herramienta de *software* de apoyo que permita recibir los datos de NEURONE, alimentar el modelo de predicción y que permita al usuario interactuar con resultados de la predicción realizada.

La componente de investigación del proyecto será guiada por la metodología Descubrimiento de Conocimiento en Base de Datos (conocido como KDD, las iniciales de *Knowledge Discovery in Databases*) (Fayyad, Piatetsky-Shapiro & Smyth, 1996), mientras que la componente de desarrollo será guiada por la metodología de desarrollo de *software* Desarrollo de Rápido de Aplicaciones (conocido como RAD, las iniciales de *Rapid Application Development*) (Martin, 1991). A continuación, se explica el uso de ambas metodologías en el trabajo propuesto.

Metodología usada en la investigación

Respecto a la componente de investigación, esta será guiada bajo la metodología KDD, la cual se define como “un proceso no trivial de identificar patrones en los datos que sean válidos, novedosos, potencialmente útiles y finalmente comprensibles” (Fayyad *et al.*, 1996, p. 5). En primer lugar, se seleccionan y limpian los datos que se deben extraer para poder realizar el modelado del comportamiento de búsqueda. Luego, se transforman los datos y se realiza minería de datos sobre ellos para buscar los patrones de interés que pueden expresarse como un modelo o que expresen dependencia de los datos. Finalmente, se identifican los patrones realmente interesantes que representan el conocimiento, usando diferentes técnicas, incluyendo análisis estadísticos para posteriormente interpretar los datos obtenidos.

Metodología usada para el desarrollo

Respecto a la componente de desarrollo de *software*, se toma en cuenta las condiciones bajo las cuales se desarrolla el proyecto, las cuales se expresan a continuación:

- El sistema es de rápido desarrollo.
- El sistema es de tamaño pequeño.
- Es un proyecto cuyos requerimientos están sujetos a cambios.
- Inicialmente no existe un número total de requerimientos especificado. Estos se irán desarrollando de forma creciente durante el avance del proyecto.
- El desarrollador no cuenta con un conocimiento profundo de la arquitectura y todas las herramientas de desarrollo, por lo tanto, se requiere un tiempo de investigación y aprendizaje.
- Se requiere documentar los aspectos fundamentales de la arquitectura, una vez que se tenga un producto estable. Esta documentación permitirá la continuidad del proyecto.
- Se requiere de varias entregas funcionales, para medir el progreso del proyecto y verificar que se cumplan los objetivos propuestos.

Dado los antecedentes mencionados anteriormente, se determina que el proyecto presenta características que se ajustan bien a un modelo de desarrollo evolutivo enfocado a la generación de prototipos. A partir de esto, se recurre a un enfoque de desarrollo inspirado en la metodología RAD, metodología de desarrollo rápido que minimiza la planificación en favor de la creación rápida de prototipos. La planificación se realiza en cada iteración, permitiendo que el *software* se desarrolle más rápido y se tenga una mayor flexibilidad con los requisitos (McConnell, 1996).

1.5.2. Herramientas de desarrollo

Las herramientas a utilizar en el trabajo de tesis, se dividen tanto en *hardware* como en *software*.

Software

El desarrollo se llevará a cabo con procesador Intel Core i7 7ma Generación *KabyLake* de 3.6 Ghz, con memoria Ram de 16 GB y 2 TB de disco duro. Además, los despliegues de prueba se realizan sobre un servidor privado virtual (VPS, por sus siglas en inglés) con el sistema operativo GNU/Linux Ubuntu Server alojado en el proveedor DigitalOcean⁷.

⁷<https://www.digitalocean.com/>

Hardware

En cuanto herramientas *software*, el desarrollo se llevará a cabo en la distribución GNU/Linux Debian en su versión 9.0. El modelo se llevará a cabo en Spark ML⁸. Para el análisis estadístico se hará uso de R. Además, cada módulo desarrollado estará contenido en contenedores de Docker para facilitar el despliegue en producción del modelo desarrollado. Todo el trabajo realizado, tanto código como documento escrito estará bajo el sistema de control de versiones Git. Finalmente, se hará uso de \LaTeX para el documento escrito.

1.6. ORGANIZACIÓN DEL DOCUMENTO

El presente documento se estructura de la siguiente forma.

Capítulo 2 Se estipulan los conceptos teóricos que se deben definir para tener una base consensuada respecto de los distintos conceptos que se tratan en este documento. En el mismo capítulo se aborda el estado del arte donde se hace una revisión bibliográfica de los últimos avances en el área.

⁸<https://spark.apache.org/>

CAPÍTULO 2. MARCO TEÓRICO

Este capítulo tiene como objetivo entregar las bases teóricas, conceptuales y empíricas que soportan cada desarrollo de esta investigación. En primer lugar, se presenta el marco conceptual donde se entregan las definiciones y conceptos necesarios para abordar esta investigación. En segundo lugar, se presenta el estado del arte relacionado con el tema.

2.1. MARCO CONCEPTUAL

En esta sección se presentan conceptos y bases teóricas respecto a la temática que conduce el desarrollo de este trabajo, el cual tiene relación con el uso de interfaces no tradicionales, específicamente con una interfaz operada con el cuerpo. Además, se indaga sobre ciertas definiciones para establecer lo que se pretende medir en este estudio, lo que involucra la experiencia de usuario y métricas de rendimiento en la realización de tareas. Finalmente, se proponen ciertas características fundamentales respecto de este tipo de proyectos relacionados con diseños experimentales con usuarios.

2.1.1. Búsqueda de información

2.1.2. Alfabetización informacional

2.1.3. Competencias de investigación en línea

2.1.4. Técnicas de minería de datos

2.2. ESTADO DEL ARTE

El propósito de esta sección es presentar los últimos trabajos realizados en la línea de investigación que se ha planteado en la sección y capítulo anterior. La búsqueda de estos trabajos considera: la exploración de trabajos con estudios que midan experiencia de usuario y/o medidas de rendimiento en la utilización de interfaces no-tradicionales operadas con el cuerpo para la realización de actividades en distintos contextos.

2.3. MARCO DE INVESTIGACIÓN

En esta sección se propone y formula el marco de investigación que guía este estudio. Para esto se desarrollan las preguntas de investigación que motivan la realización del estudio. Además se presentan las hipótesis que se desean someter a prueba en base a los resultados producto de la realización del estudio que en este documento se propone.

RQ 1 ¿De qué manera se puede estimar durante el proceso de aprendizaje de competencias informacionales la influencia de diversos factores en el desempeño de búsqueda de la información de los estudiantes?

RQ 2 ¿En qué medida es posible detectar situaciones anormales de conducta, y determinar las causas que llevan a un estudiante a fallar durante el proceso de búsqueda de información?

RQ 3 ¿De qué manera se puede implementar un módulo de clasificación y predicción del desempeño de los estudiantes en la búsqueda de información en herramientas de apoyo de la alfabetización informacional para proporcionar una retro evaluación oportuna a estudiantes y docentes?

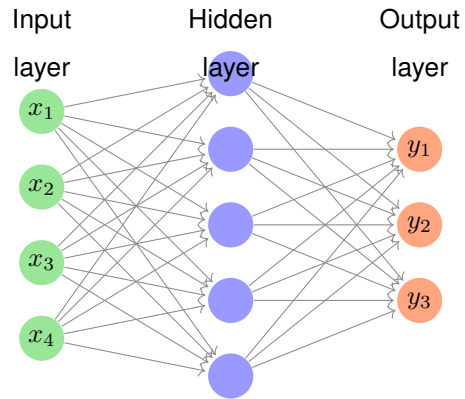
2.4. RESUMEN

BIBLIOGRAFÍA

- Association, A. L. *et al.* (2000). Information literacy competency standards for higher education. (ver pág. 1).
- Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3), 37. (ver pág. 7).
- González-Ibáñez, R., Gacitua, D., Sormunen, E. & Kiili, C. (2017). NEURONE: oNlinE inqUiRy experimentatiON systEm. En T. be included in Proceedings of the 80th Annual Meeting of the Association for Information Science & T. (2017) (Eds.). (ver pág. 3).
- Head, A. J. (2013). Project Information Literacy: What can be learned about the information-seeking behavior of today's college students? (ver pág. 1).
- Martin, J. (1991). *Rapid application development*. Macmillan Publishing Co., Inc. (ver pág. 7).
- Marzal, M. Á. & Saurina, E. (2015). Diagnóstico del estado de la alfabetización en información (ALFIN) en las universidades chilenas. *Perspectivas em Ciência da Informação*, 20(2), 58-78. (ver pág. 1).
- McConnell, S. (1996). *Rapid development: Taming wild software schedules*. Pearson Education. (ver pág. 8).
- Sormunen, E., González-Ibáñez, R., Kiili, C., Leppänen, P., Mikkilä-Erdmann, M., Erdmann, N. & Escobar-Macaya, M. (2017). A Performance-based Test for Assessing Students' Online Inquiry Competences in Schools. En E. C. in Information Literacy (ECIL) (Ed.). (ver pág. 2).
- Urrea, M. C. V. & Castro, S. O. (2016). Alfabetización en información: Estudio de su impacto en estudiantes de último año del pregrado de las facultades de educación y ciencias naturales y exactas en la Universidad de Playa Ancha de Ciencias de la Educación, 20-40. (ver pág. 1).
- Weiner, S. A. (2014). Who teaches information literacy competencies? Report of a study of faculty. *College Teaching*, 62(1), 5-12. (ver pág. 1).

APÉNDICE A. CAPÍTULO APÉNDICE

A.1. SECCIÓN DEL APÉNDICE



**Example Diagram with a Line Break in the Title
(using the text width option in the title style)**

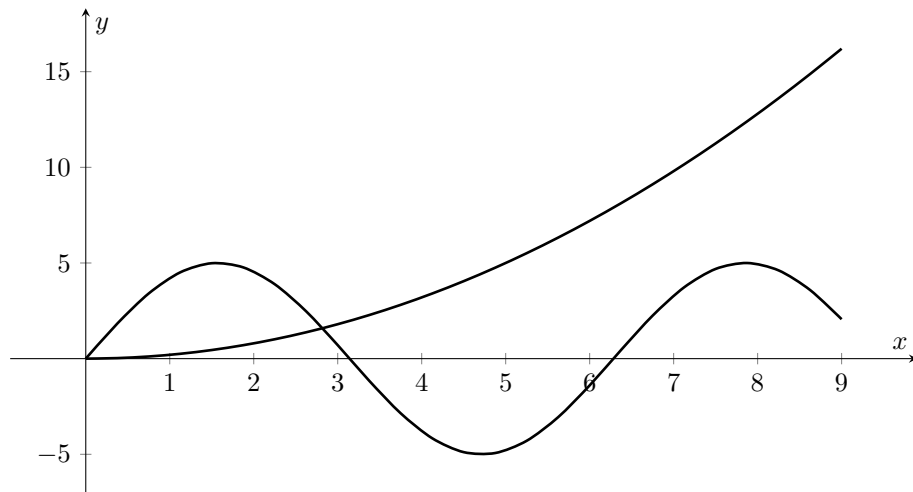


Figura A.1: A scientific diagram using the `pgfplots` package by Christian Feuersaenger using the same colors which are also used for the layout

Fuente: Elaboración propia, (2017).

A.1.1. Subseccion del apéndice

APÉNDICE B. ANOTHER APPENDIX CHAPTER

Como se puede apreciar en la Tabla B.1.

Tabla B.1: Ejemplo de una tabla.
Fuente: Elaboración propia, (2017).

header1	header2	header3
1	2	3
4	5	6
7	8	9