PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA– FACULTAD DE LETRAS – BIBLIOTECAS UC

PROPUESTA METODOLÓGICA PARA LA RECUPERACIÓN DE INFORMACIÓN, USANDO TABLAS DE CONTENIDO E ÍNDICES

RODRIGO WLADIMIR SALINAS MADARIAGA

Proyecto de Tesis para optar al Grado de
Magíster en Procesamiento y Gestión de la Información

Profesor Supervisor:
OLGA ACOSTA LÓPEZ

Santiago de Chile, Septiembre 2017.

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA– FACULTAD DE LETRAS – BIBLIOTECAS UC

PROPUESTA METODOLÓGICA PARA LA RECUPERACIÓN DE INFORMACIÓN, USANDO TABLAS DE CONTENIDO E ÍNDICES

RODRIGO WLADIMIR SALINAS MADARIAGA

Proyecto de Tesis presentado a la Comisión integrada por:

OLGA ACOSTA LÓPEZ (PROFESOR SUPERVISOR)

CESAR ANTONIO AGUILAR (PROFESOR INTEGRANTE)

MAURICIO ARRIAGADA BENÍTEZ (PROFESOR INTEGRANTE)

IRENE HERNANDEZ MORALES (PROFESOR INTEGRANTE)

Para completar las exigencias del Grado de
Magíster en Procesamiento y Gestión de la Información

Santiago de Chile, Septiembre 2017.

DEDICATORIA

A mi madre, por los años de formación y esfuerzo. A Edilberto, mi maestro, y a mi hermana Tania, que en paz y luz descansen.

AGRADECIMIENTOS

Agradezco principalmente a mi profesora Olga por la inmensa ayuda otorgada para sacar adelante todo esto, por su paciencia y voluntad. A Claudia Gilardoni, ex jefa, que fue la que me motivó a involucrarme en este desafío. También a Irene Hernández quien fue mi primera profesora guía, ayudándome mucho a estructurar buena parte de la tesis. A Leopoldo Bustos, por la ayuda en mi trabajo. Y a Liz por la paciencia, el amor y el ánimo…

TABLA DE CONTENIDO

INDICE DE TABLAS

Pág.

INDICE DE FIGURAS

Pág.

RESUMEN

En el ámbito de la búsqueda y recuperación de información, son frecuentes las dificultades que surgen en la efectividad debido a que no siempre las herramientas disponibles en bibliotecas (catálogos, bases de datos) cuentan con la factibilidad de poder dar cuenta del contenido detallado de las colecciones existentes. En términos generales los registros bibliográficos tienen entre dos y cuatro materias asignadas que reflejan los temas comprendidos en cada uno de los materiales bibliográficos incorporados. Aparte de eso, opcionalmente pueden incluir algunas palabras claves o notas de contenido y/o resumen que contribuyen a la descripción. Sin embargo, esto resulta insuficiente muchas veces para realizar búsquedas de términos, puesto que el enfoque se hace en un sentido general y no incluyen siempre términos relevantes que pueden estar contenidas en las obras.

El presente trabajo intenta presentar una solución a esto, haciendo una propuesta metodológica de un prototipo de una herramienta automatizada que permita ir más allá, a través de la búsqueda de términos en las propias tablas de contenido e índices temáticos o analíticos de los libros.

En la primera parte se aborda el tema en contexto histórico y teórico, para comprender cómo se ha intentado encontrar opciones para mejorar la precisión en los resultados de búsquedas bibliográficas. Se hace mención a los modelos de recuperación de información y a algunas tentativas de soluciones encontradas.

En la segunda parte se da cuenta de la metodología empleada para dar con una solución a la problemática expuesta, con muestra de resultados y evaluando el desempeño mostrado de los modelos utilizados. Quedan en evidencia las ventajas con respecto al actual sistema usado en la biblioteca objeto de estudio, en pos de una respuesta de mejor cantidad y calidad ante las consultas expuestas, ya que considera además un orden de relevancia de mayor a menor en la entrega de los resultados.

ABSTRACT

In the area of information research and retrieval, there are frequent difficulties in effectiveness because the tools available in libraries (catalogs, databases) which are not always feasible to account for the detailed content of existing collections. In general terms, the bibliographic records have between two and four assigned subjects that reflect the subjects. These are included in each of the incorporated bibliographic materials. Apart from that, they may optionally include some keywords, content notes and / or summary that contribute to the description. However, this is often insufficient for term searching, since the approach is taken in a general sense and does not always include relevant terms that may be contained in the items.

The present work tries to present a solution to this, making a methodological proposal of a prototype of an automated tool that allows going further, through the search of terms in the tables of content and thematic or analytical indexes in books.

In the first part, the subject is approached in a historical and theoretical context, in order to understand how there have been attempts to find options for improving the accuracy of bibliographic search results. Information fetching models and some attempts to find solutions has been mentioned.

The second part explains the methodology used to find a solution to the problem, with a sample of results and evaluating the performance shown of the models used. There is evidence of the advantages over the current system used in the library under study, for a response of better quantity and quality to the queries exposed, since it also considers an order of relevance from highest to lowest in the delivery of the results.

I. INTRODUCCIÓN

Son variadas las problemáticas que surgen en el trabajo bibliotecario en sus distintas funciones. Dentro de lo que corresponde a servicios a los usuarios, la búsqueda de información suele ser de las más complejas. Hay un volumen muy grande de datos registrados que se deben manejar y pese a que existen mecanismos y tecnologías avanzadas de recuperación de información, comúnmente hay un porcentaje de documentos que no logran rescatarse debido a que no están identificados con los términos que expresen sus contenidos y por tanto hay un grado de frustración en las búsquedas que se expresan como consulta a un sistema automatizado. Se produce una instancia compleja para el usuario, en donde existe una necesidad de información que debe formular con terminología apropiada para encontrar posibles resultados asociados. Tal disyuntiva se complejiza más dado que el o los término(s) que busca, a veces no aparece(n) como parte del título, sino que figura(n) como materia en la ficha de descripción bibliográfica del libro. Sin embargo puede que se encuentre bajo un sinónimo, lo que dificulta más la consulta pues el usuario no maneja términos de un vocabulario controlado y debe recurrir a un especialista o al bibliotecario para que lo oriente. Esa dependencia en estos casos resulta un factor limitante en la búsqueda de información.

El presente trabajo aborda esta situación y busca dar solución a un caso concreto como es el de la Biblioteca de la Universidad del Pacífico de Santiago de Chile, que permita mejorar la satisfacción en las búsquedas de información, por medio de la creación de un prototipo de buscador eficaz que permita recuperar información textual contenida en las tablas de contenido y los índices del material impreso, de modo de aumentar la relevancia en los resultados de las búsquedas y dar con documentos que sirvan para responder a los requerimientos expuestos y que den mejor cuenta del contenido real de las fuentes bibliográficas impresas existentes en la colección.

II. PLANTEAMIENTO DEL PROBLEMA

Las bibliotecas cuentan con sistemas automatizados para administrar sus colecciones. Generalmente son sistemas de gestión bibliotecaria adquiridos a proveedores externos, o bien, desarrollados dentro de las mismas instituciones. La herramienta tradicional para buscar y recuperar información es el catálogo en línea u OPAC (On Line Public Access Catalog), que da acceso a un conjunto de registros bibliográficos que dan cuenta del material existente en las bibliotecas, por medio de la descripción de sus características físicas y de su contenido.

Los catálogos permiten recuperar información por diversos puntos de acceso, siendo los más utilizados los campos de autor, título y materia. Algunos también incluyen la opción de realizar búsquedas a través de palabras claves tales como palabras en el título, en el resumen, en cualquier parte del registro o también en las materias añadidas localmente (que se diferencian de las materias normalizadas contenidas en listados de uso internacional) para aumentar el grado de precisión en la recuperación.

Sin embargo hay muchos términos, conceptos, nombres y datos que forman parte de los contenidos de los materiales bibliográficos impresos y que no son incluidos en los registros bibliográficos, por lo cual no logran recuperarse a través del catálogo. Algunas bibliotecas -de acuerdo a su política de catalogación- incorporan notas de contenido con otro tipo de información como los títulos principales de los capítulos o nombres de las partes que constituyen la obra.

Dicho esto, se puede puntualizar que el problema en particular a tratar en este trabajo se refiere a que:

Los usuarios del OPAC tienen problemas para encontrar el material que necesitan porque no está suficientemente detallada la descripción ni el contenido de los libros impresos y eso dificulta la búsqueda y recuperación de información, debido a que los registros bibliográficos del catálogo del sistema automatizado que tiene el Sistema de Bibliotecas de la Universidad del Pacífico en su mayoría sólo tienen un nivel de descripción básico de dos o tres materias, por tanto no siempre responde

óptimamente a las consultadas ejecutadas por parte de algunos usuarios que buscan información sobre un tema en particular. Esto repercute en la precisión de la recuperación de las búsquedas que se generan y en la subutilización de la colección como recurso de información.

Otro aspecto a tener en cuenta es la recuperación del contenido de monografías colectivas (actas, colecciones, antologías), donde cada capítulo o parte de la obra corresponde a autores diferentes con títulos diferentes, como trabajos intelectuales individuales que forman parte de una sola obra que los compendia. En estos casos las obras son generalmente ingresadas como un todo y el contenido de cada parte constituyente sólo es recuperable si se ingresa en una nota de contenido, de lo contrario sólo se puede encontrar si se revisa manualmente el impreso.

III. OBJETIVOS

III.1 Objetivo general

Optimizar el uso de la colección impresa del área de Humanidades y Ciencias Sociales existente en la Biblioteca Central de la Universidad del Pacífico, considerando como fuente de información las tablas de contenido e índices.

III.2 Objetivos específicos

- Identificar la colección de libros para realizar experimentos de recuperación de información.
- Digitalizar tablas de contenidos (o sumarios) y/o índices temáticos de libros seleccionados.
- Diseñar un conjunto de consultas de prueba para la realización de los experimentos.
- Evaluar el desempeño de los modelos de recuperación de información, probabilístico y espacio vectorial.

IV. JUSTIFICACIÓN DEL TEMA

Una buena opción para enriquecer los registros bibliográficos de los catálogos en línea es utilizar información contenida en las obras impresas. Fuentes ideales para ello resultan las tablas de contenido (o sumarios) y los índices que la mayoría de las monografías incorporan como herramientas de búsqueda. Esta información además tiene la ventaja de representar fielmente el contenido de las obras procesadas y de incluir la terminología que el o los autores utilizan. Cabe agregar que en el caso del índice temático (llamado también analítico) precisamente enlista y ordena aquellos términos significativos y relevantes contenidos en la obra y que no son visibles en una ficha bibliográfica para recuperarlos. Además remiten a la ubicación donde se encuentran en concreto (página).

En un estudio para ver temas de diseño de la tercera generación de OPAC, los autores Belkin y Saracevic dan cuenta de las mejoras experimentadas por los registros bibliográficos al incluir información aparecida en las tablas de contenido, ya que no sólo atendía a requerimientos de recuperación sino a aportar significativamente a la relevancia.[1]

En otro estudio sobre uso del catálogo realizado en el *Centre for Catalogue Research* de la Universidad de Bath, se recomienda incrementar las posibilidades de acceso a las colecciones más allá de la descripción bibliográfica tradicional:

> "Las oportunidades que ofrecen los nuevos medios tecnológicos implican que debe incrementarse la explotación de los materiales bibliotecarios mediante el uso de palabras claves y frases de las páginas del título, las páginas de contenido y los índices…".[2]

En ese sentido, se buscaba mejorar los registros del catálogo usando dos formas:

---

[1]BELKIN, N.J. y SARACEVIC, T. (1992) Design principles for third-generation Online Public Access Catalogs: taking account of users and library use. *Annual Review of OCLC Research, Vol. july 1991-june 1992,* 43-45.

[2]SEAL, A., BRYANT, P. y HALL, C. (1982) Full and short entry catalogues: library needs and uses. *BLRD report.* Bath University.

1. Usando las palabras de la tabla de contenidos indizadas como palabras claves y sumar así un nuevo punto de acceso.
2. Permitiendo que el buscador del catálogo recupere por las tablas de contenido en línea.

De acuerdo con la idea, según lo expuesto por Peis y Fernández-Molina en su artículo sobre aplicación de tecnologías en catálogos en línea[3] :

"la recuperación aumentó en un 300%"[4], "el flujo de materiales también aumentó"[5]*****, "y se ha observado un aumento considerable en el *recall*"[6]****.**

Para remediar la búsqueda infructuosa de temas en una consulta al sistema automatizado, una solución sería incluir una instancia nueva de búsqueda en texto completo de las tablas de contenido que permita hacer un rastreo exhaustivo de los registros ingresados y encontrar aquellos que contienen los términos buscados, permitiendo llegar a los documentos que satisfagan la consulta.

Como el sistema automatizado de la biblioteca de la Universidad del Pacífico no permite realizar búsquedas en texto completo, se hace necesario contar con una alternativa al catálogo tradicional, que permita recuperar términos no registrados en el registro bibliográfico, pero que sí se encuentran presentes en las tablas de contenidos o índices de los textos. Esta alternativa debe ser una herramienta que incorpore búsqueda en texto completo, fácil de usar y que referencie los resultados al registro bibliográfico correspondiente.

* El flujo de materiales bibliográficos también aumentó.

** Todos los documentos relevantes de la colección que fueron recuperados por el sistema posterior a una consulta.

---

[3]PEIS, E. y FERNANDEZ-MOLINA, J.C. (1998) Enrichment of Bibliographic Records of Online Catalogs through OCR and SGML Technology. *Information Technology & Libraries, Vol.17*, N°3, 167-172.

[4]BYRNE, A. and MICCO, M. (1988).Improving OPAC subject access. *College and Research Libraries*, N°49, 432-441.

[5]KNUTSON, G. (1991) Subject Enhancement: Report on an Experiment. *College and Research Librari, Vol. 52*, N°1, 65-79.

[6]DILLON, M. and WENZEL, P. (1990) Retrieval Effectiveness of Enhanced Bibliographic Records. *Library Hi Tech, Vol. 8*, N°3, 43-46.

De esta forma aumentarían los niveles de precisión en la recuperación de información dado que además de hacer un rastreo más exhaustivo de los términos buscados dentro de un conjunto mayor de términos recuperables producto de la digitalización de tablas de contenido e índices, tendría más opciones de identificar aquellos recursos más relevantes a la consulta realizada. Se enriquece además la información relacionada a los materiales impresos descritos en el catálogo en línea.

Se requiere una respuesta rápida a las consultas que se expresen, puesto que las dinámicas de hoy en día, en términos de uso y eficiencia de las tecnologías, así lo demandan (Google, Twitter, Instagram, Facebook, Youtube, etc.)

V. CONTEXTO

Las bibliotecas de la Universidad del Pacífico cuentan con colecciones multitemáticas compuestas principalmente por las bibliografías obligatorias de las carreras que imparten, además de obras generales, obras literarias y obras de referencia básicamente. Las principales demandas de información son por títulos específicos que solicitan los usuarios o bien por los temas que ellos requieren investigar, para lo que necesitan saber qué recursos bibliográficos existentes en la colección pueden ser útiles a sus necesidades de información Dado lo anterior, este trabajo se enfoca en cómo recuperar los recursos más pertinentes que den respuesta a consultas sobre temáticas específicas y que las herramientas de búsqueda tradicionales, como el catálogo, sólo responden genéricamente.

En base a la observación de demandas temáticas recibidas en el sector de circulación de la Biblioteca Central de la Universidad del Pacífico (área de préstamo de material bibliográfico), se constató que en las áreas de Humanidades y Ciencias Sociales es donde mayormente se presentan este tipo de consultas. Se realizan búsquedas sobre materias específicas (Ej.: peritaje social) o la cobertura sobre un tema amplio (Ej.: el miedo) temas que no aparecen en la descripción, pero sí pueden estar contenidos como un capítulo del libro.

Dado que el sistema automatizado de la biblioteca no permite esta búsqueda, ni en las materias de un libro en particular, se hace necesario contar con una solución que contribuya a la recuperación de registros que muestren mayor coincidencia luego de una exhaustiva revisión digital.

V.1 Sobre la institución

La Universidad del Pacífico es una institución de educación superior privada de pre y postgrado que cuenta con 41 años de existencia. Se emplaza en dos sedes: una en la comuna de Las Condes, en Santiago (Casa Central) y otra en la ciudad de Melipilla.

Posee dos facultades: Facultad de Comunicaciones y Diseño, y Facultad de Ciencias Sociales y Económicas. Además tiene tres escuelas: Escuela de Formación Técnica, Escuela de Ciencias Agropecuarias, Escuela de Ciencias de la Salud. Además de una serie de postgrados que se imparten, dentro de los cuales hay cuatro magíster y 3postítulos. Atiende a una comunidad de 3.730 alumnos y 395 profesores aproximadamente.

V.2 Sobre la biblioteca

El Sistema de Bibliotecas cuenta con dos bibliotecas, una en Las Condes y la otra en Melipilla.

El personal del Sistema de Bibliotecas de la Universidad del Pacífico (SIBUPA), está compuesto por: una directora de bibliotecas, tres bibliotecólogos (dos son jefes de cada una de las sedes), dos asistentes de bibliotecas y una auxiliar.

La Biblioteca cuenta con el programa “Horizonte” como software de almacenamiento y recuperación de información el que cuenta con un catálogo en línea (OPAC) que tiene la opción de buscar por temas. Permite en este sentido buscar ese tema dentro de : las materias ingresadas, palabras claves del título o bien como palabra clave general (donde se incorpora la búsqueda dentro de campos de notas de contenido o resúmenes en caso que se encuentren registrados en la descripción bibliográfica realizada para determinado material bibliográfico.

V.3 Sobre la colección

La colección está compuesta principalmente por material bibliográfico que atiende los requerimientos obligatorios y complementarios de las carreras de las áreas mencionadas, más la existencia de obras de referencia, publicaciones periódicas, tesis, materiales especiales (test de psicología, mapas, material didáctico, etc.), apuntes y obras de colección general. También existe una colección de material audiovisual (CD y DVD), y una colección de literatura. Aproximadamente la colección asciende a 48.370 volúmenes.

V.4 Sobre los usuarios

El perfil de los estudiantes de la universidad varía de acuerdo a las carreras en que se encuentren. La universidad cuenta con dos sedes de características distintas.

La Sede Las Condes (Casa Central) es la que acoge a las carreras de las áreas de Diseño, Comunicaciones, Negocios y Marketing, Educación y Ciencias Humanas. Esta última área anteriormente se encontraba en un campus que se cerró y que apuntaba a un segmento socioeconómico diferente. Atendía a alumnos de segmento clase media en su mayoría, a diferencia de los alumnos de Las Condes, que pertenecen a segmento de clase alta y media alta. Se encontraba en un sector más central de Santiago (Parque Bustamante) y los alumnos fueron derivados posteriormente a la sede de Las Condes.

Por otra parte, los alumnos de la Sede Melipilla son alumnos de la zona, incluyendo sectores rurales, con otras características socioeconómicas y que corresponden más a la clase media.

Cabe informar que de acuerdo a estudios institucionales de la propia universidad, se considera que el 50.91% de los alumnos de Las Condes sus familias cuentan con ingresos superiores a $300.000, mientras que en el caso

de Melipilla esa cifra sólo llega a 17.68%, concentrándose en su mayoría en un rango de ingreso entre $168.000 y $300.000 (29.57%).[7]

El tipo de formación de los alumnos en su mayoría corresponde a científico-humanista (81.57%), es primera vez que cursan una carrera en educación superior (67.17%) y aquellos que ya habían estado antes en otra institución de educación superior, un 62% lo había hecho en universidades privadas.

La situación ocupacional que presentan en un 74.75% corresponde a alumnos que sólo se dedican a estudiar.[8]

De este universo de alumnos, dado el conocimiento de las solicitudes de información que el autor tuvo en la atención de consultas en el servicio a los usuarios, se optó por elegir una de la carreras que presenta a menudo consultas más complejas debido al tipo de temáticas que se abordan en los trabajos de investigación solicitados a los alumnos y que requieren mayor respaldo bibliográfico. Esta carrera es Psicología Transpersonal.

Se abordará esta área cuya bibliografía forma parte de la colección de la biblioteca, como colección piloto para diseñar una solución a las dificultades ya mencionadas en la búsqueda de información temática. Esto se debe a que los usuarios presentan, a diferencia de otras carreras, consultas de corte más temático, sobre materias muy amplias a veces y en otras circunstancias muy específicas también. No buscan un título en particular sino una serie de libros que aborden los temas que investigan (mayor cobertura, menor precisión) o bien algo tan particular que tal vez figure presente en unos pocos textos (menor

[7]UNIVERSIDAD DEL PACÍFICO. DIRECCIÓN DE ANÁLISIS Y ASEGURAMIENTO DE LA CALIDAD. (2014). *Estudio Caracterización estudiantes 2014 en composición socioeconómica de acuerdo a ingresos.* Universidad del Pacífico. Santiago.

[8]UNIVERSIDAD DEL PACÍFICO. DIRECCIÓN DE ANÁLISIS Y ASEGURAMIENTO DE LA CALIDAD. (2015) *Resultados de encuesta de caracterización alumnos ingreso 2015.* Universidad del Pacífico. Santiago.

cobertura, mayor precisión). Esto quiere decir que en las búsquedas de bibliografía sobre un tema les resulta más válido contar con el mayor número de fuentes a citar, independiente a veces si contienen desarrollados los temas que indagan, puesto que les importa sumar títulos consultados a fin de demostrar una búsqueda exhaustiva. Por contraparte, otros casos se refieren a dar con las fuentes más precisas aunque sean menos las que contengan la respuesta a sus inquietudes.

Los usuarios de esta carrera se caracterizan por ser más asiduos a la biblioteca que el resto de los otros alumnos de la universidad. Solicitan títulos más específicos pero cuando buscan por temas, consultan varios libros antes de pedir uno determinado.

Actualmente la cifra de alumnos asciende a 350 alumnos de pregrado y 6 de postgrado.

V.4.1 Comportamiento

Los estudiantes, para desarrollar sus trabajos de investigación, presentan los requerimientos de búsquedas de información de un tema en específico en la biblioteca, basándose en la forma que se lo plantean los docentes. Al no existir un servicio de referencia (área de la biblioteca donde en forma personalizada se atienden consultas y se realizan búsquedas de información), debido a la falta de personal profesional, se encuentran en una incertidumbre de cómo abordar esas búsquedas si no aparecen los términos en el catálogo. Esto conlleva a que necesariamente presenten sus consultas al personal de biblioteca para que pueda indicarles qué libros les pueden servir. El problema es que tampoco el asistente o bibliotecario sabe responder siempre a estas demandas ya que la mayoría de las veces sólo hace uso de las herramientas de búsqueda del sistema de información bibliotecaria que responde a lo descrito en el catálogo, o bien al conocimiento personal que tenga sobre algunos materiales bibliográficos en su memoria.

En lo que respecta a los docentes, pese a manejar un mayor número de términos que pueden ingresarse en la búsqueda, también se encuentran con los obstáculos antes descritos por lo que sus limitaciones se asemejan en alguna medida a lo que presentan los alumnos, en términos de acceder al contenido de los libros.

El proyecto se da dentro de un entorno de usuarios con limitaciones en la búsqueda de información, especialmente en lo referido a temáticas. Por eso que continuamente estas consultas las derivan al mesón de atención para tener una respuesta más rápida, pero que va en contra de sus propio desarrollo de habilidades para el establecimiento de estrategias de búsqueda informacional.

Son múltiples las consultas que pueden presentarse en una unidad de información. Desde asuntos muy específicos que no dan ni siquiera para escribir un subcapítulo de un libro hasta temas muy amplios que sin embargo dadas las aristas que tienen cuesta dar con la que se requiere en particular abordar.

V.4.2 Las necesidades de información de los usuarios

Basándose en la experiencia de años en atención al público en la biblioteca de la universidad, se establecieron cuáles podrían ser el tipo de solicitudes que los usuarios tendrían al hacer consultas y que situaciones podrían generarse, teniendo en cuenta las acepciones de los términos, las complejidades de encontrar temáticas o palabras específicas que figuran en los textos, las búsquedas por frases, la recuperación de títulos de capítulos, entre otros aspectos. Sumado a ello se estipularon los requerimientos operativos del administrador para poder resolver las consultas en forma efectiva.

Las consultas que se reciben en la unidad de información que se presenta como objeto de interés en este trabajo de tesis, varían en complejidad de acuerdo a las áreas de las carreras que se imparten en la Universidad del Pacífico. En ese sentido, las carreras de las áreas de negocios y marketing generalmente resultan

más abordables que las que se puedan dar en otros casos. Son temas más amplios y que cuentan con la posibilidad de recuperarlos con las materias asignadas o con palabras claves ingresadas en la descripción de las fuentes de información. Los mismos títulos de los libros dan solución ya que se enfocan a aspectos concretos de sus disciplinas, y truncando los términos se pueden obtener buenas respuestas.

Las carreras de las áreas de diseño ya tienen un grado de complejidad mayor ya que a veces no sólo se refieren a conceptos sino a imágenes en particular, por tanto escapan de lo que aquí pretende abordarse. No por eso dejan de presentarse a veces consultas que requieren de búsqueda exhaustiva debido a un léxico particular que usan en las diferentes gamas del diseño, tecnicismos lejos del alcance de un vocabulario normal. Se pueden incluir aquí las carreras relacionadas con el área del mundo digital que aparte de lo mencionado hacen uso común de anglicismos lo que complica algo más el universo de términos.

Por último, las que provienen de áreas de ciencias humanas o relacionadas con comunicaciones, resultan de mayor complejidad, porque tiene un campo de teoría más vasto y un manejo de conceptos de mayor riqueza. Es dentro de estas áreas que se produce el número mayor de demandas de información no sólo por términos en particular, sino que enlazando dos o más temas que se contengan en una obra. Y aquí se produce entonces que más allá de buscar un concepto, lo que se requiere a veces que se puedan juntar varios de ellos o ir agregando condicionantes para dar con resultados atingentes.

Debido a esto, se eligió trabajar con la carrera de Psicología Transpersonal ya que es una de las más asiduas a presentar requerimientos de información de este tipo, o sea búsquedas de conceptos bien específicos, temas relacionados o temas amplios poco frecuentes. Presentan campos de acción en ese sentido más atractivos de indagar. Es decir, pueden presentarse consultas tan amplias como buscar sobre el tema del “duelo” (amplio en el sentido de que es posible abordarlo por connotaciones psicológicas o socioculturales, tradiciones, etc. del cual no existen libros en concreto dedicados a este tema dentro de la colección

con que cuenta la biblioteca de la universidad) o consultas sobre un tema muy específico, como “dispraxias” por ejemplo.

Las consultas pueden darse en búsqueda de un asunto tan amplio como el mencionado recién. Podría de partida considerarse como un término de doble significado: una afrenta, o un dolor por la pérdida de alguien. Ya presenta un punto de partida confuso que debe aclararse en la consulta. Si tiene que ver con lo primero podría buscarse también por sinónimos como “combate” o “reto”. Si se trata de lo segundo, podría buscarse por “luto”. Si lo llevamos a términos relacionados puede darse que aparezcan “desafío” o “condolencia”. Como se ve las complejidades por un concepto pueden darse fácilmente.

Dada la experiencia del suscrito en la atención de consultas, esta carrera se puede decir que es una de las más recurrentes en hacer consultas sobre temas que demandan mayor búsqueda bibliográfica ya sea a través de palabras claves del título o materias en el catálogo tradicional que cuenta la biblioteca. Como no siempre se resuelven por esa vía, cabe entonces indagar en los contenidos más detallados de las fuentes bibliográficas que pudiesen tener relevancia.

Un término como “holográfico” o “reencuadre” son términos específicos de estas disciplinas y que pueden estar contenidos en algún capítulo o parte de algún libro o documento y que sin embargo es tan poco lo que se abarca sobre ello a veces, que no da para asignar una palabra clave al registro bibliográfico ya que daría cuenta de algo muy menor en comparación a otros temas que sí tienen mayor presencia evidente.

Palabras compuestas como “alteraciones de la conciencia de identidad” son ejemplos de que se requiere buscar por la suma de estos términos en donde se encuentre presencia de ellos tres necesariamente para dar con documentos relevantes. Si aparece “Alteraciones de la conciencia” podría tratarse de un libro sobre uso de drogas, pero no necesariamente con pérdida o alteración de identidad tal vez. Si fuese “conciencia de identidad” podría tratarse de un libro donde se trata el tema de la formación de identidad en una fase inicial del desarrollo humano, en una de esas. Pero si aparecen los tres términos en una

frase, se estaría más cerca de dar con algo realmente asociado a los cambios que se pueden producir en la noción de uno mismo en determinadas circunstancias. Acotar la proximidad de la presencia de los términos a la mínima expresión sería lo más acertado para reducir la dispersión de los términos.

Si se presenta la consulta sobre encontrar algún documento sobre “atención psicológica infanto-juvenil en Chile”, no estamos trabajando con términos complejos, pero sí se trata de una consulta específica que debe tener presencia de todos estos factores, deseablemente con un grado alto de proximidad en el texto para que sea mayor la precisión de los resultados que se obtengan en la búsqueda bibliográfica.

Para efectos del sistema este tipo de consulta puede resultar la más desafiante de resolver ya que la aparición de los términos asoma como más dispersa y eso afecta la precisión, ya que las menciones pueden variar de contexto dentro del texto.

En menor grado pueden darse consultas que apunten a la raíz de un término y que requieran rescatar todo lo relacionado a ella para revisar los resultados más convenientes. Ejemplo: “trans*” para recuperar “transpersonal”, “transfiguración”, “transformación”, “transición”, etc.

Otro tipo de requerimiento que se identificó en la experiencia de la atención de consultas y que esta herramienta bien podría dar una respuesta favorable es aquella que dice relación con los títulos de capítulos de libros, ya que en determinadas ocasiones algunos usuarios llegan solicitando ese título y desconocen el del libro del que forma parte. Por tanto puede ser una ayuda para resolver ese tipo de requerimientos también.

Estos son algunos tipos de posibles consultas que pudiesen presentarse y que el sistema debiese abordar eficazmente para arrojar resultados favorables.

V.4.3 Los requerimientos del operador

En cuanto a los requerimientos del operador de sistema, es necesario que al ver los resultados de la consulta, aparte de los términos ingresados puedan figurar los números de las páginas en donde se encuentran dentro de los libros (para eso se incluyeron las tablas de contenido e índices precisamente).

Además se desea que los términos y números aparezcan lo más legibles y completos posibles, libre de basura como líneas de puntos o similares.

También es deseable que los resultados figuren por separado, visualmente fáciles de identificar. Con el término resaltado o destacado de alguna manera para poder ubicarlo inmediatamente.

V. 5 Sobre medidas de solución aplicadas

Como consecuencia de la situación descrita anteriormente, se entiende que se invisibilizan y subutilizan parte importante de los recursos de información que dispone la biblioteca, ya que se presume que ciertos temas pueden estar contenidos en algunos libros, pero sin tener clara idea de cuáles son o cómo acceder a ellos.

La medida técnica que se ha tomado para este tipo de situaciones ha sido incorporar en algunos registros bibliográficos una(s) materia(s) no autorizada(s) o normalizada(s) (campo 690 en formato MARC) basándose en las consultas directas del usuario sobre ciertos temas que requieren de un medio de recuperación. Son los propios usuarios en estos casos quienes entregan el tema o materia por el que hacen la consulta y que se contrasta con lo que figura en los textos afines. Generalmente está tratado de esa manera y por tanto se ingresa como materia normalizada en el campo señalado anteriormente.

Sólo se agrega en algunos debido a que precisamente son pertenecientes a las áreas que demandan más por temas que por títulos específicos.

El procedimiento es identificar qué libros incluyen determinadas materias no registradas, después de una consulta reiterada y luego se crea una palabra clave dentro del campo mencionado.

VI. MARCO TEÓRICO

Dentro de las funciones esenciales de una biblioteca podemos mencionar la que corresponde a los servicios. Luego, en la diversa gama de servicios se incluye el servicio de referencia y el de búsquedas bibliográficas. Ambos tienen en común que a partir de una consulta de información planteada por el usuario se debe hacer uso de los recursos bibliográficos existentes para satisfacer dicha necesidad. Ante esto y dado que estamos hablando de un volumen considerable de fuentes a revisar, se requiere establecer un filtro para buscar información en un grupo particular de documentos que pueden resolver esta situación. Por tanto, dentro de toda la información registrada en un sistema, a través de las descripciones bibliográficas correspondientes, se debe recuperar aquella que resulte pertinente. Es sumamente destacable la relevancia que tiene el registro bibliográfico por contener los datos y las temáticas de los materiales bibliográficos, lo que implica que la recuperación se realiza en base a lo ingresado en el proceso de descripción bibliográfica y de los contenidos.

Al hablar de "recuperación de información" debemos remontarnos hacia 1950 cuando el Sr. Calvin N. Mooers usó ese término dentro de la literatura de documentación, definiéndola como:

"la búsqueda de información en un stock de documentos, efectuada a partir de la especificación de un tema" [9].

Posteriormente incluiría dentro de este proceso la descripción de información y especificaciones para la búsqueda dentro de cualquier tipo de sistema que se utilice para tales efectos.

Frederick W. Lancaster proporciona una definición similar a la anterior pero agrega otros aspectos significativos, planteando que:

"en la actualidad la recuperación de información "convencional" significa la búsqueda online en bases de datos electrónicas, de forma interactiva y en tiempo

[9] MOOERS, C.N. (1950) The theory of digital handling of non–numerical information and its implications to machine economics. *Association for Computing Machinery*, Rutgers University, 29 march 1950, New Brunswick, New Jersey.

real. Normalmente, esto implica que el usuario construye una estrategia de búsqueda usando términos con distintas relaciones lógicas (booleanas) y que el programa de búsqueda simplemente divide la base de datos en dos conjuntos: elementos recuperados y elementos no recuperados" [10]

Croft la define como:"el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc.". [11]

Baeza-Yates y Ribeiro-Neto mencionan que "la Recuperación de Información trata con la representación, el almacenamiento, la organización y el acceso a ítems de información". [12]

Definiciones como la anterior abundan, sobre lo que significa "recuperación de información"; algunas involucran más tareas que otras, pero hoy en día apuntan en su mayoría a relaciones lógicas que realiza un sistema computacional para encontrar resultados relevantes a las necesidades de información presentadas por los usuarios.

Hay una relación directa entre el almacenamiento y la recuperación de información. El almacenamiento a través del registro organizado y representativo de los contenidos de los materiales bibliográficos, y la recuperación por medio de estrategias adecuadas que permitan acceder a la información requerida. Por tanto, el concepto de relevancia surge en esta instancia.

Al respecto Tolosa y Bordignon, plantean*:* "la relevancia como similitud, de manera de poder comparar documentos con consultas y – bajo ciertos criterios – definir una medida de distancia entre ambos. Por lo tanto, se puede plantear la idea de que "un documento es relevante a una consulta si son similares", donde la medida de similitud puede estar basada en diferentes criterios (coincidencias de términos,

[10]LANCASTER, F.W. (2001) Sistemas avanzados de recuperación de información. En: Wilfrid Lancaster y María Pinto (Coords.), *Procesamiento de la información científica*, Arco/Libros. Madrid.

[11]CROFT, W.B. (1987) Approaches to intelligent information retrieval. *Information Proccesing & Management, Vol.23*, N°4, 249-254.

[12] BAEZA-YATES, R. y RIBEIRO-NETO, B. (1999*) Modern Information Retrieval*. ACM Press. Addison Wesley. New York.

significado de éstos, frecuencia de aparición de términos y distribución del vocabulario, entre otros).”[13]

La relevancia es una medición subjetiva ya que es el propio usuario el que evalúa la pertinencia de los resultados en relación a sus necesidades de información. Dependiendo de este criterio exclusivo se puede determinar si un texto o documento resulta relevante o no, pudiendo inclusive existir diferentes grados de relevancia ya que ésta puede reflejarse en la totalidad o una parcialidad del documento.

Por otra parte, el almacenamiento de información se refiere al resguardo de datos en diferentes soportes o dispositivos (internos o externos) que aportan a la conservación de estos dentro de un sistema estructurado.

Estos conceptos (recuperación de información, almacenamiento de información y relevancia) han llamado la atención en distintas áreas relacionadas a la información, a saber: informática, biblioteconomía, comunicación, periodismo, lingüística, entre otras, conformando a la vez, según Salvador y Arquero:

“…el tema central de la recuperación de información. El cómo lograrlo ha atraído el interés de investigadores de diversas disciplinas y provocado una actividad multi e interdisciplinar en las tres direcciones señaladas anteriormente, conformando a la vez tres características esenciales y que consideramos el leitmotiv de la evolución y existencia del campo de la recuperación de información: Interdisciplinariedad, estrecha relación con la tecnología de la información, y participación activa en la evolución de la sociedad de la información”.[14]

Cabe mencionar que para efectos del presente trabajo el enfoque del concepto de información se hará dentro del contexto de la bibliotecología principalmente, con algunos enlaces a elementos de la informática, principalmente respecto al uso de herramientas que colaboran con la recuperación automática de información.

---

[13] TOLOSA, GABRIEL H. y BORDIGNON, FERNANDO R.A. (2008) *Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos.* Universidad Nacional de Luján, Argentina.

[14]SALVADOR OLIVÁN, JOSÉ ANTONIO y ARQUERO AVILÉS, ROSARIO. (2006) Una aproximación al concepto de recuperación de información en el marco de la ciencia de la documentación. *Investigación bibliotecológica*, *Vol. 20*, N°41, 13-43.

La recuperación de información ha evolucionado con el paso del tiempo incorporando el uso de tecnologías que facilitan esta labor. Los lenguajes que permiten restringir búsquedas en las bases de datos, la aplicación de algoritmos para ejecutar búsquedas en grandes cantidades de registros bibliográficos, la inclusión de índices y control de términos, el uso de operadores en tratamiento de palabras (para truncar, agregar, separar términos, entre otros), son ejemplos de las variadas formas utilizadas en el proceso de recuperación. El factor común de estas operaciones es que se presentan dentro de una dinámica que implica la participación de personas en el procesamiento de datos: léase asignación de marcadores, ejecución de combinaciones de términos, aplicación de algoritmos para procesamiento de datos, etc.

Por otra parte existe otro tipo de tecnología empleada para estas labores de recuperación, que tiene la particularidad de procesar y reproducir el contenido textual presentado en el libro o documento, una imagen tal cual viene en el original y que se procesa por reconocimiento de caracteres o símbolos. En estos casos los elementos que entregan información son datos no estructurados, sin formato (documentos en texto libre). OCR (Optical Character Recognition) es el programa de digitalización de caracteres (palabras, números, símbolos) a través de un proceso de digitalización, para convertirlos como datos, almacenarlos y que sean posibles de recuperar con algún programa editor de textos.

Según Ballesteros, Morales y Cedillo: "Es importante anotar que el funcionamiento exitoso del proceso OCR radica esencialmente en la clase de imagen a la que se le apliquen estos procesos, es decir, una buena imagen supone dos cosas fundamentales: primera, que el texto en el documento original sea legible, exento de roturas o manchas, con letras uniformes y bien impresas (elementos comunes en una publicación moderna); segunda, que la representación digital que se obtenga de ella sea nítida, encuadrada, sin perspectiva o deformaciones por curvatura y a una resolución suficiente que permita la captura fiel del texto impreso"[15]

[15] BALLESTEROS ESTRADA, SILVIA; MORALES ROMERO, GUILLERMO Y CEDILLO PÉREZ, PAVEL. (2012) Los problemas de identificación de caracteres OCR para la recuperación de texto en el libro antiguo: un análisis de caso en el Fondo Antiguo de la Biblioteca Central. *Biblioteca Universitaria*, V*ol. 15*, N°1, 25-34.

Para Peis y Fernández-Molina: “el uso eficiente de tecnología OCR es un elemento esencial, no sólo por la conversión de las imágenes capturadas dentro del texto (preservando información esencial), sino también para la identificación de la fuente monográfica. Para ello se utiliza un, método de reconocimiento de cadena de texto basado en el título o ISBN de la monografía que posteriormente, por comparación, nos permite grabar la información necesaria ya etiquetada.”[16]

En este proceso no hay demasiada intromisión de la labor humana, más allá del hecho de capturar la imagen pero sin interceder en la interpretación de los contenidos, ni tampoco implica tareas de tipeo, lo que disminuye el riesgo de error, excepto si se presentan problemas de legibilidad de la imagen original. Una letra mal identificada hará perder la recuperación de ese término debido a que al tratarse de una parte de una palabra completa y no de una letra, con ese tipo de errores el término no es reconocido por el sistema y queda al margen.

Este tipo de tecnologías han favorecido el proceso de recuperación de información y su intervención en procesos dentro de las bibliotecas y centros de documentación han traído una serie de beneficios adicionales como el ahorro de tiempo en proceso y respuesta, la gestión de recursos y la valoración positiva por parte de los usuarios.

Byrne y Micco mencionan que: “los términos extraídos de la tabla de contenido, incrementaron la recuperación en un 300%”. [17]

Si se dirige la mirada hacia lo tradicional, en términos de recuperación de información, las posibilidades que otorgan los OPAC (On line public access catalogue), se enfocan a entradas tipificadas como campos recuperables: datos de autor, título, materia y serie son los más comunes.

---

[16]PEIS, EDUARDO Y FERNÁNDEZ-MOLINA, J. CARLOS. (1998) Enrichment of bibliographic records of online catalogs through OCR and SGML technology. *Information Technology & Libraries, Vol. 17*, N°3, 161-172.

[17]BYRNE, A., & MICCO, M. (1988).Improving OPAC subject access: The ADFA experiment. *College & Research Libraries, Vol.49, N*°5, 432-441.

La falta de información específica sobre una referencia bibliográfica, causa que al usuario sólo le quede la opción de hacer búsquedas por materia o palabras claves. Al respecto, estos sistemas brindan la posibilidad de recuperar registros que se encuentren ingresados ya sea bajo las materias o palabras claves asignadas por criterio del catalogador por un lado, o bien recuperar palabras que figuren en el título o en algún resumen que se pueda haber incluido. Adicionalmente también puede darse una búsqueda dentro del campo de nota de contenido en caso que se haya creado y registrado.

Los resultados por tanto se restringirán a cuánta información se haya registrado en el momento del procesamiento del material bibliográfico. Sin desmerecer la labor que se ejecute, ciertamente hay un tema de criterios y exhaustividad en el análisis temático que puede perjudicar en parte la recuperación de registros, ya que si bien puede haber una mayor cobertura (*recall*) puede afectar la precisión por la disminución en aciertos efectivos. Al fin y al cabo se trata de una labor subjetiva la de asignar descriptores temáticos, por tanto tiene una importantísima incidencia en la recuperación de información. Cabe mencionar que las mediciones en recuperación de información se determinan por el *recall* y precisión que se determine. El *recall* es el porcentaje de documentos relevantes de una colección recuperados después de una consulta. Precisión es el porcentaje de documentos recuperados que son relevantes.

Otro aspecto digno de considerar, son las obras de autores colectivos donde cada capítulo cuenta con título individual, como una antología o unas actas de un congreso por ejemplo. Si no está creada en detalle la nota de contenido esa información difícilmente se recupera.

La nula incorporación de elementos que favorezcan aún más la precisión en la búsqueda por temas, se produce que se gana en la diversidad - o *recall* - de fuentes bibliográficas a revisar, pero con el inconveniente de gastar tiempo en ver si responden a los reales requerimientos. Se hace necesario entonces contar con más elementos descriptivos que aporten a dar con los contenidos que se andan buscando.

Ese es un problema que podría suceder en general en varias bibliotecas en el trabajo de procesamiento de fuentes bibliográficas, donde cabe mencionar que la asignación de materias se restringe a listas de encabezamientos o tesauros poco flexibles, limitados, desactualizados a veces, asuntos que obviamente no cooperan a acertar en las búsquedas bibliográficas.

Este inconveniente repercute en otro, de suma importancia. Se refiere a la experiencia del usuario al hacer una búsqueda por materia o palabra clave en un sistema de recuperación de información.

Ya se enunciaba al comienzo sobre la dinámica que se produce cuando no se cuenta con datos concretos sobre una referencia bibliográfica y sólo queda hacer una búsqueda por el tema (ya sea por materia o palabra clave). Los usuarios tienden a ingresar términos muy extremos al hacer una búsqueda. O bien algo muy genérico que le pueda mostrar todos los aciertos referidos a un tema (por ejemplo buscan “drogadicción”, y luego por su cuenta empiezan a revisar un largo listado de resultados y eligen los que más le parezcan), o bien ingresan un tema tan específico que difícilmente exista un libro referido a eso en particular (ejemplo, buscan por “lateralidad” que es un concepto relacionado con la psicomotricidad). Puede que exista un capítulo, subcapítulo o se haga una mención menor tal vez dentro de un texto, pero ¿cómo dar con ello entonces?

Además súmese a esto que los usuarios no conocen la terminología usada para registrar materias y las referencias “véase” da la impresión que no siempre están presente en un catálogo determinado, como para ingresar el término correcto al sistema para que lo recupere. Ya que al tratarse de bibliotecas universitarias, el tratamiento y uso de terminología es más bien general, pues no se está considerando el uso de tesauros especializados para cada área temática que componen la colección.

Entonces la búsqueda por materias se reduce dada esta dificultad y aumentan las consultas directas al referencista, siempre que esté presente tal servicio en la unidad de información. De lo contrario queda la opción de seguir probando palabras en el catálogo, pedir varios libros para revisar y tal vez terminar desistiendo en la

búsqueda siendo que perfectamente puede existir información en las fuentes bibliográficas pero que no existe el vínculo entre sistema automatizado y formato físico, a través de un texto recuperable por una herramienta adecuada.

Para confirmar esta situación se hace mención a la ponencia presentada en el III Encuentro Internacional de Catalogadores que al analizar una serie de catálogos de bibliotecas del Mercosur (universitarias, públicas, nacionales y especializadas) se determina que dentro de las funcionalidades o atributos, las relacionadas con la búsqueda son las que presentan mayores porcentajes, desglosándose de la siguiente forma:

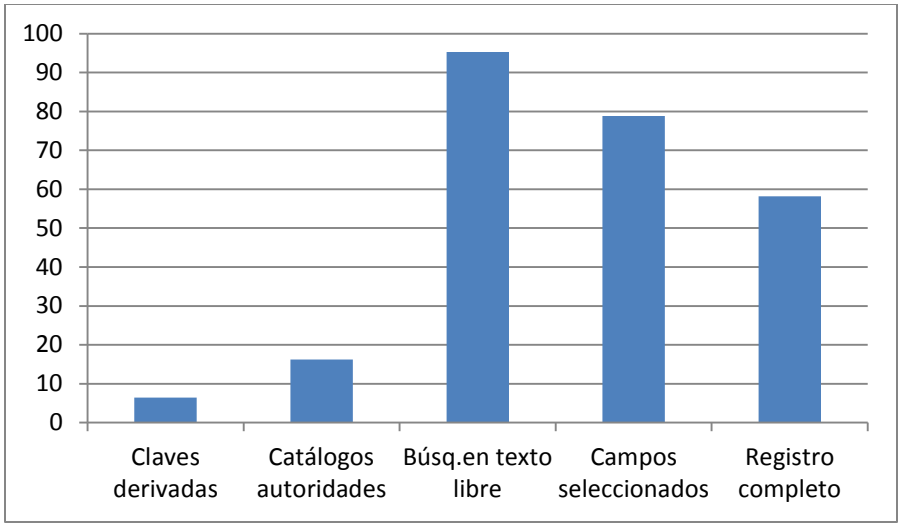
Tabla 1. Atributos de búsqueda en bibliotecas Mercosur.[18]

Funcionalidades	**Porcentaje**
Búsqueda en texto libre	97.70%
Búsqueda en texto libre en campos seleccionados	81.80%
Puntos de acceso por autor	90.10%
Puntos de acceso por título	89.20%
Puntos de acceso por materia	84.80%

Para refrendar esto, las mismas autoras un año después en un artículo relacionado, demostraron que dentro de los catálogos OPAC del Mercosur es de uso general la opción de búsqueda en texto libre con un 95%:

[18]BARBER, E. [et al] (2007) La recuperación de la información bibliográfica en los catálogos en línea de acceso público del Mercosur. III Encuentro Internacional de Catalogadores. Universidad de Buenos Aires, 28-30 noviembre 2007, Buenos Aires.

Tabla 2. Control de formulación de la búsqueda.[19]



Un tercer problema se deriva de esta situación. El referencista o algún sustituto en su defecto, debe ante este panorama y en el afán de asistir el requerimiento de sus usuarios, acudir a revisar aquellos textos u otros materiales bibliográficos para ver si contienen la información que se pide. Lo normal es que se revise la tabla de contenido o sumario, resumen e índices en caso que existan, hasta dar con aquellos materiales que puedan dar respuesta efectiva a lo presentado.

Lógicamente, muchas veces esto toma demasiado tiempo y la memoria del personal bibliotecario no da abasto para tanto volumen de información registrada.

Por último, se debe mencionar que también se produce otro problema que tiene relación con la colección misma, ya que no se podrá sacar el provecho deseable si no se cuenta con el detalle más acabado posible del contenido de las fuentes mismas. Eso produce por cierto un reporte negativo en la inversión económica que se ha hecho ya que los recursos no se usan en la medida esperada muchas veces. A la larga también por supuesto en la imagen misma de la biblioteca si no satisface los

[19]BARBER, E.[et al] (2008) Los catálogos en línea de acceso público del Mercosur disponibles en entorno web. Información, cultura y sociedad, N°18, 37-55.

requerimientos de sus usuarios. En las mismas encuestas que se les hacen responden que no asisten a ella porque “no tienen lo que busco”.

Tenemos así entonces distintos escenarios que requieren de soluciones para mejorar efectos a posterior. Imprecisiones, pérdida de tiempo, frustración, devaluación de colección, pérdida de imagen positiva no son temas menores y que requieren de una solución eficaz y efectiva para mejorar la recuperación de información y la satisfacción del usuario.

VII. ESTADO DEL ARTE

La recuperación de información cuenta con modelos asociados a ese proceso. Existen varias opiniones al respecto.

De partida se puede determinar, según Baeza Yates[20], una diferenciación de modelos de acuerdo a las características estructurales o de conformación de los documentos. Así, él menciona que se dividen en documentos estructurados y no estructurados.

Dentro de los estructurados (los que cuentan con estructura semántica definida, como las bases de datos relacionales o aquellos con marcas o etiquetas html, pdf, entre otros) tenemos el modelo basado en álgebra de regiones. En este modelo se guardan las ocurrencias de los términos a indizar en estructuras de datos distintos, según figuren en alguna parte de la estructura (o región) o en otro como capítulos, secciones, subsecciones, etc. Esta conformación se rige con la idea de favorecer la búsqueda de información.

Por otra parte, los documentos no estructurados (los de texto libre, sin formato alguno o texto plano) se ajustan principalmente a los modelos más tradicionales que son tres:

VII.1 Modelo booleano

Es un modelo de recuperación y clasificación simple, que se basa en la teoría de conjuntos y álgebra de Boole. Ha sido principalmente adoptado por sistemas comerciales de búsquedas bibliográficas. Incluye en el cuadro de búsqueda operadores entre los términos a consultar. Estos pueden ser AND, OR y NOT (intersección, unión y diferencia respectivamente) permitiendo usar más de uno de ellos para efectuar operaciones más complejas.

En este modelo, cada documento es constituido por un conjunto de términos, luego cada uno de ellos se trata como una variable en donde si el término está

[20] BAEZA-YATES, R. y RIBEIRO-NETO, B., op.cit.

en el documento se toma como verdadero y si no, como falso. Cada término tiene una ponderación de existencia (presencia o ausencia), lo que no indica si el término es frecuente o no dentro de la colección. La frecuencias término-documento en la matriz término-documento son todas binarias.

El modelo predice si cada documento es relevante o no relevante. No hay opción de coincidencia parcial a los requerimientos de una consulta específica.

VII.1.1 Ventajas del modelo booleano

- Considera un formalismo claro y simple.
- Utiliza una semántica intuitiva y precisa
- Es flexible, porque permite usar combinaciones de operadores y términos para hacer más precisa una consulta.
- Funciona bien con vocabulario controlado, logrando recuperar con efectividad por frases exactas.
- Considera un desarrollo sencillo, sólo requiere el uso de un índice invertido y una interfaz de consulta para realizarlas con expresiones booleanas.

VII.1.2 Desventajas del modelo booleano

- El grado de respuesta a una consulta suele ser muy breve (pocos documentos) o demasiado extensa (muchos documentos).
- No hay clasificación, pues no entrega los resultados con algún orden de relevancia de los documentos.
- Las consultas formuladas pueden resultar muy simplistas y tienden a confundir en cómo plantearlas por tener que expresarlas con uso de operadores booleanos.
- Todos los términos tienen el mismo peso.

VII.2 Modelo probabilístico

El modelo probabilístico fue desarrollado por Karen Spärck Jones y Stephen Robertson entre 1977 y 1979.

En el marco del modelo probabilístico clásico se propone que de acuerdo a la consulta realizada, los documentos de la colección se clasifican en dos conjuntos, uno de documentos relevantes y otro de documentos irrelevantes. En este sentido, es un modelo de tipo binario, existencia o inexistencia de los términos expresados en la consulta. Posteriormente, existe un conjunto de documentos relevantes que dan respuesta a una consulta (conjunto de respuesta ideal).La probabilidad de relevancia depende de la consulta y de la representación de los documentos, que es la información disponible para el sistema. El conjunto de respuesta ideal corresponde a los documentos relevantes y los que no se encuentren incluidos se asume que no lo son. Este supuesto no es del todo cierto, dado que existen factores externos al sistema que pueden afectar la relevancia.

La “consulta ideal”, se determina por los términos indizados dentro de un conjunto de documentos relevantes. Como en la realidad esa “consulta ideal” no es conocida por el usuario, el objetivo es procesar la consulta para ser reformulada hasta lograr el conjunto de respuesta ideal, usando para ello la ponderación de los términos. La ponderación de los términos empleados en la consulta pasa por calcular la probabilidad de que figure el término en el conjunto de los documentos relevantes (peso positivo) y la probabilidad que figure además en el conjunto de los documentos irrelevantes (peso negativo).

Dada una consulta q, este modelo da a cada documento d, como medida de similitud, la relación P (d_j relevante a q) / P (d_j no relevante a q), para calcular las probabilidades que el documento d_j sea relevante para la consulta q. Se deben minimizar las probabilidades de error.

Para ello se debe calcular el peso de los términos de la consulta por medio del ratio Odds:

$$Wti = \frac{P(Ti/R)}{P(Ti/R\neg)}$$

donde:

Wti = peso de un término

P = probabilidad

(Ti/R) = término presente en conjunto de docs.relevantes

$(Ti/R\neg)$ =término presente en conjunto de docs.irrelevantes

Como no se sabe a priori si los términos usados en la consulta son los indicados (buen descriptor) o no (mal descriptor) se habla de una "hipótesis inicial" respecto al peso o valor de esos términos, surgiendo esta dificultad dentro de este tipo de modelo de recuperación de información. Para aquello se asignan valores iniciales llamados de "máxima incertidumbre", que van entre 0 y 1 (0.5) que indica que la probabilidad que el término usado en la consulta esté dentro o fuera de los documentos relevantes, sea la misma.

Sin embargo dado que otorga una ponderación a los documentos de respuesta, los puede ordenar de acuerdo a ese valor de probabilidad en la relevancia. Esto se conoce como **"correspondencia parcial**" ya que no actúa sobre los términos de la colección sino sobre los términos utilizado en la consulta.

Para mejorar la descripción probabilística del total de respuestas ideales se incluye una interacción con el usuario, a fin que determine cuáles documentos son relevantes y cuáles no. Así el modelo refina la descripción de la respuesta ideal, repitiendo el proceso varias veces para mejorar la precisión de ésta.

Un ejemplo de ello es el algoritmo de función de ranking **Okapi BM25** (BM significa mejores coincidencias "*best matching*"). Pone énfasis en la frecuencia del término (número de veces que aparece en el documento) y en la extensión del documento, para poder clasificar los documentos coincidentes según la relevancia que tengan ante una consulta hecha. Mostrará primero el documento cuya relevancia sea más alta.

La fórmula utilizada para ponderar la frecuencia de términos en el documento es:

$$sim_{BM25}(d_j, q) \sim \sum_{k1[q,dj]} B_{i,j} \text{ x } \log\left(\frac{N - n_{i+0.5}}{n_{i+0.5}}\right)$$

VII.2.1 Ventajas del modelo probabilístico

- Presenta los documentos en orden decreciente de probabilidad de relevancia.
- Tiene retroalimentación por relevancia, es decir usa información expresada en consultas anteriores para que se analice la relevancia de los documentos que envía como respuesta. Por tanto hay ahí un mejoramiento en el proceso por mayores probabilidades de acierto.
- Asigna pesos a los términos de la consulta, permitiendo recuperar documentos que pueden ser relevantes.
- Obtiene resultados satisfactorios con colecciones reales y corpus de entrenamiento, lo que lo lleva a constituirse dentro de los mejores modelos de recuperación de información.
- La recuperación es mediante un método de correspondencia parcial, superando al método de correspondencia exacta del modelo booleano.

VII.2.2 Desventajas del modelo probabilístico

- Al ser un modelo binario de recuperación de información, no incluye frecuencia de aparición de los términos en los documentos como ocurriría en el modelo vectorial.
- Al otorgar pesos a los términos, igual se incluye en la recuperación documentos que puede que sean irrelevantes.
- Falta de normalización de la longitud del documento.
- Requiere que se adivine la separación inicial de documentos relevantes y no relevantes.

- La probabilidad de relevancia se basa sólo en la presencia de los términos de la consulta en los documentos.

VII.3 Modelo espacio vectorial

En 1971 el informático estadounidense Gerald Salton propuso un sistema basado en la construcción de un espacio vectorial para la indización y clasificación de documentos para su posterior recuperación. Este modelo lo implementó en el sistema SMART (System for Manipulation and Retrieval Text) que había desarrollado en 1968.

Con modelo de espacio vectorial se refiere que al contar con un índice invertido se registran por cada término los números de identificación de los documentos que lo contienen, más el número de apariciones que tienen dentro de éstos. Este par de datos (el n° del documento donde figura determinado término) más el número de frecuencia (o apariciones del término) forma un vector. Hay algunas funciones implementadas para uso de vectores, lo que constituye un requisito fundamental para que estas resulten y se obtengan ponderaciones de resultados, ya sean de las búsquedas realizadas o de los términos de la consulta ingresados.

El peso asociado al par término-documento no es negativo ni binario. Se supone que todos los términos del índice son independientes y se representan en un espacio t-dimensional, en donde t es el número total de términos del índice.

Las representaciones del documento (dj) y de la consulta (q) son vectores t-dimensionales dados por:

$\overrightarrow{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$

$\vec{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$

donde $w_{i,q}$ es el peso asociado al par término-consulta (k_i, q), con $w_{i,q} \geq 0$.

El modelo espacio vectorial evalúa el grado de similitud entre la consulta y cada documento de la colección, como la correlación entre los vectores que se cuantifica por el coseno del ángulo entre ambos vectores:

$$sim_{(\overrightarrow{d_j},\vec{q})} = \frac{\overrightarrow{d_j} \bullet \vec{q}}{|\overrightarrow{d_j}| x |\overrightarrow{q}|} = \frac{\sum_{i=1}^{t} w_{ij\,x\,} w_{iq}}{\sqrt{\sum_{i=1}^{t} w_{ij}^2}\ x \sqrt{\sum_{i=1}^{t} w_{iq}^2}}$$

donde $\overrightarrow{d_j}$ y $\vec{q}$ son los vectores de la consulta y vectores $\overrightarrow{d_j} \bullet \vec{q}$ es el producto interno de dos vectores. El factor $\vec{q}$ es el mismo para todos los documentos, el factor $\overrightarrow{d_j}$ indica la normalización de la longitud del documento. Ya que los pesos de la consulta y los documentos son iguales o mayor a 0, la similitud varía de 0 a 1. Por tanto un documento puede recuperarse si tiene una similitud parcial con la consulta.

Además, la ponderación se hace por los términos de los documentos con un peso que se le otorga para dar cuenta de la relevancia que posee.

Para ponderar la importancia de un término se calcula un peso, que considera la frecuencia con que un término figura dentro de una colección de documentos (IDF) y la aparición dentro de un documento específico (TF). El peso de un término aumenta si ocurre con más frecuencia dentro de un documento y baja si ocurre más a menudo en todo el resto de los documentos. El cálculo del factor de peso para un término dentro de un documento se define como combinación de la frecuencia de término (TF), y la frecuencia inversa del documento (IDF).

Cuando ya contamos con el grado de similitud entre cada documento de la colección y la consulta, el sistema ordena todos los documentos de la colección en orden decreciente de acuerdo a este grado de similitud con la consulta, incluyendo a los que parcialmente responden a los términos de la consulta.

La fórmula para calcular el peso de un término es:

$$W_{d=0.5+\frac{0.5*tf}{\max(tf)}*} \log\frac{N}{ni}$$

donde:

$\max(t,f)$ = frecuencia máxima de un término en el documento

tf = frecuencia del término dentro de un documento

N = número de documentos de la colección

ni = número de documentos de la colección en donde está presente el término

VII.3.1 Ventajas del modelo espacio vectorial

- El esquema de ponderación de términos mejora a calidad de la recuperación.
- La estrategia de correspondencia parcial permite la recuperación de documentos que se aproximan a las condiciones impuestas por la consulta.
- La fórmula de clasificación por coseno ordena los documentos de acuerdo al grado de similitud con la consulta.
- La normalización de la longitud de los documentos ya se encuentra normalizada.

VII.3.2 Desventajas del modelo espacio vectorial

- Asume que los términos del índice son mutuamente independientes, lo cual no siempre es el caso.

VII.4 Antecedentes históricos

El enriquecimiento de catálogos en línea desde finales de la década de los 80 se empezó a producir por razones económicas en respuesta a las disminuciones de presupuesto para bibliotecas y el aumento de los costos de recursos. Esto llevó a aprovechar mejor los recursos existentes a través de distintas modalidades como aumentar el número de encabezamientos de materia, aplicar indizaciones, creación de resúmenes o sugerir cambios de clasificación. La inclusión de tablas de contenido dentro de los catálogos en línea pasó a conformar uno de los mayores aportes en ese sentido.

Han existido estudios y proyectos relacionados al uso de las tablas de contenido en línea, para poder aportar a la recuperación de información y el incremento del uso de colecciones de libros. Pese a no existir mucha literatura al respecto se dará cuenta de algunas experiencias al respecto.

Cabe mencionar que inicialmente se trató de incorporar términos tomados desde las tablas de contenido para sumarlos a los encabezamientos de materia existentes en los registros bibliográficos, como palabras claves. También se adoptó en algunos casos la modalidad de incorporarlos dentro de campos específicos de los softwares utilizados en bibliotecas, para poder ser recuperados en una búsqueda.

El primer caso se reporta en 1976 con el proyecto de enriquecimiento temático de registros bibliográficos conocido como SAP (Subject Access Project) de Pauline Atherton. Este fue desarrollado en las bibliotecas de la Universidad de Toronto bajo el auspicio del Concilio en Recursos para Bibliotecas. Abarcaba aproximadamente 2.400 títulos de humanidades y ciencias sociales y en 90 búsquedas realizadas se analizaron los resultados alcanzados. Se obtuvo que:

> "… aumentó el acceso, la precisión fue mayor, fue menos costosa la búsqueda on line que la búsqueda en MARC, y tuvo la capacidad para responder a preguntas que la búsqueda en MARC no podía abordar." [21]

[21] MORRIS, R.C. (2001) Online tables of contents for books: effect on usage. Bulletin of the Medical Libraries Association. Vol.89, N° 1, 29-36

En su estudio, el 90% de los títulos seleccionados vieron enriquecidos sus registros bibliográficos con un promedio de 30 términos tomados de las tablas de contenido e índices. Tan solo 50 títulos se recuperaron de las bases de datos usando los campos MARC, mientras que 130 títulos relevantes fueron recuperados de la base de datos enriquecida. Usando ambas sólo 40 títulos se recuperaron. La precisión en búsqueda usando MARC fue de un 35% mientras que en la búsqueda en la base de datos fue de un 46%. En otro estudio de esta autora en conjunto con Markey, sobre uso de OPAC, encontraron que los usuarios deseaban:

"mejoras en búsquedas temáticas, listas en línea de palabras relacionadas y la capacidad para buscar tablas de contenido de libros, sumarios o índices"[22].

Por otra parte, en 1977 el Consejo de Investigaciones de Suecia para la Información Científica y Técnica encargó a Irene Wormell desarrollar una nueva versión del catálogo de la biblioteca que pudiese dar un mejor acceso a las colecciones de esta institución. Ella se afilió a la biblioteca de la Universidad de Lund para diseñar y desarrollar una base de datos piloto que aumentase la descripción temática del catálogo de libros. Dado que las temáticas abordadas eran especializadas en *protección ambiental* y en métodos para la *producción de energía alternativa* se requería algo más que una lista de materias asignadas por el indizador. Se privilegió usar los términos incluidos en los documentos en vez del uso de terminología de vocabularios controlados o de tesauros creados por bibliotecarios. Se constituyó así un fichero SAP (Subject Access Project) que permitía incluir aspectos específicos de temas incluidos en los libros, incorporando tablas de contenido e índices. Estos términos indizados se ordenaron y procesaron, lo que permitió localizar el título del libro y las páginas donde el tema se trataba. Los términos se pudieron

[22] COCHRANE, P.A. Y MARKEY, K.(1983) Catalog use studies – since the introduction of online interactive catalogs : impact on design for subject access. Library Information Science Research, Winter, Vol.5, N°4: 337-63.

recuperar en su totalidad mediante el modo de búsqueda en texto libre. Se destacó la ventaja que los términos usados para la identificación de contenidos estaban representados en cadenas contextuales que le otorgaban mayor significancia y restaba la vaguedad que puede implicar en ocasiones el uso de términos sueltos o frases cortas (palabras claves, descriptores). Esto implicaba que la búsqueda y la relevancia eran optimizadas para el usuario.

Dados los buenos resultados en 1979 este proyecto avanzó a una segunda fase de ejecución para aumentar el acceso a una serie de reportes del gobierno sueco, de tipo factográfico (datos, gráficos, tablas) llamados SOU (*Statens Offentliga Utredningar = Swedish Government Official Reports*). La base de datos que contenía estos reportes alcanzaba los 1.200 ítemes. Se usó la indización SAP que según Wormell:

> "había sido reconocida como una nueva forma prometedora para producir descripciones temáticas detalladas y sus principios se convirtieron en una directriz para el desarrollo de bases de datos domésticas en varios ambientes"[23]

Todos estos proyectos con diferentes metodologías establecían criterios para seleccionar términos relevantes tomados desde las tablas de contenidos e incorporarlos al registro bibliográfico correspondiente.

A comienzos de la década de los 80 Mandel y Herschman,[24] en un artículo publicado para *The Journal Academic of Librarianship,* describen que se comenzaron a implementar OPACs en bibliotecas y se buscaron las formas de mejorar los registros bibliográficos agregando más encabezamientos de materia, incluyendo términos de tesauros especializados, búsquedas por clasificación, mejoramiento de interfaces y estudios relacionando registros bibliográficos y búsquedas temáticas. En 1990 es Richard Van Orden (como

[23] WORMELL, I. (1994) Indización SAP para la exploración del amplio contexto temático de libros y para el acceso a entidades semánticas más pequeñas. *Ciencias de la Información, Vol. 25*, N° 4: 178-186.

[24] MANDEL C. Y HERSCHMAN, J. (1983) Online subject access: enhancing the library catalog. *Journal of Academic Librarianship, Vol. 9*, N° 3:148–55.

director del programa de bibliotecas de investigación y académicas en la OCLC) quien recalcó la importancia de mejorar el contenido al que se accede por información electrónica.

Estimaba que: "Con los continuos aumentos en el tratamiento informático y las capacidades de almacenamiento, las barreras y los beneficios del acceso electrónico más información de contenido se está convirtiendo en un serio problema en la investigación de ciencias de la información. Componentes de contenido bien seleccionados y materiales de texto completo en sistemas electrónicos deben vincularse con metodologías de búsqueda mejoradas, mejores interfaces con el computador, y una mayor comprensión de la estructura y uso del conocimiento."[25]

Como respuesta a esto algunas bibliotecas más visionarias implementaron indizaciones SAP, otras agregaron tablas de contenido como un todo para sus colecciones, mientras otras eligieron parte de sus colecciones para el enriquecimiento de contenidos. La mayoría se enfocó en optar por la inclusión de tablas de contenido como una opción. Tanto cualitativa (impacto en la precisión) como cuantitativamente (costos de enriquecer registros, porcentaje de colección enriquecida) se evaluó la inclusión de este metodología.

El experimento de la biblioteca de la ADFA (The Australian Defence Force Academy) en 1986, incorporó términos tomados desde las tablas de contenido de casi 6.000 libros, que fueron agregados al campo 653 del formato MARC. También esta iniciativa destacó la utilidad de usar este método de bajo costo para la mejorar el acceso a materias aumentando la relevancia.

Como ya se mencionó anteriormente en la enunciación del problema, el estudio hecho por Belkin y Saracevic[26] también resaltó la mejora de los registros bibliográficos al incluir información de las tablas de contenido, con la idea de contribuir tanto a la recuperación como a la relevancia al ejecutar las consultas.

---

[25] VAN ORDEN, R. (1990) Content-enriched access to electronic information: summaries of selected research. *Library High Technology Vol.8,* N° 3:27-32.

[26] BELKIN, SARACEVIC, op.cit

Otro proyecto que se debe mencionar al respecto es el EIS (Engineering Information System) de la Universidad de Purdue (Indiana, EEUU) iniciado en 1987.

Se trataba de un sistema de información desarrollado en su biblioteca de Ingeniería, en donde a los registros bibliográficos aparte de incluírseles los campos tradicionales en la base de datos, se les copiaban las tablas de contenido por cada libro, donde un bibliotecario las digitalizaba todas y personal asistente ayudaba a la edición posterior dentro del archivo. En 1986 la colección contaba con más de 20.000 libros monográficos donde para la mayoría de los títulos se había hecho este procedimiento. El número de términos de búsqueda se había incrementado notoriamente y la indización era una gran ventaja que se lograba con esto. Los usuarios podían hacer sus consultas en su propio lenguaje técnico de ingenieros sin tener que hacer la adecuación al lenguaje del catalogador que incluía términos de materias asignadas por la Library of Congress. Los resultados eran instantáneos y mucho más atinados. Además esta medida permitía hacer una evaluación precisa de los libros involucrados.

En 1990, Belkin et al.[27], notaron en un estudio enfocado al comportamiento de búsqueda del usuario, que éstos tomaban los libros de estantería para revisar las tablas de contenido de los libros para comprobar fehacientemente que tratasen sobre los temas que andaban averiguando. A partir de aquello sugirieron que los OPAC pudiesen incorporar las tablas de contenido o índices para la exploración de los usuarios.

[27] BELKIN, N,J., CHANG, S,J., DOWNS, T., SARACEVIC, T., ZHAO, S. (1990) Taking account of user tasks, goals and behavior for the design of online public access catalogs. *Proceedings of the 53rd Annual Meeting of the American Society for Information Science, Vol. 27*: 69– 73.

En 1992, Wittenbach confirma que el enriquecimiento por medio de la inclusión de tablas de contenido era "*un altamente exitoso método de enriquecimiento de registros*" [28]

En 1997 Morris[29], en un estudio hecho en la Biblioteca del Centro de Ciencias de la Salud de la Universidad de Nuevo México, intentó determinar si los libros que contenían tablas de contenidos en los OPAC eran más usados que los que no las tenían. Se tomaron 3.823 registros aproximadamente de un total de colección que se elevaba por los 55.000 títulos. Para evaluar qué factores impactarían en la oportunidad de que un libro se usase, se definieron algunas características que pudiesen influir en ello:

Año (1991 a1997), tabla de contenido (con o sin), circulación (se presta o no se presta), ubicación del libro, número de clasificación, uso previo (hasta dos meses previo al estudio). Los resultados indicaron que la probabilidad de usar un libro con tabla de contenido incluida en su registro en el OPAC aumentaba en un 45%, privilegiándose el uso de los libros más recientes, cabe indicar. Se concluyó que aparte de incrementar el uso tanto en consulta como préstamo a domicilio, había una correlación en el tamaño del efecto que se basaba en la circulación de los libros y la historia del uso previo que tuviesen. Si un título tenía mucho uso anterior es más probable que se elija si además tiene presente la tabla de contenido. El estudio indica además del valor e importancia que tendría incorporar más tablas de contenido en retrospectiva a sus colecciones aplicando los filtros necesarios.

---

[28]WITTENBACH, S. (1992) Building a better mousetrap : enhanced cataloguing and access for the online catalog. En: M. Ra (ed) *Advances in online public access catalogs: v.I.,* Meckler Publishing. Westport, Connecticut.

[29] MORRIS, R. (2001) Online tables of contents for books: effect on usage. *Bulletin of the Medical Libraries Association, Vol.8,* N°1 : 29-36

Cabe mencionar que este estudio se desarrolló en un área (medicina) que presenta alta circulación. No se puede extrapolar a todas las bibliotecas ya que incide también el tipo de colección y las áreas temáticas a las que atienden.

Sin embargo es perentorio incluir también las limitaciones y dificultades que presenta esta esta medida. En 1991 un estudio hecho por G. Knutson[30] quiso determinar el impacto de incluir tablas de contenido en línea como punto de acceso. Puede decirse que fue la primera evaluación formal de enriquecimiento temático del catálogo a través de esta modalidad. Este estudio se produjo en la biblioteca de la Universidad de Illinois en Chicago. Se trabajó con una colección de ensayos y actas de conferencias sobre ciencias sociales, las cuales no presentaban circulación. Se monitoreó por espacio de un año un total de 291 títulos. Se conformaron tres grupos con esta colección. Un primer grupo con registros que tenían encabezamientos de materia extra y datos de tablas de contenido en línea. Un segundo grupo que no tenía encabezamientos de materia o datos de tablas de contenido en línea. Y un tercer grupo que fue enriquecido con datos tablas de contenido en línea pero sin encabezamientos de materia extras.

En el análisis de los resultados 57 de los 291 títulos (19.6%) circularon al menos de una vez (98 préstamos). De los 50 títulos que se prestaron 23 títulos (46%) correspondía al grupo de enriquecimiento temático mejorado, en comparación al 28% de los registros sin mejoras y al 26% del grupo con sólo datos de tablas de contenido en línea. La combinación de agregar datos de tablas de contenido sumado a agregar más encabezamientos de materia en los registros bibliográficos, aumentaron la circulación de los libros notoriamente.

Por tanto la sola inclusión de tablas de contenido en línea no pareció suficiente para mejorar el uso de las colecciones, apuntando que era necesario incluir además palabras claves en el catálogo. Tal vez el tamaño de la muestra puede haber afectado en los resultados finales.

[30]KNUTSON, op.cit.

Similares resultados se produjeron en un estudio hecho por Dillon y Wenzel en la OCLC (Online Computer Library Center). Al incluir tablas de contenido sumadas a las entradas de títulos y materias como medio de recuperación de información en su catálogo en línea, se demuestra que "el *recall* mejora a medida que información de contenido adicional se agrega a los registros. Sin embargo este incremento es acompañado por una baja en la precisión."[31]

Cuestiones como el gasto adicional que implica procesar colecciones de libros (en términos de tiempo, y pago de horas hombre) o limitaciones metodológicas (tipo de colecciones e ítemes a incluir, tipo de procesamiento para enriquecer contenidos), se suman a la posible sobrecarga de datos en los registros bibliográficos y pertinencia. No todo lo incluido en una tabla de contenido necesariamente responde a una relevancia y basta con revisar el contenido del libro donde se indica que figura tal tema o nombre para ver si es significante o no. Esto pudiese ocasionar entonces una generación de "ruido" innecesario en los registros que afectaría la efectividad en la recuperación. Esto también depende del área temática del libro en cuestión por cierto, en ciencia y tecnología debiese haber menor "ruido" que en ciencias sociales por ejemplo.

Otro aspecto que puede ser limitante es el hecho que algunas tablas de contenido no cuentan con calidad representativa al no expresar los temas que trata el libro en cuestión. Por ejemplo incluir asuntos tales como: "Antecedentes", "Historia", "Proyectos asociados", "Nuevas tendencias", etc.

También hay términos que revisten características especiales en términos recuperativos. Se trata de nombres propios, nombres comerciales, fechas, acrónimos, siglas y topónimos. Estas situaciones como se mencionaba, presentan limitaciones en el tratamiento.

[31]DILLON, MARTIN Y WENZEL, PATRICK (1990) Retrieval effectiveness of enhanced bibliographic records. *Library Hi Technology, Vol.31,* N°3: 43-47.

Para Wittenbach[32] una limitación de las primeras experiencias en este tipo de metodologías es la mala relación entre costo-eficacia del tiempo de trabajo involucrado, y para Van Orden[33] las dificultades de aplicar criterios de selección de términos significativos a las bases de datos (refiriéndose cuando se incluían como palabras claves en los registros bibliográficos).

VII.5 Situación a nivel nacional

Para constatar si existen experiencias similares a nivel nacional en cuanto a la inclusión de tablas de contenido e índices de libros impresos en los catálogos en línea, se realizó el siguiente procedimiento:

- Revisión de diversos sitios web correspondientes a universidades públicas y privadas e inclusión además de bibliotecas de otro tipo, como el catálogo de la red de bibliotecas públicas (dentro de la cual está la Biblioteca de Santiago), la Biblioteca Nacional, la Biblioteca de la CORFO, la Biblioteca del Congreso y la Biblioteca del Centro Gabriela Mistral. En total fueron revisados los catálogos de 25 universidades públicas, 22privadas y 5 de instituciones no académicas.
- Ejecución de búsquedas para ver el despliegue de las fichas bibliográficas con el fin de constatar si existía detalle del contenido de las obras. Se constata que **ninguna permite hacer búsquedas sobre la información contenida en los índices.** La mayoría no cuenta con información sobre las tablas de contenido, siendo pocos los casos de catálogos en línea que sí las incluyen.

  Algunos casos incluyen enlaces al despliegue de la imagen de la tabla de contenido (tal cual viene en el impreso original).Sin embargo, lo más frecuentes que incluyan un campo con los nombres de los capítulos

[32]WITTENBACH, op.cit.

[33] VAN ORDEN, op.cit.

principales incluidos en las tablas de contenido y en pocas ocasiones el detalle completo del contenido.

Sobre esta información incluida en un campo de contenido, se pueden recuperar términos desde el cuadro principal de búsqueda del catálogo. La diferencia reside en que algunos buscadores sólo recuperan si la frase es exacta (se deben incluir palabras de la lista de palabras no relevantes -*stopword*, en inglés- como artículos y preposiciones por ejemplo), mientras que otros permiten encontrar los términos por separado sin necesidad de incluir palabras de la mencionada lista.

- A fin de indagar con mayor precisión respecto a la realidad de este asunto, se intentó tomar contacto con estas instituciones, formulándoles cuatro preguntas en concreto para que pudiesen dar cuenta de su situación. Las preguntas fueron:

  1. ¿Cuál es el sistema que usan en su biblioteca para el catálogo en línea?
  2. ¿Cómo completan el campo de la tabla de contenido?, ¿digitan la información desde el libro o hacen una conversión de imagen a txt?
  3. Si lo digitan, ¿incluyen todo el contenido de la tabla o solo los títulos de los capítulos principales?
  4. ¿Han hecho algún estudio sobre la recuperación de información en vuestro catálogo? para ver cuánto es el nivel de satisfacción en las búsquedas.

A continuación se da cuenta de las 15 instituciones que incluyen en sus catálogos de biblioteca la presencia de tablas de contenido. De estas, hay 10 que respondieron las consultas que se les envió.

a) Universidad de Chile:

El sistema que usa la biblioteca es Symphony y el descubridor ILibrary. En el despliegue de la ficha catalográfica incluye enlaces a tablas de contenido. Al pinchar este enlace se despliega la imagen del impreso original. No permite

recuperar los términos incluidos en estas imágenes a través de su buscador. En el caso de libros electrónicos tampoco se logra recuperar por palabras incluidas en las tablas de contenido, sólo el despliegue de estas.

En la colección de tesis resulta la búsqueda al encontrarse como colección digitalizada por tanto funciona la recuperación en texto completo. Lo mismo en la colección de revistas, que también está digitalizada la información en el servicio “Al día” que reúne el detalle de los artículos publicados en las revistas suscritas por la universidad.

Digitan la información de la tabla de contenido desde el libro y en otros casos las digitalizan y las agregan en un campo 856 del formato MARC. Incorporan sólo los títulos de los capítulos principales. No han realizados estudios sobre niveles de satisfacción en búsquedas por catálogo.

b) Biblioteca del Congreso:

El catálogo está en sistema Horizonte y usando descubridor Plone. Al ver la información relacionada con el recurso encontrado, se despliega la tabla de contenido permitiendo todo tipo de búsquedas para poder recuperar términos o frases exactas.

Si la tabla de contenido está disponible en la web, se copia desde ahí. Si no está, se digitaliza y luego se pega en el campo asignado. Solo se incluyen los títulos principales de los capítulos.

No se han hecho estudios sobre la recuperación de información del catálogo público.

c) Universidad Diego Portales:

Usa el sistema Horizonte. En las fichas del catálogo se desglosa la tabla de contenido del libro. Se recuperan términos individuales y por frases al hacer una búsqueda. (No respondieron las preguntas)

d) Universidad Federico Santa María:

Usa el sistema Koha con descubridor VuFind. En las fichas del catálogo se desglosa la tabla de contenido del libro que es copiado de otros registros. En algunos libros se registra sólo los capítulos principales y en otros de arte y arquitectura en detalle completo. Se recuperan términos individuales y por frases al hacer una búsqueda.

El catálogo permite recuperar información del texto mediante su ISBN, lo cual permite ubicarlo a a través de Google Books (siempre que ahí esté incorporado) y entrega una vista previa en la que podemos consultar el texto con su tabla de contenido. Pero esto sólo funciona como enlace, no se pueden recuperar términos de estos libros de Google Books a través del catálogo.

e) Red de Bibliotecas Públicas (incluye Biblioteca de Santiago y todas las bibliotecas públicas regionales y comunales):

Usa el sistema Aleph. Sólo en algunos registros se desglosa la tabla de contenido y se recuperan términos individuales y por frases al hacer una búsqueda. (No respondieron las preguntas)

f) Biblioteca Nacional:

Usa el sistema Aleph. En algunos registros se desglosa la tabla de contenido y se recuperan términos individuales y por frases al hacer una búsqueda. En ocasiones se traspasa la información digitando el contenido en un campo 505 del formato MARC y en otras oportunidades se digitaliza la imagen de la tabla de contenido creando un enlace que se incorpora en un campo 856. Esto varía según políticas consensuadas de la institución y de acuerdo al tipo de documento.

Al ser una institución depositaria de la producción bibliográfica nacional y aquella que se produzca sobre Chile, el nivel de descripción en esos casos se describe en el campo de contenido con los subcampos de títulos y autor cuando amerite.

g) Instituto DUOC:

Usa el sistema Portfolio. Incluye tablas de contenido en el despliegue de sus registros y permite todo tipo de búsquedas. La información se copia de otros registros o se digita en un campo asignado. Generalmente se incluyen sólo los nombres de los principales capítulos.

h) Universidad Tecnológica Metropolitana:

Usa el sistema Horizonte con el descubridor Enterprise. Incluye tablas de contenido en la ficha catalográfica y permite todo tipo de búsquedas. Esa información la digitan o bien se copia de fichas del libro creadas por sitios web de librerías o de editoriales.
Se intenta incluir en detalle el contenido, si es demasiado se anotan los nombres de secciones o capítulos solamente. En caso que los nombres de los capítulos sean muy genéricos (Introducción, Marco Teórico, Ejercicios, etc.) se prefiere agregar una reseña del libro.

i) Corporación de Fomento de la Producción (CORFO):

Usa el sistema Koha Incluye tablas de contenido en la ficha catalográfica y permite todo tipo de búsquedas. La información del contenido se digita en un campo asignado, sólo los capítulos principales.

j) Universidad Católica de Temuco:

Usa el sistema Aleph. Incluye algunas tablas de contenido, en la actualidad no se realiza ese registro. Sólo permite búsquedas por términos individuales y frases exactas.
Lo que permite es conectar los registros bibliográficos desde el OPAC a Google Books, pudiendo el usuario tener acceso a información extendida sobre los títulos que busque y que tengan presencia en Google Books donde eventualmente se puede revisar la tabla de contenido.

k) Universidad Autónoma:

No se pudo determinar qué sistema usan. Incluye algunas tablas de contenido, permite cualquier tipo de búsqueda. (No respondieron las preguntas)

l) Universidad Las Américas:

Usa el sistema Symphony con el descubridor Portfolio. Incluye tablas de contenido las que son digitadas en un campo asignado. Sólo se incluye los nombres de los capítulos principales. Permite todo tipo de búsquedas dentro de este campo de contenido.

m) Universidad San Sebastián:

Usa el sistema Janium. Incluye tablas de contenido que se digitan en campo asignado.
Si es muy extensa la tabla de contenido se digita solo los capítulos o partes principales. De lo contrario se incluyen todos los capítulos, partes y secciones que tenga. En el caso de los libros que pertenecen a la Colección Patrimonial, se hace un detalle mucho más exhaustivo, por lo tanto a este tipo de material se

incorpora toda la información posible, incluyendo su tabla de contenido completa. Permite todo tipo de búsquedas.

n) Universidad INACAP:

Usa el sistema Millenium. Incluye tablas de contenido y permite todo tipo de búsquedas. (No respondieron las preguntas)

o) Universidad Santo Tomás:

Usa el sistema Alma con descubridor Primo. Incluye tablas de contenido y permite todo tipo de búsquedas. (No respondieron las preguntas)

En líneas generales se desprende que son pocos los catálogos en línea que consideran la inclusión de las tablas de contenido. Y los que las incluyen, en su gran parte crean un campo donde hacen la descripción de contenido de la obra a un nivel básico (registrando en su mayoría sólo los títulos de los capítulos principales). En cuanto a la recuperación, la mayoría se realiza sobre cualquier palabra o frase contenida en este campo.

No hay antecedentes sobre búsquedas en los índices contenidos en los libros.

Tampoco sobre estudios sobre satisfacción en las búsquedas usando catálogos en línea dispuestos en los sitios web de las bibliotecas.

Cabe mencionar que la gran mayoría de los sistemas de recuperación utilizadas en bibliotecas, por no aventurarse a decir la totalidad, funcionan con la modalidad de búsqueda booleana. Horizonte, que es el programa usado en la Biblioteca de la Universidad del Pacífico, también lo hace. Las búsquedas pueden resultar insuficientes muchas veces por no contar con un vocabulario más extenso para recuperar términos de interés. También por realizarse sobre registros con campos circunscritos que describen los contenidos en forma más sucinta, ya sea porque cuentan con una cantidad limitada de caracteres por

asunto de usabilidad o limitantes técnicas del sistema usado; o bien por decisiones humanas en el nivel de descripción bibliográfica.

VII.6 Posibles soluciones

En respuesta a la problemática presentada, la opción propuesta es ampliar las capacidades de recuperación que se puedan ejecutar en línea a través de la búsqueda en texto libre sobre las tablas de contenidos e índices incluidos en los materiales impresos. Para aquello, resulta necesario digitalizar los correspondientes a la colección elegida, conformar una colección de documentos en donde realizar las búsquedas, con los códigos necesarios para que el buscador pueda encontrar el o los términos consultados y obtener la referencia a los documentos donde figuren (no descuidando por cierto el tema de los derechos de autor), para después poder recuperarlos físicamente en estanterías ubicando el texto original.

Para abordar técnicamente esta situación, existen algunos programas y herramientas que, de diferentes formas, entregan posibles soluciones que remedian el hecho que el sistema automatizado que usa la biblioteca tenga la limitación de no hacer búsquedas en texto libre. Se debe contar con una herramienta tecnológica que soporte una colección de archivos digitalizados, genere un vocabulario para indizar y pueda ser compatible con el sistema automatizado que usa la biblioteca, para identificar los resultados con los registros que se le asocian.

Más allá de su funcionamiento y propiedades técnicas de desempeño, entran en juego otros factores asociados que se deben tomar en cuenta al momento de elegir la opción adecuada.

Por supuesto está el costo asociado, tanto en términos económicos como de tiempos. Hay soluciones que implican pago y en términos temporales algunas demoran más en implementarse debido al proceso de diseño y configuración.

También está el tema de la dependencia de otros agentes (proveedores, dueños de marca registrada), las compatibilizaciones y dificultades técnicas que presente la solución, los recursos adicionales que requiera tener para poder implementar, entre otros factores (alojamiento en servidores por ejemplo).
A continuación se exponen algunas posibles soluciones, presentando sus características y atributos. Se optó por incluir tres que actualmente cuentan con presencia en este tipo de iniciativas digitales y el respaldo de comunidades de desarrolladores y usuarios.

VII.6.1 Dspace

Programa de código abierto que permite gestionar contenidos de colecciones digitales en sus distintos formatos (libros, tesis, archivos audiovisuales, imágenes, etc.). Es un proyecto conjunto de las bibliotecas del MIT (Massachussets Institute of Technology) y Hewlett-Packard Company. Es usado como plataforma para repositorios institucionales y gracias a una licencia de software libre BSD (Berkeley Software Distribution) puede personalizarse o ampliarse según lo determinen sus usuarios. Este software fue desarrollado utilizando las normas y estándares existentes lo que le permite integrarse fácilmente a otros sistemas de información. Permite organizar los documentos en colecciones y comunidades, además de contar con sistema de indización.
Aparte de hacer búsquedas en metadatos y en texto completo, permite enviar contenidos (con inclusión de metadatos) vía interfaz web para compartir documentos con otros usuarios, sacar estadísticas, gestionar permisos (para usuarios individuales o grupales), administración de ítems del repositorio, entre otras funciones. Dspace crea URL´s permanentes para los materiales almacenados y permite la generación de copias de seguridad automáticamente de los archivos entre instituciones.

Presenta un modelo de desarrollo comunitario a través de un staff de sus propios usuarios con algunas atribuciones de desarrolladores, que en conjunto con un grupo de técnicos van creando y ejecutando mejorías continuamente, contribuyendo por cierto a su expansión.

Tiende hacia la preservación funcional, para que los documentos se mantengan accesibles con formatos actuales, mientras se desarrollen nuevos o se actualicen formatos existentes.

Está construido con Java y es compatible con el protocolo Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) y por tanto con manejos de datos en Dublin Core.

Actualmente va en la versión 5.3. Es administrado por la organización Dura Space (independiente y sin fines de lucro).

VII.6.1.1 Ventajas

- Programa libre de acceso abierto.
- Opción popular para uso de varias instituciones (más de 100 instituciones en el mundo). A nivel nacional es elegido por bibliotecas de instituciones importantes para conformar y administrar sus repositorios digitales. A nivel internacional es el más usado en repositorios institucionales de código abierto.
- Herramienta multilingüe.
- Fácil de instalar.
- Cuenta con un depurado sistema de búsqueda que incluye uso de filtros para delimitar las búsquedas.
- Continuas versiones mejoradas.
- Maneja cualquier formato de documento.
- Cuenta con documentación técnica actualizada en caso de requerimientos.

- Recibe alertas de protocolo RSS.
- Cada colección de documentos puede ser personalizada en cuanto a restricciones de acceso y uso de metadatos.
- Por ser un programa de código abierto no tiene limitaciones de almacenamiento, concurrencia o descargas de objetos o documentos, como lo presentan algunos software comerciales que permiten sólo cierta cantidad preestablecida.
- La visualización de resultados de búsqueda se pueden abrir en un navegador web o algún programa convencional.

VII.6.1.2 Desventajas

- La administración de un repositorio requiere de un trabajo con tiempo y detallista.
- Presenta cierto grado de dificultad técnica pues tiene una instalación y configuración compleja.
- Obliga a crear y mantener estructuras estáticas, en desmedro de atributos más flexibles. Solicita crear comunidades y colecciones (opcionalmente también crear subcomunidades).

VII.6.2 Omeka

Programa de acceso libre y código abierto desarrollado para publicar en la web los repositorios digitales de instituciones, administración de contenidos web y de colecciones de museos y sistemas de exhibición on line. Es accesible desde cualquier equipo fijo o móvil conectado a la web.

También cumple con el protocolo OAI-PMH (para interoperabilidad) y cuenta con comunidades de desarrolladores de código abierto. Tiene una comunidad

de usuarios que va en aumento (para desarrollo de documentación y soporte técnico).

Independiente del ámbito que se use, el objetivo es proporcionar una plataforma para el desarrollo de proyectos que trabajen con contenidos digitales.

Dentro del ámbito bibliotecario permite conformar un complemento a los catálogos en línea y exhibiciones en línea en el sentido de poder constituir secciones relacionadas con los documentos de cada colección. Permite compartir y marcar colecciones como públicas y destacadas (o lo contrario), documentos y crear archivos digitales con contenido generado por los propios usuarios. El diseño está basado en un sistema hecho en base a plantillas simples y manejables parecidas a las usadas en Wordpress, Joomla o Drupal.

Cuenta con estándares de metadatos de aceptación internacional (Dublin Core, COinS, Simple Vocab, entre otros) y herramientas de migración de datos.

Tiene las propiedades de ser escalable, extensible y flexible. Puede trabajar con archivos de formato TIFF, JPEG, BMP, GIF, PNG y PDF.

Este programa es de propiedad del Roy Rosenzweig Center for History and New Media de la Universidad de George Mason (creadores del gestor bibliográfico Zotero). Se puede implementar el programa en algún servidor particular o bien recurrir a los planes pagados que Omeka ofrece en su nube en omeka.net.

Está desarrollado en PHP. Va en la versión 2.3 con modelo de web 2.0

VII.6.2.1 Ventajas

- Programa libre de acceso abierto.
- Permite crear publicaciones y exposiciones virtuales en torno a sus colecciones digitales almacenadas (archivos, bibliotecas, escuelas, museos, etc.)
- Fácil y rápido de descargar e instalar.

- Factible de ser personalizado.
- Gestiona y almacena cualquier tipo de archivos.
- Permite crear perfiles de acuerdo a los usuarios.
- Permite la creación y uso de vocabularios controlados propios.
- Está hecho para personas sin mayores conocimientos en programación ni en manejo de tecnologías de información (TIC).
- Sistema estable y sostenible.
- Alta usabilidad.
- El programa puede implementarse en servidor propio o bien usar los planes que Omeka ofrece en la nube.
- Cuenta con varios puntos de acceso a los contenidos del repositorio (mediante motor de búsqueda, por geolocalización, por colecciones, por medio de la exposiciones u objetos del repositorio)
- Permite contribuciones al repositorio mediante comentarios o subiendo archivos.
- Actualización mediante plugin propio MyOmeka.
- Se pueden crear suscripciones a los contenidos mediante RSS por ejemplo.
- Permite búsquedas avanzadas en cualquiera de los campos del esquema de Dublin Core.
- Integración con el gestor bibliográfico Zotero.

VII.6.2.2 Desventajas

- No es un programa tan conocido a nivel masivo todavía.
- No está diseñado para un manejo integral de ningún tipo de colección, se focaliza en la capa de comunicación pública.
- Tiene pocas templates (temas) disponibles y no son muy atractivas.

- No tiene estrategia de preservación digital.

VII.6.3 Modelos de R.I. implementados en librerías de Python

El lenguaje de programación Python incluye unas librerías para búsqueda y recuperación de información (RI) llamadas Whoosh y Gensim. Ambas presentan la opción de poder también hacer consultas aplicando extractores de raíces para aumentar la recuperación de resultados relevantes. Aunque la visualización de estos se muestra diferente (en Gensim se ordenan por orden de relevancia según ponderación que hace la librería con un algoritmo) los resultados son similares. A continuación se hace una reseña de ambas.

VII.6.3.1 Whoosh

Es una librería de Python que permite la indización y búsqueda de texto. Funciona para que el programador pueda crear un motor de búsqueda, permitiendo personalizar este tipo de desarrollo. Resulta útil para ser usado en aplicaciones y sitios web. Es factible de implementarlo también en otros lenguajes como Java para construir aplicaciones más completas.

Whoosh fue creado y mantenido por el Sr. MattChaput. Se creó originalmente para uso en el sistema de ayuda on line del software de animación 3D SecondEffects, cuya empresa dueña abrió el código fuente posteriormente para uso libre de quienes requiriesen un motor de búsqueda flexible.

En relación al modelo que usa, corresponde el modelo probabilístico. Pero además puede usar el modelo booleano, que permite a través de algunas funciones específicas otorgar valores binarios a los resultados (verdadero o falso, 1 o 0, sí o no). También formular consultas con los operadores convencionales para aunar, excluir o alternar términos (AND, NOT, OR). Permite usar dos no tradicionales (ANDNOT, ANDMAYBE) uno que excluye y el otro marca la posibilidad o eventualidad de incluir determinado término en la respuesta a una consulta.

Por defecto usa la función de ranking BM25F para clasificar por relevancia la lista de resultados ante una consulta.

VII.6.3.1.1 Ventajas

- Es un recurso de libre acceso.
- La mayoría de sus funciones son personalizables, reemplazables y extensibles.
- Responde rápidamente a las consultas que se ejecuten, incluyendo correcciones ortográficas a éstas.
- Viene con campos predefinidos, aunque también acepta otros nuevos que se puedan generar.
- Permite hacer búsquedas por términos unitarios, por frases o por N gramas (sílabas).
- Destaca los términos de búsqueda en extractos de los documentos originales.
- Amplía los términos de la consulta sobre la base de los mejores documentos encontrados.
- Reconoce palabras claves, lo que permite usarlas para futuras consultas y como aporte a la descripción por materias en el registro bibliográfico.
- Resuelve consultas que estén con términos difusos (errores ortográficos o palabras similares), sugiriendo términos de reemplazo en estos casos.
- Permite el uso de comodines para truncar palabras y recuperar algunas similares.
- Puede realizar consultas binarias, a través de instrucciones que logran hacer un cruce entre dos consultas para obtener un resultado final.
- Al trabajar con texto y no imágenes se disminuyen las opciones de riesgo de recuperar términos, ya que se pueden sortear mejor las dificultades de legibilidad de caracteres que pudiesen presentarse.

VII.6.3.1.2 Desventajas

- Si los documentos indizados están con errores ortográficos las sugerencias derivadas también presentarán errores por consiguiente.
- Presenta inconvenientes con el uso de acentos; los términos pueden contener acentos en el texto original, pero no en la consulta del usuario, o viceversa. No se reconocen como similares.
- Debido a que la implementación de las consultas depende de los documentos primarios y secundarios, no se puede actualizar / borrar un solo documento secundario. Sólo se puede actualizar / borrar todo un documento de alto nivel a la vez. Ejemplo si su jerarquía es "Capítulo - Sección - Párrafo ", sólo se puede actualizar o eliminar los capítulos completos, no a una sección o párrafo.
- Hay idiomas con los cuales no funciona (coreano, japonés, chino).
- La lista de resultados muestra siempre el principio de las páginas y no el lugar donde fue / fueron encontrados la palabra clave (s)
- No es compatible con derivados (por ejemplo búsqueda de "casas" en la búsqueda de "casa"). Se debe usar el signo asterisco “*” para hacerlo efectivo.
- No es compatible con las palabras vacías (palabras comunes que haciendo caso omiso de lo que la búsqueda de ellos es inútil, por ejemplo, "el", "ella", "una").

VII.6.3.2 Gensim

Gensim es una librería de Python que efectúa indización de documentos (corpus), modelado de temas y la recuperación por similitud (opera con modelo de **espacio vectorial**), diseñado para operar en grandes colecciones de documentos de textos no estructurados y apuntando a favorecer la recuperación de información y el procesamiento de lenguaje natural.

Gensim comenzó como un conjunto de scripts de Python para la Biblioteca Checa de Matemáticas Digitales en 2008 para poder generar un listado de artículos publicados más similares a uno en particular, de ahí procede su nominación (GenSim = Generar Similares). Los fundamentos de esta librería se basan en la claridad, escalabilidad y eficiencia.

En el 2011 se comenzó a usar Github para alojar el código y se trasladó su sitio web a su actual dominio. En el 2013 ya adquirió su actual diseño web y es soportado y mantenido en los grupos de Google.

Gensim es de uso libre personal y comercial, redistribuible aunque no se puede hacer cambio de licencia.

En sus funcionalidades, Gensim funciona con TF-IDF (*term frequency* o frecuencia de término, e *inverse document frequency* o frecuencia inversa de documento), es decir el primer factor pondera el valor del término en cuanto a más veces aparece dentro de un documento específico. El segundo factor se refiere al número de ocurrencias del término dentro de una colección de documentos. Mientras menos figure en ese conjunto, más importancia adquiere. Por el contrario, más figuraciones indicarían que no tiene una calidad de término preciso y determinante. Esta característica de Gensim es de suma importancia, pues crea un ranking de documentos que nos determina un orden de relevancia en cuanto a la aparición del término consultado dentro de ellos.

Otra de las funcionalidades son las proyecciones aleatorias y aprendizaje profundo con word2vec También aplica análisis semántico latente (en inglés, Latent Semantic Analysis o LSA) y asignación de Dirichlet latente (LDA) para dar cuenta de la similitud de datos en un grupo de documentos.

VII.6.3.2.1 Ventajas

- Escalabilidad: Procesa grandes corpus en escala web, usando algoritmos incrementales de entrenamiento en línea.
- No depende de plataforma: Se ejecuta en Lynux, Windows y OS X, así como también en plataformas que soporten Python y NumPy.

- Robusto: Ha sido usado por varios sistemas de personas y organizaciones por más de cuatro años. Son más de 400 aplicaciones comerciales. Hoy en día se presenta como una herramienta robusta, eficiente y sin problemas de software para realizar el modelado de temas.
- Código abierto: La licencia GNU LGPL v.2.1 permite uso personal como comercial y requiere que cualquier modificación que se implemente también quede en código abierto.
- Implementaciones eficientes: Los algoritmos básicos en Gensim utilizan rutinas matemáticas altamente optimizadas. Gensim también contiene una versión distribuida de varios algoritmos, destinados a acelerar el procesamiento y recuperación en clusters de máquinas.
- Convertidores y formatos de E/S: Contiene implementaciones eficientes de memoria para varios formatos de datos populares. Éstas se pueden utilizar para la entrada, la salida, o para convertir entre uno otro.
- Consultas similares: Contiene código para la indexación rápida de documentos en su representación semántica y recuperación de documentos tópicamente similares.
- Soporte: Gensim es apoyado y mantenido por medio del esfuerzo de la comunidad a través de cuentas de correos. Cuenta con tutoriales de la lista de correo, FAQ, alojamiento de código e instrucciones para los contribuyentes. Existe también un foro de soporte en Google Groups y otro en Gitter.

VII.6.3.2.2 Desventajas

- Instalación algo compleja para quien no conoce Python.
- No incluye búsqueda ni análisis de relaciones entre las entidades de texto.
- No cuenta con cualquier funcionalidad de visualización.

VIII. HIPÓTESIS

Los modelos probabilísticos y espacio vectorial presentan mejor desempeño en recuperación de información que el uso del modelo tradicional booleano.

IX. METODOLOGÍA

A continuación se da cuenta de la metodología empleada para el desarrollo del experimento donde se compara el desempeño de estos dos modelos usados en recuperación de información: el probabilístico y el espacio vectorial, implementados como librerías de Python en Whoosh y Gensim, respectivamente.

La idea fue probar ambas con un conjunto idéntico de consultas para evaluar cuál presentó el mejor desempeño en términos de precisión y cobertura.

Se trabajó con un conjunto de 65 documentos donde cada uno comprende la tabla de contenido y el índice temático (si estaba presente) de idéntica cantidad de libros. Con estos documentos sometidos a un pre-procesamiento de textos se buscó obtener los mejores resultados, en la medida de tratar de recuperar la mayor cantidad de los que son efectivamente relevantes.

La metodología se divide en: conformación de la colección documentos, pre-procesamiento de la colección, herramientas computacionales, resultados del procesamiento, evaluación y conclusiones. Se presenta la serie de pasos implicados en forma secuencial de estas partes en la línea de tiempo para dar cuenta de los distintos elementos que se fueron integrando a medida que se desarrolló el trabajo.

IX.1 La colección de documentos

IX.1.1 Conformación de la colección de trabajo

En primer término, cabe mencionar que se hizo una selección de libros para conformar una colección de prueba. Esta selección se realizó con el criterio de contar con textos que incluyesen temas de compleja resolución inmediata al momento de realizarse la consulta en el mesón de atención de usuarios en biblioteca. Temas que tanto en una vertiente genérica o específica demandasen una búsqueda bibliográfica más precisa y exhaustiva que lo que el catálogo común de la biblioteca pudiese entregar como respuesta, que se ve limitada en estas instancias.

Para ello se trabajó con la bibliografía oficial de la carrera de Psicología Transpersonal, impartida en la Universidad del Pacífico. Como se mencionó anteriormente, esto se debe a que es una de las carreras que presenta, dadas sus características de temáticas e investigación, aspectos más complejos en las consultas que hacen en biblioteca.

Dicho esto, lo siguiente fue acotar qué textos iban a considerarse para conformar esta colección de prueba. Dado lo extenso de cada una de ellas se optó por incluir solamente los textos incluidos en la bibliografía obligatoria, que en el caso de Psicología Transpersonal fue de 65 textos como se mencionó previamente.

IX.1.2 Digitalización de tablas de contenido e índices

Como principales fuentes de información que dan cuenta de los temas y términos incluidos en los textos, se digitalizaron las tablas de contenido e índices de cada libro seleccionado y se crearon carpetas para cada texto que las contenían. La idea principal es que se referencien con exactitud dónde se encuentra ubicada la información de cada tema o término mencionado en ellos con su(s) respectivas página(s) asociadas.

Se incluyeron además criterios en la selección de aquellos libros que reportasen utilidad para poder desarrollar la idea.

Hay algunos libros que sólo venían con tabla de contenido, y en otros casos sólo se consideró el índice, por lo que no todos los libros fueron digitalizados por igual.

Hubo algunos casos especiales en donde las tablas de contenido no eran lo suficientemente útiles, en el sentido que incluían términos muy genéricos (ver Fig.1) como “facetas”, “fronteras” , “diagnóstico” o “biografías”, o bien incluían sólo nombres de pila en relatos personales (ver Fig.2). Se optó por no incluirlos dentro de la colección dado su escaso aporte.

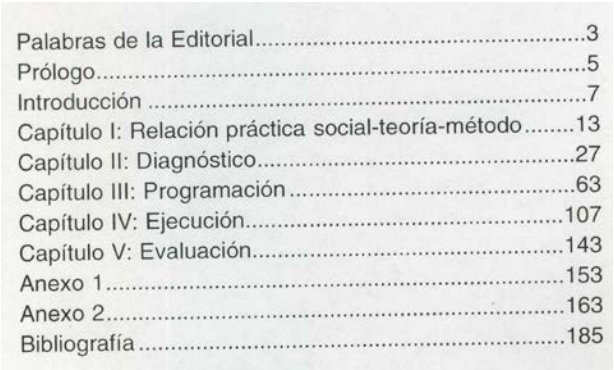


Figura 1. Ejemplo de tablas de contenido con títulos genéricos.

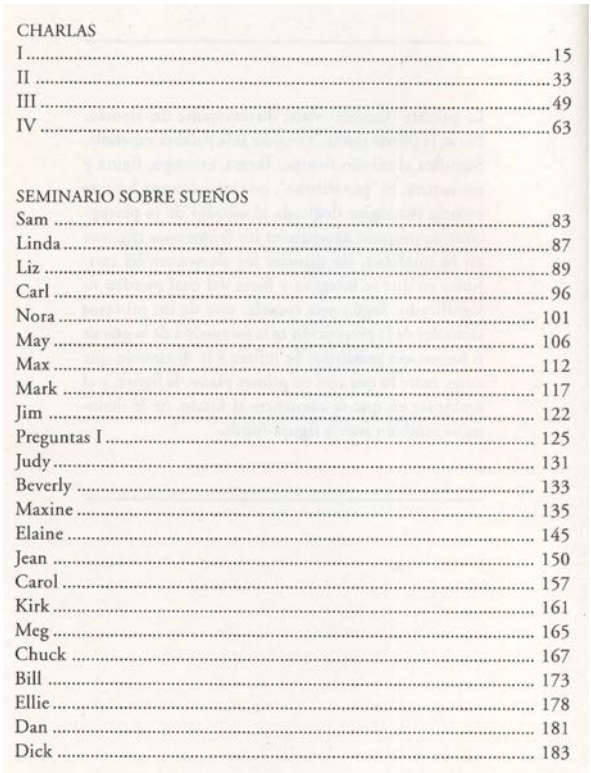
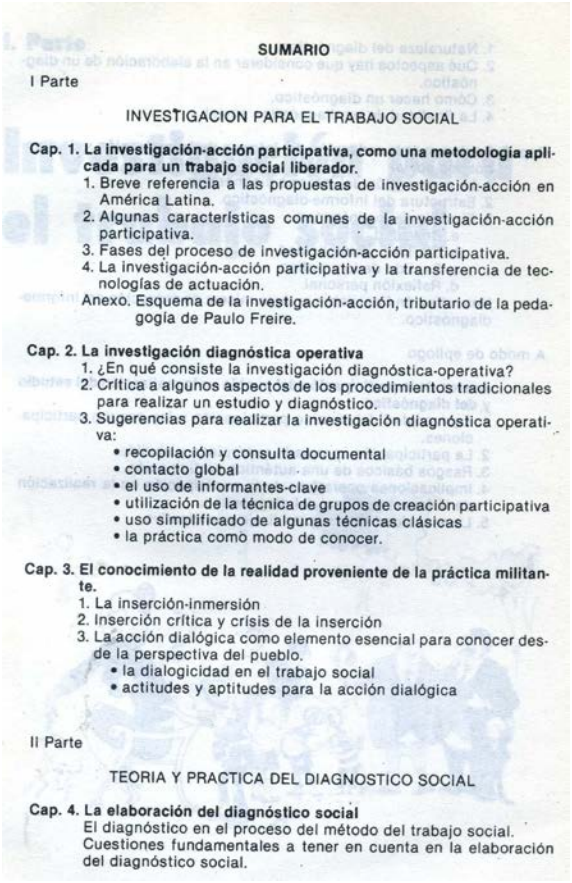


Figura 2. Ejemplo de tablas de contenido con títulos de capítulos usando nombres personales.

Hubo otros casos en que sí había bastante información contenida y útil para responder consultas pero lamentablemente sin páginas asociadas (ver Fig.3). Es decir, en el diseño del libro en sí no se incluyeron las páginas en la tabla de contenido, aspecto fundamental para que un sumario pueda tener sentido. Por tanto tampoco fueron considerados estos casos.



SUMARIO

I Parte

INVESTIGACION PARA EL TRABAJO SOCIAL

Cap. 1. La investigación-acción participativa, como una metodología aplicada para un trabajo social liberador.
1. Breve referencia a las propuestas de investigación-acción en América Latina.
2. Algunas características comunes de la investigación-acción participativa.
3. Fases del proceso de investigación-acción participativa.
4. La investigación-acción participativa y la transferencia de tecnologías de actuación.

Anexo. Esquema de la investigación-acción, tributario de la pedagogía de Paulo Freire.

Cap. 2. La investigación diagnóstica operativa
1. ¿En qué consiste la investigación diagnóstica-operativa?
2. Crítica a algunos aspectos de los procedimientos tradicionales para realizar un estudio y diagnóstico.
3. Sugerencias para realizar la investigación diagnóstica operativa:
   - recopilación y consulta documental
   - contacto global
   - el uso de informantes-clave
   - utilización de la técnica de grupos de creación participativa
   - uso simplificado de algunas técnicas clásicas
   - la práctica como modo de conocer.

Cap. 3. El conocimiento de la realidad proveniente de la práctica militante.
1. La inserción-inmersión
2. Inserción crítica y crisis de la inserción
3. La acción dialógica como elemento esencial para conocer desde la perspectiva del pueblo.
   - la dialogicidad en el trabajo social
   - actitudes y aptitudes para la acción dialógica

II Parte

TEORIA Y PRACTICA DEL DIAGNOSTICO SOCIAL

Cap. 4. La elaboración del diagnóstico social
El diagnóstico en el proceso del método del trabajo social.
Cuestiones fundamentales a tener en cuenta en la elaboración del diagnóstico social.

Figura 3. Ejemplo de tablas de contenido sin páginas asociadas.

Para efectos de esta tarea se usó en principio un escáner Epson Perfection 3490 y posteriormente se adquirió un escáner Fujitsu Scansnap SV600 con el que se terminó de trabajar la mayoría de los textos.

IX.2 Preprocesamiento

Antes de realizar las consultas se realizó esta fase previa que consiste en depurar la información transformada desde imagen a texto. Aquello implica revisión, corrección, eliminación o edición de una serie de detalles con tal de contar con los términos lo más correctos y normalizados posibles para tratar de dejar el mínimo de margen de error a las eventuales consultas que se realizarán. Dicha etapa contempló lo siguiente:

IX.2.1 Conversión de imágenes a archivos de texto

Una vez que se obtuvo la serie de documentos en pdf producto del trabajo de digitalización, se procedió a convertir esos documentos pdf a documentos de texto. Para ello se usó un convertidor on line llamado online-convert: (http://documento.online-convert.com). Esto permitió en gran medida transformar toda la información a un archivo txt aunque no estuvo libre de algunas dificultades. Con el segundo escáner la tarea fue más sencilla dada las características técnicas del equipo, mejorando la precisión significativamente.

Aun así igual hubo casos en que se tuvieron que intervenir textos manualmente para editar y corregir errores producto que la conversión no fue bien asimilada. Principalmente la diagramación del texto original, como textos con imágenes en las tablas de contenido fueron complejas de convertir; o la tipografía usada también (las cursivas y negritas no siempre fueron bien asimiladas). Implica un trabajo permanente de edición para ordenar la disposición de los textos cuando la conversión no se realiza tal cual figura el orden original diseñado.

Hay otros casos que se identifican en esta fase que presentaron algún tipo de problemas, a saber:

No reconocimiento de caracteres (lo más común, por ejemplo por dificultades para abrir un libro para digitalizarlo se tiende a distorsionar la imagen y afecta la conversión por estar las letras muy pegadas una al lado de la otra), que haya una mancha en la página digitalizada puede convertirla en un caracter inexistente, separación de letras producto de la conversión, palabras cortadas, palabras entrecortadas, no figuración de la página referenciada en un índice o sumario por un tema de diagramación original, palabras que quedan unidas al transformarse el documento a txt, entre otras.

Esta fase retarda la ejecución y puede considerarse como la primera y una de las mayores dificultades con que se encuentra el trabajo de procesamiento de textos.

IX.2.2 Corrección de textos

Obtenida la serie de documentos de textos producto de la conversión se procedió a revisar y corregir todos los casos que lo requiriesen para tener archivos coherentes, legibles y libres de errores. Esta etapa es sumamente necesaria para poder trabajar con información limpia y que se encuentre bien ingresada en los archivos a fin de que pueda recuperarse al hacer las consultas y sea visible en forma correcta. De otra forma varios términos pudieran quedar excluidos por no estar expresados en la forma que corresponde. Por lo mismo hubo que corregir varios que presentaban errores y uniformar términos, como por ejemplo poner acento gráfico a aquellos que no lo tenían para dejar la palabra de una manera ingresada y no dos o más.

En algunos casos hubo que sencillamente digitar líneas completas del texto original, como se mencionaba en el punto anterior. Por tanto el trabajo en esta fase fue directo con el texto original en mano la mayoría de las veces. Es una fase que demanda tiempo y exactitud. Se abordaron todas las fallas mencionadas en el punto anterior para dejar corregidos los textos.

IX.2.3 Tratamiento de palabras con particularidades

Se generó un archivo para detectar aquellos casos de palabras que incluyesen guion o bien se encontrasen truncadas por guion de continuación en otra línea. Se remediaron en este último sentido pero en el primero no fue necesario ya que en Python igualmente se pudo identificarlas y recuperarlas.

IX.2.4 Configuración actualizada de la colección de trabajo

La colección de textos de Psicología Transpersonal se aunó en un archivo de texto por cada libro, es decir a diferencia de lo anterior, se juntaron en un solo archivo de texto las diferentes partes que componían cada texto, dependiendo de la cantidad de páginas digitalizadas. Cada archivo fue identificado con el

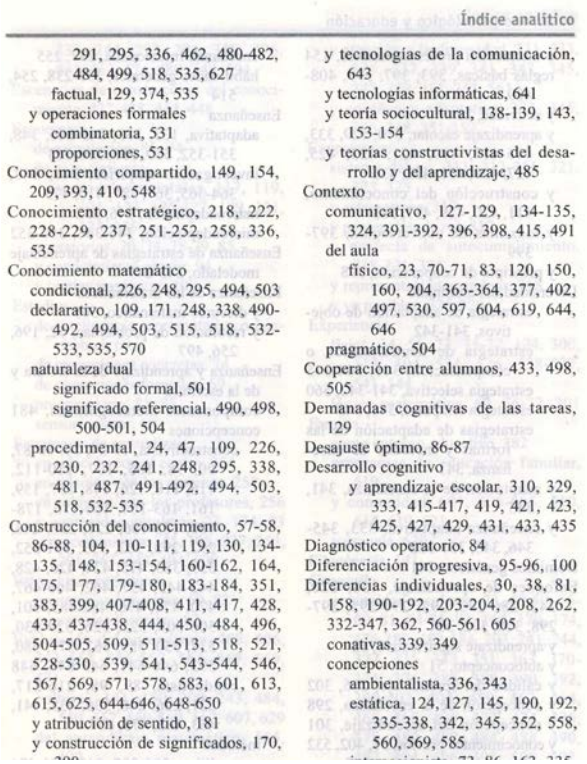
nombre de la autoridad (autor o palabra inicial del título, en caso de más de tres autores). En caso de textos con autor repetidos se optó por incluir un carácter diferenciador como guion. Si hablamos de 65 libros en la colección digitalizada de Psicología Transpersonal, pues a partir de este procedimiento hay 65 archivos de texto para trabajarlos.

Esto benefició en poder -ante una consulta- remitir al usuario a un solo documento identificado e ingresado al sistema, y no a varios archivos fragmentados del mismo. Por ejemplo supongamos que el término “histeria” sólo aparece en un libro, cuyo autor es Freud y se llama “Obras”. El resultado entonces sólo sería un archivo (que comprende toda la tabla de contenido) y no varios documentos generados del mismo libro (es decir, un archivo por página de la tabla de contenido digitalizada). Además contribuye a poder estimar con mayor precisión qué libro en su totalidad resulta más relevante a la consulta formulada, dado que el cálculo se va a producir sobre los documentos en concreto que semánticamente presenten más presencias del término consultado. O bien también al grado de importancia que tenga la aparición de un término dentro del documento. Por ejemplo: al buscar “alcoholismo” el término pudiera figurar como mención dentro de un documento en un contexto de violencia intrafamiliar, o bien como un capítulo completo dedicado al tema, lo que indica mayor importancia dado que puede abordar muchas más aristas.

Cabe mencionar que la actualización de la colección implicaría necesariamente incorporar nuevos documentos y la posible eliminación de otros. Esta situación repercute en la generación del índice de la colección, ya que la idea es que no se genere con cada consulta - como acontece en el modo de prueba usado - sino una vez cada cierto tiempo se requiera la incorporación de un conjunto de nuevos documentos. En este caso eso será dado por la modificación de bibliografías de la carrera, asunto que se da generalmente por la renovación de planes de estudios, aunque no es imperativo. Pueden incorporarse modificaciones mínimas en la colección pero igualmente relevantes de incluir.

IX.2.5 Identificación de casos complejos en documentos originales

Se ejemplificaron casos particulares de tablas de contenido e índices con numeraciones complejas (ver Fig.6) o uso de sangrías que afectan la visualización de los números de páginas o la recuperación de hipónimos (ver Fig. 4 y 5). Esto con el fin de poder abordar de alguna manera estas problemáticas.



Índice analítico

Figura 4. Ejemplo de índice con términos con muchas páginas asociadas donde figuran.

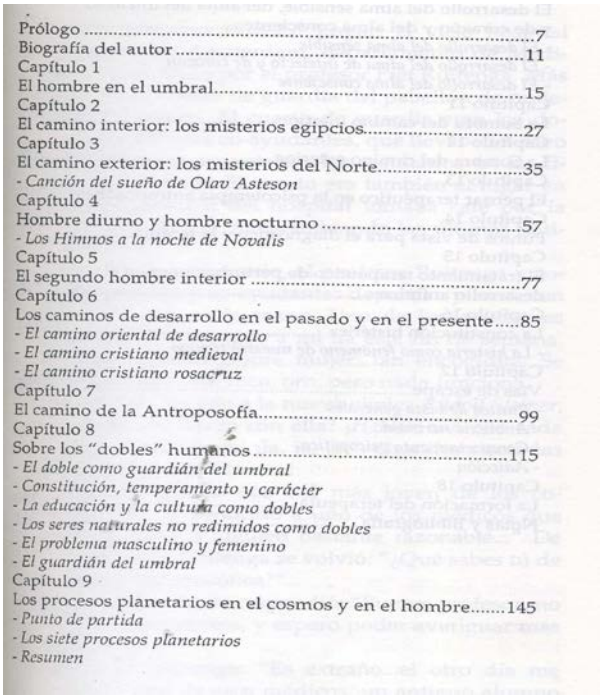


Figura 5. Ejemplo de tabla de contenido con subcapítulos sin numeración asociada.

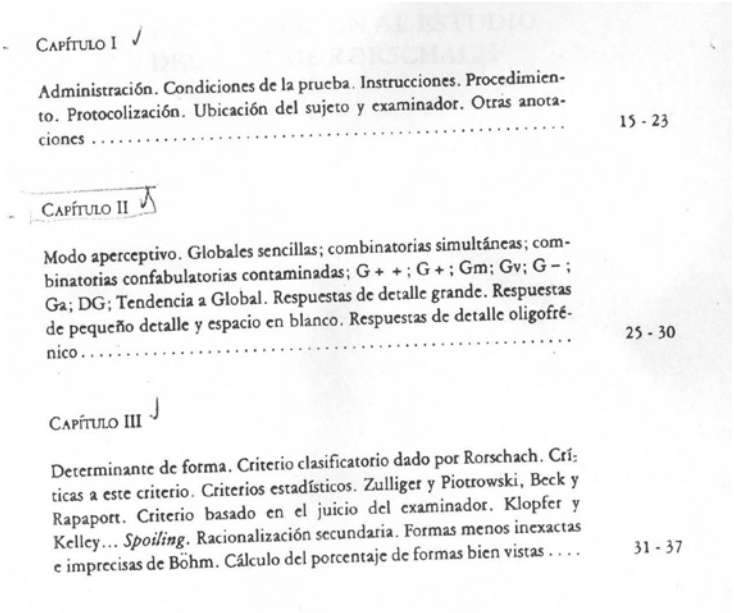


Figura 6. Ejemplo de tabla de contenido con rangos de números por capítulos.

IX.2.6 Generación de vocabulario

Una vez ya corregidos preliminarmente los textos en cuanto a coherencia y orden en el proceso de conversión de PDF a archivo de texto y sumado a esto la detección de otro tipo de situaciones complejas (como existencia de términos truncados sin guion), se procedió a generar un vocabulario con exclusión de lo que se denomina clase de palabras funcionales. *** Esta lista incluye todas las palabras que pudiesen tener alguna significancia en la respuesta a consultas. Es lo que para estos efectos van a constituirse como tokens. ****

El procesamiento de texto implica segmentación en palabras, oraciones y frases con vocablos y signos de puntuación. Esta fase se produce antes del procesamiento de texto propiamente tal.

En el presente trabajo, se ordenaron alfabéticamente los tokens en un archivo Excel y se corrigieron los que presentaban fallas ortográficas producto de la conversión. Casos como términos que presentaban leves diferencias como acentos, se normalizaron para no crear duplicados y dejar una sola forma aceptada. También aquellos casos en que se truncaron los términos se revisaron y completaron. Se reitera que todas estas fases de corrección demandan una exhaustiva revisión y correcciones asociadas.

*** *Se refiere a palabras no relevantes en una búsqueda con alta presencia en los textos. Carecen de significado por sí solas, cumplen una función gramatical en el texto. Incluye artículos, preposiciones, conjunciones.*

**** *Serie de caracteres que tienen un significado. Puede tratarse de palabras, signos, números que se representan unívocamente.*

IX.2.7 Determinación de heurísticas

Entendiendo las heurísticas como el conjunto de técnicas o métodos para resolver un problema, se puntualizaron cuáles fueron las salidas planteadas para resolver los inconvenientes que se fueron presentando a medida que se desarrolló el trabajo de digitalizar las tablas de contenido e índices cuyo fin es poder recuperar información exhaustivamente en documentos procesados. En esa fase de procesamiento se produjo una serie de inconvenientes que afectarían el rendimiento y los resultados del buscador.

A continuación se mencionan las medidas tomadas para optimizar las vistas de resultados con ánimo de favorecer la comprensión:

- Eliminación de líneas de puntos y espacios en blanco entre tema y número(s) de página(s) en forma automática.
- Ampliación de caracteres para resolver problemas de visualización incompleta por saltos de líneas en texto original. En el caso de asociación entre un término consultado que figure como hipónimo presente en el índice y su correspondiente hipónimo, fue posible mostrar el vínculo sólo cuando la distancia entre las palabras era menor ya que al existir mayor cantidad de términos entre el término consultado y el hiperónimo no se alcanza a visualizar, ya que implicaría mostrar textos demasiado extensos que afectan negativamente en la claridad del despliegue de la respuesta en la interfaz.
- Lista de tokens con guion para completar palabras que se encontraban truncadas por continuación en otra línea del texto original, de manera de regularizar estos términos.
- Reunión de archivos de texto en un solo documento extenso para poder contener las diferentes partes íntegras de un libro digitalizado a fin de evitar problemas de continuidad de frases.
- Edición de textos que presentaban problemas de diagramación, de tipografía, de inconsistencia (palabras cortadas, palabras pegadas, palabras entrecortadas) corrigiendo caso por caso los errores presentes en cada archivo, si es que los tuviesen.

- El principal problema que aconteció fue que varias tablas de contenido e índices no figuraban con las páginas asociadas correspondientes por diagramaciones originales que se distorsionaron al hacer la conversión de la imagen pdf a documento de texto. De ahí que se tuvo que completar esos datos cuando así aconteció.

IX.3 Herramientas computacionales

Para realizar el proceso de recuperación de información en Python, se tuvo que trabajar con diferentes herramientas que permitieron la optimización de resultados. Se detallan a continuación cuáles fueron utilizadas, comenzando por las dos librerías que se evaluaron y usaron para la formulación de consultas, donde se hizo una investigación documentada sobre sus características, funciones y potencialidades a fin de generar una breve descripción de ambas. Hacia el final se mencionan otros recursos adicionales implicados en el pre-procesamiento también.

IX.3.1 Whoosh

Ya se mencionó anteriormente la descripción de esta librería.

IX.3.2 Gensim

Ya se mencionó anteriormente la descripción de esta librería.

IX.3.3 **NLTK** (Natural Language Toolkit)

Es una plataforma que permite trabajar con datos de lenguaje natural. Es de código abierto y comunitario. Permite contar con interfaces sencillas para hacer uso sobre los 50 corpus y recursos léxicos. De esta manera se puede contar con la posibilidad de procesar textos para clasificación, creación de tokens,

stemming, etiquetados, análisis semántico, entre otras. Se encuentra disponible para Windows, Mac OS X y Linux.

Se parte haciendo un llamado a esta librería en el cuadro de comando de Python para empezar el proceso de trabajo con los documentos de la colección elegida.

IX.3.4 Stopword (Listas de paro)

También llamadas listas de palabras vacías; se refiere a la lista que incluye artículos, conjunciones, adverbios, algunos verbos, preposiciones y pronombres. Son todas las palabras que no revisten significancia en las búsquedas. El filtro de ellas se realiza antes o después del procesamiento de texto, aunque en este caso se realizó antes de la obtención del vocabulario. Quedan fuera de los procesos de indización de términos.

En Python viene integrada esta lista y se debe dar la instrucción en el comando para activarla.

De todas formas no siempre deben aplicarse en totalidad pues puede que para ciertas temáticas sea mejor incorporarlas, como por ejemplo en análisis literario.

IX.3.5 Stemming

Se refiere a reducir una palabra a su raíz (stem). Si bien permite ampliar los resultados en una búsqueda en proceso de recuperación de información ya que incluye varias palabras que parten del mismo origen morfológico, por otra parte afecta la precisión al hacerse más dispersa la búsqueda.

El stemming se realiza a través de la aplicación de algoritmos especializados. Python al llamar a nltk funciona con el algoritmo de Porter o con Lancaster (que sólo eliminan sufijos). También trabaja con el lenguaje Snowball.

Permite reducir el tamaño de los índices ya que elimina los afijos del término que queda unitario como stem.

IX.3.6 Grep

Herramienta original para Unix usada dentro de línea de comandos. Significaría “Global Regular Expression Print” (imprimir expresión regular global).

Grep toma una palabra o expresión regular en forma literal, lee la colección de archivos, y muestra el resultado con las líneas que tengan coincidencias para la palabra o expresión consultada. Como la recuperación la realiza por una cadena de caracteres, si se busca por ejemplo el término “autismo” también va a recuperar “bautismo”.

A modo de sólo mención se usaron otras herramientas como full convert (para convertir la imagen digitalizada a texto) y las funciones de tokenizador y ampliación de consultas incluidas en Python.

X. RESULTADOS

Luego de un proceso de evaluación de las tres opciones propuestas, se optó por elegir una de ella que pudiese dar respuesta a la instancia de realizar búsquedas en texto completo para poder recuperar términos incluidos en el conjunto de documentos que conforman la colección procesada.

En el caso de Omeka, se descartó este programa ya que no contaba con un buscador integrado de texto libre por tanto hubo que dar con un pluggin que pudiese añadírsele para efectuar esa función. Sin embargo, pese a que se encontró un plugging creado dentro de la comunidad de usuarios, no dio resultados al probarlo.

En el caso de Dspace, es un programa que pudiese dar resultados al transformar los metadatos MARC de los registros bibliográficos en metadatos Dublin Core. Sin embargo el desarrollo del repositorio donde alojar los documentos digitales toma tiempo dada la envergadura técnica que reviste crearlo, sumado al conocimiento experto que se requiere para poder generar los enlaces necesarios entre lo que se haga en el repositorio y los registros existentes en el software Horizonte, en caso de que se trabaje para una vinculación entre ambos sistemas a futuro.

Finalmente se optó por elegir las librerías de Python, dado que las condiciones técnicas son mucho más favorables para el desarrollo de un prototipo de herramienta que se desea diseñar. Presenta una opción más inmediata y accesible de implementar, usando documentos de texto para procesarlos con las diversas funciones que incluyen estas librerías. El hecho de estar descargables en código abierto, de contar con manual de ayuda en línea y de funcionar con el lenguaje Python, que se trabajó en la asignatura “Recuperación de Información” del magíster, contribuyó a elegir esta opción. En la evaluación de los pro y contras, son muchas más las ventajas que tiene versus las escasas desventajas.

Con el propósito de poder indizar la información plasmada en las tablas de contenido e índices de los libros, se pretende que los modelos de recuperación de información usados en Python entreguen ventajas comparativas frente al modelo tradicional booleano.

A continuación se da cuenta de los resultados obtenidos usando las librerías de Python, Whoosh y Gensim. En ambos casos se realizaron las mismas consultas con y sin aplicación del extractor de raíces, para ver las posibles significancias en los resultados que implica el uso de esa función. El objetivo del uso de extractores de raíces es recuperar más resultados que pudiesen ser relevantes dado que tienen el mismo origen de la palabra. De este modo se amplían las opciones de poder encontrar documentos que respondan a la consulta, aunque puede repercutir en la precisión. El desglose que se ve a continuación, está en el orden secuencial en que se ejecutaron las diferentes etapas.

X.1 Ejecución de consultas con Whoosh

Las consultas simples o individuales (por un término) realizadas fueron las siguientes:

- fobias
- Pavlov
- vivienda
- Husserl
- kundalini
- prejuicios
- acupresión
- incineración
- enmascaramiento
- redacción
- purificación
- histérica
- autotélico
- esquizoide
- bipolar
- rostros
- vulnerabilidad
- simetría
- leucina
- sadismo

Se escogieron los términos al azar teniendo algunas consideraciones. A saber: incluyendo palabras en singular y plural, nombres propios, palabras con acento gráfico. Se constató por supuesto a priori la existencia de estos términos en los documentos de la colección procesada. La primera prueba se realizó utilizando la librería Whoosh y este es un ejemplo de los resultados (ver Fig.7), donde se da cuenta del nombre del archivo (en este caso los apellidos de autores) y las líneas textuales donde sale el término con su(s) correspondiente(s) página(s):

Figura 7. Resultados con Whoosh sin raíz para consulta simple por un término.

X.2 Detección de fallas en archivos

Pese a haberse hecho una corrección preliminar de los textos, se tomaron de la colección digitalizada veinte documentos al azar para detección de fallas y constatar qué tipo de errores se encontraban presentes, para ir corrigiendo la bolsa de palabras a trabajar. Entre estos errores estaban los siguientes que perjudicaban la visualización correcta de los resultados:

* Casos donde no venían los números de página del sumario y/o índices asociados al término consultado.
* Casos en que la página referenciada no correspondía al término consultado, sino al del término anterior que figuraba en el documento original.
* Casos en que había presencias del término dentro de determinado documento que sencillamente no eran detectadas, o que salían mencionadas algunas páginas referenciadas pero no la totalidad existente;
* Casos en que el uso de acentos afectaba la recuperación en algunas veces.
* Casos en que figuraban las palabras truncadas por guiones.

La corrección es un proceso reiterativo que demanda este tipo de operaciones, por tanto no basta una sola revisión, dado que la cantidad de textos y situaciones particulares es amplia.
Debido a esta serie de problemas hubo que revisar las causas para mejorar los resultados de las consultas y el desempeño de ambas librerías.

X.3 Funciones de ampliación de consultas

Dentro de las funciones que trae Whoosh, se buscaron unas que permitiesen ampliar las consultas por frases exactas, con la finalidad de lograr mayor precisión en las respuestas que arroja la herramienta al buscar por más de un solo término. Estas funciones que no vienen por defecto son:
Whoosh.fields.ngram (que incluye los espacios y signos)
Whoosh.fields.ngramwords (sólo considera palabras)
Además se encontró una función para buscar una frase exacta que va entre comillas con la función: **Whoosh.qparser.SequencePlugin**.
También se indagó para poder contar con una función que permitiese ampliar la cantidad de caracteres en la visualización de los resultados, a fin de optimizar la comprensión de la información desplegada. La función es:
results=mysearcher.search(myquery)
results.fragmenter.charlimit=100000

Estas funciones se identificaron para una futura utilidad hasta ese momento. Posteriormente se usarían algunas implementándolas dentro de un script.

X.4 Consultas compuestas

Ya realizadas las consultas simples, se procedió a ejecutar las consultas con palabras compuestas o frases para ver el comportamiento de la herramientas en estos casos. Whoosh por defecto incluye el operador booleano AND al ingresar dos o más términos.

Se observaron problemas similares a los de las consultas simples, con el adicional que se presenta la dificultad de relacionar los términos ingresados con los hiperónimos (término genérico) e hipónimos (sus derivaciones o acepciones). Debido a que un término, como por ejemplo “comunicación no verbal” figura como “comunicación” como hiperónimo y en un desglose a continuación sale “no verbal” como hipónimo, es difícil así de recuperar, a no ser que la extensión de caracteres a visualizar en pantalla sea muy amplia, cosa que tampoco es muy amistosa de ver en la interfaz de las respuestas.

También surgieron casos de respuestas en que no figuraba el número de la página asociada, o que sólo se recuperaba un solo término, o más de uno dentro de mismo documento pero por separado, lo que hacía perder relevancia, puesto que se perdía el sentido de coherencia de la frase original.

En Whoosh para ejecutar búsquedas en el índice se requiere de un Objeto Buscador.

Se ejecutan consultas directamente o bien se usa un analizador de consultas. El que viene por defecto es el QueryParser.

QueryParser funciona con un tipo de lenguaje de consulta similar a la API Lucene. Enlaza términos con los operadores AND y OR, descarta a los que le indica el NOT, agrupa a los que están entre paréntesis, y permite el uso de comodines y prefijos. El analizador tokeniza los términos del documento y luego los filtra, dejando afuera las palabras de la lista de paro y normaliza a minúscula los tokens.

Las consultas compuestas (por dos o más términos) o por frases realizadas fueron las siguientes:

- Acto fallido
- Hemisferios cerebrales
- Anorexia nerviosa
- Ritmos circadianos
- Aprendizaje por descubrimiento
- Teoría de Kohlberg
- Comunicación no verbal
- Cognitivo-conductual
- Abuso de sustancias
- Grupos de discusión

También se escogieron al azar teniendo algunas consideraciones, a saber: incluyendo palabras en singular y plural, nombres propios, palabras con acento gráfico, con guion intermedio, y con más de dos palabras. Se constató por supuesto a priori la existencia de estos términos en los documentos de la colección procesada.

Al no aplicarse la lista de paro (stoplist) hubo que reformular las consultas, omitiendo incluir preposiciones ya que recuperaba partes del documento donde sólo figuraban estas y por tanto carecían de relevancia, generando "ruido" en la respuesta. En concreto los casos de "aprendizaje por descubrimiento", "abuso de sustancias" y "grupos de discusión".

En casos como "comunicación no verbal" se ingresó como "comunicación verbal" porque incluir el "no" implicaba aumentar en demasía la cantidad de resultados dado que recuperaba muchos documentos donde figuraba el "no" más la presencia de los otros dos términos, todos por separado.

X.5 Consultas con el operador booleano AND

En adición a lo anterior, se realizaron diez consultas usando expresamente en la consulta el operador AND para encontrar resultados donde figuren ambos términos. En la mayoría de los casos hubo presencias por separado de los términos y pocas veces aparecieron asociados, sin embargo igualmente pueden revestir una relevancia preliminar a constatar posteriormente con el texto original en mano. Fueron 10 consultas en total, de las cuales 8 resultaron positivas y 2 negativas.

Los términos consultados fueron los siguientes:

- Bullying AND prevención
- Cannon-Bard AND Stanford-Binet
- CIE AND 10
- DSM-IV AND CIE-10
- Especies AND Darwin
- Ética AND deontología
- Información AND GAP
- Insight AND creatividad
- Madre AND hijo
- Stanford AND Binet

Dentro de las positivas resolvió bien consultas con términos en otro idioma (bullying AND prevención, insight AND creatividad), entre apellidos (Stanford AND Binet), y uso de siglas (información AND GAP, CIE AND 10).
En las respuestas negativas, hubo dificultades con el uso de guiones en términos compuestos (Cannon-Bard AND Stanford-Binet, DSM-IV AND CIE-10) que hizo que entregasen resultados de documentos pero en blanco. (Sólo se indica el nombre de los documentos, sin fila ni número).

X.6 Consultas con el operador booleano OR

Se ejecutaron algunas consultas simples de prueba, con la opción booleana de alternativa “OR” para ver el desempeño de la librería Whoosh en estos casos. Fueron 10 consultas en total, de las cuales 7 resultaron positivas y 3 negativas.
Los términos consultados fueron los siguientes:

- ADN OR DNA
- Anorexia OR bulimia
- Autismo OR autism
- DSM-IV OR CIE-10
- EC OR EI
- Jueces OR jurado
- Nosotros OR ello
- Rolland OR Jones
- Transmutación OR sublimación
- Traqueidas OR traqueofitas

Dentro de las positivas cabe resaltar que resolvió bien cuando la consulta se trató sobre acrónimos (ADN OR DNA), apellidos (Rolland OR Jones), palabras con acentos (transmutación OR sublimación) y palabras que en el documento original llevaban un signo de interrogación (¿jueces o miembros del jurado? - que logró recuperar al buscarlas aunque sólo por una de ellas "jurado").

En cambio, los resultados negativos se produjeron en los casos de término en dos idiomas (autismo OR autism), en palabras con guion intermedio (DSM-IV OR CIE-10) y en términos que figuraban entre comillas en el documento original ("nosotros", "ello").

En el caso de los términos en dos idiomas ni siquiera arrojó resultados, mientras que en los otros dos hubo resultados con los nombres de los textos pero sin figuración de las referencias al número de las páginas donde aparecen.

X.7 Consultas con extractor de raíces

Usando los mismos términos iniciales de las consultas simples y por frases, se ejecutaron las consultas aplicando esta vez una función de Python alojada en su módulo NLTK (SnowballStemmer), que opera como extractor de raíces para ampliar las posibilidades de respuesta (ver Fig.8). Esto significa que las búsquedas se realizan en todos los términos que tengan una raíz común. Esto aumenta la recuperación de documentos aunque no significa necesariamente que ayude a encontrar aquellos más relevantes pues el espectro se hace mayor y difuso, y las respuestas pierden en precisión.

Los problemas se volvieron a reiterar (figura el texto donde sale contenido el término pero sin página asociada) aunque esta vez la efectividad bajó notoriamente a un 17%.

Tras hacer algunas mejorías correctivas en los textos, los resultados comparativamente hablando arrojaron el mejor desempeño a Whoosh sin raíz para consultas simples por un término y un bajo desempeño del Whoosh con raíz en la obtención de resultados. Además se constata que es mejor incluir

términos en singular para mayor recuperación, pues incluye además los plurales del término consultado.

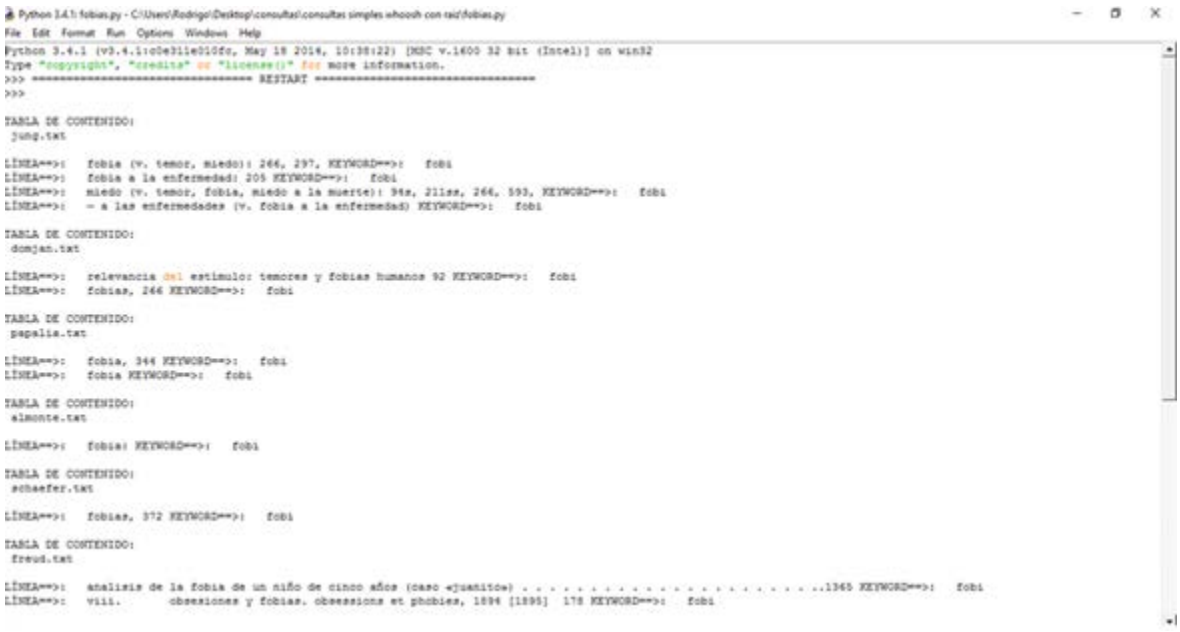


Figura 8. Resultados con Whoosh con raíz para consulta simple por un término.

X.8 Consultas por frases con sustantivos y preposiciones

Se realizaron consultas para frases que estaban compuestas de sustantivo+ preposición+sustantivo para ver el desempeño de la librería Gensim sin extractor de raíces.

En términos generales se destaca la altísima cobertura, pero muy baja precisión, lo que no resulta conveniente.

En la visualización de resultados en estos casos, no siempre el documento mejor ponderado es el más pertinente a la consulta. Los resultados serían más pertinentes si mostrasen niveles de proximidad entre los términos de la consulta, pues en la mayoría de ellos figuran por separado.

La cobertura alcanzó a un 100%, la precisión a un 18% y la medida F a un 30%.

Las consultas realizadas fueron las siguientes:

- “experiencias referidas” como práctica
- Unidad del yo
- Trastornos del sueño
- Movimiento de doble sentido
- Muerte, miedo a la
- Psicopatología del autismo
- Motivación de logro
- Alma según Platón
- Memoria a corto plazo
- Padres, influencia de los

En la primera, la idea era probar la recuperación para la expresión entre comillas. Resultó más eficiente usar sólo sustantivos y adjetivos en la consulta, omitiendo varios términos de la lista de paro (“de”, “del”, “a la”) para que los resultados sean más acotados y simples de revisar. De lo contrario se ampliaba demasiado la lista con mucha irrelevancia por la recuperación que se producía de sólo palabras de la lista de paro.

X.9 Uso de librería Gensim

Ya se mencionó anteriormente que Python cuenta con esta librería de búsqueda que tiene la particularidad de hacer una ponderación de los resultados de búsqueda en base al número de ocurrencias del término consultado dentro de una colección de documentos. Está diseñado para trabajar con grandes colecciones. Trabaja con algoritmo de espacio vectorial para producir los resultados más relevantes.

Se efectuaron las mismas consultas con Gensim para ver el rendimiento que presentaba y compararlo con el de Whoosh (ver Fig.9 y 10).

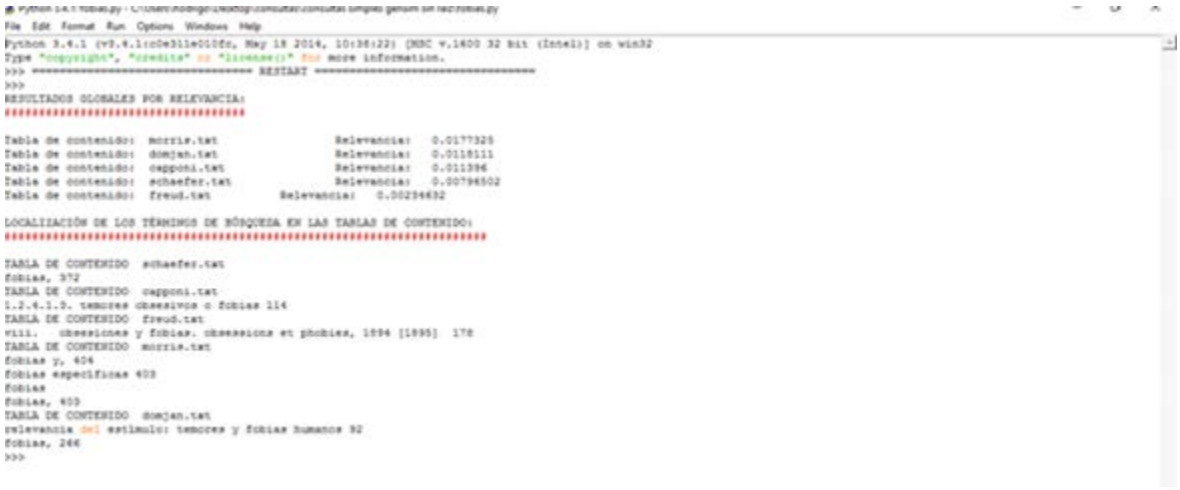


Figura 9. Resultados con Gensim sin raíz para consulta simple por un término.

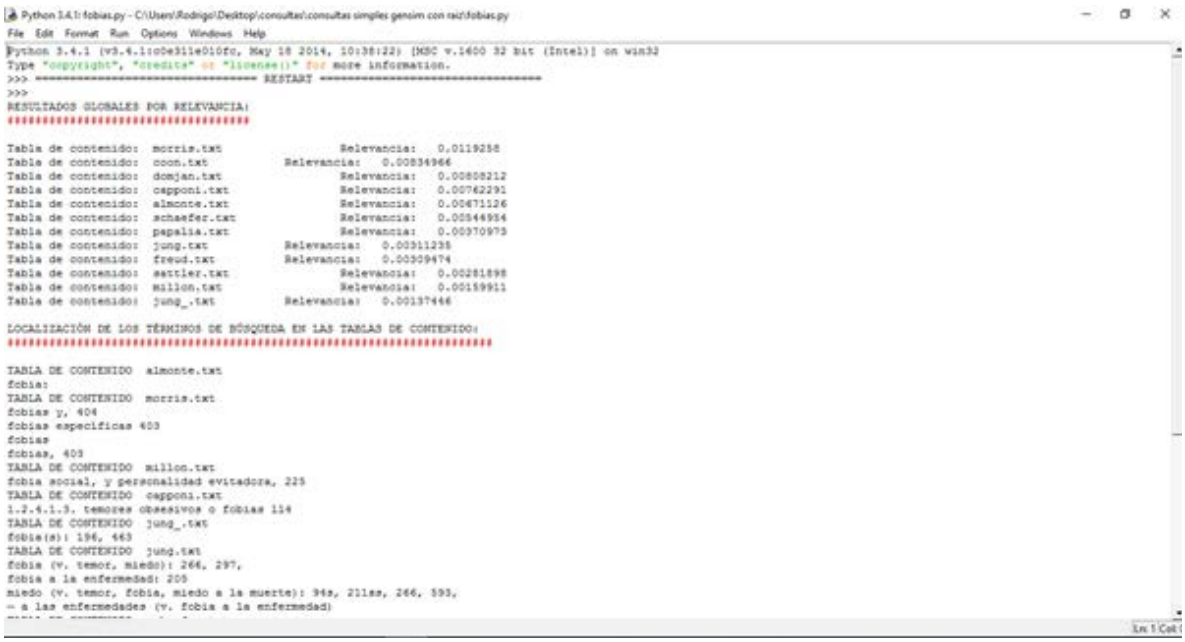


Figura 10. Resultados con Gensim con raíz para consulta simple por un término.

X.10 Corrección de visualizaciones de resultados

Inicialmente se utilizó un código o “script” en Python para poder eliminar líneas de puntos y espacios vacíos que estaban causando “ruido” en la visualización de resultados. Tanto en Whoosh como en Gensim se amplió la cantidad de caracteres a desplegar en pantalla por lo mismo, una expansión de las respuestas para ver más completas las líneas de texto tal como figuraba en el texto original. La idea es mostrar la existencia concreta del término con su página asociada y que fácilmente pudiera constatarse donde ubicarla dentro del libro original.

El principal problema que se detectó con esta modificación es la demora en los tiempos de respuesta en general, lo que repercute demasiado en la efectividad de la herramienta. Bueno, esto también depende de las características técnicas de rendimiento del computador donde se ejecute.

Sin embargo una vez obtenidos los resultados, mejoró el rendimiento en efectividad, recuperación e interfaz para ir aproximándose a una idea final de visualización.

X.11 Consultas por frases exactas usando comillas

Usando una función específica de Whoosh, se realizaron las mismas 10 consultas por frases pero esta vez en forma de búsqueda exacta usando la expresión entre comillas.

Los resultados en lo que respecta a precisión fueron óptimos, lográndose un 100%, y un 68% en cobertura (ver Fig.11). Esto se debe a que recupera los términos tal cual se expresan en la consulta pero por lo mismo deja fuera otros resultados que son relevantes pero que incluso por una letra de diferencia pueden quedar excluidos.

Es destacable que logra el mejor rendimiento en términos de precisión/cobertura con un 81% de efectividad.

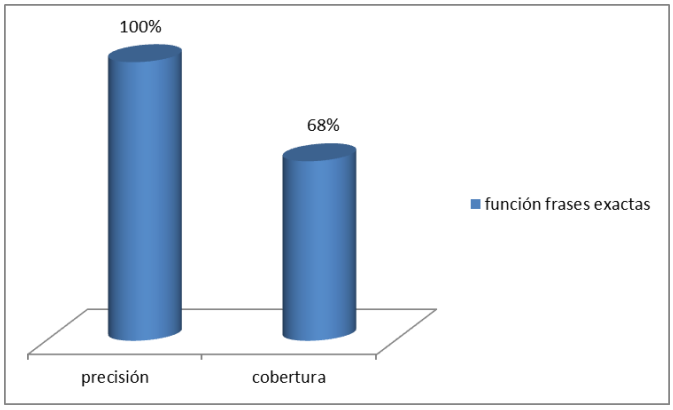


Figura 11. Rendimiento de Whoosh buscando frases exactas.

X.12 Uso de Grep

Para la constatación dentro de la colección, de la existencia de determinados términos expresados en las consultas, se usó Grep. (ver Fig.12)

La idea era constatar a priori el número de apariciones de determinado término y la figuración dentro de algunos documentos, para contrastarlo después con los resultados que arrojasen las consultas en Whoosh y Gensim, y determinar si la recuperación fue completa o parcial. La dificultad que tiene es que no trabaja con acentos gráficos, por tanto los términos que lo tenían hubo que buscar por la raíz.

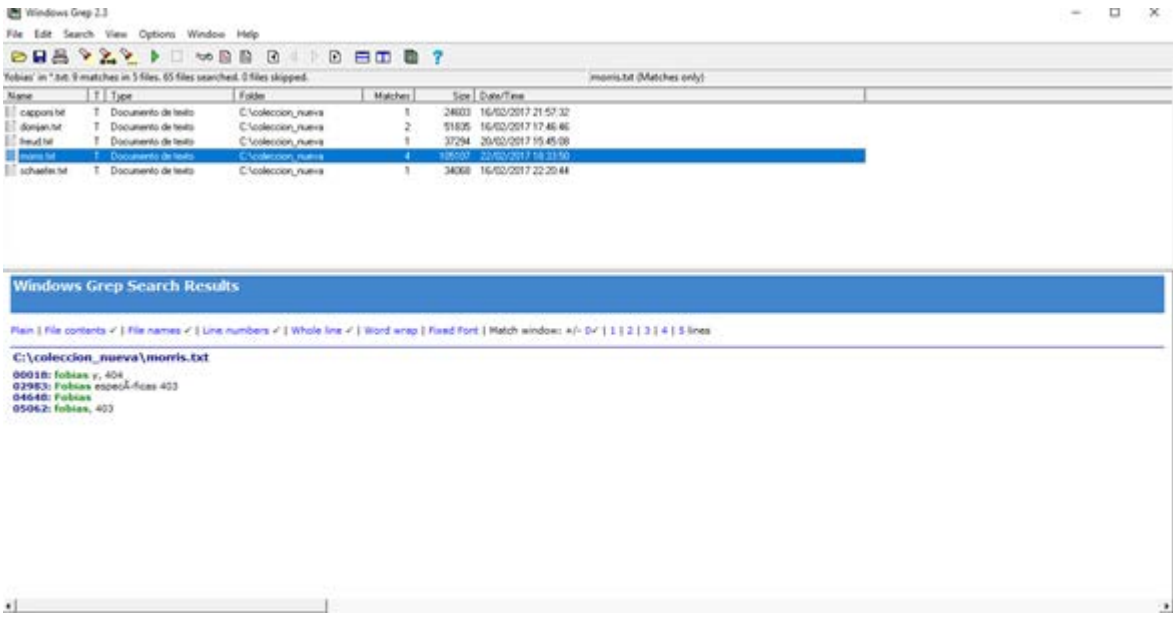


Figura 12. Resultados con Grep al buscar por un término.

XI. EVALUACIÓN

XI.1 Comparación de consultas entre ambas librerías

En esta sección se presentan los resultados de la comparación entre Whoosh y Gensim. Para ello se realizaron las mismas consultas simples y compuestas con y sin extractor de raíces en ambas librerías, para comparar desempeño en términos de cobertura y precisión.

Las consultas se realizaron con los códigos mejorados como se mencionó en puntos anteriores (ver Fig.13 a 16).

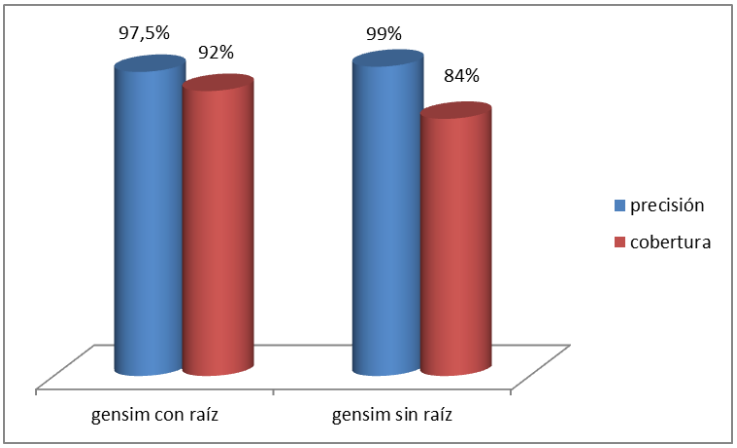


Figura 13. Rendimiento Gensim consultas simples.

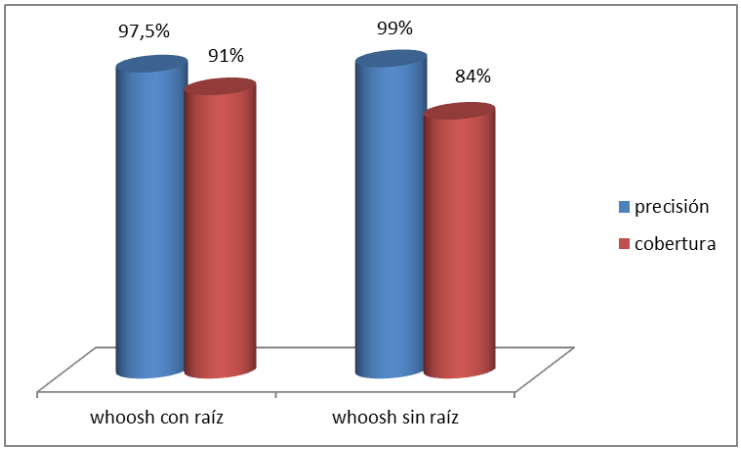


Figura 14. Rendimiento Whoosh consultas simples.

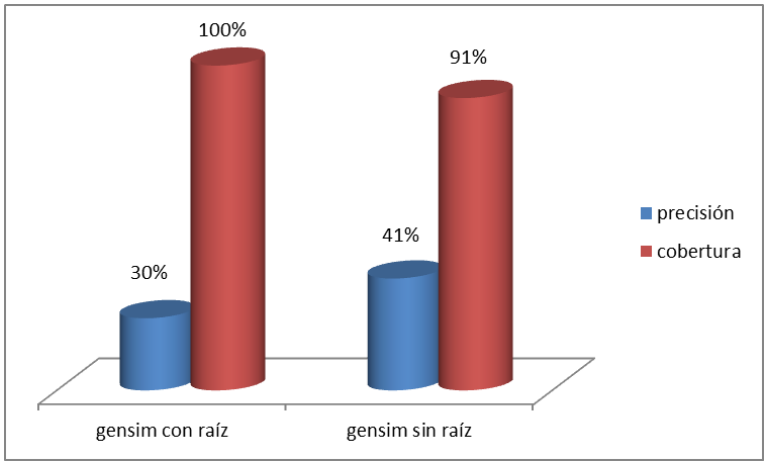


Figura 15. Rendimiento Gensim consultas por frases.

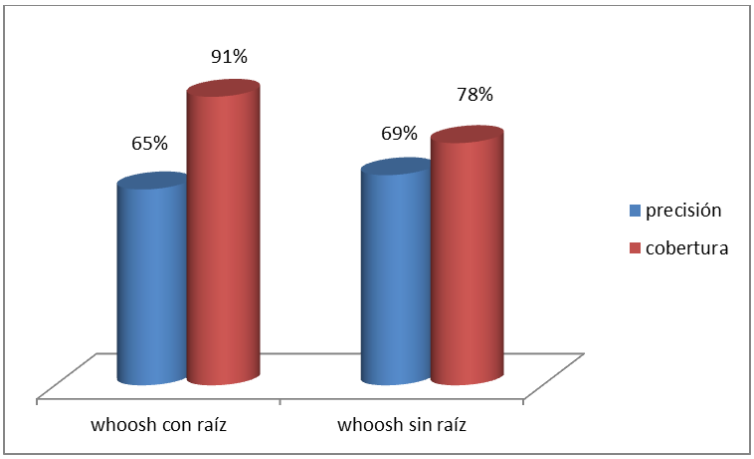


Figura 16. Rendimiento Whoosh consultas por frases.

XI.2 Ponderación de resultados con medida F

Una vez obtenidos los resultados de la comparación se aplicó un algoritmo que mide la exhaustividad (o cobertura) y la precisión para obtener un valor único, expresado en un rango de 0 a 1 (aunque más comúnmente expresado en porcentajes en una escala de 0 a 100). En estadística esto se conoce como **medida o valor F, que se define como la media armónica entre precisión y cobertura**, usada especialmente en la búsqueda y recuperación de documentos y la clasificación de documentos. Se pondera por igual ambos factores por ser

medidas de relevancia que indican cuán cerca del ideal de documento (cercana a 1, con alta precisión y cobertura) se encuentran los correspondientes aun conjunto dado.

El cálculo de la medida es:

$$F1 = 2 * \frac{precisión * cobertura}{precisión + cobertura}$$

En base a aquello e incluyendo consultas simples con y sin extractor de raíces en ambas librerías, se obtuvo nuevamente que el de mejor rendimiento fue Gensim con raíz con el mismo 95% para consultas simples aunque Whoosh con extractor de raíz apenas le sucede con un 94% de rendimiento. En este tipo de consultas la precisión es uniforme y la cobertura es mejor al incluir extractor de raíces. En segundo orden de efectividad, las consultas sin aplicación de extractor de raíces resultaron con un 91,5% en ambos casos. (ver Fig.17). En cuanto a las consultas con palabras compuestas, bajó el rendimiento en general, siendo Whoosh con extractor de raíz el de mejor desempeño con un 76%. En estas consultas se gana en cobertura pero se pierde en precisión. (ver Fig.18)

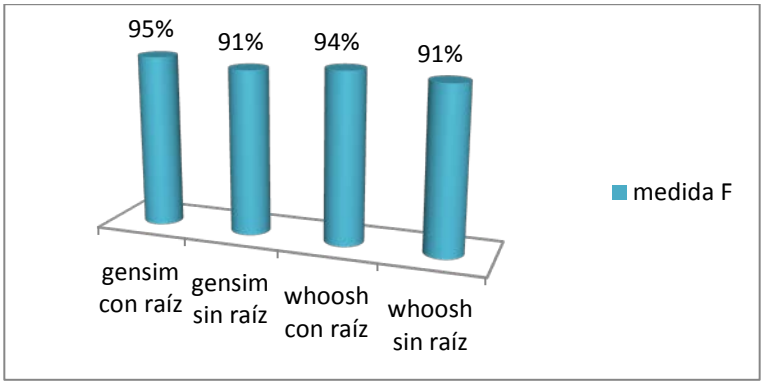


Figura 17. Comparativo precisión/cobertura con consultas simples.

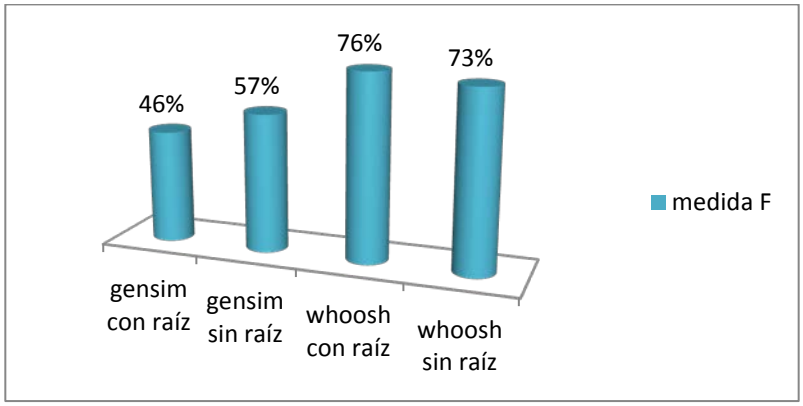


Figura 18. Comparativo precisión/cobertura con consultas por frases.

XI.3 Umbrales de resultados

Evaluando la cantidad de resultados entregados por cada librería con y sin extractor de raíces, se pueden determinar diferencias que dan cuenta del grado de "ruido" o irrelevancia que hay en las respuestas a las consultas expresadas.

Se hace notoria la gran cantidad de resultados que entrega Gensim al buscar por frases con extractor de raíces (222) y sin extractor de raíces (159), si lo comparamos con el que le sucede, que es Gensim con extractor de raíces para consultas simples (76).(ver Fig.19)

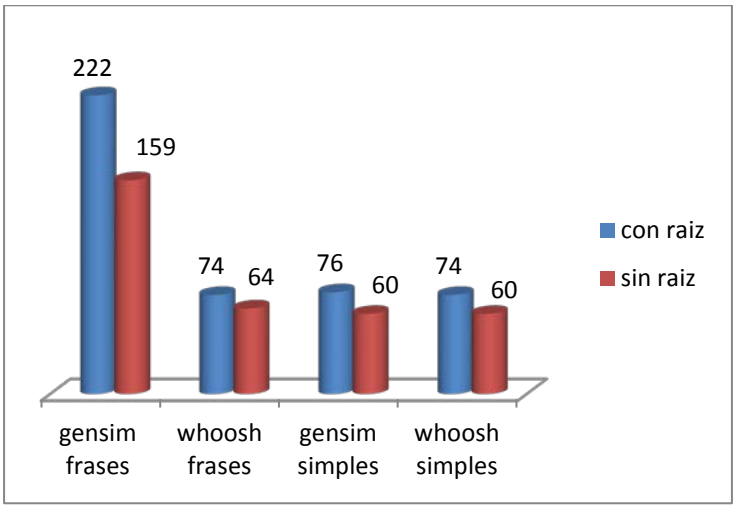


Figura 19. Comparativo en cantidad de resultados, según modo de búsquedas.

Si lo mismo se expresa considerando promedio de resultados, las respuestas de Gensim buscando frases, alcanza 22 con extractor de raíces y 16 sin extractor. (ver Fig.20)

Una gran cantidad de resultados repercute en la precisión de la respuesta pero amplía la cobertura.

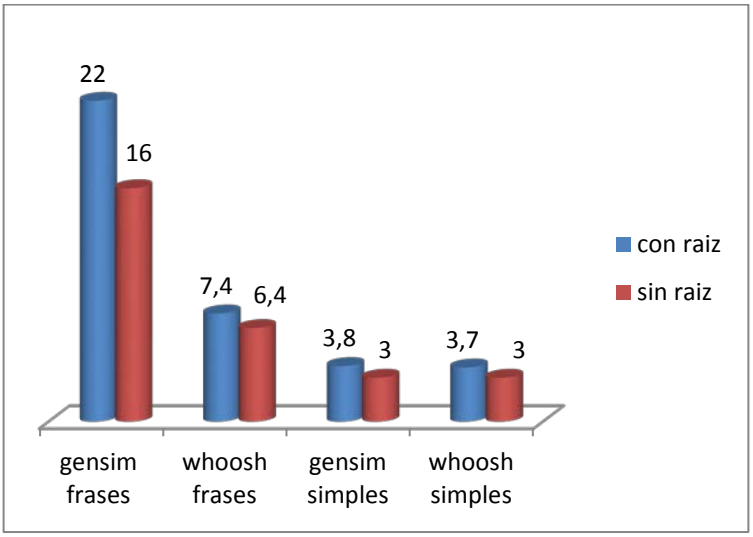


Figura 20. Comparativo en cantidad de resultados, según modo de búsquedas.

XI.4 Comparativo con sistema Horizonte

Teniendo presente que ante una consulta el sistema Horizonte no entrega los resultados basándose en lo expresado en las tablas de contenido e índices (que corresponde a vocabulario no controlado), sino en campos de autor, título, materias, notas, entre otros; pese a aquello, se quiso hacer una comparación a nivel de cantidad y aciertos en los resultados usando las mismas mediciones de cobertura y precisión. Son resultados complementarios en estricto rigor, pero resulta interesante evaluar la respuesta de los buscadores ante un mismo set de consultas y la ganancia en recuperación de información contrastando las

diferentes plataformas aunque, se reitera, sin igualdad de condiciones. Para ello se usó la opción de búsqueda por palabra clave general.

En las consultas simples sin extractor de raíces, Horizonte arrojó respuestas en sólo 11 casos de 20 (55%) y con extractor de raíces en 13 casos (65%). En las consultas por frase lo hizo sólo en 4 de 10 casos (40%) y con extractor de raíces aumenta a 6 casos (60%).

Si bien es cierto en las consultas simples arroja mayor cantidad de resultados con y sin extractor de raíces (ver Fig.21), muchos de ellos responden a que sólo entre dos términos consultados: “vivienda” y “redacción”, concentran 80 resultados de 101 (sin raíz) y 121 resultados de 170 (con raíz). Esto debido a que ambos términos también se usan como encabezamiento de materia, por tanto aparecen varios registros asociados como resultados.

Sin aquellos, y remitiendo los aciertos a la figuración de los términos consultados dentro del área de título principalmente, la diferencia se marca inmediatamente, quedando Horizonte relegado a una menor cantidad de resultados válidos, en comparación a lo que arrojan las librerías de Python.

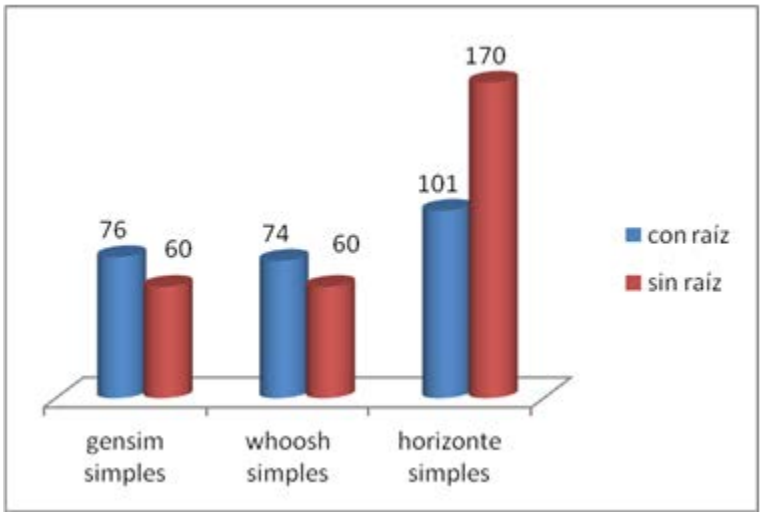


Figura 21. Cantidad de resultados por consultas simples según plataforma usada.

En el caso de las búsquedas por frases la situación varía. Horizonte disminuye significativamente la cantidad en comparación a ambas librerías de Python (ver Fig.22). La mayoría de los resultados se debe a que los términos forman parte del título principalmente. Disminuye la presencia de éstos como parte de los encabezamientos de materias asignados a los registros recuperados.

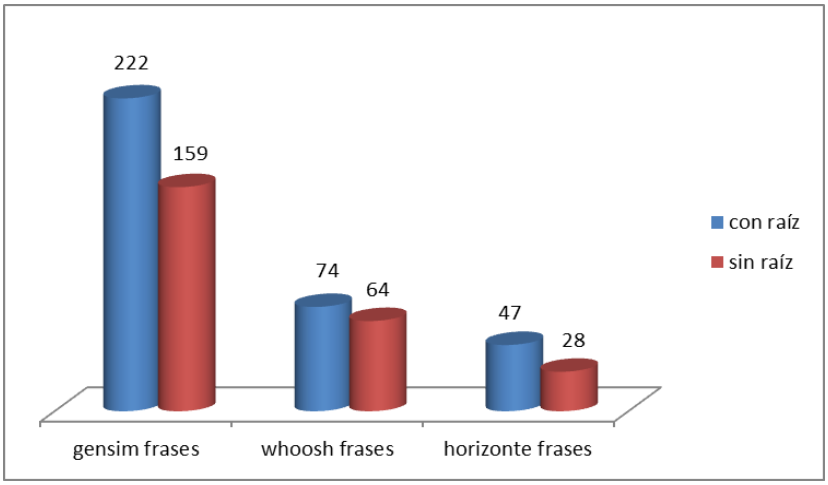


Figura 22. Cantidad de resultados por consultas por frases según plataforma usada.

Resumiendo los resultados anteriores, se plantea que la mejor opción a considerar en un primer prototipo de buscador es el modelo probabilístico de la librería Whoosh que haga uso de un extractor de raíces, ya que obtuvo el mejor desempeño en términos de precisión y cobertura en el comparativo final (ver tabla 3 y Fig.23). En consultas simples obtiene buena respuesta; principalmente en lo que respecta al desempeño en consultas por frases, es significativa la diferencia entre los modelos evaluados.

Además tiene la gran ventaja de realizar búsquedas por frases exactas con una precisión total, y un buen promedio con la medida F (81%) donde se pondera igual precisión y cobertura.

Las búsquedas en texto libre (en este caso en tablas de contenido y/o índices) amplían la posibilidad de recuperar información, al haber un mayor número de

términos disponibles en un sistema de búsqueda, y con el beneficio de aumentar la precisión.

Los modelos de recuperación de información utilizados (probabilístico y espacio vectorial) demostraron a través de los resultados obtenidos con las pruebas que se realizaron, mejor desempeño que el modelo booleano que es el que utiliza el sistema Horizonte. Fue mayor la cantidad de documentos relevantes recuperados y se logra cumplir con la idea de hacer visibles los recursos bibliográficos que pueden ser útiles a resolver un requerimiento de información, y que a través del catálogo tradicional no hubiese podido cumplirse. De ahí la importancia de incorporar en las búsquedas términos presentes en las tablas de contenido e índices de los libros que conforman la colección a consultar.

Además está el factor adicional que tanto Whoosh como Gensim presentan los resultados por orden de relevancia, pudiéndose así identificar aquellos que presentan más probabilidades de ser consultados inicialmente para resolver la consulta hecha. Es otra ventaja que presenta en el comparativo final y que reporta beneficios a la larga al usuario que usa un sistema de búsqueda más funcional.

Tabla 3. Cuadro resumen de porcentaje de rendimiento entre plataformas de recuperación de información

plataforma	consulta simple						consulta compuesta					
	simple con raíz			simple sin raíz			compuesta con raíz			compuesta sin raíz		
	precisión	cobertura	medida F	precisión	cobertura	medida F	precisión	cobertura	medida F	precisión	cobertura	medida F
horizonte	47%	65%	55%	41%	55%	47%	60%	60%	60%	40%	40%	40%
gensim	98%	92%	95%	99%	84%	91%	30%	100%	46%	41%	91%	57%
whoosh	98%	91%	94%	99%	84%	91%	65%	91%	76%	69%	78%	73%

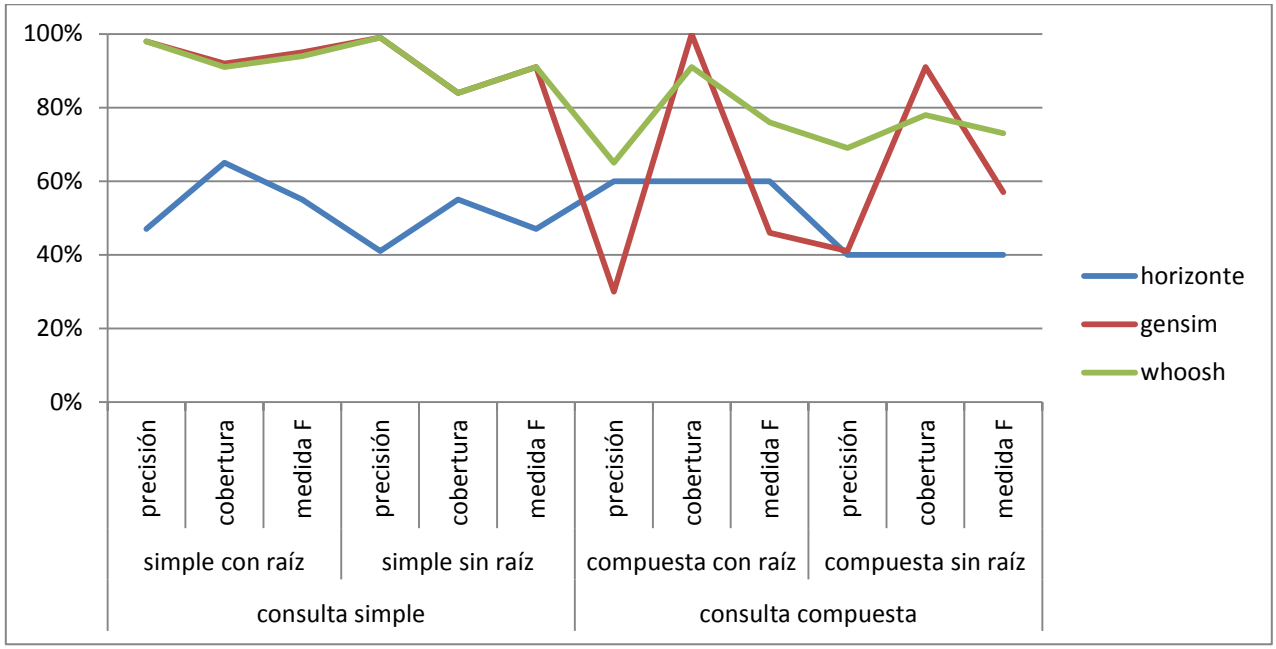


Figura 23. Comparativo de rendimiento entre plataformas de recuperación de información.

XII. CONCLUSIONES

Después del extenso proceso realizado en pos de lograr una solución lo más efectiva posible para mejorar la recuperación de información usando los libros que conforman una colección de biblioteca, queda la inmediata sensación de cuán arduo trabajo es desarrollar una herramienta de búsqueda que proporcione buenos resultados en términos de precisión y cobertura, ya que son varios los elementos que deben estar en corrección (ortografía, espacios, diagramación, visualización de datos completos, etc.) para que el sistema pueda funcionar bien.

En términos de rendimiento, es óptimo realizar búsquedas con extractor de raíces para ampliar la cantidad de resultados favoreciendo la cobertura aunque se pierda precisión, pero da mayores opciones para elegir.

También resulta decisiva la manera en que se expresan las consultas, por el detalle si el término está consultado en singular o plural, o si incluye signos por ejemplo. Cuestiones así pueden ser determinantes en el éxito de una consulta.

A modo de sugerencia como mejoramiento, sería muy interesante contar con la posibilidad de tener una función incorporada en Whoosh para trabajar con sinonimias, de modo que las consultas puedan arrojar resultados relevantes aunque no aparezca exactamente el término consultado. No se indagó en detalle, pero existe el módulo lang.wordnet que contiene una base de datos de sinónimos llamada Thesaurus para el idioma inglés - aunque también existe la opción en español llamada SpanishWordnet-que permite buscar sinónimos y ampliar las consultas; sin embargo aparentemente sólo es para sugerir los sinónimos del término consultado e iniciar una nueva consulta, no para incluirlos como parte de la consulta por el término inicial.

También sería deseable contar con un método automatizado de preprocesamiento de textos para acortar los tiempos de disponer de una colección de documentos que permita una vista simple y comprensible en pantalla, ante las consultas que se realicen.

Dada la situación actual en cuanto a procesos de recuperación de información, se estima que esta modalidad propuesta se constituye en una interesante contribución para implementar en los servicios de referencia de bibliotecas y en un elemento contundente para aprovechar más las colecciones existentes, en cuanto a poder desentrañar varios contenidos no visibles a través de medios tradicionales.

La posibilidad de incluir en las búsquedas los términos presentes en las tablas de contenido e índices de los textos, contribuye significativamente a ponderar la relevancia de los documentos ante consultas específicas y determinar así cuáles son los que pudiesen resolver requerimientos de información, que por medio de los catálogos tradicionales no se alcanza a vislumbrar en efectividad.

Las dificultades encontradas en el trayecto de la tesis, inicialmente referido a la fase del preprocesamiento de textos, y después con la estructura compleja y estado irregular de la base de datos con la que se pensó hacer el cruce de datos, llevó a ajustar la propuesta hacia un planteamiento metodológico solamente.

Si bien es cierto que con el código desarrollado en Whoosh actualmente se logran obtener resultados concretos, hubiese sido ideal haber desarrollado el prototipo para ponerlo en funcionamiento con enlace al sistema de información Horizonte utilizado en la biblioteca. En definitiva, instalar un sistema buscador que se hubiese establecido como alternativa al catálogo tradicional, ampliando las posibilidades de búsqueda.

Queda como paso siguiente a considerar, el incremento de la colección de documentos con más libros digitalizados para ampliar el índice a consultar. Aquello, unido al desarrollo del prototipo en una plataforma web que permita en una interfaz simple poder ejecutar las consultas y visualizar ordenadamente los registros recuperados, contribuiría mejor al aprovechamiento de los recursos de información existentes en definitiva.

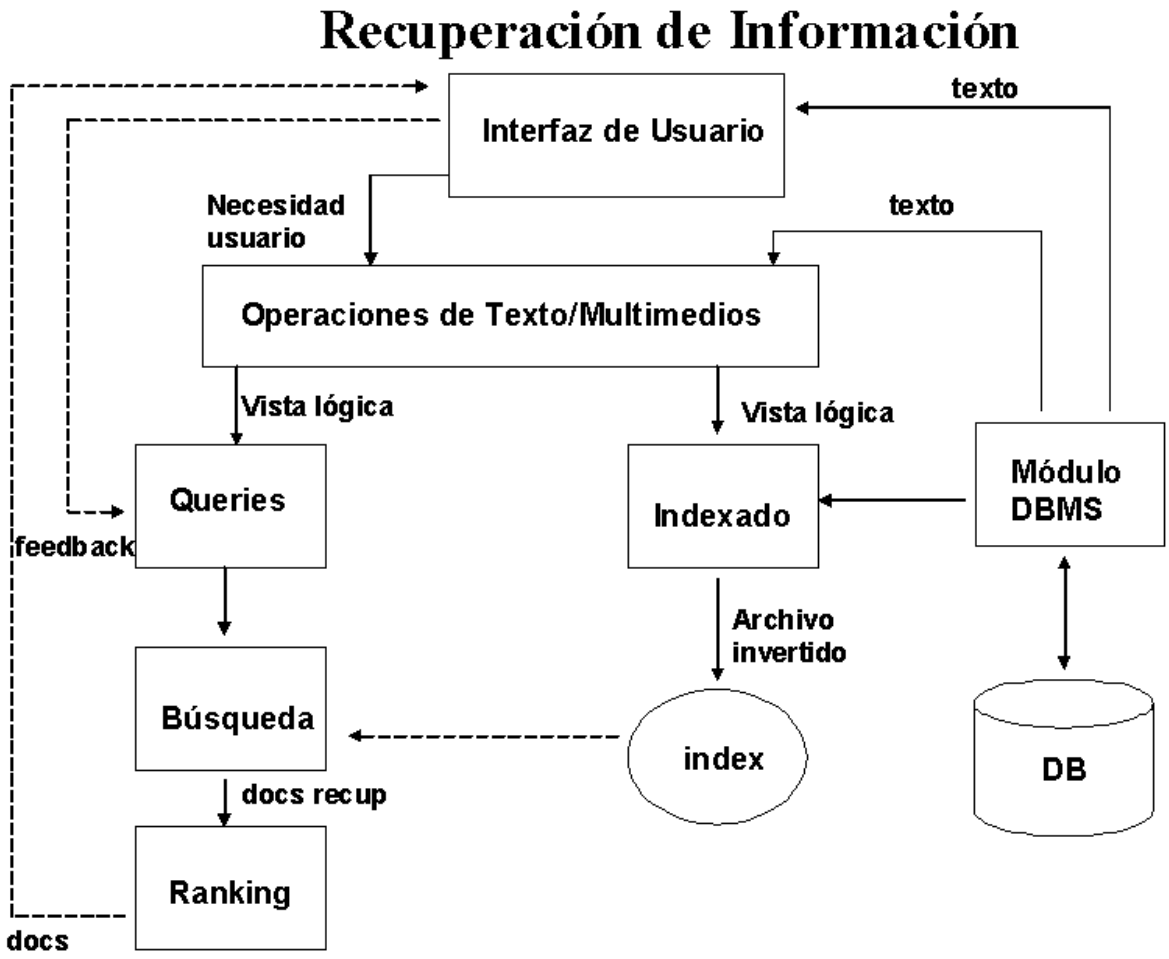


Figura 24. Diagrama de flujo del proceso de recuperación de información según Baeza - Ribeiro.

BIBLIOGRAFÍA

BAEZA-YATES, R. y RIBEIRO-NETO, B. (1999) *Modern Information Retrieval*. ACM Press, Addison Wesley. New York.

BALLESTEROS ESTRADA, SILVIA; MORALES ROMERO, GUILLERMO Y CEDILLO PÉREZ, PAVEL. (2012) Los problemas de identificación de caracteres OCR para la recuperación de texto en el libro antiguo: un análisis de caso en el Fondo Antiguo de la Biblioteca Central. *Biblioteca Universitaria, Vol. 15*, N°1, 25-34.

BARBER, E.[et al] (2007) La recuperación de la información bibliográfica en los catálogos en línea de acceso público del Mercosur. *III Encuentro Internacional de Catalogadores*, Universidad de Buenos Aires, 28-30 noviembre 2007. Buenos Aires.

BARBER, E.[et al] (2008) Los catálogos en línea de acceso público del Mercosur disponibles en entorno web. *Información, cultura y sociedad*, N°18, 37-55.

BELKIN, N,J., CHANG, S,J., DOWNS, T., SARACEVIC, T., ZHAO, S. (1990) Taking account of user tasks, goals and behavior for the design of online public access catalogs. *Proceedings of the 53rd Annual Meeting of the American Society for Information Science, Vol. 27*: 69– 73.

BELKIN, N.J. y SARACEVIC, T. (1992) Design principles for third-generation Online Public Access Catalogs: taking account of users and library use. *Annual Review of OCLC Research, Vol. july 1991-june 1992*, 43-45.

BLÁZQUEZ OCHANDO, M. (2012).Modelo probabilístico. *Técnicas avanzadas de recuperación de información* [en línea]. 19 diciembre 2012. Disponible en: http://ccdoc-tecnicasrecuperacioninformacion.blogspot.cl/2012/12/modelo-probabilistico.html [19 diciembre 2012]

BYRNE, A. and MICCO, M. (1988).Improving OPAC subject access. *College and Research Libraries,* N°49, 432-441.

COCHRANE, P.A. Y MARKEY, K.(1983) Catalog use studies – since the introduction of online interactive catalogs : impact on design for subject access. *Library Information Science Research, Winter, Vol.5*, N°4: 337-63.

CROFT, W.B. (1987) Approaches to intelligent information retrieval. *Information Proccesing & Management, Vol.23*, N°4, 249-254.

DATTA, J.(2010). *Ranking in Information Retrieval*. Mumbai: Department of Computer Science and Engineering, Indian Institute of Technology [en línea]. Disponible en: https://www.cse.iitb.ac.in/internal/techreports/reports/TR-CSE-2010-31.pdf [16 abril de 2010]

DILLON, M. and WENZEL, P. (1990) Retrieval Effectiveness of Enhanced Bibliographic Records. *Library Hi Tech*, *Vol. 8,* N°3, 43-46.

ELSHEIKH, A. (2013). Exploring Syntactic Relations for Literature-based Discovery. University of Calgary, Calgary.

HERNANDEZ, R., FERNANDEZ, C. y BAPTISTA, P. (2014). Metodología de la investigación. McGraw-Hill, México.

ISHIKAWA, Y. (2016). Spark ranking algorithms [en línea]. Disponible en: https://github.com/yu-iskw/spark-ranking-algorithms/blob/master/docs/okapi-bm25.md [26 agosto 2016]

KNUTSON, G. (1991) Subject Enhancement: Report on an Experiment. *College and Research Libraries*, *Vol. 52,* N°1, 65-79.

LANCASTER, F.W. (2001) Sistemas avanzados de recuperación de información. En W. Lancaster y M. Pinto (Coords.), Procesamiento de la información científica. Arco/Libros, Madrid.

MANDEL C. Y HERSCHMAN, J. (1983). Online subject access: enhancing the library catalog. *Journal of Academic Librarianship*. *Vol. 9*, N° 3:148–55.

MANNING, C., RAGHAVAN, P. y SCHÜTZE, H. (2008) Introduction to information retrieval. Cambridge University Press, New York.

MARTÍNEZ COMECHE, J. (2006). Los modelos clásicos de Recuperación de información y su vigencia. *III Seminario Hispano-Mexicano de investigación en Bibliotecología y Documentación.* Universidad Complutense de Madrid, 29-31 Marzo 2006, México DF.

MARTÍNEZ MÉNDEZ, F. (2004). *Recuperación de información: modelos, sistemas y evaluación*. KIOSKO JMC, Murcia.

MOOERS, C.N. (1950) The theory of digital handling of non–numerical information and its implications to machine economics. *Association for Computing Machinery*, Rutgers University, 29 marzo 1950, New Brunswick, New Jersey.

MORRIS, R.C. (2001) Online tables of contents for books: effect on usage. *Bulletin of the Medical Libraries Association. Vol.89*, N° 1, 29-36.

NLTK Project (2017).*NLTK 3.2.4 documentation* [en línea]. Disponible en :<http://www.nltk.org/> [21 mayo 2017]

PEIS, E. y FERNANDEZ-MOLINA, J.C. (1998) Enrichment of Bibliographic Records of Online Catalogs through OCR and SGML Technology. *Information Technology & Libraries, Vol.17*, N°3, 167-172.

PYTHON SOFTWARE FOUNDATION (2017).*Python package index* [enlínea]. Disponible en: https://pypi.python.org/pypi/Gensim [2017]

REHUREK, R. (2017) *RaReconsulting*: *machine learning and data mining expert* [en línea].Disponible en: https://radimrehurek.com/Gensim [2017]

RAMÍREZ BENAVIDES, K. (2012). *Stemming-lematización*. [en línea]. Disponible en: http://www.kramirez.net/wp-content/uploads/2012/02/Stemming.pdf [2012]

SALVADOR OLIVÁN, JOSÉ ANTONIO y ARQUERO AVILÉS, ROSARIO. (2006) Una aproximación al concepto de recuperación de información en el marco de la ciencia de la documentación. *Investigación bibliotecológica, Vol. 20*, N°41, 13-43.

SEAL, A., BRYANT, P. y HALL, C. (1982) *Full and short entry catalogues: library needs and uses*. Bath University Library, Centre for Catalogue Research. Bath.

TOLOSA, GABRIEL H. y BORDIGNON, FERNANDO R.A. (2008) *Introducción a la Recuperación de Información: Conceptos, modelos y algoritmos básicos.* Universidad Nacional de Luján, Argentina.

UNIVERSIDAD DEL PACÍFICO. DIRECCIÓN DE ANÁLISIS Y ASEGURAMIENTO DE LA CALIDAD. (2014). Estudio Caracterización estudiantes 2014 en composición socioeconómica de acuerdo a ingresos. Universidad del Pacífico. Santiago.

UNIVERSIDAD DEL PACÍFICO. DIRECCIÓN DE ANÁLISIS Y ASEGURAMIENTO DE LA CALIDAD. (2015) Resultados de encuesta de caracterización alumnos ingreso 2015.Universidad del Pacífico. Santiago.

VAN ORDEN, R. (1990) Content-enriched access to electronic information: summaries of selected research. *Library High Technology Vol.8*, N° 3: 27-32.

WITTENBACH, S. (1992) Building a better mousetrap: enhanced cataloguing and access for the online catalog.En: M. Ra (ed) *Advances in online public access catalogs: v.I*., Meckler Publishing. Westport, Connecticut.

WORMELL, I. (1994) Indización SAP para la exploración del amplio contexto temático de libros y para el acceso a entidades semánticas más pequeñas. *Ciencias de la Información. Vol. 25*, N° 4: 178-186.