Análisis de Planes de Estudio Mediante Determinación Automática de Pertinencias Sintáctico-Temáticas en Carreras de Informática

Caso de Estudio: Carrera Licenciatura en Ciencias de la Computación

Raúl Oscar Klenzi, Laura Viviana Gutierrez, Alejandra Malberti, Graciela Beguerí, Tamara Pinto, Jorge Matías Araya

Departamento de Informática – Facultad de Ciencias Exactas Físicas y Naturales.
Universidad Nacional de San Juan (DI-FCEFN-UNSJ)

Ignacio de la Roza y Meglioli. CUIM. Rivadavia - San Juan (SJ) - Argentina

{rauloscarklenzi, gutierrez.laura, amalberti, grabeda,tamara932,jorgemaraya}@gmail.com

Abstract. *This paper proposes an automated methodology for determining the syntactic-themed belongings between the minimum contents of careers computer area with standards established according to different degrees assigned to them. The automatic process groups the minimum contents of the careers establishing the degree of syntactic similarity of these, with the areas or thematic objectives resolutions or appropriate regulatory frameworks, thus favoring an initial analysis of curricula that may lead to the decision to make corrections on them. The application was carried out using modules modeling and text mining (TextMining -TM-) tool free software RapidMiner (RM) version 5.3.15. The case study considered is the Bachelor em career Computer Science Department of Informatics Science Faculty of Exact, Physical and Natural Sciences of the National University of San Juan.*

Resumen. *El presente trabajo propone una metodología automática para determinar las pertinencias sintactico-temáticas entre los contenidos mínimos de carreras de informática con las normas establecidas según las diferentes titulaciones asignadas a las mismas. El proceso automático agrupa los contenidos mínimos de las carreras estableciendo el grado de similitud sintáctica de estos, con las áreas u objetivos temáticos de las resoluciones o marcos regulatorios correspondientes, favoreciendo así, un análisis inicial de planes de estudio que pueda llevar a la decisión de hacer correcciones sobre los mismos. La aplicación se llevó adelante mediante la utilización los módulos de modelado y minería de texto (TextMining -TM-) de la herramienta de software libre RapidMiner (RM) versión 5.3.15. El caso de estudio considerado es la carrera Licenciatura en Ciencias de la Computación del Departamento Informática, Facultad de Ciencias Exactas Físicas y Naturales de La Universidad Nacional de San Juan.*

1. Introducción

Durante el último lustro las diferentes carreras universitarias del área informática de la República Argentina, se han visto sometidas a un proceso de evaluación del cual se deriva su posible acreditación. El marco regulatorio se basa en la Resolución ministerial 786/09 del Ministerio de Educación de la Nación, Secretaría de Políticas Universitarias.

El citado acto administrativo aprueba los **contenidos curriculares básicos**, la carga horaria mínima, los criterios de intensidad de formación práctica, los estándares y la nómina de Actividades Profesionales Reservadas para las carreras correspondientes a los título de: LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN, LICENCIADO EN SISTEMAS / SISTEMAS DE INFORMACIÓN / ANÁLISIS DE SISTEMAS, LICENCIADO EN INFORMÁTICA, INGENIERO EN COMPUTACIÓN e INGENIERO EN SISTEMAS DE INFORMACIÓN / INFORMÁTICA.

Un punto relevante que puede determinar si una carrera del área informática, acredita o no, está supeditado al análisis de los **contenidos curriculares básicos** realizado por pares evaluadores que se encargan de contrastar la información elevada por la Unidad Académica que contiene a una determinada carrera, con el marco regulatorio definido en RM 786/9.

En particular desde la perspectiva de los pares evaluadores a los efectos de analizar si una determinada carrera cumple con lo establecido en el anexo IV respecto de los estándares de acreditación y específicamente lo referido a la **Dimensión II) PLAN DE ESTUDIO Y FORMACIÓN**, una "primer aproximación", que estos realizan, es ir "a la caza" de contenidos básicos curriculares. Así, la información en cuanto a planes de estudios, contenidos mínimos etc., informados o provistos por las Unidades Académicas en la que se enmarca una determinada carrera, se contrastan con los requerimientos establecidos en la mencionada Dimensión II del anexo IV de la RM 786/9. En la medida que la información contrastada coincide, a criterio de los pares evaluadores, se dará por aprobada la mencionada dimensión.

De igual modo, y según las diferencias que se observen, se generarán recomendaciones de corrección, cuando estas sean menores, o los déficits que deberán ser solucionados por la carrera, cuando las diferencias sean mayores. Es de destacar que el conocimiento de los pares evaluadores es necesario por el hecho de que si bien la primera aproximación que se menciona, radica esencialmente en una búsqueda de coincidencias sintácticas, la evaluación final surge de un análisis semántico de mayor profundidad, de acuerdo al conocimiento y experiencia del evaluador. La presente propuesta tiene como objetivo automatizar el proceso de la "primera aproximación" realizada por los pares evaluadores respecto de la Dimensión II, de tal manera de facilitar esta tarea a las carreras que se sometan a esta instancia de evaluación o pretendan modificar planes de estudio tratando de adaptarse a nuevas resoluciones que establezcan el marco regulatorio de las carreras.

En particular como caso de estudio, mediante métricas, similitudes sintácticas y algoritmos de segmentación del área de minería de texto, se contrastarán los contenidos

curriculares básicos de la carrera LCC (plan 2006 y plan 2011), respecto de los requerimientos establecidos en la RM 786/09 tratando de evidenciar las mejoras que se sucedieron en el plan de estudio de la carrera del DI-FCEFN-UNSJ.

Este trabajo tiene como objetivo profundizar lo desarrollado en "Pertinencias De Planes De Estudio De Carreras De Informática Con Normativas Establecidas Por CONEAU",[1] presentado en WICC 2013. El mismo se enmarca en el proyecto trianual 2011-2013 "**MINERÍA DE DATOS EN LA DETERMINACIÓN DE PATRONES DE USO Y PERFILES DE USUARIO**" código 21/E889 que se desarrolló en el ámbito de la FCEFN-UNSJ, aprobado por el Consejo de Investigaciones Científicas Técnicas y de Creación Artística (CICITCA), financiado por la propia Universidad y ajustado a evaluación externa y actualmente bajo el proyecto bianual 2014-2015 "**EXTRACCIÓN DE CONOCIMIENTO EN DATOS MASIVOS"** aprobado por Resol 018/14 – CS de la UNSJ.

Los datos sobre los que se trabaja son generados en el marco de la acreditación de las carreras del DI y planes de estudios de las carreras del citado departamento de la FCEFN. Los contenidos mínimos de aquellos planes serán procesados observando las diferencias sintácticas entre los contenidos establecidos en la 786/09 marcando así, aquellos que son factibles de modificar en las diferentes unidades temáticas.

Así mismo se desarrollarán tareas de agrupamiento sintáctico entre contenidos mínimos con el objeto de ver si las unidades o áreas temáticas generadas coinciden con las propuestas en la normativa vigente.

En éste contexto, toda posible mejora en los planes de estudio de la carrera LCC del DI, será considerada positivamente.

2. Marco Teórico

La Minería de Texto (Text Mining –TM-) surge como un conjunto de funcionalidades destinadas a construir tecnología de análisis de texto; el texto es el método más común de intercambio de información.

Mientras los sistemas de recuperación de texto comerciales tradicionales se basan en índices con las ocurrencias de palabras en documentos, TM va un poco más allá y encuentra palabras claves. TM detecta patrones semánticos dentro del texto y se define como el proceso de análisis de texto que permite extraer información y conocimiento no trivial, de utilidad para determinados propósitos. Kantardzic, M (2003)

El texto es la forma más natural de almacenamiento de información, de esta manera TM tiene un potencial de aplicabilidad, superior a otras técnicas de Minería de Datos (Data Mining –DM-) sobre datos estructurados, dado que recientes estudios revelan que el 80% de la información de las compañías está en formato de texto.

2.1. Objetivos del DM y Fases involucradas.

El análisis automático de información textual puede usarse para diferentes propósitos generales; entre otros podemos mencionar:

- Proveer un resumen de los contenidos de grandes colecciones de documentos (semiestructurados) y organizarlos de la manera más eficiente.
- Identificar estructuras ocultas entre documentos o grupos de documentos.
- Incrementar la eficiencia y eficacia de los procesos de búsqueda para encontrar similitudes o información relacionada.
- Detectar documentos duplicados, o información duplicada en grandes archivos.

El proceso de text minig, representado gráficamente por la Figura 1, consta globalmente de dos fases [1].

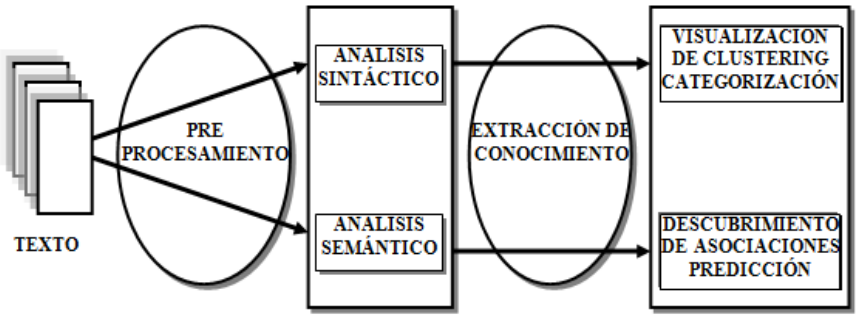

Figura 1: Fases del Text Mining

En el análisis de texto se usarán los términos análisis sintáctico, que en esencia comprende un estudio de la gramática del documento, y análisis semántico, que comprende el estudio del significado que se otorga a la asociación de palabras en una sentencia y depende del contexto en que se ubica y del conocimiento del usuario.

Una Forma Intermedia (Intermediate Form IF), obtenida tras el preprocesamiento puede ser semiestructurada, tal como una representación de un grafo-conceptual, o estructurada tal como una base de datos relacional. Las formas intermedias pueden variar en grados de complejidad, dando resultados satisfactorios a diferentes propósitos de búsqueda. Estas IF, se clasifican en basadas en documentos o basadas en conceptos.

En las IF basadas en documentos, cada entidad representa un objeto o concepto de interés en un dominio específico. En este caso se deducen patrones y relaciones a través del documento. Clustering, visualización y categorización de documentos son ejemplos de IFs basadas en documento.

Para un análisis más fino, en un dominio específico, la tarea de descubrir conocimiento requiere de un análisis semántico y manejar una representación

suficientemente rica que permita capturar la relación entre objetos o conceptos descriptos en el documento. La búsqueda basada en conceptos deriva patrones y relaciones a través de objetos y conceptos. Estos análisis semánticos son computacionalmente costosos y se trabaja tratando de hacerlos más eficientes y escalables para textos de gran extensión.

Las operaciones de DM, como modelos predictivos y descubrimiento de asociaciones, caen en este tipo de categoría. Una IF, basada en documento, se puede transformar en una basada en conceptos mediante el reordenamiento y extracción de información relevante a un dominio específico. De esta manera una forma intermedia basada en documento es independiente del dominio, mientras que la basada en conceptos brinda representaciones dependientes del dominio.

2.2 Como interpretar los Documentos de Texto

Para interpretar en detalle los documentos de texto, se puede buscar en ellos palabras claves o categorizarlos según su contenido semántico. Cuando se identifican palabras claves, se observan definiciones o detalles característicos de esos documentos que pueden usarse para buscar relaciones, conexiones o parecidos con otros documentos.

En el área de recuperación de información (Information Retrieval IR) los documentos se han representado tradicionalmente en un modelo de espacio vectorial. Dichos documentos se muestrean usando reglas sintácticas simples, delimitadores como espacios en blanco o puntos permiten extraer palabras y frases claves dentro de un mismo contexto del documento respectivamente. Luego las muestras se transforman a formas canónicas (ej: leyendo por leer, es, son, fue por el verbo ser). Cada forma canónica representa un eje en el espacio euclideano. Los documentos se pueden representar como vectores en un espacio n-dimensional. Si un cierto valor **t** ocurre **n** veces en un documento **d**, entonces la **t**-ésima coordenada del documento **d** es simplemente **n**. Se puede seleccionar normalizar la longitud del documento a 1, usando normas L1, L2 o L∞.

$$\| d_1 \| = \sum_t n(d,t) \quad ; \quad \|d_2\| = \sqrt{\sum_t n(d,t)^2} \quad ; \quad \|d_\infty\| = max_t\, n(d,t)$$

Donde **n(d,t)** es el número de ocurrencias del término **t** en un documento **d**. Esta representación no rescata que algunos términos, los llamados palabras claves, (ej: algoritmo) son más representativos que otros (ej: El, la, es…). Si **t** ocurre **nt** veces en **N** documentos **nt/N**, indica cuan común es la aparición de **t** en los documentos. De aquí la importancia del término. La frecuencia inversa del documento (Inverse Document Frequency IDF) = **1 + log (nt/N)** se usa para estirar las diferencias en los ejes del espacio vectorial. De este modo la **t**-ésima coordenada del documento **d** se representa con el valor (**n(d, t)/‖ d1‖)×IDF (t)** en el modelo de espacio vectorial pesado. A pesar de ser extremadamente duro y sin capturar absolutamente nada de la semántica del lenguaje, este modelo trabaja bien en definidos contextos. Aún con ciertas variaciones, todos estos modelos de análisis de textos consideran a los documentos como múltiples conjuntos de términos, sin prestar atención al orden entre los mismos, por lo que comúnmente se los denomina modelos de valijas de palabras.

El problema de afinidad o similitud entre documentos es de mayor complejidad, pues implica la subdivisión de documentos basada en un análisis de contenidos. Dependiendo del algoritmo particular para generar el mapeo, el resultado del mapa topográfico puede reforzar similitudes entre documentos en términos de distancia euclídea. Si bien hay diferentes alternativas a la distancia euclidea, una muy utilizada, es la distancia del coseno. Bing Liu, (2007) que permite clasificar documentos de acuerdo al grado de relevancia de la consulta. La distancia del coseno calcula la importancia de la similitud entre consulta q y cada documento **dj** en la colección de documentos D. Esta medida es el coseno del ángulo entre el vector consulta q y el documento dj.

$$\text{Coseno (dj, q)} = \frac{< dj \bullet q >}{\| dj \| x \| q \|} = \frac{\sum_{i=1}^{|V|} Wijx Wiq}{\sqrt{\sum_{i=1}^{|V|} W_{ij}^2} x \sqrt{\sum_{i=1}^{V} W_{iq}^2}}$$

En la anterior expresión **Wij** representa la cantidad de veces que una cadena i aparece en el documento dj (componente i del vector W asociado a dj), **Wiq** es la cantidad de veces que la cadena i aparece en la consulta q. V representa la totalidad de cadenas consideradas. Si bien se habla de q como una consulta, la misma puede ser un documento de la colección en la búsqueda de similitudes con otros documentos de la colección. A mayor valor del coseno, mayor afinidad o similitud entre la consulta q y el documento que la contiene.

3. Desarrollo de la Aplicación

La aplicación se ha llevado adelante mediante la utilización de la herramienta de software libre GPL RapidMiner 5.3.15. La misma consta de una tarea de preprocesamiento de los documentos de texto involucrados y posteriormente de las tareas de segmentación que determina en forma automática la pertenencia de cada contenido mínimo del plan de estudio a un área temática y la correspondencia de ésta con la de la Res 786/09. En la Figura 2 se observa el esquema de la aplicación.

Se destaca que las áreas temática para carreras de Licenciatura en Ciencias de la Computación según la Res. 786/09 son: Algoritmos y lenguajes, Arquitectura. Sistemas Operativos y Redes, Ciencias Básicas, Formación Socio Profesional, Ingenieria de Software, Base de Datos y Sistemas de Información y Teoría de la Computación.

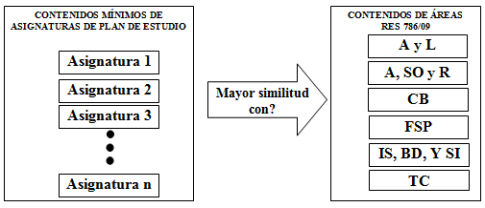

Figura 2 Esquema de la Aplicación

3.1. Preprocesamiento y Análisis de Documentos de Texto

En la Figura 3 se puede observar el conjunto de módulos que permite la implementación, en RM, de las tareas programadas.

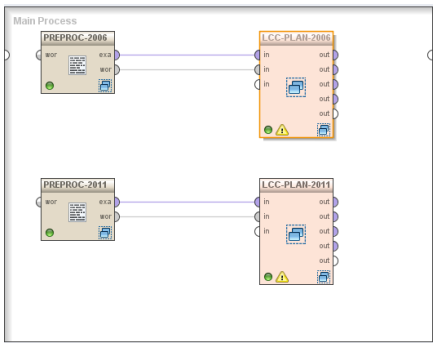

Figura 3: Módulos de RM que permite realizar una comparación entre planes de estudio.

Los módulos **PREPROC-2006** y **PREPROC-2011** se encargan de realizar el preprocesamiento de los documentos asociados a los planes de estudio de Licenciatura en Ciencias de la Computación 2006 y 2011 respectivamente, como así también las áreas de conocimientos y sus contenidos especificados en la Res. 786/09. Los documentos, y planes de estudio de carreras de los diferentes departamentos de la FCEFN son preprocesados por un módulo de RM.

En esta instancia de preprocesamiento, para cada documento y mediante la secuencia de cinco pasos, se realizan sucesivamente la separación en palabras (Tokenize), la eliminación de palabras carentes de significado (Filter Stopwords), el filtrado de palabras de cierta cantidad de caracteres (Filter Token), se reducen los términos a una forma base o raíz (Stem), y por último se regeneran los documentos con cadenas de hasta una cierta cantidad de palabras (Generate n-Grams), generándose como salida el vector de palabras desde el cálculo de TF-IDF . En este caso aún cuando se constató un aumento del espacio de búsqueda (cantidad de tokens) en aproximadamente un 10%, no se realizaron tareas de steeming para que el resultado final fuera de mejor comprensión por parte de los usuarios atento a que las expresiones visualizadas reflejan, en una primera instancia de la aplicación, posibles contenidos faltantes en los contenidos mínimos de las carreras respecto de los establecidos en el marco regulatorio.

3.2. Visualización de Contenidos Faltantes

En esta etapa se realiza una comparación sintáctica entre contenidos mínimos de los planes de estudio y lo establecido en el marco regulatorio. Así, para el plan de estudios correspondiente a la carrera Licenciatura en Ciencias de la Computación del año 2006, se observa lo siguiente:

Tabla 1. Contenidos faltantes en plan de estudio 2006

Row No.	word	CONT_MINIMOS_ASIGNATURAS_PLAN2006	CONT_MINIMOS_786
1	CONCURRENCIA	0	4
2	ARQUITECTURA	0	3
3	CASO	0	3
4	CONCEPTOS_ARQUITECTURAS	0	3
5	DATOS_SISTEMAS	0	3
6	PROCESO	0	3
7	BÁSICAS	0	2
8	COMPUTABLES	0	2
9	ELEMENTOS	0	2
10	FUNCIONAL	0	2
11	MÁQUINAS	0	2
12	OPERATIVOS_REDES	0	2
13	PROFESIONAL	0	2
14	PRUEBAS	0	2
15	REAL	0	2
16	SIMBÓLICA	0	2
17	SISTEMAS_OPERATIVOS_REDES	0	2
18	TIEMPO_REAL	0	2

La Tabla 1 permite visualizar contenidos que figuran en la resolución del marco regulatorio y no figuran en los contenidos mínimos del plan de estudio. Se observa que hay cadenas de caracteres como “CONCURRENCIA”, que aparece 4 veces en el marco regulatorio y “COMPUTABLES” “SIMBÓLICA” y “TIEMPO_REAL” que aparecen 2 veces cada una y ninguna de ellas en los contenidos mínimos de las asignaturas del plan de estudio. Esta situación debería generar un llamado de atención hacia quienes realizan el análisis del plan de estudio con la finalidad de acercarse a lo establecido en el marco regulatorio y continuar acreditando.

Tras este primer análisis se tratará de ver si los diferentes contenidos mínimos se ubican adecuadamente, desde el análisis sintáctico, en la diferentes área de conocimiento establecidas en el marco regulatorio. A modo de ejemplo se presenta en la Tabla 2 los contenidos correspondientes a dos asignaturas del plan de estudio LCC-2006 “Estructura y funcionamiento de computadoras II” y “Compiladores” como así también, los contenidos mínimos exigidos por la Res 786/09 para el área de conocimiento “Arquitecturas, Sistemas Operativos y Redes” correspondientes a la carrera licenciatura en Ciencias de la Computación. El objetivo, de esta parte de la aplicación, es tratar de determinar la afinidad sintáctica entre los contenidos mínimos de las diferentes asignaturas que conforman el plan de estudio, con cada una de las áreas de conocimiento establecidas en el marco regulatorio. En esta tarea se utiliza la métrica de similitud del coseno explicitada con anterioridad.

Tabla 2. Ejemplos de contenidos mínimos de asignaturas y áreas temáticas Res 786/09

estructura y funcionamiento de las computadoras II .
set de instrucciones . lenguaje de máquina . lenguaje ensamblador .
estructura y funcionamiento de la cpu . registros . pipelining . microoperaciones .
control cableado . microprogramación . entrada / salida .
componentes . técnicas . canales y procesadores de e/s .
controladores . interfaces . sistemas operativos . funciones . estructuras . tipos .
administración del procesador y los procesos, memoria, e/s y archivos . protección .

compiladores .
estructura de un compilador .
especificación de lenguajes . sintáxis y semántica . gramática .
fases de un compilador . pasadas . un compilador sencillo de una pasada .
análisis léxico . construcción de un analizador léxico . uso de autómatas finitos .
análisis sintáctico . función del analizador sintáctico . construcción del analizador sintáctico .
métodos ascendentes y descendentes . manejo de errores . traducción dirigida por la sintaxis .
comprobación de tipos . sistemas de tipos . especificaciones de un comprobador de tipos .
conversiones . sobrecarga . ambientes para el momento de ejecución .
organización de la memoria . asignación de memoria . tabla de símbolos .
generación de código . código intermedio . lenguajes intermedios .
la máquina objeto . administración de memoria en ejecución .
un generador de código simple . optimización de código . validación de un compilador .

	arquitectura .
	arquitectura arquitectura y organización de computadoras .
	representación de los datos a nivel máquina . error . lenguaje ensamblador .
	jerarquía de memoria, organización funcional .
	circuitos combinatorios y secuenciales .
	máquinas algoritmicas .
	procesadores de alta prestación .
	arquitecturas no von neumann .
	arquitecturas multiprocesadores . conceptos de arquitecturas grid .
	conceptos de arquitecturas reconfigurables .
	conceptos de arquitecturas basadas en servicios .
	sistemas operativos .
	sistemas operativos . concepto de proceso . planificación de procesos .
	concurrencia de ejecución . interbloqueos .
Arquitectura	administración de memoria .
Sistemas Operativos	sistema de archivos . protección .
y	sistemas operativos: de tiempo real, embebidos (embedded), distribuidos .
Redes	comunicación, sincronización manejo de recursos y sistemas de archivos en sistemas distribuidos .
	memoria compartida distribuida .
	control de concurrencia en sistemas distribuidos . transacciones distribuidas: seguridad en sistemas distribuidos .
	redes .
	redes y comunicaciones .
	técnicas de transmisión de datos, modelos, topologías, algoritmos de ruteo y protocolos .
	sistemas operativos de redes .
	seguridad en redes . elementos de criptografía .
	sistemas cliente/servidor y sus variantes . el modelo computacional de la web .
	administración de redes . computación orientada a redes .

3.3. Determinación de pertinencias sintácticas de Asignaturas con Res.786/09.

La Figura 4 muestra la secuencia de módulos, que en la herramienta RM, permite encontrar la afinidad sintáctica entre los contenidos mínimos de las diferentes asignaturas del plan de estudio respecto de lo establecido en el marco regulatorio de la Res 786/09

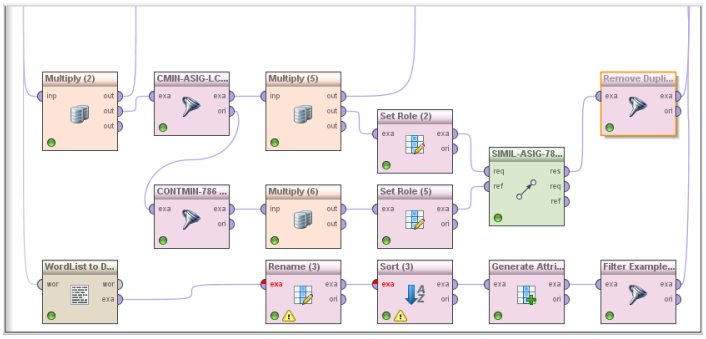

Figura 4 Secuencia de módulos utilizados en RM

Se observa que los contenidos mínimos de las asignaturas son la consulta (**req**) y los contenidos de las diferentes áreas de conocimiento de la 786/09 son la referencia (**ref**) en el módulo **SIMIL-ASIG-786** que mediante la métrica de similitud del coseno calcula la aproximación sintáctica entre ambas entradas. La siguiente tabla presenta la similitud sintáctica entre las asignaturas del plan de estudio LCC 2006 con las áreas de conocimiento especificadas en la Res 786/09, destacando que se presenta la asignatura y el mayor valor de similitud encontrado con algún área temática.

Tabla 3. Grado de similitud sintáctica entre asignaturas plan 2006 y áreas temáticas Res 786/09

Row No.	request	document	similitud
1	estructuras de datos y algoritmos .	algoritmos y lenguajes .	0.22481
5	algoritmos y resolución de problemas .	algoritmos y lenguajes .	0.12855
13	inteligencia artificial .	algoritmos y lenguajes .	0.06325
14	programación procedural .	algoritmos y lenguajes .	0.06283
15	arquitecturas avanzadas de computadoras .	algoritmos y lenguajes .	0.06092
18	programación orientada a objetos .	algoritmos y lenguajes .	0.05179
20	paradigmas de lenguajes .	algoritmos y lenguajes .	0.04244
24	compiladores .	algoritmos y lenguajes .	0.02901
6	sistemas distribuidos .	arquitectura sistemas operativos y redes .	0.12282
12	redes .	arquitectura sistemas operativos y redes .	0.06631
16	estructura y funcionamiento de las computadoras ii .	arquitectura sistemas operativos y redes .	0.05981
22	introducción a los sistemas digitales ii .	arquitectura sistemas operativos y redes .	0.03329
25	estructura y funcionamiento de las computadoras i .	arquitectura sistemas operativos y redes .	0.02861
28	introducción a los sistemas digitales i .	arquitectura sistemas operativos y redes .	0.01977
10	probabilidad y estadística .	ciencias básicas .	0.07604
17	matematica discreta .	ciencias básicas .	0.05403
26	análisis matemático i .	ciencias básicas .	0.02678
2	fundamentos profesionales y legaleS .	formación socio profesional .	0.18205
23	aspectos profesionales y sociales .	formación socio profesional .	0.02959
29	inglés i .	formación socio profesional .	0.01331
3	tópicos de ingeniería de software y de requerimientos .	ingeniería de software base de datos y sistemas de informacion .	0.16888
4	información y sistemas .	ingeniería de software base de datos y sistemas de informacion .	0.14453
7	base de datos avanzadas .	ingeniería de software base de datos y sistemas de informacion .	0.10803
8	calidad de software .	ingeniería de software base de datos y sistemas de informacion .	0.09194
11	diseño de software .	ingeniería de software base de datos y sistemas de informacion .	0.07185
19	base de datos .	ingeniería de software base de datos y sistemas de informacion .	0.05062
21	teoría de la información .	ingeniería de software base de datos y sistemas de informacion .	0.04131
9	teoría de la computación .	teoría de la computación .	0.08981
27	matemática básica .	teoría de la computación .	0.02419
30	análisis matemático ii .	teoría de la computación .	0.00909
31	inglés ii .	teoría de la computación .	0.00879

En la Tabla 3 se aprecia la mayor similitud sintáctica entre contenidos mínimos de asignaturas (request) y áreas temáticas establecidas en la Res 786/09 (ref). Se puede observar, a modo de ejemplo, que la asignatura “Estructura y Funcionamiento de Computadoras II”, en función de sus contenidos mínimos ha sido correctamente asignada al área de conocimiento “Arquitectura, Sistemas Operativos y Redes”. A su vez se puede observar también, que hay asignaturas como “Compiladores” que debiera estar contenida en el área de “Teoría de la computación”, sin embargo en este caso, el mayor valor de similitud sintáctica le ha correspondido con el área temática “Algoritmos y Lenguajes”. De la Tabla 3 la autoridad encargada de hacer el análisis y posibles modificaciones al plan de estudio comienza a “tener pistas”, de cuáles son las asignaturas sobre cuyos contenidos mínimos hay que prestar mayor atención.

Seguidamente se realiza el análisis de contenidos faltantes y similitudes sintácticas para el nuevo plan de estudio (2011) de Licenciatura en Ciencias de la Computación.

Tabla 4. Contenidos faltantes en plan de estudio 2011

Row No.	word	CONT_MINIMOS_ASIGNATURAS_PLAN2011	CONT_MINIMOS_786
1	CASO	0	3
2	CONCEPTOS_ARQUITECTURAS	0	3
3	DATOS_SISTEMAS	0	3
4	BÁSICAS	0	2
5	OPERATIVOS_REDES	0	2
6	PROFESIONAL	0	2
7	SISTEMAS_OPERATIVOS_REDES	0	2

En la Tabla 4, se puede observar que muchos de los elementos sintácticos faltantes se han eliminado en los contenidos mínimos del nuevo plan. Es de destacar que el análisis sintáctico automático sólo contempla asignaturas, que en carácter de obligatorio, deben ser cursadas por los alumnos. En el caso particular de las carreras del DI-FCEFN-UNSJ cuenta, además de las asignaturas obligatorias, con un par de asignaturas optativas en donde los planes de estudio tratan de seguir, al menos parcialmente, la dinámica del marco regulatorio en cuanto a contenidos mínimos faltantes.

Tabla 5. Grado de similitud sintáctica entre asignaturas plan 2011 y áreas temáticas Res 786/09

request	document	similitud
estructuras de datos y algoritmos .	algoritmos y lenguajes .	0.43447
algoritmos y resolución de problemas .	algoritmos y lenguajes .	0.36003
paradigmas de lenguajes .	algoritmos y lenguajes .	0.24045
programación procedural .	algoritmos y lenguajes .	0.20555
algoritmos numéricos .	algoritmos y lenguajes .	0.13588
sistemas distribuidos .	arquitectura sistemas operativos y redes .	0.42032
estructura y funcionamiento de las computadoras ii .	arquitectura sistemas operativos y redes .	0.40749
redes .	arquitectura sistemas operativos y redes .	0.34955
arquitecturas de computadoras .	arquitectura sistemas operativos y redes .	0.26399
estructura y funcionamiento de las computadoras i .	arquitectura sistemas operativos y redes .	0.11503
sistemas digitales .	arquitectura sistemas operativos y redes .	0.06479
probabilidad y estadística .	ciencias básicas .	0.19211
matemática básica .	ciencias básicas .	0.15668
análisis matematico ii .	ciencias básicas .	0.15596
álgebra lineal .	ciencias básicas .	0.12166
lógica y optimización aplicadas .	ciencias básicas .	0.07199
análisis matematico i .	ciencias básicas .	0.03512
fundamentos profesionales y legales .	formación socio profesional .	0.31498
aspectos profesionales y sociales .	formación socio profesional .	0.20792
inglés i .	formación socio profesional .	0.03191
ingeniería de software i .	ingeniería de software base de datos y sistemas de informacion .	0.50112
sistemas de información .	ingeniería de software base de datos y sistemas de informacion .	0.35193
base de datos .	ingeniería de software base de datos y sistemas de informacion .	0.31668
ingeniería de software ii .	ingeniería de software base de datos y sistemas de informacion .	0.29469
ingeniería de software iii .	ingeniería de software base de datos y sistemas de informacion .	0.25021
proyectos de innovación tecnológica .	ingeniería de software base de datos y sistemas de informacion .	0.10446
teoría de la información .	ingeniería de software base de datos y sistemas de informacion .	0.07761
inglés ii .	ingeniería de software base de datos y sistemas de informacion .	0.04145
teoría de autómatas y computabilidad .	teoría de la computación .	0.34338
matemática discreta .	teoría de la computación .	0.25864
inteligencia artificial .	teoría de la computación .	0.20863
programación orientada a objetos .	teoría de la computación .	0.17537
compiladores .	teoría de la computación .	0.10400

Como se observa en la Tabla 5 existe una mejor correspondencia entre contenidos mínimos de asignaturas y las áreas definidas en el marco regulatorio en este plan de estudio (2011) que en el anterior (2006).

Tabla 6. Correspondencia entre asignaturas plan 2011 y áreas temáticas Res 786/09 elaborada por comisión de seguimiento de plan de estudio

	LICENCIATURA EN CIENCIAS DE LA COMPUTACIÓN (Carrera que se propone)		ÁREAS											
			CB		IS-BD-SI		FSP		AyL		A-SO-R		TC	
		HS TOT	%	HS	%	HS	%	HS	%	HS	%	HS	%	HS
1	Algoritmos y Resolución de Problemas	105							100	105				
2	Matemática Básica	90	100	90										
3	Estructura y Funcionamiento de las Computadoras I	75									100	75		
4	Sistemas de Información	75			100	75								
5	Programación Procedural	135							90	121,5			10	13,5
7	Aspectos Profesionales y Sociales	60					100	60						
6	Algebra Lineal	105	100	105										
8	Sistemas Digitales	45									100	45		
9	Programación Orientada a Objetos	105							90	94,5			10	10,5
10	Matemática Discreta	120	20	24					20	24			60	72
11	Análisis Matemático I	105	100	105										
12	Análisis Matemático II	90	100	90										
13	Estructura de Datos y Algoritmos	120							80	96			20	24
14	Estructura y Funcionamiento de las Computadoras II	90							13	12	87	78		
15	Inglés I	45					100	45						
16	Paradigmas de Lenguajes	120							80	96			20	24
17	Probabilidad y Estadística	90	100	90										
18	Bases de Datos	120			90	108							10	12
22	Algoritmos Numéricos	60	40	24					50	30			10	6
20	Redes	90							5	4,5	95	85,5		

Tabla 6 (Continuación)

21	Ingeniería de Software I	105			100	105								
19	Inglés II	45					100	45						
23	Teoría de Autómatas y Computabilidad	90											100	90
24	Lógica y Optimización Aplicadas	120							20	24			80	96
25	Ingeniería de Software II	105			100	105								
26	Inteligencia Artificial	120							50	60			50	60
27	Arquitecturas de Computadoras	75							20	15	80	60		
28	Compiladores	135											100	135
29	Teoría de la Información	105									70	73,5	30	31,5
30	Sistemas Distribuidos	75							30	22,5	70	52,5		
32	Ingeniería de Software III	105			90	94,5			10	10,5				
33	Fundamentos Profesionales y Legales	45					100	45						
35	Proyectos de Innovación Tecnológica	60			100	60								
	HORAS POR AREA			528		547,5		195		715,5		469,5		574,5
			3030 hs.											
31	OPTATIVA I		A DETERMINAR 90 hs											
34	OPTATIVA II		A DETERMINAR 90 hs											
36	TRABAJO FINAL		150 hs											
	REQUISITO		60 hs											
	HORAS TOTALES		**3360 hs + 60 (requisito) = 3420 hs**											

La Tabla 6 es la generada por la comisión de expertos que entendió sobre la modificación del plan de estudios 2006 que deriva en la versión actual 2011, en ella se aprecia no solo la correspondencia de contenidos mínimos y áreas temáticas, sino también la carga horaria correspondiente. A modo de ejemplo la asignatura "Arquitecturas de Computadoras", tiene un 20% de su contenido asociada al área "Algoritmos y Lenguajes" y un 80% al área "Arquitecturas, Sistemas Operativos y Redes". Se puede apreciar que la Tabla 5 obtenida mediante similitudes sintácticas respecto de las áreas temáticas del marco regulatorio, tiene una mejor aproximación a la Tabla 6 que lo evidenciado en la Tabla 3 correspondiente al plan 2006.

4. Conclusiones

Con los objetivos iniciales de comparar los planes de estudio de la carrera LCC, vigente hasta el 2011 y el plan de estudio actual de dicha carrera del DI, con los contenidos de las áreas establecidas por Resolución Ministerial N' 786/2009, utilizando los módulos citados anteriormente de RM, se ha logrado determinar la pertinencia entre el contenido mínimo del plan de estudio de la carrera LCC plan 2006 con las áreas establecidas en Resolución N' 786/2009, utilizando el cálculo de similitud, cuya ponderación indica lo cercano de los contenidos mínimos con las áreas consideradas en Resolución Ministerial Nº 786/2009.

Cuando el análisis se realizó para el plan de estudio 2011, se obtuvieron mejores resultados debido a modificaciones y agregados de contenidos mínimos realizados por comisiones de docentes conformadas a tal fin.

Se destaca que el procedimiento automático llevado adelante a dado una excelente aproximación en cuanto a correspondencia de contenidos mínimos de asignaturas respecto de contenidos mínimos de áreas temáticas del marco regulatorio y lo desarrollado manualmente por la comisión de plan de estudio.

La realización de este trabajo permitirá sugerir a las autoridades del DI, posibles cambios que deberían realizarse a los contenidos mínimos de las asignaturas que forman el plan de estudio de la carrera LCC con la finalidad de llegar cada vez más a lo requerido por el marco regulatorio para el proceso de acreditación.

Así mismo y como una tarea futura se tratará de ver, atento a que cada área temática de la Res 786/09 posee su carga horaria semestral, si los valores de similitud de cada asignatura con un área temática se pueden traducir en su carga horaria semestral dando como propuesta no sólo una indicación de contenidos faltantes y posibles asignaturas a modificar, si no también la carga horaria de las mismas.

Este mismo análisis fue realizado para la carrera Licenciatura en Sistemas de Información (LSI), que también depende del DI de la FCEFN de la UNSJ y que también ha estado en proceso de acreditación.

Cabe destacar que tanto la carrera LCC plan 2011 y la carrera LSI plan 2011, han obtenido la acreditación por parte de la Comisión Nacional de Evaluación y Acreditación Universitaria, y son los planes que en la actualidad están en vigencia en el DI

5. Referencias

Gutiérrez, L. (2013) "Pertinencias De Planes De Estudio De Carreras De Informática Con Normativas Establecidas Por CONEAU". En XV Workshop de Investigadores en Ciencias de la Computación. WICC 2013. ISBN: 9789872817961.

Kantardzic, M (2003) "Data Mining: Concepts, Models, Methods, and Algorithms"ISBN:0471228524 John Wiley & Sons © (343 pages)

Klenzi, R.(2008) Tesis de posgrado de maestría "Aplicación de minería de datos a la gestión bibliotecaria". Biblioteca FCEFN-UNSJ.

Klenzi, R ,Gutiérrez L., Villafañe V. (2012) "Técnicas de recuperación de información en la determinación de pertinencias bibliográficas".

Liu B. (2007) "Web DataMining. Exploring Hyperlinks, Contents, and Usage Data"Springer-Verlag Berlin Heidelberg.

Manning C, Prabhakar R. Hinrich & Hinrich Schütze. (2009) "An Introduction to Information Retrieval ", Cambridge University Press.

Min, S; Yi-Fang B. (2009) Handbook of Research on Text and Web Mining Technologies -Information science reference- Editorial Advisory Board.

Rapid-I. http://rapid-i.com ver.5.3.015 de 2014.