Aplicando Minería de Textos en el Análisis de Mallas Curriculares de Carreras de Computación en el Perú

Nils Murrugarra-Llerena Fernando Alva-Manchego

[1] Instituto de Ciências Matemáticas e de Computação – ICMC
Universidade de São Paulo – USP
Caixa Postal: 668 – CEP: 13560-970 – São Carlos – SP

{nineil,falva}@icmc.usp.br

Resumen

La comparación de mallas curriculares permite identificar y evaluar la calidad de los programas de carrera en una universidad; sin embargo, esta comparación casi siempre es realizada manualmente. En este trabajo, se propone una metodología de minería de textos para ayudar en la tarea de comparar mallas curriculares. Fue realizado un estudio de caso utilizando mallas curriculares de programas de pre-grado en Computación de universidades peruanas, aplicando diferentes algoritmos de agrupamiento jerárquico aglomerativo. Los resultados obtenidos muestran que la metodología propuesta permite descubrir relaciones ocultas entre las mallas curriculares analizadas.

1. Introducción

La minería de textos (MT) puede ser definida como la aplicación de métodos computacionales y técnicas sobre datos textuales para encontrar información intrínseca y relevante, así como conocimiento previamente desconocido (Do Prado and Ferneda, 2007). Existen numerosas aplicaciones de la MT, que incluyen investigación pionera en análisis y clasificación de noticias, *email* y filtro de *spam*; extracción jerárquica de tópicos de páginas web; extracción y gestión automática de ontologías; e inteligencia competitiva (Srivastava and Sahami, 2009).

Existen muchas áreas en las cuales es necesario analizar y comparar documentos, siendo una de ellas el estudio comparativo de mallas curriculares. (Biddle and Tempero, 1996) efectúan una comparación de la malla curricular del programa de Computación de la Universidad de *Wellington* con recomendaciones de la ACM/IEEE, específicamente del núcleo básico. En América Latina, los trabajos de (Pereira et al., 2010) y (Prietch and Pazeto, 2010) examinan mallas curriculares para realizar un análisis de la oferta y propuestas de estandarización de mallas curriculares para el contexto brasilero. A pesar de que los objetivos de esos trabajos sean distintos, poseen una característica en común: el proceso de comparación es manual.

El análisis de mallas curriculares es importante porque permite evaluar la calidad de los programas correspondientes. Sin embargo, no siempre se cuenta con el tiempo necesario para realizar este tipo de estudio de forma completamente manual.

En este artículo, se propone emplear um enfoque basado en minería de textos para ayudar en la tarea de comparación y análisis de mallas curriculares. Más específicamente, fueron probados métodos de agrupamiento jerárquico aglomerativo de documentos para permitir una comparación en diferentes niveles de agrupamiento. Fue realizado un estudio de caso comparando archivos de texto que contienen información de las mallas curriculares de 40 cursos de pre-grado en Computación peruanos. En nuestro conocimiento, no han sido publicados estudios que pretendan realizar este tipo de análisis comparativo para el contexto peruano.

Los resultados obtenidos demuestran que el método propuesto permite encontrar relaciones ocultas entre las mallas curriculares. Por ejemplo, estas se reflejan en los agrupamientos de los cursos que tienen como foco alguna área específica de la Computación (ciencia de la Computación y/o sistemas de información).

La organización del artículo corresponde a las etapas del proceso de minería de datos indicadas en (Ebecken et al., 2003): en la sección 2 son descritos conceptos principales del área de minería de textos; los pasos de pre-procesamiento y extracción de patrones son explicados en la sección 3; en la sección 4 son presentados y discutidos los resultados obtenidos como parte de la etapa de pos-procesamiento; finalmente, en la sección 5, se indican algunas conclusiones y se describen áreas a ser exploradas en el futuro.

2. Minería de Textos

La minería de textos se refiere al proceso de descubrir conocimiento en grandes bases de datos textuales y puede ser considerada como una especialización del proceso de minería de datos. Mientras esta última trabaja con datos que poseen estructura definida, la minería de textos trabaja con datos no estructurados.

El agrupamiento de textos, una de las técnicas de minería de textos, consiste en la organización de un conjunto de documentos, basados en una medida de similaridad, en la cual los documentos de un mismo grupo son altamente similares entre si, pero diferentes cuando comparados con documentos de otros grupos (Manning et al., 2008).

Dentro de los métodos de agrupamiento, existen los denominados métodos jerárquicos, los cuales generan un conjunto de grupos anidados que son organizados en un árbol. Cada vértice (grupo) en el árbol (excepto los vértices hoja) es la unión de sus hijos (sub-grupos), y la raíz del árbol es el grupo que contiene todos los objetos (Tan et al., 2005).

Para obtener una representación visual del agrupamiento jerárquico es utilizado un dendograma. Esta estructura es un árbol con N hojas y altura $N - 1$, en la cual los documentos son dispuestos en el eje horizontal, mientras que el eje vertical indica la distancia (o la similaridad) con que los agrupamientos son creados.

El nodo raíz del dendograma representa todo el conjunto de datos, y cada nodo hoja es considerado un punto de los datos. Los nodos intermedios, por lo tanto, indican cuán próximos los objetos están unos de otros y la altura del dendograma usualmente expresa la distancia entre cada par de puntos de los datos o grupos, o entre un punto de datos y un grupo (Xu and Wunsch, 2008). Un ejemplo de dendograma es presentado en la Figura 1.

De acuerdo con (Zheng et al., 2006), el agrupamiento de textos se puede dividir en 3 categorías: basado en palabras, en conocimiento y en información. El agrupamiento basado en palabras representa los documentos por sus palabras clave. El basado en conocimiento considera un conocimiento creado manualmente. Finalmente, el agrupamiento basado en información es sensible al contexto, y considera frases, segmentos de texto y la semántica.

Los métodos más representativos para obtener agrupamientos jerárquicos son: *single-link*, *complete-link* y *average-link* (Tan et al., 2005). El método *single-link* define la proximidad de *clusters* como la distancia entre los dos puntos más próximos que están en diferentes *clusters*. Por otro lado, el método *complete-link* considera la distancia entre los puntos más alejados en diferentes *clusters* como la proximidad de *clusters*. Finalmente, el método *average-link* define la proximidad de *clusters* como la distancia media de pares de puntos de *clusters* diferentes. Una ilustración de las proximidades de estos métodos es presentada en la Figura 2.

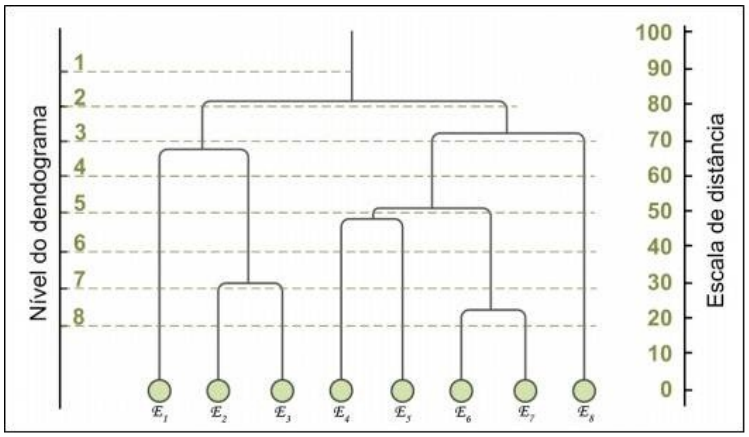

Figura 1: *Clustering* Jerárquico representado en un dendograma. Fuente (Metz, 2006).

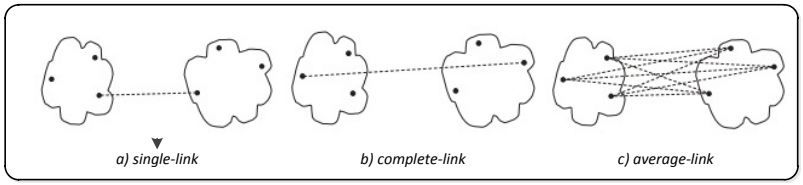

Figura 2: Criterios de proximidad de *clustering*. Fuente (Tan et al., 2005).

3. Metodología Propuesta para Comparación Automática de Mallas Curriculares y Estudio de Caso

Se propone utilizar una metodología de minería de textos para identificar similaridades entre mallas curriculares casi automaticamente y, de esta manera, facilitar un análisis entre ellas. En esta sección, se presenta una aplicación de esta metodología para un conjunto de mallas curriculares de carreras de Computación correspondientes a Perú. En la Figura 3 se presenta un diagrama general de este proceso. Los detalles del *corpus* de documentos utilizado y los pasos de la metodología propuesta para este proceso son descritos a seguir.

3.1. Pre-procesamiento

En esta sección son descritas las actividades utilizadas en el procesamiento de documentos para obtener una matriz atributo-valor [1] y, de esta manera, dejarlos en un formato adecuado para el proceso de extracción de patrones.

3.1.1. Extracción e Integración El *corpus* de documentos está conformado por 40 mallas curriculares correspondientes a carreras peruanas en Computación en las áreas de: Ciencias de

[1]Este tipo de matriz permite representar datos não estructurados como el caso de un documento de texto.

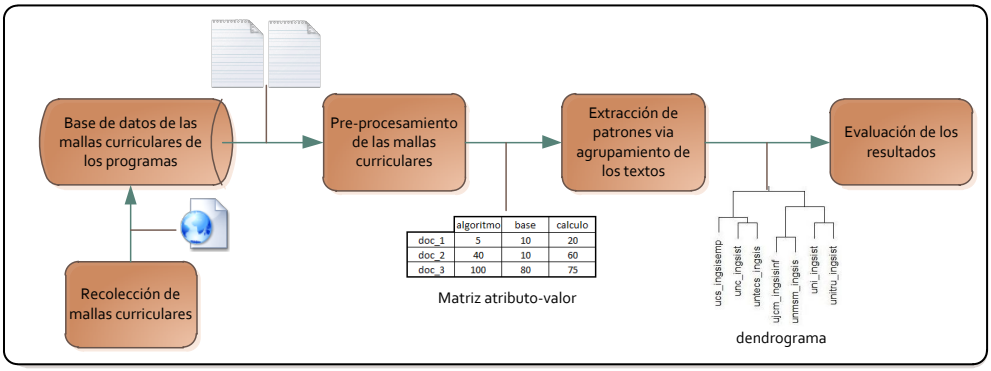

Figura 3: Metodología propuesta para comparación automática de mallas curriculares.

la Computación, Sistemas de Información e Ingeniería de Software[2].

Para cada malla curricular fue creado un archivo de texto con los nombres de los cursos ofrecidos (obligatorios y electivos). Cada archivo fue nombrado considerando la sigla de la universidad de la carrera y la sigla del nombre de la carrera. Por ejemplo, en el archivo uni_cc, uni es la sigla de la Universidad Nacional de Ingeniería y cc es la sigla de la carrera de Ciencia de la Computación.

3.1.2. Transformación Para transformar texto no estructurado en una tabla atributo-valor, fue utilizado el enfoque *bag-of-words*, en el cual cada palabra es considerada como un atributo y la frecuencia de la palabra en el texto es un valor del atributo. Para este proceso, primero se realizó una *tokenización* del contenido de cada malla curricular para separarla en palabras aisladas (sin considerar los signos de puntuación). Después, la técnica de *stemming* permitió transformar cada palabra de los textos en el radical que las originó, a través de eliminación de sufijos, siguiendo algunas reglas lingüísticas pre-establecidas.

3.1.3. Limpieza El enfoque *bag-of-words* presenta el problema que, al considerar cada palabra como un atributo, la dimensionalidad (tamaño) de la matriz atributo-valor es alta. Para tratar con este problema, se realizó una *eliminación de stopwords* para remover aquellas palabras que son muy comunes (preposiciones, artículos, etc.) y, por lo tanto, no significativas para el algoritmo de aprendizaje.

3.1.4. Selección y Reducción de datos Para sólo considerar en el análisis aquellas palabras (o *stems*) más representativos de las existentes y reducir aún más la dimensionalidad, fue realizada una selección de palabras considerando su frecuencia. Sólo aquellas palabras en la matriz atributo-valor con una frecuencia mayor a 10 y menor a 100 fueron seleccionadas.

En las actividades de transformación, limpieza y reducción de datos, fue utilizada la herramienta PreTexT (Soares et al., 2008), que implementa estas funcionalidades.

[2]Las mallas curriculares peruanas pertenecen a estas áreas a pesar de que las carreras sean llamadas de forma distinta.

3.2. Extracción de Patrones a través de Agrupamiento de Textos

El objetivo de esta sección es comparar mallas curriculares. De esta forma, fueron consideradas las siguientes actividades:

3.2.1. Elección de la Tarea La comparación de textos es considerada como una actividad descriptiva ya que se desea identificar conjuntos de mallas curriculares similares. Por este motivo, fue empleado un algoritmo de agrupamiento para identificar y analizar las similaridades entre las carreras.

3.2.2. Elección del Algoritmo Fueron seleccionados algoritmos de *clustering* jerárquico para poder visualizar los diferentes conjuntos de mallas curriculares en 2, 3,...,k grupos. Los algoritmos empleados fueron *single-link*, *complete-link* y *average-link* implementados en R (R Development Core Team, 2011).

3.2.3. Extracción de patrones Empleando los algoritmos seleccionados, fueron generados los dendogramas correspondientes a los experimentos planeados. En la Figura 4, son presentados los grupos generados considerando el algoritmo *average-link*; en la Figura 5 fue considerado el algoritmo *complete-link* y, finalmente, en la Figura 6 son presentadas las relaciones obtenidas usando el algoritmo *single-link*.

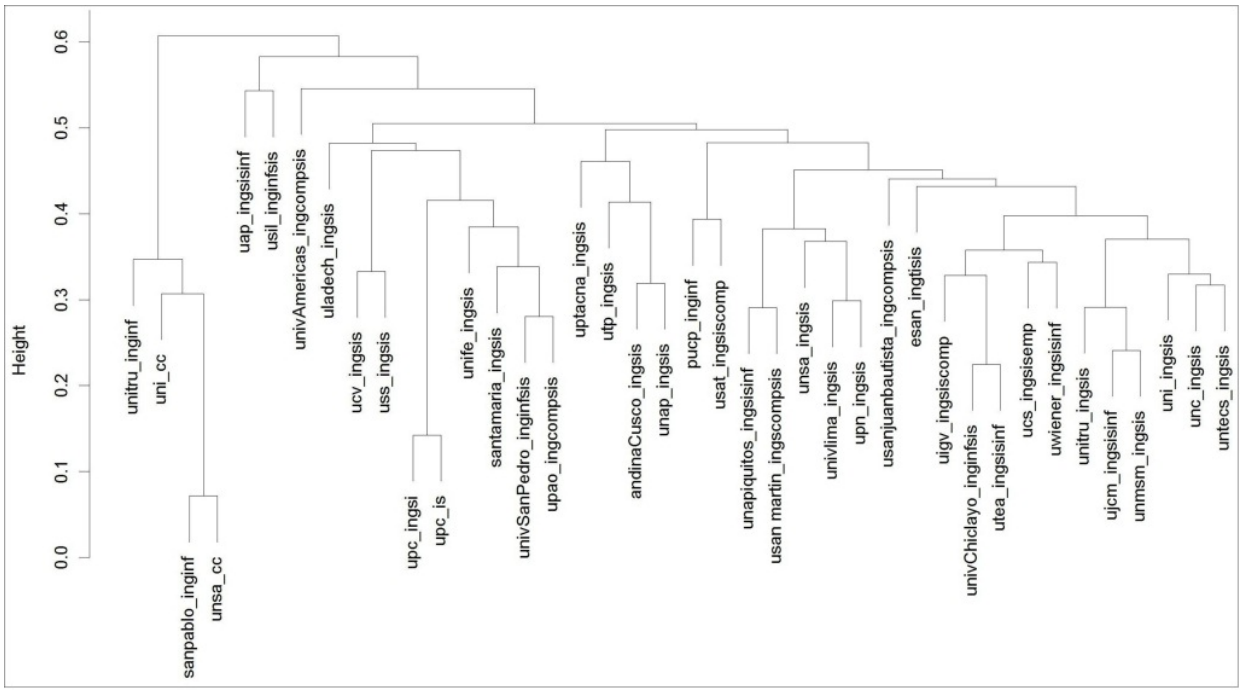

Figura 4: Dendograma de las mallas curriculares del Perú usando el algoritmo *average-link*

4. Pos-procesamiento, Análisis y Discusión de Resultados

En la Figura 4 se presentan dos grupos de mallas curriculares: el primero (de la izquierda) formado por los programas unitru_inginf, uni_cc, sanpablo_inginf y unsa_cc; y el segundo (de la

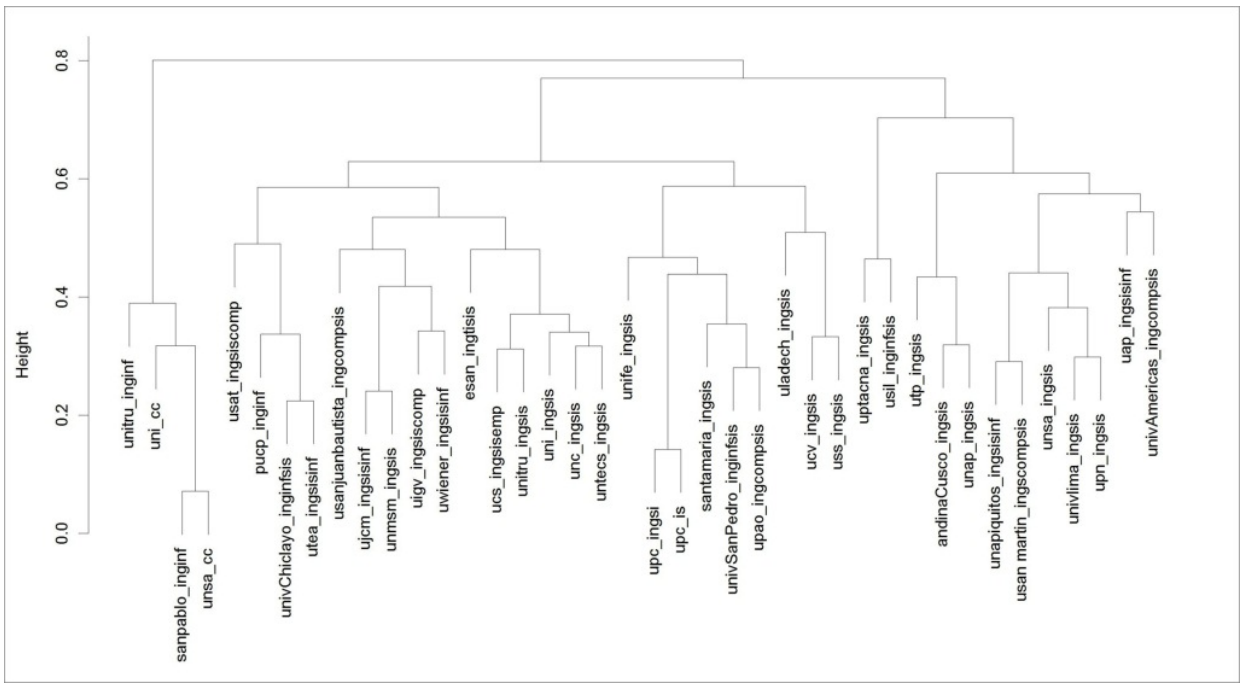

Figura 5: Dendograma de las mallas curriculares del Perú usando el algoritmo *complete-link*

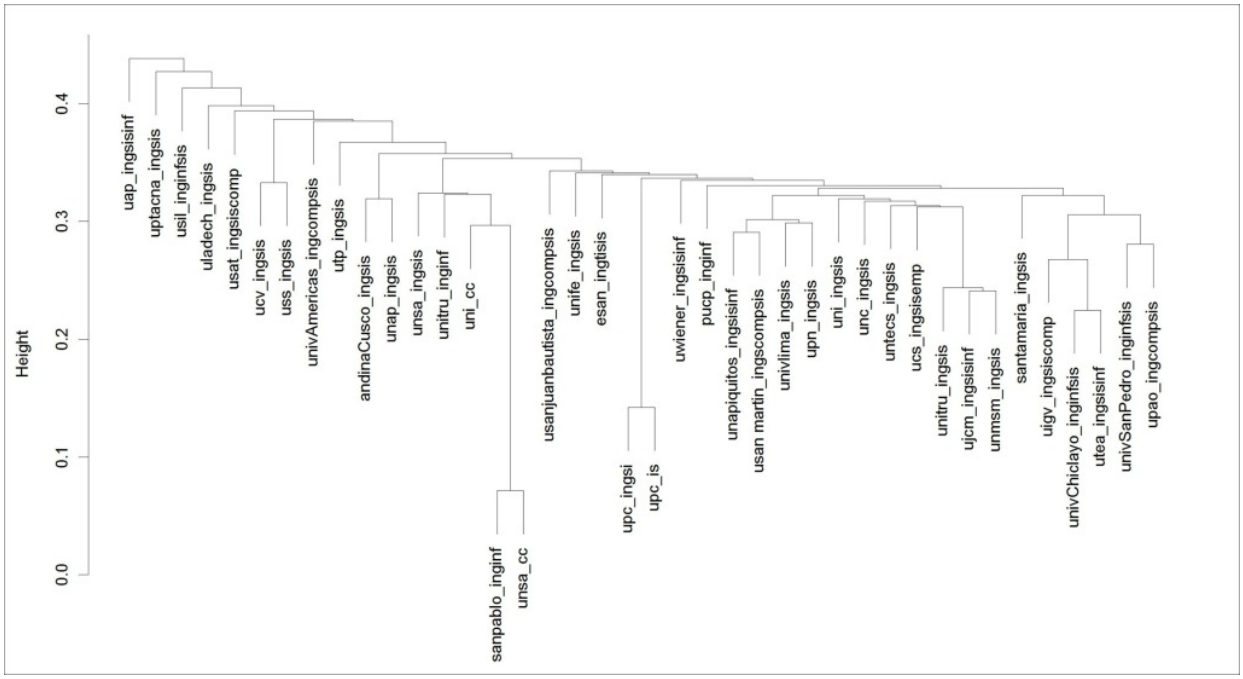

Figura 6: Dendograma de las mallas curriculares del Perú usando el algoritmo *single-link*

derecha) formado por todos los demás.

Por su parte, la Figura 5 confirma los agrupamientos anteriormente señalados. Las mismas 4 universidades del grupo de la izquierda se mantienen fuertemente unidas, tanto así que conforman su propio grupo, separado de las demás. Ya la Figura 6, si bien no diferencia esos 4 programas del resto, sí las muestra juntas en el centro del gráfico.

Las cuatro primeras mallas curriculares, regresando a la Figura 4, mantienen una fuerte relación, que es evidente por su proximidad en el primer grupo, como fue descrito previamente. Después de la revisión manual de los cursos ofrecidos en los programas de cada grupo, se puede observar que el primero tiene un mayor porcentaje de cursos orientados a Ciencia de la Computación, y el segundo esta más enfocada a Sistemas de Información.

5. Conclusiones y Trabajos Futuros

En este trabajo fue aplicado un método de agrupamiento jerárquico de textos para comparar mallas curriculares. El método fue empleado para descubrir relaciones ocultas entre las mallas curriculares de cursos de pre-grado en Computación, que corresponden a universidades del Perú. Los resultados obtenidos son evidencia de que este tipo de análisis ayuda en el estudio de mallas curriculares.

El método propuesto está basado en el proceso de minería de datos, con ciertas especificaciones para el manejo de textos como mallas curriculares, pero sin perder su generalidad. Es por esto que el método propuesto puede ser modificado en muchas secciones para adecuarse al tipo de estudio que se desee realizar. Dado que el objetivo del trabajo realizado era explorar la forma en que los programas de carrera se agrupaban cuando representados por sus mallas curriculares, fueron utilizados métodos de agrupamiento aglomerativo. Sin embargo, el método es lo bastante flexible como para permitir modificaciones en los algoritmos empleados y, así, obtener diferentes relaciones y resultados.

Pueden existir muchas más relaciones ocultas entre los diferentes grupos formados. Para hallarlas, se necesitaría de una revisión más profunda de las mallas curriculares de cada uno de los programas que representa cada grupo. Para ello, se requería del conocimiento de un especialista en diseño de mallas curriculares. Dado que los autores de este documento no poseen este conocimiento, no se puede presentar un mayor análisis de los resultados. Sin embargo, queda evidenciado que el método propuesto es útil para comparar programas de universidades, cuando representados por sus mallas curriculares.

El último paso en el proceso de minería de datos consiste en el **uso del conocimiento** extraído. En el caso de estudios como el aquí presentado, los resultados pueden servir, por ejemplo, para evidenciar problemas en el diseño de las mallas curriculares y proponer cambios en los programas.

Es conocido que no siempre el nombre de un curso corresponde con el contenido enseñado en el mismo. Por esto, para mejorar el proceso de comparación, sería mejor considerar el contenido total de los cursos, i.e, sus sílabos. Desafortunadamente, no todas las universidades publican este tipo de información en sus páginas web, de manera tal que sean de fácil acceso. Construir un corpus con estas características y aplicar el método descrito para descubrir nuevas relaciones puede considerarse como un trabajo futuro.

Los métodos jerárquicos aglomerativos explorados tienen la desventaja que al ingresar un nuevo documento para ser analizado es necesario realizar todo el proceso de extracción de patrones nuevamente. Cuando el corpus de documentos es pequeño, esto no constituye un gran problema, pero si el corpus fuese de un tamaño mayor, se presentarían dificultades en el tiempo de procesamiento. Para solucionar este problema, una alternativa es usar un enfoque jerárquico incremental, que permite actualizar el dendograma sin la necesidad de procesar todo el conjunto de textos nuevamente. Incorporar el uso de este tipo de algoritmos a la metodologia propuesta, se presenta como un camino a ser explorado.

Agradecimientos

Este trabajo fue realizado con el apoyo financiero de las agencias CAPES y CNPq.

Referencias

Biddle, R. L. and Tempero, E. D. (1996). Comparing a Computing Curriculum with the ACM/IEEE-CS Recommendations. In *Proceedings of the 1996 International Conference on Software Engineering: Education and Practice*, SEEP '96, pages 263–270, Washington, DC, USA. IEEE Computer Society.

Do Prado, H. A. and Ferneda, E. (2007). *Emerging Technologies of Text Mining: Techniques and Applications*. Information Science Reference (an imprint of IGI Global), Hershey, PA.

Ebecken, N. F. F., Lopes, M. C. S., and de Aragão Costa, M. C. (2003). Mineração de Textos. In Rezende, S. O., editor, *Sistemas Inteligentes - Fundamentos e Aplicações*, chapter 13, pages 338–370. Manole.

Manning, C. D., Raghavan, P., and Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Metz, J. (2006). Interpretação de clusters gerados por algoritmos de clustering hierárquico. Master's thesis, Instituto de Ciências Matemáticas e de Computação - USP - São Carlos.

Pereira, L. Z., de Albuquerque, J. P., and de S. Coelho, F. (2010). Uma Análise da Oferta e Abordagem Curricular dos Cursos de Bacharelado em Sistemas de Informação no Brasil. In *XVIII Workshop de Educação em Computação (WEI 2010), Anais do XXX Congresso da Sociedade Brasileira de Computação - CSBC 2010*, pages 897–906.

Prietch, S. S. and Pazeto, T. A. (2010). Mapeamento de Cursos de Licenciatura em Computação seguido de Proposta de Padronização de Matriz Curricular. In *XVIII Workshop de Educação em Computação (WEI 2010), Anais do XXX Congresso da Sociedade Brasileira de Computação - CSBC 2010*, pages 921–930.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Soares, M. V., Prati, R. C., and Monard, M. C. (2008). PreTexT II: Descrição da Reestruturação da Ferramenta de Pre-Processamento de Textos. Technical Report 333, ICMC-USP, São Carlos - SP.

Srivastava, A. and Sahami, M. (2009). *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC, 1st edition.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005). Cluster Analysis: Basic Concepts and Algorithms. In *Introduction to Data Mining*, chapter 8, pages 487–568. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition.

Xu, R. and Wunsch, D. C. (2008). Hierarchical Clustering. In *Clustering*, chapter 3, pages 31–62. John Wiley & Sons, Inc.

Zheng, Y., Cheng, X., Huang, R., and Man, Y. (2006). A Comparative Study on Text Clustering Methods. In *Proceedings of the Second International Conference on Advanced Data Mining and Applications*, ADMA 2006, pages 644–651, Xi'An, China.