Comparação de Grades Curriculares de Cursos de Computação Baseada em Agrupamento Hierárquico de Textos

Nils Murrugarra-Llerena, Fernando Alva-Manchego, Solange Oliveira Rezende

[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,solange}@icmc.usp.br

Abstract. *Comparing curricula allows an evaluation of the quality of the courses and programs in a university; however, this comparison is often done manually. In this paper, it is proposed an approach based on hierarchical clustering of texts to help in the task of comparing curricula. A case study was done using curricula from computing undergraduate courses from Peruvian and Brazilian universities. The results show that the approach proposed allows to discover hidden relations between curricula.*

Resumo. *A comparação de grades curriculares permite identificar e avaliar a qualidade dos cursos e programas em uma universidade; contudo, essa comparação é quase sempre feita manualmente. Neste trabalho, propõe-se uma abordagem baseada em agrupamento hierárquico de textos para auxiliar na tarefa de comparação de grades curriculares. Um estudo de caso foi realizado utilizando grades curriculares de cursos de graduação em computação de universidades peruanas e brasileiras. Os resultados obtidos mostram que a abordagem proposta permite descobrir relações ocultas entre grades curriculares.*

1. Introdução

A análise de grades curriculares é um ponto importante que permite avaliar a qualidade dos cursos correspondentes. Por exemplo, [Biddle and Tempero 1996] fazem uma comparação da grade curricular do curso de computação da Universidade de *Wellington* com recomendações da ACM/IEEE, especificamente do núcleo básico. Já no contexto brasileiro, os trabalhos de [Pereira et al. 2010] e [Prietch and Pazeto 2010] examinam grades curriculares para fazer uma análise de oferta e propostas de padronização de matriz curricular. Embora os objetivos desses trabalhos sejam distintos, eles possuem uma característica comum: o processo de comparação é manual.

Prém, nem sempre pode-se contar com o tempo necessário para realizar este tipo de estudo comparativo de maneira completamente manual e, portanto, seria útil contar com um recurso computacional que auxile à tarefa. Nesse contexto, as técnicas de mineração de textos se apresentam úteis pois permitem encontrar automaticamente relações ocultas entre distintos documentos.

A mineração de textos (MT) pode ser definida como a aplicação de métodos computacionais e técnicas sobre dados textuais para encontrar informação intrínseca e relevante, assim como conhecimento previamente desconhecido [Do Prado and Ferneda 2007]. Existem numerosas aplicações da MT, que incluem

pesquisa pioneira em análise e classificação de notícias, *email* e filtro de *spam*, extração hierárquica de tópicos de paginas web, extração e gestão automática de ontologias e inteligência competitiva [Srivastava and Sahami 2009].

O agrupamento (*clustering*) de textos é uma das técnicas usadas para MT, que é útil quando se quer agrupar documentos similares e torna a navegação entre eles mais fácil para usuários finais [Ebecken et al. 2003]. Existem dois tipos de agrupamento: particional e hierárquico. Este último, de interesse neste artigo, permite obter uma hierarquia, que é uma estrutura mais informativa do que o conjunto não estruturado de grupos obtidos pelos métodos particionais. Uma vantagem do agrupamento hierárquico é que não precisa da especificação do número de grupos e a maioria dos algoritmos hierárquicos que têm sido usados em recuperação de informação são determinísticos [Manning et al. 2008].

Para auxiliar na tarefa de comparação e análise de grades curriculares, neste artigo propõe-se empregar uma abordagem baseada em métodos de agrupamento de documentos. Mais especificamente, foi escolhido o agrupamento hierárquico aglomerativo para permitir uma comparação em diferentes níveis de agrupamentos. Para testar o método, foi realizado um estudo de caso comparando as grades curriculares de 40 cursos de graduação de computação peruanos e 33 brasileiros. Foram gerados arquivos de textos contendo informações de cada grade curricular em português.

Os resultados obtidos demonstram que é possível encontrar relações ocultas entre as grades curriculares usando o método proposto. Por exemplo, no caso do Peru, mostram-se agrupamentos entre cursos que têm como foco alguma área específica da computação (ciência da computação e/ou sistemas da informação). No caso das grades curriculares do Brasil, verificou-se que os cursos que pertencem a uma mesma instituição possuem uma maior relação entre eles do que com cursos de outras universidades ou institutos. E no caso das relações entre os documentos do Peru e do Brasil, apresentam-se semelhanças entre programas de computação peruanos e brasileiros com a mesma orientação na área de computação.

Na Seção 2 são descritos alguns conceitos de agrupamento de textos. O corpus de documentos relacionados à grades curriculares de cursos de computação, processo de pré-processamento e a extração dos padrões são descritos na Seção 3. A Seção 4 apresenta a discussão dos resultados. Na Seção 5 são descritas algumas situações nas quais uma análise comparativa deste tipo poderia ajudar na toma de decisões. Por fim, na Seção 6 descrevem-se as conclusões e trabalhos futuros.

2. Agrupamento de Textos

A mineração de textos refere-se ao processo de descobrir conhecimento em grandes bases de dados textuais e pode ser considerada como uma especialização do processo de mineração de dados. Enquanto a mineração de dados trabalha com dados com estrutura definida, a mineração de textos trabalha com dados não estruturados.

O agrupamento de textos, uma das técnicas de mineração de textos, consiste na organização de um conjunto de documentos, baseada em uma medida de similaridade, na qual os documentos de um mesmo grupo são altamente similares entre si, mas dissimilares em relação aos documentos de outros grupos [Manning et al. 2008].

Dentre os métodos de agrupamento consideram-se os métodos hierárquicos, os

quais geram um conjunto de grupos aninhados que são organizados em uma árvore. Cada vértice (grupo) na árvore (exceto os vértices folha) é a união de seus filhos (sub-grupos), e a raiz da árvore é o grupo contendo todos os objetos [Tan et al. 2005].

Para uma representação visual do agrupamento hierárquico é utilizado o dendrograma. Esta estrutura é uma árvore com N folhas e altura $N - 1$, na qual os documentos são dispostos no eixo horizontal, enquanto que o eixo vertical indica a distância (ou a similaridade) com que os agrupamentos são criados.

O nó raiz do dendrograma representa todo o conjunto de dados, e cada nó folha é considerado um ponto de dados. Os nós intermediários, portanto, indicam quão próximos estão os objetos uns dos outros e a altura do dendrograma usualmente expressa a distância entre cada par de pontos de dados ou grupos, ou entre um ponto de dados e um grupo [Xu and Wunsch 2008]. Um exemplo de dendrograma é apresentado na Figura 1.

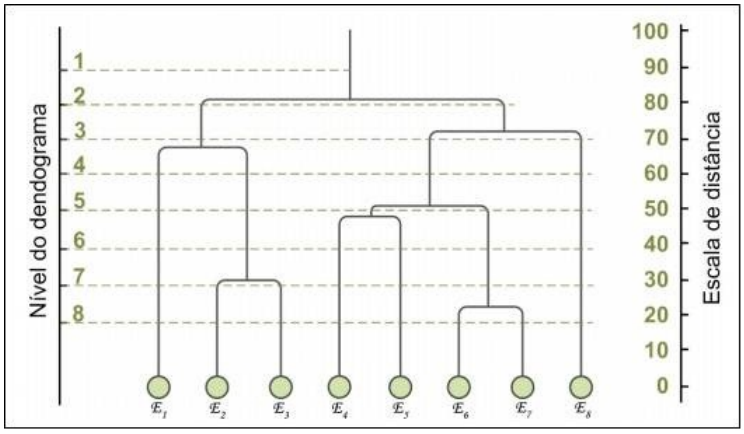

Figura 1. Clustering Hierárquico representado num dendrograma. Fonte [Metz 2006].

De acordo com [Zheng et al. 2006], o agrupamento de textos pode-se dividir em três categorias: agrupamentos baseado em palavras, em conhecimento e em informação. O agrupamento baseado em palavras representa os documentos pelas suas palavras-chaves. O agrupamento baseado em conhecimento considera um conhecimento manualmente criado. Finalmente, o agrupamento baseado em informação é sensível ao contexto, este considera frases, segmentos de texto e a semântica.

Os métodos mais representativos para obter agrupamentos hierárquicos são: *single-link*, *complete-link* e *average-link* [Tan et al. 2005]. O *single-link* define a proximidade de *clusters* como a proximidade entre os dois pontos mais próximos que estão em diferentes *clusters*. Enquanto, o *complete-link* considera a proximidade entre os pontos mais afastados em diferentes *clusters* como a proximidade de *clusters*. Finalmente, o *average-link* define a proximidade de *clusters* como a proximidade média de pares de pontos de *clusters* diferentes. Uma ilustração das proximidades desses métodos é apresentada na Figura 2.

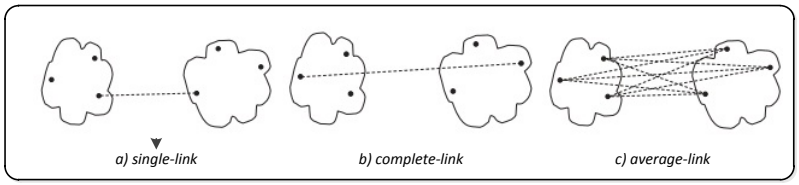

Figura 2. Critérios de proximidade de *clustering*. Fonte [Tan et al. 2005].

3. Abordagem Proposta para Comparação Automática de Grades Curriculares e Estudo de Caso

Para identificar similaridades entre grades curriculares e, assim, realizar uma análise entre elas de forma quase automática, propõe-se utilizar agrupamento hierárquico de textos. Nesta seção, apresenta-se a aplicação dessa técnica para um conjunto de grades curriculares de cursos de computação correspondentes a dois países: Peru e Brasil. Na Figura 3 apresenta-se um diagrama geral do processo seguido. Os detalhes do corpus de documentos utilizado, assim como os passos envolvidos na abordagem proposta para esse processo são descritos a seguir.

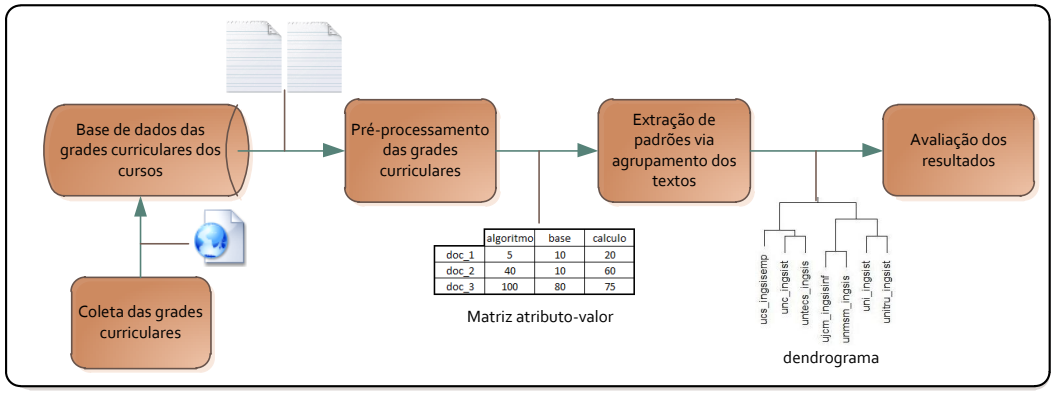

Figura 3. Abordagem Proposta para Comparação Automática de Grades Curriculares

3.1. Corpus de documentos

O corpus está conformado por 73 grades curriculares correspondentes a cursos de graduação em computação nas áreas de: ciências da computação, sistemas de informação e engenharia de software. De todas elas, 40 correspondem a universidades peruanas (totalidade de cursos com informação disponível na web) e 33 a universidades brasileiras (instituições com cursos de pós-graduação em computação - avaliação CAPES ≥ 3.

Para cada grade curricular foi criado um arquivo de texto com os nomes das disciplinas oferecidas (obrigatórias e optativas). Cada arquivo foi nomeado considerando

a sigla da universidade e/ou instituto do curso e a sigla do nome de curso. Por exemplo, no arquivo usp-icmc_cc, usp-icmc é a sigla do Instituto de Ciências Matemáticas e de Computação (ICMC) da Universidade de São Paulo (USP) e cc é a sigla do curso de Ciência da Computação.

Adicionalmente, considerando as diferenças entre as línguas dos países, foi necessário traduzir os nomes das disciplinas peruanas para o português. A tradução foi feita manualmente, pois a tradução automática realizada de forma direta (palavra por palavra) em muitos casos não permitiria uma comparação apropriada. Por exemplo, a disciplina de Trabalho de Conclusão de Curso no Brasil, possui diferentes nomes no Peru, como: Trabajo de Tesis, Trabajo de Graduación, Tesis e Proyecto de Fin de Carrera. Para poder obter concordância na tradução manual, foi criada uma tabela de equivalências entre os nomes de várias disciplinas que apresentam esse tipo de problema. Na Tabela 1 são apresentadas algumas dessas equivalências.

Nome em português	**Nome em espanhol**
Laboratório	Taller Laboratorio
Pesquisa Operacional	Investigación Operacional Investigación de Operaciones
Trabalho de Conclusão de Curso	Trabajo de Graduación Trabajo de Tesis Tesis Proyecto de Fin de Carrera
Estagio Pré-profissional	Práctica Pre profesional

Tabela 1. Algumas equivalências entre nomes de disciplinas

3.2. Pré-processamento das Grades Curriculares

Um dos passos de maior importância na mineração de textos consiste em processar os documentos para obter uma matriz atributo-valor e, assim, deixá-los no formato adequado para o processo de extração de padrões.

Para transformar texto não estruturado numa tabela atributo-valor, foi utilizada a abordagem *bag-of-words*, na qual cada palavra é considerada como um atributo e a frequência da palavra no texto é o valor do atributo. Para este processo, primeiro realizou-se a *tokenização* do conteúdo de cada grade curricular para separá-lo em palavras isoladas (sem considerar as sinais de pontuação). Depois disso, a técnica de *stemming* permitiu transformar cada palavra dos textos para o radical que o originou, através da remoção de sufixos, seguindo umas regras linguísticas pré-estabelecidas na ferramenta.

Um problema com a abordagem *bag-of-words* é que ao considerar cada palavra como um atributo a dimensionalidade (tamanho) da matriz atributo-valor é alta. Para lidar com esse problema primeiro realizou-se a *remoção de stopwords* para remover aquelas palavras que são muito comuns (preposições, artigos, etc.) e, portanto, não significativas para o algoritmo de aprendizado quando consideradas isoladamente. Finalmente, para só empregar na análise aquelas palavras (ou *stems*) mais representativas dentre as existentes e reduzir ainda mais a dimensionalidade, foi realizada uma seleção de palavras baseada

em frequência. Só aquelas palavras na matriz atributo-valor com uma frequência maior a 10 e menor do que 100 foram selecionadas. Neste processo, foi utilizada a ferramenta PreTexT [Soares et al. 2008], que implementa as funcionalidades descritas previamente.

3.3. Extração de Padrões via Agrupamento de Textos

A comparação de textos é considerada como uma atividade descritiva já que se deseja identificar conjuntos de grades curriculares similares. Por este motivo, foi empregado um algoritmo de agrupamento para identificar e analisar as similaridades entre os cursos.

Foi selecionado um algoritmo de *clustering* hierárquico para poder visualizar os diferentes conjuntos de grades curriculares em 2, 3,...,k grupos. O algoritmo empregado foi o *average link* (implementado no R [R Development Core Team 2011]), pois, em dados textuais, avaliações experimentais têm mostrado que ele é uma boa opção entre os algoritmos que adotam estratégias aglomerativas [Zhao et al. 2005].

Empregando o algoritmo selecionado, foram obtidos os dendrogramas correspondentes aos experimentos planejados. Na Figura 4, são apresentados os grupos gerados considerando só as grades curriculares dos cursos peruanos; na Figura 5 foram considerados só os cursos brasileiros e, finalmente, na Figura 6 são apresentadas as relações obtidas de todo o conjunto de documentos considerado no estudo de caso.

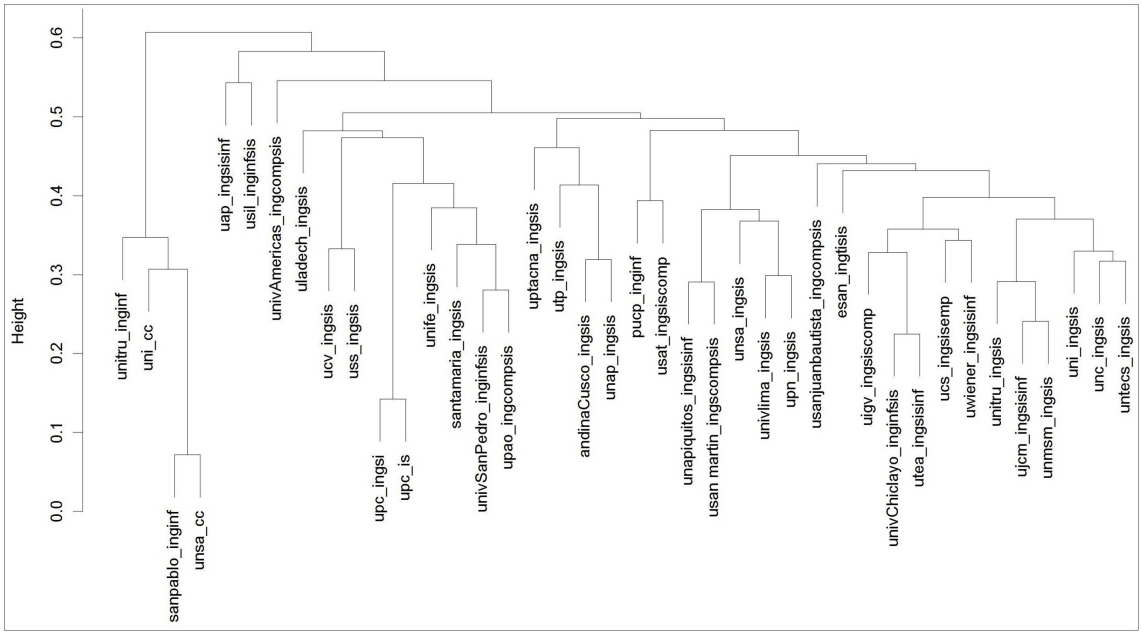

Figura 4. Dendrograma das grades curriculares do Perú

4. Análise e Avaliação dos Resultados

Na Figura 4 são apresentados dois grupos de grades curriculares: o primeiro (da esquerda) formado pelos programas unitru_inginf, uni_cc, sanpablo_inginf e unsa_cc, e o segundo (da direita) formado por todos os demais. Depois da revisão manual das disciplinas oferecidas nos programas de cada grupo, pode-se observar que o primeiro tem uma maior porcentagem de disciplinas orientadas a Ciências da Computação, e o segundo está mais focado em Sistemas de Informação.

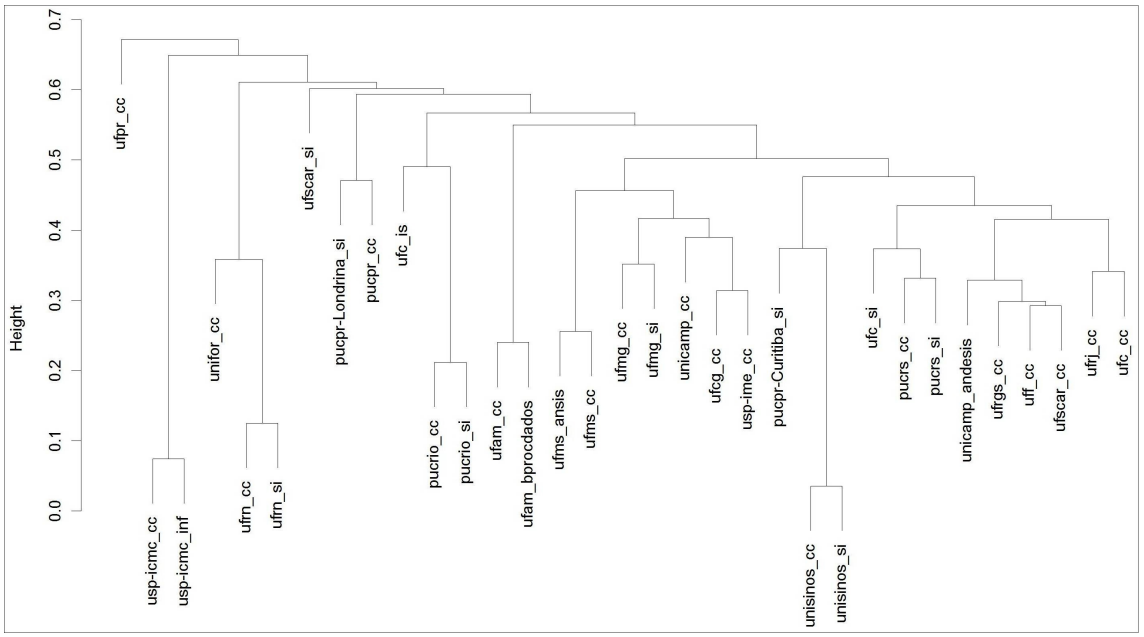

Figura 5. Dendrograma das grades curriculares do Brasil

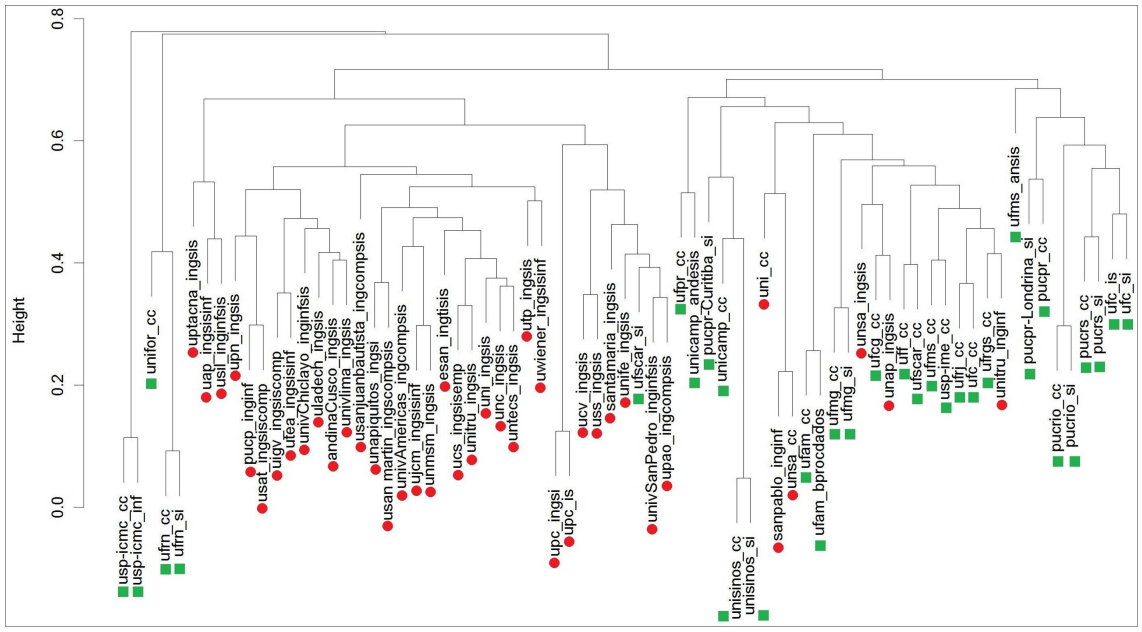

Figura 6. Dendrograma das grades curriculares do Perú (•) e do Brasil (■)

Na Figura 5, os grupos formados pelas grades curriculares brasileiras possuem a característica que os cursos que pertencem à mesma universidade ou instituto estão fortemente relacionados. Por exemplo: usp-icmc_cc com usp-icmc_inf, pucrio_cc com pucrio_si, entre outras. Quando for buscado o mesmo padrão na Figura 4, isso só seria possível com as grades curriculares da universidade *upc*, mas os cursos das universidade *unitru* e *uni* não possuem esse mesmo relacionamento. Analisando as disciplinas dos pares de cursos que apresentam esta característica, é apreciado que eles compartilham várias disciplinas dos primeiros anos e, depois, os estudantes recebem matérias de

especialização na área escolhida. Isto não acontece na *unitru* ou na *uni* e por isso essa separação entre os seus cursos.

Finalmente, na Figura 6, as grades curriculares formam quatro grandes grupos: o primeiro formado pelos dois cursos de computação do *icmc-usp*; o segundo, pelo curso de computação da *unifor* e os dois cursos da *ufrn*; no terceiro, 34 dos 40 cursos de universidades peruanas; e finalmente, no extremo direito, têm-se 27 dos 33 cursos brasileiros.

Considerando só o grupo do extremo direito, na Figura 6, pode ser observado que somente seis cursos de universidades peruanas estão relacionadas com os brasileiros: uni_cc, sanpablo_inginf, unsa_cc, unsa_ingsis, unap_ingsis e unitru_inginf. As quatro primeiras, voltando para a Figura 4, mantém uma forte relação, evidenciada pela sua proximidade no primeiro grupo, como foi descrito previamente. A característica que as grades curriculares destas universidades apresentam é que as disciplinas estão completamente orientadas à área de Ciências de Computação.

5. Como Utilizar o Conhecimento Extraído das Grades Curriculares

De modo geral, o conhecimento extraído automaticamente a partir das grades curriculares dos cursos pode ser utilizado para analisar situações e apoiar tomadas de decisões.

Particularmente, o conhecimento extraído pode ser usado para testar modificações nas grades curriculares atuais ou avaliar novas propostas de grades. Os encarregados em projetar a nova grade curricular (ou propor as modificações de uma existente) de um determinado curso poderia usar a comparação da sua proposta com as existentes atualmente. Uma análise como apresentada neste artigo permite contrastar o "novo" modelo com o "antigo", assim como as diferenças com outros cursos mais consagrados.

Outro uso pode ser a comparação das grades curriculares de um país com as de cursos de referência internacional, o que permite avaliar a situação desses cursos e, talvez, a qualidade dos mesmos. No estudo de caso realizado com cursos de universidades peruanas e brasileiras, por exemplo, pode-se observar que só quatro universidades peruanas possuem cursos de computação que têm muita semelhança com os cursos brasileiros, e que a grande maioria dos cursos peruanos é muito distante dos brasileiros em termos de grades curriculares. Uma possível interpretação deste resultado é que o perfil do egresso desses cursos é diferente do perfil brasileiro, i.e., pode estar mais relacionado com a realidade do mercado de trabalho peruano. Estes resultados podem servir, por exemplo, para evidenciar problemas no projeto das grades curriculares e propor mudanças nestes cursos. Comparações deste tipo também poderiam ser realizadas considerando propostas de grades curriculares de instituições amplamente renomadas, como o caso da ACM/IEEE.

Adicionalmente, este tipo de análise também pode ser aproveitado para realizar um estudo de evolução dos cursos na linha do tempo. Gráficos como os aqui apresentados, poderiam ser gerados periodicamente para analisar a progressão dos cursos e apoiar a tomada de decisões estratégicas.

6. Conclusões e Trabalhos Futuros

Neste trabalho foi aplicado um método de agrupamento hierárquico de textos para comparar grades curriculares em cursos de computação. O método foi aplicado para descobrir relações ocultas entre as grades curriculares de cursos de graduação, correspondentes a

universidades do Peru e do Brasil. Os resultados obtidos evidenciam que este tipo de análise ajuda na exploração das relações embutidas nas grades curriculares.

Foram descritas algumas situações nas quais este tipo de análise poderia ser aproveitada como suporte na toma de decisões sobre as características de uma grade curricular. O método proposto permite comparar quase automaticamente uma grade contra outras, o que poderia ser utilizado na estimação da qualidade de um curso. Contudo, sempre é necessário o conhecimento de um especialista na área para poder interpretar apropriadamente os resultados obtidos.

É sabido que nem sempre o nome de uma disciplina corresponde ao conteúdo ensinado na mesma. Por isso, para melhorar o processo de comparação, seria melhor considerar o conteúdo total das disciplinas, i.e., as suas ementas. Infelizmente, nem todas as universidades e/ou institutos disponibilizam este tipo de informação publicamente na web para que possa ser aproveitada. Construir um corpus com estas características e aplicar o método descrito para descobrir novas relações constitui-se num trabalho futuro.

Os métodos hierárquicos aglomerativos usados não são incrementais e isso indica que ao ingressar um novo documento é necessário realizar todo o processo novamente. Quando o corpus de documentos é pequeno isso não é um grande problema, mas se o corpus fosse de um tamanho maior apresentariam-se dificuldades no tempo de processamento. Utilizar uma abordagem hierárquica incremental para o auto-agrupamento de novas grades curriculares sem a necessidade de processar todo o conjunto de textos de novo, apresenta-se como um caminho a ser explorado.

Agradecimentos

Este trabalho foi realizado com apoio financeiro da CAPES e CNPq.

Referências

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.

Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical Clustering Algorithms for Document Datasets. *Data Mining Knowledge Discovery*, 10:141–168.

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.