

Ontologias Aplicadas ao Problema de Correlação Litológica no Domínio da Geologia do Petróleo*

Luan Fonseca Garcia, Joel Luis Carbonera, Mara Abel

¹ Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Porto Alegre – RS – Brasil

{lfgarcia, jlcarbonera, marabel}@inf.ufrgs.br

Abstract. *In this work we apply a domain ontology for developing a computational approach for the task of lithologic correlation, within the Petroleum Geology domain. In this context, a domain ontology is applied for imposing a rich and homogeneous structure to the visual descriptions of the domain objects that are the targets of this task. In our approach, we combine the use of ontologies with clustering techniques and sequence alignment algorithms, which are typically applied in DNA sequencing. A domain ontology with a vocabulary sufficiently expressive for allowing rich visual descriptions of the domain objects is a key aspect of our proposal.*

Resumo. *Neste trabalho, exploramos o uso de uma ontologia de domínio para o desenvolvimento de uma abordagem computacional para a tarefa de correlação litológica, no domínio da Geologia do Petróleo. Neste contexto, uma ontologia de domínio é utilizada para impor uma estrutura rica e homogênea às descrições visuais dos objetos de domínio que são o foco desta tarefa. Além da ontologia de domínio, a abordagem também combina técnicas de clusterização e algoritmos de alinhamento de sequências, tipicamente utilizados para realizar o sequenciamento de DNA. A disponibilidade de uma ontologia, com um vocabulário expressivo o suficiente para proporcionar uma descrição visual rica dos objetos do domínio é um aspecto chave desta proposta.*

1. Introdução

Domínios visuais são aqueles em que a resolução de problemas é fortemente baseada na aplicação de *conhecimento visual* dos especialistas. Consideramos conhecimento visual como sendo o conjunto de modelos mentais que suportam o processo de raciocínio sobre informação relacionada ao arranjo espacial e outros aspectos visuais das entidades de domínio [Lorenzatti et al. 2009, Carbonera et al. 2011]. Este trabalho insere-se no contexto do projeto *Obaitá*, desenvolvido pelo grupo *BDI* (grupo de bancos de dados inteligentes da UFRGS). Neste projeto, investigamos abordagens integradas para aquisição, modelagem, representação e raciocínio sobre conhecimento visual. Um dos resultados esperados é uma ontologia para o domínio (visual) da *Estratigrafia Sedimentar*, que viabilize o desenvolvimento de diversos sistemas baseados em conhecimento, que operem sobre uma mesma conceitualização deste domínio. Em [Lorenzatti et al. 2009] são apresentados os passos iniciais em direção a este resultado, enquanto em [Carbonera 2012] esta ontologia é expandida, utilizando-se a abordagem descrita em [Carbonera et al. 2012].

*Este trabalho foi desenvolvido com recursos do CNPq e do programa PRH PB-217, mantido pela Petrobras e Agência Nacional do Petróleo (ANP). Também gostaríamos de agradecer à Endeeper pela disponibilização dos dados utilizados para a realização do trabalho.

Atualmente, investigamos abordagens para a tarefa de correlação litológica, no domínio da Estratigrafia Sedimentar, que se beneficiem da ontologia desenvolvida. A Estratigrafia Sedimentar é uma sub-área da Geologia que estuda as camadas que compõem a Terra e busca determinar como ocorreu a formação dessas camadas. Neste domínio, na tarefa de correlação litológica o geólogo busca reconhecer a mesma fácies sedimentar (Figura 1) em duas ou mais seções estratigráficas diferentes, mesmo que espacialmente distantes entre si. Ou seja, nesta tarefa o geólogo investiga a continuidade lateral de fácies sedimentares em subsuperfície, onde não é possível realizar observação direta destas unidades. A correlação permite determinar a distribuição espacial e o volume das rochas que subsidiam a avaliação de economicidade dos reservatórios de petróleo.

Para alcançar este objetivo, o geólogo inicia descrevendo visualmente *corpos de rocha*, tal como o *testemunho de sondagem* apresentado na Figura 1. A descrição destes corpos envolve discretizá-los em *fácies sedimentares* e descrever todos os atributos visuais que caracterizam cada uma delas. Neste contexto, a fácies sedimentar é uma dada porção de um corpo de rocha, visualmente distinguível das porções adjacentes. Além dos atributos que as caracterizam, as fácies sedimentares possuem uma ou mais *estruturas sedimentares*, que correspondem a padrões geométricos externamente visíveis, que indicam padrões de arranjos espaciais internos dos grãos que constituem uma fácies.

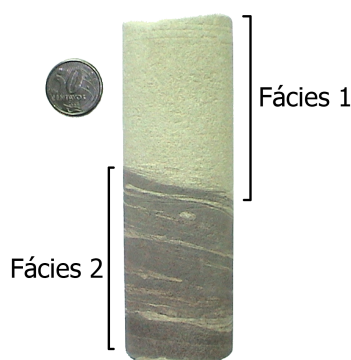


Figura 1. Trecho de testemunho de sondagem, com duas fácies distintas. Adaptado de [Lorenzatti 2009].

A continuidade das unidades de rochas é julgada pela similaridade entre fácies, visto que elas são as porções discretizadas de rocha passíveis de observação direta e visualmente distinguíveis das demais. A correspondência entre corpos de rocha pode ser parcial, assim como a similaridade entre as unidades relacionadas; como exemplificado na Figura 2, onde a fácies 4 está presente nas seções *A* e *B*, mas não está presente na seção *C*. É importante notar que, a correlação é estabelecida não apenas com base na identificação do mesmo tipo de rocha nos diferentes poços, mas principalmente pelo sequenciamento semelhante de diferentes tipos de rochas em cada um dos poços.

Atualmente, a correlação é realizada a partir de múltiplos registros textuais distintos, sem uma estrutura padrão, capturados por geólogos diferentes, sem apoio de um vocabulário padrão. Essas condições fazem com que, no atual estado da arte, a tarefa de correlação litológica careça de métodos automáticos para processamento em larga escala. Neste trabalho, nós propomos que a correlação pode ser realizada com métodos computacionais de correlação automáticos, desde que as descrições de rochas sejam orientadas

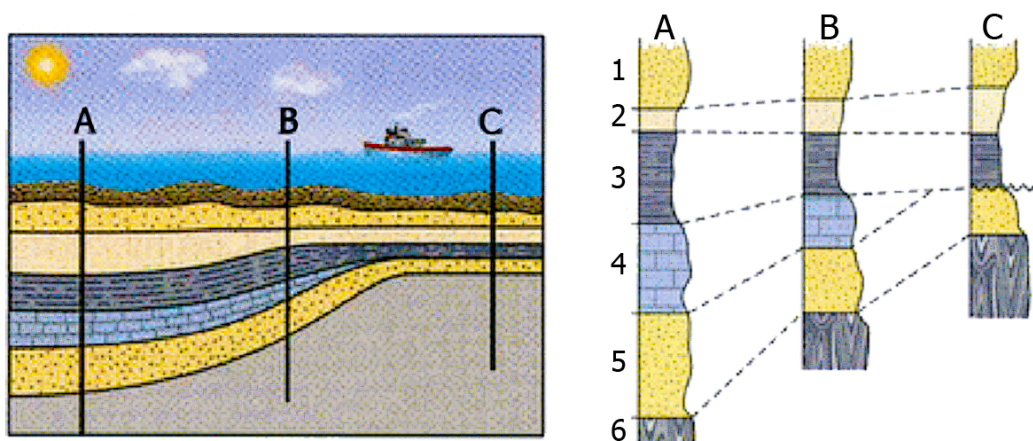


Figura 2. Representação de uma correlação litológica entre três seções estratigráficas distintas (A, B e C), envolvendo seis fácies. Adaptada de [Parsons 2013].

por uma ontologia de domínio bem fundamentada, tal como a incorporada no software Strataledge^{®1} e que é apresentada em detalhes em [Carbonera 2012]. A utilização de uma ontologia de domínio expressiva, permite impor uma estrutura formal homogênea às descrições dos objetos de domínio, viabilizando a descrição de um conjunto rico de informações, através de um vocabulário formal bem definido. Essas informações, capturadas de forma uniforme e não ambígua, permitem a comparação entre porções descritas das unidades de rocha espacialmente distintas, que não são influenciadas pelo uso de diferentes vocabulários e estilos descritivos, suportando a correlação entre elas. Além disso, considerando que a ontologia especifica a conceitualização compartilhada no domínio, o seu uso para descrição dos objetos do domínio permite que os sistemas processem as informações acerca desses objetos de um modo que se aproxime da forma como os geólogos os concebem. Assim, partimos da hipótese de que abordagens automáticas para correlação podem se beneficiar do uso de ontologias, oferecendo resultados geologicamente mais significativos.

2. Abordagem proposta

Diversas abordagens têm sido propostas para lidar computacionalmente com o problema da correlação litológica. Uma abordagem que tem se revelado promissora neste sentido, tal como a adotada em [Waterman and Raymond Jr 1987], envolve a aplicação de *algoritmos de alinhamento de seqüências*. Estes algoritmos vêm sendo utilizados com sucesso na tarefa de alinhamento de seqüências de DNA no domínio da bioinformática. Entre estes algoritmos, destaca-se o algoritmo de programação dinâmica *Smith-Waterman*, que possui resultado ótimo para o alinhamento de seqüências locais.

Segundo [Chao and Zhang 2009], o algoritmo de Smith-Waterman parte de uma seqüência $A = a_1 a_2 \dots a_m$ e uma seqüência $B = b_1 b_2 \dots b_n$, que podem ter tamanhos diferentes. De modo geral, o alinhamento entre estas seqüências é obtido pela inserção de lacunas (representadas pelo caractere “-”) em ambas, representando deslocamentos entre os segmentos similares, de tal modo que o tamanho final de ambas seja idêntico, sendo

¹<http://www.endeeper.com/products/software/strataledge>

que não pode haver alinhamentos entre lacunas. A Figura 3 apresenta dois exemplos de alinhamentos de sequências de DNA resultantes da aplicação deste algoritmo. O funcionamento detalhado deste algoritmo foge ao escopo deste artigo, mas pode ser encontrado em [Chao and Zhang 2009].

| | |
|--|--|
| <pre> -ATACATGTC--T G-TAC--GTCGG- </pre> | <pre> -----AATGCCATTGAC----GG CAGCC--T--C---G-CTTAG-- </pre> |
| (a) | (b) |

Figura 3. Dois exemplos (a e b) de alinhamentos de pares de sequências de DNA realizados pelo algoritmo de Smith-Waterman.

Para aplicar o algoritmo de Smith-Waterman sobre duas sequências, deve haver uma maneira de comparar elementos de ambas, determinando quando eles são equivalentes. Quando aplicado no alinhamento de sequências de DNA, este algoritmo opera sobre strings construídas a partir de um alfabeto finito – as quatro letras que representam as bases nitrogenadas básicas do DNA. Neste caso a comparação entre elementos de duas sequências é trivial, bastando verificar se os dois elementos são a mesma letra (representando o mesmo tipo de base nitrogenada). Por outro lado, quando aplicado ao problema de correlação litológica, o algoritmo deve ser capaz de operar sobre sequências de fácies sedimentares. Em relação a este ponto, [Griffiths and Bakke 1990] afirma que aplicações convencionais deste algoritmo para o problema em foco devem determinar uma forma de codificar a informação das fácies sedimentares de um modo análogo ao que ocorre no caso do sequenciamento de DNA, utilizando um conjunto finito de símbolos bem definidos, que podem ser comparados pelo algoritmo. Em nossa abordagem, adaptamos o algoritmo para que sejam comparados os clusters aos quais as duas fácies comparadas pertencem. Para isso, antes de aplicar o algoritmo de alinhamento, utilizamos um algoritmo de clusterização² sobre um *dataset* formado pelo conjunto de descrições das fácies que se pretende correlacionar. Lembrando que estas descrições são realizadas com suporte da ontologia de domínio. A partir desse passo, é obtido um modelo *clusterizador* que é capaz de classificar instâncias de fácies (incluindo instâncias não consideradas durante o treinamento do clusterizador). Este clusterizador, por sua vez, é utilizado pelo algoritmo de alinhamento para comparar se duas fácies pertencem à um mesmo *cluster*. Se as duas fácies, em corpos de rocha distintos, estão no mesmo cluster, consideramos que elas são equivalentes e que podem ser alinhadas. Atualmente, o treinamento do clusterizador é realizado através da API Weka³, aplicando o algoritmo *EM* (*expectation-maximization*) [Witten et al. 2011]. A Figura 4 representa esquematicamente a abordagem proposta.

A ontologia de domínio utilizada [Carbonera 2012] descreve o conceito de fácies através de 23 atributos. Na fase de *conversão* das descrições de corpos de rocha para o dataset de treinamento do clusterizador (Figura 5), cada instância f do conceito fácies na ontologia é convertida em um vetor de características v . Cada posição p neste vetor representa um atributo descritivo a que caracteriza o conceito de fácies na ontologia, de modo que cada valor v_p do vetor v representa o valor específico que f possui para o respectivo

²De acordo com [Witten et al. 2011], clusterização é uma técnica de mineração de dados utilizada para determinar um conjunto de categorias ou agrupamentos (*clusters*) a partir de um conjunto de dados sem classificação prévia.

³<http://www.cs.waikato.ac.nz/ml/weka/>

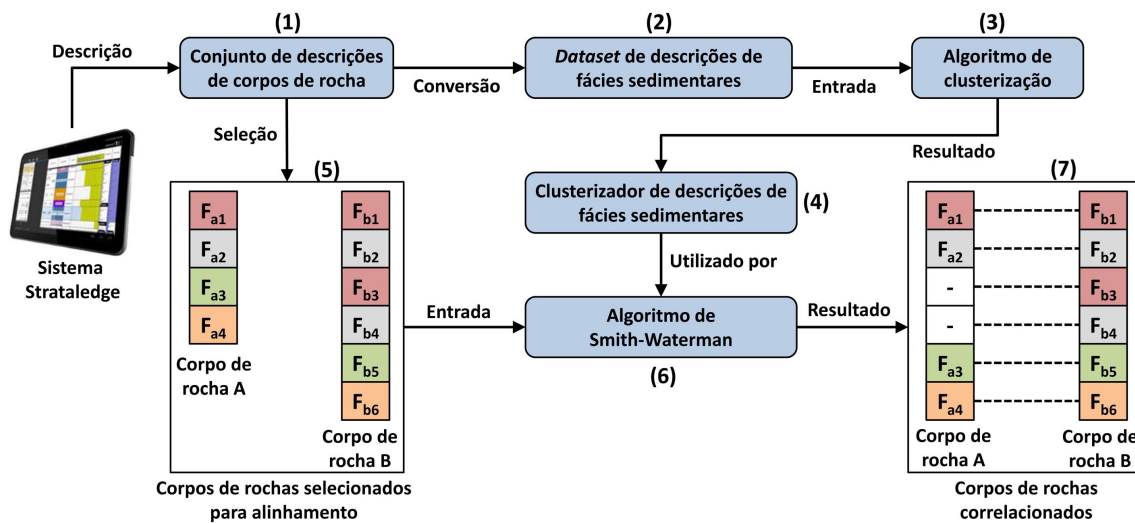


Figura 4. Representação dos procedimentos realizados na abordagem proposta. Cada F_{ij} representa uma fácies j no corpo de rocha i .

atributo. Além disso, em nosso caso, o vetor v também possui uma posição especial que representa a relação *temEstrutura* entre a instância de fácies e uma instância de estrutura sedimentar. Esta posição especial recebe como valor o tipo específico da instância e de estrutura sedimentar relacionada à fácies f . Assim, o dataset de treinamento é um conjunto V de vetores de características v , cada qual representando uma instância f do conceito de fácies, descrita pela ontologia.

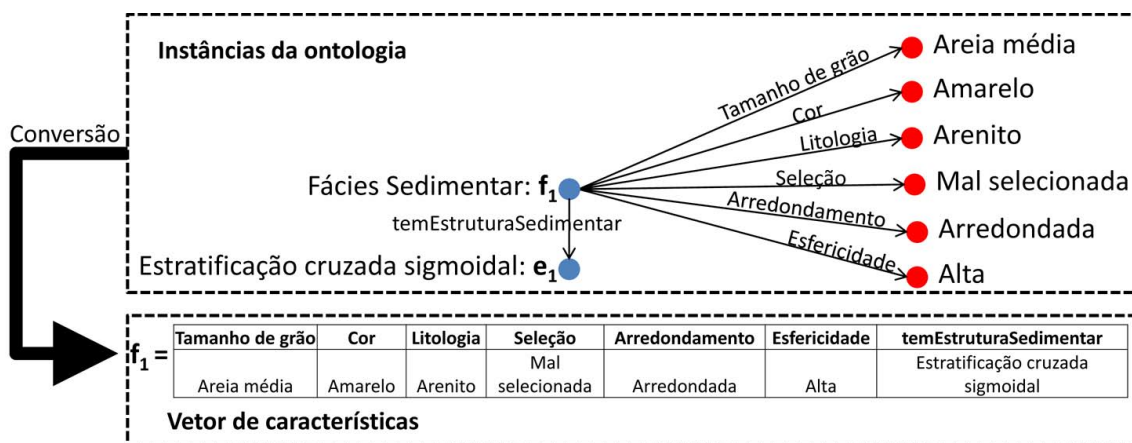


Figura 5. Representação do processo de conversão de instâncias de fácies sedimentares em vetores de características, considerando um conjunto reduzido de atributos.

3. Considerações finais

Neste trabalho, apresentamos uma abordagem computacional para correlação litológica automática no domínio da Estratigrafia Sedimentar. Esta abordagem está alinhada às abordagens mais promissoras oferecidas pela literatura. A nossa principal contribuição reside no uso de uma ontologia de domínio para conferir uma estrutura formal homogênea às descrições dos objetos do domínio. Assim, considerando que a ontologia captura de

modo formal e explícito a conceitualização compartilhada pela comunidade, ela permite que os usuários descrevam os objetos do domínio de um modo padrão, formal e com uma estrutura rica de informações. Isto viabiliza o tratamento computacional destas descrições e permite que nossa abordagem processe informações sobre os objetos do domínio de um modo que se aproxime da forma como os geólogos os conceitualizam.

Na fase atual deste trabalho, com o auxílio de especialistas do domínio, estamos coletando um conjunto de descrições de corpos de rocha reais. Nos próximos passos deste projeto será investigado como considerar a importância relativa dos atributos da fácies durante a clusterização, de um modo que seja possível determinar similaridades geologicamente mais significativas entre fácies sedimentares.

Referências

- Carbonera, J., Abel, M., dos Santos Scherer, C. M., and Bernardes, A. (2012). Abordagem para aquisição de conhecimento visual e refinamento de ontologias para domínios visuais. In Vieira, R., Guizzardi, G., and Fiorini, S. R., editors, *Proceedings of Joint V Seminar on Ontology Research in Brazil and VII International Workshop on Metamodels, Ontologies and Semantic Technologies*, volume 776.
- Carbonera, J. L. (2012). Raciocínio sobre conhecimento visual: Um estudo em estratigrafia sedimentar. Master's thesis, Universidade Federal do Rio Grande do Sul (UFRGS).
- Carbonera, J. L., Abel, M., Scherer, C. M. S., and Bernardes, A. K. (2011). Reasoning over visual knowledge. In Vieira, R., Guizzardi, G., and Fiorini, S. R., editors, *Proceedings of Joint IV Seminar on Ontology Research in Brazil and VI International Workshop on Metamodels, Ontologies and Semantic Technologies*, volume 776.
- Chao, K.-M. and Zhang, L. (2009). *Sequence comparison: theory and methods*, volume 7. Springer.
- Griffiths, C. and Bakke, S. (1990). Interwell matching using a combination of petrophysically derived numerical lithologies and gene-typing techniques. *Geological Society, London, Special Publications*, 48(1):133–151.
- Lorenzatti, A. (2009). Ontologia para domínios imagísticos: Combinando primitivas textuais e pictóricas. Master's thesis, UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL.
- Lorenzatti, A., Abel, M., Nunes, B. R., and Scherer, C. M. S. (2009). Ontology for imagistic domains: Combining textual and pictorial primitives. In Heuser, C. A. and Pernul, G., editors, *ER Workshops*, volume 5833 of *Lecture Notes in Computer Science*, pages 169–178. Springer.
- Parsons, S. B. (2013). Historical geology. Disponível em: <http://www.ocean.odu.edu/~spars001/geology_112/laboratory/session_04/handout.html>. Acesso em: 14 de Julho de 2013.
- Waterman, M. S. and Raymond Jr, R. (1987). The match game: new stratigraphic correlation algorithms. *Mathematical geology*, 19(2):109–127.
- Witten, I. H., Frank, E., and Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier.