Similaridade entre documentos semi-estruturados

Rodrigo Gonçalves, Ronaldo dos Santos Mello

[1]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88040-900 – Florianópolis – SC – Brasil

{rodrigog,ronaldo}@inf.ufsc.br

Resumo. *Neste artigo faz-se uma breve revisão da literatura sobre um tema atual de pesquisa em banco de dados: similaridade entre documentos semi-estruturados. Primeiramente, introduz-se conceitos básicos e algoritmos relacionados ao tema, sugerindo uma taxonomia para os trabalhos existentes. Alguns trabalhos relacionados são revisados, destacando-se sua abordagem geral e particularidades. Conclui-se com uma comparação entre os trabalhos, analisando suas contribuições e limitações, além de sugerir alguns tópicos para pesquisa futura.*

1. Introdução

A análise de similaridade envolve comparar dados com o objetivo de determinar o quão similares eles são. As justificativas para a realização desta análise são inúmeras, como a integração de instâncias heterogêneas que representam o mesmo dado e a eliminação de dados redundantes em um repositório de dados.

Este artigo apresenta uma revisão a respeito de similaridade entre documentos semi-estruturados, enfatizando documentos XML. Ele está organizado em duas grandes partes: a primeira introduz o tema similaridade em dados semi-estruturados e apresenta uma taxonomia sobre o assunto. A segunda revisa e compara alguns trabalhos relacionados, estabelecendo um *estado da arte* e sugerindo algumas idéias para trabalhos futuros derivados dos tópicos em aberto encontrados. As contribuições deste artigo são a taxonomia proposta para o assunto, o comparativo de trabalhos existentes e a discussão de questões em aberto.

2. Conceitos e Métricas

Em documentos semi-estruturados, têm-se dois tipos básicos de similaridade: estrutural e por conteúdo. Na estrutural, abstrai-se o conteúdo e considera-se apenas a estrutura. Na similaridade por conteúdo, além da estrutura, a informação contida é usada para determinar a similaridade. Existe ainda uma terceira abordagem, onde analisa-se apenas o conteúdo dos documentos, ignorando aspectos relacionados à estrutura dos mesmos (faz-se uma *"planificação"* dos documentos).

A maioria das funções de similaridade na literatura são métricas que tratam instâncias de dados em um espaço métrico, estabelecendo distâncias neste espaço para definir um grau de similaridade. As cinco principais categorias de métricas utilizadas para documentos semi-estruturados são:

1. ***Espaço Vetorial*** - Esta métrica é usada principalmente para comparar objetos complexos. Os atributos dos objetos são convertidos em um ou mais vetores baseados

nos valores. Então, através de fórmulas como a *distância Euclidiana*, um grau de similaridade é calculado [Carvalho and da Silva 2003];

2. ***Comparação entre strings*** - Uma das abordagens mais utilizadas é a *distância de edição*. Nesta métrica, operações como inserção e remoção de caracteres são executadas sobre uma *string*, com o objetivo de convertê-la em outra *string*. O custo em termos de operações para a transformação define a similaridade [Navarro 2001]. Outra métrica popular é a *Jaccard Similarity*, que transforma cada *string* em um conjunto de *tokens* e estabelece a similaridade por meio de operações entre os conjuntos;
3. ***Comparação entre árvores*** - Objetos estruturados e semi-estruturados podem ser representados como estruturas de dados em árvore. Uma das métricas mais utilizadas para comparar árvores é a *distância de edição entre árvores*. Esta métrica segue o mesmo princípio da distância de edição para *strings*, definindo um conjunto de operações necessário para transformar uma árvore na outra. As operações podem ter pesos específicos ou não [Nierman and Jagadish 2002];
4. ***Séries Temporais*** - Nesta métrica, a idéia básica é converter a estrutura de um objeto em uma série numérica. Então, um sinal (*onda*) é gerado a partir desta. Os sinais representando os objetos analisados são então comparados, determinando um grau de similaridade [Buttler 2004];
5. ***Freqüência de valores*** - Nesta abordagem, a freqüência de ocorrência dos valores contidos nos objetos sendo comparados determina sua similaridade. Uma métrica bastante empregada é a IDF (*Inverse Document Frequency*), onde a freqüência dos valores incomuns é usada para definir a similaridade [Weis and Naumann 2004].

3. Taxonomia sobre similaridade de dados

A análise de similaridade de dados pode ser classificada de acordo com o tipo de dado que ela trata. Uma primeira categoria é a de *dados simples*, relacionada às métricas que tratam de valores atômicos (*strings*, números, etc.). Como exemplos, temos a *distância de edição* e os MAVs (*Metrics for Atomic Values*) [Dorneles et al. 2004].

Dados complexos são aqueles compostos por dados simples e/ou outros dados complexos. Incluem-se aqui tuplas de um banco de dados relacional, estruturas de lista, etc. Um bom exemplo de métricas para valores complexos são as MCV (*Metrics for Complex Values*) [Dorneles et al. 2004].

A *comparação estrutural* procura por similaridade entre objetos estruturados ou semi-estruturados, analisando apenas quais valores (nome de um livro, idade de uma pessoa, etc.) são armazenados e a forma como estão organizados, sem considerar o conteúdo dos valores. Este tipo de comparação possui várias implementações, dependendo dos aspectos considerados e dos objetivos (clusterização, detecção de duplicatas, etc).

Por último, trabalhos que realizam uma *comparação completa* consideram, além da estrutura, o conteúdo dos valores contidos nos objetos comparados. As abordagens variam ao tratar a estrutura e valores de forma independente ou em conjunto.

Em todas as categorias os aspectos semânticos dos dados comparados podem ou não ser levados em conta. Nas abordagens atuais, verificou-se que muitos consideram aspectos sintáticos, porém não identificou-se a consideração de semântica nos trabalhos analisados.

4. Estado da arte

Nesta seção revisa-se brevemente alguns trabalhos a respeito de análise de similaridade que direta ou indiretamente tratam de dados semi-estruturados. A apresentação dos trabalhos está organizada de acordo com a taxonomia sugerida na seção anterior.

4.1. Dados simples e complexos

Na proposta de A. Broder [Broder 1998] introduz-se duas notações matemáticas para o domínio de documentos textuais: *similaridade* e *contenção*. Os documentos são tratados como conjuntos de seqüências de *tokens*, denominados *shingles*. Através de operações aplicadas aos conjuntos de *shingles* dos documentos, um grau de similaridade é determinado. Um exemplo de operação sugerida é determinar a relação entre o tamanho dos conjuntos intersecção e união dos conjuntos. Quanto mais próximo de 1 for o valor encontrado, mais similares os documentos são.

No trabalho de Joyce Carvalho et al. [Carvalho and da Silva 2003] propõe-se uma abordagem para identificar objetos similares usando um espaço vetorial. Quatro métodos para estabelecer e comparar os objetos no espaço vetorial são introduzidos. No primeiro método, o vetor representando o objeto é construído levando em conta todos os valores de atributos do objeto. A similaridade é calculada pelo *cosseno* entre os vetores. No segundo método, apenas um subconjunto dos atributos é utilizado para construir o vetor. O terceiro método constrói um vetor independente para cada atributo. A similaridade é calculada pela soma das similaridades dos atributos dos objetos encontrada em cada vetor. O último método segue o mesmo princípio do terceiro, porém leva em conta apenas um subconjunto dos atributos dos objetos.

Carina Dorneles et al. [Dorneles et al. 2004] propõe um conjunto de métricas para manipular coleções de dados em documentos XML. Dois tipos de métricas são definidas: *MAVs (Metrics for Atomic Values)* e *MCVs (Metrics for Complex Values)*. As métricas *MAVs* são aplicadas a valores atômicos considerando o tipo dos dados (*strings*, números, etc.). As MCVs definem métricas para valores complexos: tuplas, coleções e conjuntos.

4.2. Similaridade estrutural

Na proposta de Sudarshan Chawathe et al. [Chawathe and Garcia-Molina 1997] apresenta-se uma técnica para detectar modificações em dados estruturados em árvores. A técnica compara dois documentos como um grafo bipartido (cada documento ocupando metade do grafo). Conecta-se então cada nodo de um lado do grafo a pelo menos um nodo do outro lado do grafo, indicando, no rótulo da aresta, o custo da operação de converter um nodo no outro. No caso de nodos incluídos ou excluídos, nodos especiais representando tais operações são incluídos. O conjunto de arestas que estabelece a ligação entre as partes com um custo mínimo (representado nos rótulos das arestas) é o custo para converter um documento no outro, o que indica a similaridade entre os mesmos.

Andrew Nierman et al. [Nierman and Jagadish 2002] usa a distância de edição entre árvores representando documentos XML como uma métrica para a similaridade entre os mesmos. Considera aspectos como elementos repetidos e opcionais. [Nierman and Jagadish 2002] também restringe as seqüências de operações permitidas, reduzindo o custo para obter a menor distância de edição.

O trabalho de Sergey Melnik et al. [Melnik et al. 2002] define um algoritmo para gerar um mapeamento entre os nodos de dois grafos baseado na sua similaridade. Basicamente ele considera que dois nodos são similares se seus vizinhos correspondentes (ligados por arestas de mesmo rótulo) são similares. O algoritmo propõe uma propagação da similaridade entre os elementos para seus vizinhos, em uma abordagem chamada *"similarity flooding"*. Ela leva em conta o seguinte princípio: se dois elementos *a1* e *b1* estão conectados por arestas a *a2* e *b2*, e as arestas têm o mesmo rótulo *l1*, e *a2* e *b2* são similares, então presume-se que *a1* e *b1* também são similares.

David Buttler [Buttler 2004] usa o conceito de *shingles* [Broder 1998] para comparar a estrutura de documentos XML. Ele adota, como *shingles* dos documentos, os *paths* dos documentos XML (como por exemplo `/autor/cidade/nome`).

Em Karin Kailing et al. [Kailing et al. 2004] propõe-se um "filtro" para determinar com custo reduzido se compensa comparar dois grafos de acordo com um valor mínimo de similaridade. O filtro baseia-se no fato que a diferença na distribuição de atributos nos grafos impacta na similaridade. Se os atributos estão distribuídos de forma similar, então é possível que os nodos sejam similares.

4.3. Comparação completa

Na proposta de Melanie Weis et al. [Weis and Naumann 2004] trata-se o problema de detectar objetos duplicados em um documento XML. Sua solução é baseada em uma análise iterativa *top-down* na hierarquia dos elementos (objetos) do documento, que identifica duplicatas em cada nível. Com o objetivo de acelerar o processo, três filtros são estabelecidos para evitar a comparação entre *strings* (por distância de edição) em dados que não serão similares.

Após as filtragens, a similaridade entre os objetos é calculada e os objetos similares são *clusterizados*. Cada *cluster* origina um objeto que substitui, no documento original, todos os outros objetos do *cluster*. Desce-se então um nível na hierarquia do documento e executa-se uma nova iteração do processo. Este conclui-se quando todos os níveis do documento tenham sido analisados.

5. Análise Comparativa e Considerações Finais

Esta seção apresenta algumas considerações e tendências de pesquisa que tomam como base as abordagens de análise de similaridade de dados semi-estruturados discutidas anteriormente, em especial documentos XML.

Com relação ao trabalho de [Dorneles et al. 2004], um aspecto não claro é como usar a abordagem certa para os dados (por exemplo, lista e coleção). [Dorneles et al. 2004] foca sua abordagem no contexto de um banco de dados XML, onde a semântica dos documentos provavelmente está disponível através de seus esquemas.

Na proposta de Andrew Nierman et al. [Nierman and Jagadish 2002], um aspecto importante a ser levado em conta é como a similaridade acusada pela métrica é menos próxima da similaridade ideal em uma distância de edição geral, que permita qualquer seqüência de operações. Além disso, os pesos associados a cada operação poderiam variar conforme o texto e a estrutura sendo analisada, para melhor respeitar os aspectos semânticos da representação das informações (para algumas estruturas a eliminação de um elemento pode ter menos impacto que a inclusão).

Com respeito ao contexto das informações, no trabalho de Melanie Weis et al.[Weis and Naumann 2004] não tratam-se casos onde objetos similares podem estar localizados em diferentes contextos do documento XML. A comparação dos elementos descendentes é insatisfatória, pois não leva em conta como os dados estão organizados. Isto pode ser problemático se existe uma semântica relacionada a sua organização [Dorneles et al. 2004].

Tabela 1. Comparativo dos trabalhos relacionados.

Trabalho	Foco	Considera dados e/ou estrutura?	Considera semântica?	Principais métricas	Ontologias
[Dorneles et al. 2004]	Dados estruturados e semi-estruturados	Ambos	Não	*tupleSim, listSim, setSim*	Não usa
[Broder 1998]	Frases e blocos de texto	Dados	Não	*shingsem, shingcon*	Não usa
[Carvalho and da Silva 2003]	Objetos complexos	Dados	Não	*Espaço vetorial*	Não usa
[Nierman and Jagadish 2002]	Documentos XML	Estrutura	Não	*Distância de edição entre árvors*	Não usa
[Chawathe and Garcia-Molina 1997]	Dados estruturados	Ambos	Não	*Edge cover*	Não usa
[Buttler 2004]	Documentos XML	Estrutura	Não	*3Shingles*	Não usa
[Melnik et al. 2002]	Grafos	Estrutura e/ou dados	Não	*Similarity flooding*	Não usa
[Weis and Naumann 2004]	Documentos XML	Ambos	Não	*IDF, distância de edição*	Não usa

Na Tabela 1 comparam-se os trabalhos apresentados levando em conta alguns aspectos relevantes à comparação entre documentos XML para fins de determinação de similaridade. Buscou-se um conjunto de trabalhos que analisassem documentos semi-estruturados em vários níveis de complexidade, indo desde frases e blocos de textos[Broder 1998] até uma comparação específica para documentos XML [Weis and Naumann 2004]. Em relação aos trabalhos, vê-se que tratam tanto da estrutura quanto do conteúdo dos documentos, porém em nenhum caso o aspecto semântico relacionado aos mesmos é considerado, assim como também o uso de ontologias na comparação.

Em relação as métricas utilizadas, conforme a complexidade dos elementos envolvidos na comparação (desde blocos de texto até a estrutura de um documento XML), as métricas adotadas tendem a ser variações de métricas tradicionais para grafos, usando-se também métricas para dados simples/complexos como suporte à comparações mais complexas [Weis and Naumann 2004].

O tópico similaridade entre dados semi-estruturados (especialmente XML) tem despertado a atenção da comunidade de pesquisa em banco de dados e apresenta muitas questões em aberto. Com base no comparativo feito, uma das contribuições deste artigo é levantar aspectos relacionados a este tópico que possam ser foco de novos trabalhos. Um primeiro aspecto é o uso de ontologias como apoio semântico na compararação de documentos, visto que nenhum trabalho o considera. Com ontologias seria possível estabelecer melhor similaridades entre elementos individuais como também entre relacionamentos entre elementos. As ontologias podem também ser úteis na determinação do grau de importância de um elemento, considerando a sua relevância no domínio.

Outro aspecto que os trabalhos analisados não consideram é como comparar elementos de documentos que são blocos de textos. Com relação a isto, pode-se pensar na utilização de uma abordagem baseada em *shingles*, que podem indicar textos que tratam

do mesmo assunto, mesmo sendo a estrutura dos mesmos diferente.

De acordo com a Tabela 1, cada trabalho segue uma métrica específica e foca em um tipo de comparação. Entretanto, seria interessante comparar documentos XML com base no conceito de *plugins* [Gamma et al. 1995]. Aqui, baseado no tipo de dado comparado, determinados recursos (estratégias de comparação e métricas) poderiam ser utilizados, em uma abordagem adaptativa. Outra sugestão é detectar se dois documentos são versões de um mesmo documento, de forma a utilizar métricas mais adequadas (como [Chawathe and Garcia-Molina 1997]), podendo melhorar a comparação.

A consulta à esquemas para auxiliar a comparação é outra abordagem a experimentar, como suporte a comparação. Também seria relevante o estabelecimento de regras, por um especialista, que possam ajudar a determinar quais elementos são ou não importantes na comparação. Dependendo do domínio considerado para os dados, regras de conversão de valores como temperaturas, distâncias, etc. também seriam úteis.

Como um último aspecto referente à organização de documentos XML, a ordem dos elementos geralmente não influi na semântica dos mesmos. Portanto, reorganizar os elementos pode facilitar ou mesmo acelerar o processo de comparação.

Referências

Broder, A. (1998). On the resemblance and containment of documents. In *SEQS: Sequences '91*.

Buttler, D. (2004). A short survey of document structure similarity algorithms. In *International Conference on Internet Computing*, pages 3–9.

Carvalho, J. C. P. and da Silva, A. S. (2003). Finding similar identities among objects from multiple web sources. In *WIDM*, pages 90–93.

Chawathe, S. S. and Garcia-Molina, H. (1997). Meaningful change detection in structured data. *SIGMOD Record (ACM Special Interest Group on Management of Data)*, 26(2):26–37.

Dorneles, C. F., Heuser, C. A., Lima, A. E. N., da Silva, A. S., and de Moura, E. S. (2004). Measuring similarity between collection of values. In *WIDM*, pages 56–63.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley Pub Co.

Kailing, K., Kriegel, H.-P., Schönauer, S., and Seidl, T. (2004). Efficient similarity search for hierarchical data in large databases. In *EDBT*, pages 676–693.

Melnik, S., Garcia-Molina, H., and Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *ICDE*, pages 117–128.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys*, 33(1):31–88.

Nierman, A. and Jagadish, H. V. (2002). Evaluating structural similarity in XML documents. In *WebDB*, pages 61–66.

Weis, M. and Naumann, F. (2004). Detecting duplicate objects in XML documents. In *IQIS*, pages 10–19.