Melhorando a Precisão do Processo de Integração por meio da Ferramenta VER (Visual Entity Resolution)

Mariângela Cardoso Miguel, Bruno Tomazela, Cristina Dutra de Aguiar Ciferri
Instituto de Ciências Matemáticas e de Computação (ICMC), USP/São Carlos
Departamento de Ciências de Computação, Grupo de Bases de Dados e Imagens
mariangelacm@gmail.com, btomazela@gmail.com, cdac@icmc.usp.br

Resumo

Na integração de dados heterogêneos, é impossível garantir 100% de precisão na identificação de ocorrências de diferentes referências à mesma entidade do mundo real. Este artigo descreve a ferramenta VER, a qual oferece funcionalidades de visualização, cópia e sincronização de dados que permitem melhorar a precisão do processo de integração por meio da geração de entidades integradas.

1. Introdução

Em muitas aplicações, é necessário integrar dados de fontes heterogêneas [4]. Um desafio relacionado à integração dos dados consiste em identificar ocorrências de diferentes referências à mesma entidade do mundo real. Este desafio requer a resolução de ambigüidades, as quais surgem devido ao fato de que entidades em conjuntos de dados do mundo real são geralmente identificadas por meio de seus valores de atributos ao invés de identificadores únicos (chaves primárias). Este desafio tem sido recentemente referenciado na literatura por reconciliação de referências e resolução de entidades, e consiste em automaticamente determinar referências a uma mesma entidade do mundo real e reconciliá-las [1, 2].

Desde que fontes heterogêneas podem possuir diversas formas de se referir à mesma entidade do mundo real, o objetivo das técnicas no estado da arte em reconciliação de referências (e.g., [1, 2]) consiste em gerar agrupamentos (i.e., *clusters*) de entidades que têm um grau de similaridade e, portanto, têm uma alta probabilidade de ser a mesma. Estes agrupamentos são chamados, neste artigo, de *agrupamentos de entidades reconciliadas.* Entretanto, os agrupamentos gerados enfrentam dois problemas principais. Primeiramente, eles podem não ser 100% precisos, desde que suas entidades são determinadas com base em similaridade. Depois, embora cada agrupamento contenha entidades similares, ele não especifica qual é a entidade integrada que melhor o representa.

Surge, então, a necessidade de se requerer a intervenção humana como uma maneira de produzir um processo de integração de alta qualidade. A intervenção humana é um aspecto chave em integração de dados, porque ela pode validar os agrupamentos de entidades reconciliadas, melhorando a qualidade destes agrupamentos. Além disso, a intervenção humana também pode determinar qual a entidade integrada que melhor representa cada agrupamento.

2. Objetivos

Este artigo apresenta a ferramenta VER (Visual Entity Reconciler), a qual tem como objetivo ajudar os usuários a melhorar a precisão de agrupamentos de entidades relacionadas por meio da geração de entidades integradas.

3. Integração de Dados

Reconciliar referências é uma tarefa desafiadora em diferentes contextos [1,2]. A entidade *pessoa* é um conceito fundamental em aplicações de gerenciamento de informação pessoal e redes sociais. Assim, esta entidade deve ser agrupada nestas aplicações para resolver ambigüidades. Em bibliotecas digitais, a ambigüidade é relacionada a *publicação*, *autor*, *instituição*, *editora* e *conferência*. Cada uma destas entidades deve ser agrupada, por exemplo, para reconciliar referências bibliográficas. Exemplos comuns de aplicações que requerem resolução de ambigüidades são comércio eletrônico e qualquer outro portal

baseado na Web que integre dados de fontes pré-existentes.

Exemplos de *entidade* no contexto deste artigo são o Artigo 1, Artigo 2 e Artigo 3, cada um representando um *artigo publicado em periódico* (Figura 1). Uma entidade é composta por vários *atributos*. Para artigos publicados em periódicos, exemplos de atributos incluem título, natureza, meio de divulgação, série, local de publicação, título do periódico ou revista e nomes completos dos autores.

Artigo 1:
Título: Topological dynamics of retarded functional differential equations
Natureza: Completo
Meio de divulgação: Meio Digital
Série: “”
Local publicação: California - EUA
Título do periódico ou revista: Journal of Differential Equations
Nome completo do autor: Márcia Cristina Anderson Braz Federson

Artigo 2:
Título: Topological dynamics of retarded functional differential equations
Natureza: Completo
Meio de divulgação: Meio Digital
Série: 2
Local publicação: USA
Título do periódico ou revista: Journal of Differential Equations
Nome completo do autor: Márcia Cristina Anderson Braz Federson

Artigo 3:
Título: Large time behaviour of linear functional differential equations
Natureza: Completo
Meio de divulgação: Impresso
Série: 47
Local publicação: “”
Título do periódico ou revista: Integral Equations and Operator Theor
Nome completo do autor: Miguel Vinícius Santini Frasson

Figura 1. Exemplos de artigos publicados em periódicos.

Baseado na análise da similaridade entre as entidades da Figura 1, técnicas de reconciliação de referências existentes na literatura (e.g., [1, 2]) produzem dois agrupamentos de entidades relacionadas, como ilustrado na Figura 2.

Agrupamento 1: {Artigo 1, Artigo 2}
Agrupamento 2: {Artigo 3}

Figura 2. Exemplos de agrupamentos.

A ferramenta VER descrita neste artigo usa como entrada agrupamentos de entidades relacionadas. Portanto, qualquer trabalho existente na literatura que tenha esta finalidade pode ser usado para a geração desses agrupamentos.

Outra característica da VER é que ela está sendo desenvolvida como extensão da ferramenta RDA (Reconciliador de Dados Acadêmicos), a qual é usada para a reconciliação de dados acadêmicos de currículos de docentes [5]. Uma primeira diferença entre RDA e VER é que a RDA compara entidades de apenas dois documentos de entrada, enquanto que VER usa como entrada agrupamentos de entidades relacionadas. Além disso, VER também oferece funcionalidades para a geração de uma entidade integrada a partir dos agrupamentos de entrada. Essa funcionalidade não está presente na ferramenta RDA.

4. Ferramenta VER

A ferramenta VER (Visual Entity Reconciler) usa como entrada um arquivo de agrupamentos de entidades relacionadas em formato XML (eXtensible Markup Language). Outra característica é que a VER exibe os dados em sua interface de acordo com arquivos de configuração, os quais especificam como dados de uma aplicação base são visualizados e manipulados na interface da ferramenta, independentemente da forma como estes dados são armazenados nesta aplicação. A utilização dos arquivos de configuração visa garantir maior flexibilidade à ferramenta.

A ferramenta VER mostra agrupamentos de entidades relacionadas e, adicionalmente, para cada agrupamento, também mostra uma entidade integrada que representa este agrupamento. Na ferramenta VER, as

entidades que pertencem ao agrupamento são chamadas de entidades-fonte.

A Figura 3 ilustra a interface da ferramenta VER. Nesta figura:

(1) o lado esquerdo mostra a entidade integrada.

(2) o lado direito exibe as entidades-fonte do agrupamento.

(3) o contador logo abaixo de cada painel exibe o número de entidades. Mais especificamente, o lado esquerdo exibe o número de agrupamentos gerados, enquanto que o lado direito exibe o número de entidades-fonte que compõem o agrupamento sendo exibido.

(4) é possível alternar entre todas as entidades-fonte do agrupamento selecionado por meio da barra de navegação localizada abaixo do painel direito.

(5) é possível alternar entre todas as entidades integradas por meio da barra de navegação localizada abaixo do painel esquerdo**.**

Outras funcionalidades disponibilizadas na ferramenta VER são:

(6) existe a opção de sempre selecionar (ou não) as mesmas linhas nos dois painéis. Por exemplo, se um usuário posiciona o cursor no lado esquerdo no atributo *série*, se esta opção estiver selecionada, o cursor no lado direito também será posicionado sobre o atributo *série*. Na Figura 3, esta opção não está selecionada e, portanto, o cursor no lado direito está posicionado sobre o atributo *local*, ao passo que o cursor no lado esquerdo está posicionado sobre o atributo *idioma*.

(7) a cópia pode ser feita de uma entidade-fonte do agrupamento para a entidade integrada.

(8) o usuário também tem a possibilidade de validar a entidade integrada, indicando que ela é a entidade que melhor representa o agrupamento. No caso de agrupamentos com apenas uma entidade-fonte, a validação é feita automaticamente pela ferramenta.

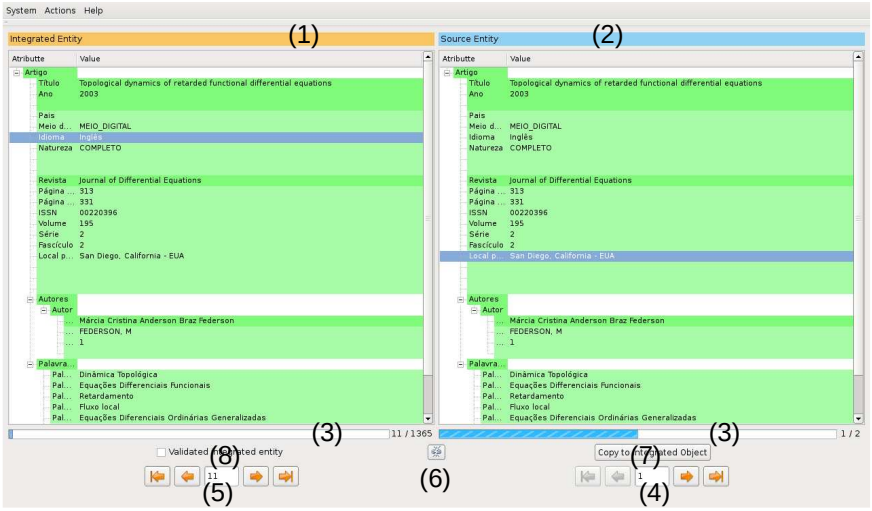

Figura 3. Interface da ferramenta VER.

No exemplo da Figura 3, a entidade integrada e a entidade-fonte são as mesmas. Um exemplo de sincronização quando a entidade integrada é diferente da entidade-fonte é ilustrado na Figura 4. Nesta figura:

(1) marcas em verde nos atributos das entidades indicam que eles possuem os mesmos valores;

(2) marcas em amarelo nos atributos das entidades indicam que eles possuem valores diferentes; e

(3) marcas em amarelo na entidade indicam que as entidades possuem alguns valores de atributos diferentes.

Na Figura 4, a opção de sempre selecionar as mesmas linhas nos dois painéis está ativada.

Portanto, o cursor no lado esquerdo e o cursor no lado direito estão posicionados sobre o atributo *natureza*.

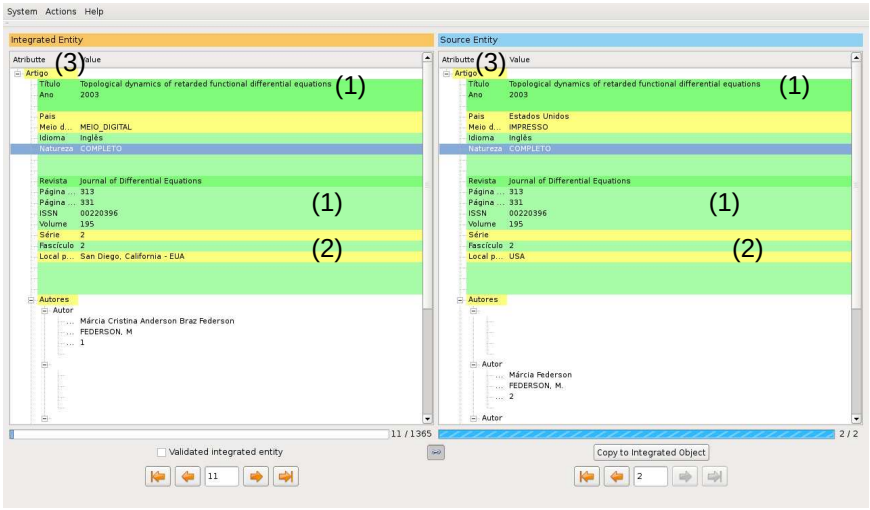

Figura 4. Sincronização de entidades.

Foram propostos três métodos para identificar automaticamente a entidade integrada que melhor representa o agrupamento. Esses métodos indicam que a entidade integrada é: (i) a primeira do agrupamento; (ii) a com maior freqüência no agrupamento: neste caso, apenas alguns atributos da entidade são considerados para se fazer uma comparação exata entre os valores desses atributos, e para se contar quantas vezes uma 'mesma' entidade aparece no agrupamento; e (iii) a que tem mais atributos preenchidos, com base em diferentes pesos atribuídos a atributos mais importantes. Ao usar a ferramenta VER, o usuário pode escolher qual método deseja aplicar.

A ferramenta VER está implementada nas linguagens C++ e XML, além de usar o framework de desenvolvimento Qt [3].

4. Conclusões

Agrupamentos gerados por técnicas de reconciliação de referências na literatura podem não ser 100% precisos, desde que suas entidades são determinadas usando similaridade. Ademais, essas técnicas não identificam qual é a entidade integrada que melhor representa cada agrupamento. A ferramenta VER supre essas limitações, oferecendo funcionalidades de visualização, diferentes métodos para identificar automaticamente a entidade integrada que melhor representa o agrupamento. A análise dos métodos desenvolvidos demonstrou que o método que considera atributos mais importantes encontra uma entidade integrada mais representativa do que os outros métodos.

Como trabalho futuro, a ferramenta também permitirá que entidades de um agrupamento sejam movidas para outro agrupamento.

5. Referências

[1] Benjelloun, O.; et al. Swoosh: a Generic Approach to Entity Resolution. *The VLDB Journal*, 18(1): 255-276, 2009.

[2] Bhattacharya, I.; Getoor, L. Collective Entity Resolution in Relational Data. *ACM TKDD*, v. 1, article 5, 36p, 2007.

[3] Blanchette, J.; Summerfield, M. *C++ GUI Programming with Qt 3*. Prentice Hall, 2004.

[4] Halevy, A. Y.; et al. Data Integration: The Teenage Years. In *Proc.* VLDB Conference, p. 9-16, 2006.

[5] Tomazela, B.; et al. Reconciliando Dados de Cunho Acadêmico. In Proc. XXIII SBBD, p.283-297, 2008