Coleta, Integração e Pré-processamento de Dados de Múltiplas Fontes

Natércia A. Batista[1], Michele A. Brandão[1], Michele Brito[1], Daniel H. Dalip[2], Mirella M. Moro[1]

[1]Universidade Federal de Minas Gerais – Belo Horizonte – Brasil

[2]Centro Federal de Educação Tecnológica de Minas Gerais – Belo Horizonte – Brasil

{natercia,micheleabrandao,mibrito,mirella}@dcc.ufmg.br,
hasan@decom.cefetmg.br

Resumo. *Dados extraídos da Web são cada vez mais heterogêneos e não estruturados, representando desafios para atividades de coleta, integração e pré-processamento de dados. Existem estudos que são "orientados a dados", i.e., com base nos dados disponíveis, mas seus resultados ficam restritos aos respectivos dados. Em contraponto, vários problemas existem antes de se identificar quais dados são necessários para solucioná-los, e muitas vezes, são necessários dados de múltiplas fontes. Nesse contexto, o primeiro problema ao lidar com dados provenientes da Web é definir a estratégia de coleta, que pode ser classificada de acordo com o período e a forma de buscar a semente. Outro problema é definir uma estratégia para integrar os dados de diferentes fontes de forma a ter uma visão uniforme para usuários ou aplicações, além de armazená-los de maneira a permitir uma consulta eficiente. Finalmente, pode ser necessário realizar o pré-processamento de dados, que acontece antes ou depois da integração de dados, e envolve resolver dados faltantes e duplicados, normalização, etc. Este tutorial aborda esses três problemas de forma integrada com foco em questões práticas e de pesquisa.*

1. **Coleta de Dados.** Esta parte do tutorial cobre as principais estratégias de coleta de múltiplas fontes, bem como os principais desafios. Especificamente: uma visão geral sobre os três principais tipos de coleta considerando o período de realização da mesma (contínua, periódica e/ou ocasional) e a busca da semente -- entidade alvo de coleta ou ponto inicial (busca em largura, caminhamento aleatório e caminhamento aleatório Metropolis-Hastings [Gjoka et al. 2010]). Ademais, são apresentados exemplos práticos e desafios para cada estratégia.
2. **Estratégias para Integrar Dados de Múltiplas Fontes.** Após os dados serem coletados de múltiplas fontes, eles precisam ser integrados. Note que é possível coletar dados de cada fonte e armazená-los de forma separada para posterior integração, ou já armazenar todos os dados em um único local de forma integrada à medida que cada coleta é realizada. Nesta etapa, são abordadas vantagens e desvantagens dessas duas estratégias e das diferentes formas de armazenamento (planilhas, arquivos CSV, banco de dados relacionais, etc).

3. **Pré-processamento de Dados.** Nesta etapa do tutorial são abordados problemas geralmente encontrados nos dados após a realização da coleta: (*i*) valores faltantes - quando nenhum valor é armazenado para uma variável; (*ii*) veracidade dos dados - refere-se a viés, anormalidades e ruídos presentes nos dados; (*iii*) remoção de dados duplicados - dados iguais inseridos múltiplas vezes; (*iv*) falta de normalização - processo de reestruturar os dados a uma forma comum; e (*v*) redução dos dados - processo de minimizar a quantidade de dados a serem armazenados.
4. **Aplicações Reais.** Existem diferentes estudos que combinam múltiplas fontes de dados com objetivos distintos [Batista et al. 2017, Brandão et al. 2017, Dalip et al. 2013, Farnadi et al. 2018]. Por exemplo, Batista et al. (2017) medem a força dos relacionamentos entre desenvolvedores através da combinação de dados originados de um banco de dados relacional disponível na Web e dados coletados utilizando a API (*Application Programming Interface*) do GitHub. Brandão et al. (2017) utilizam dados de múltiplas fontes (DBLP e páginas Web) para fornecer visualizações com informações mais completas sobre pesquisadores. Ademais, Dalip et al. (2013) geram um ranking de respostas para perguntas no Stack Overflow por meio de indicadores. Dentre tais indicadores, utilizou-se a comparação entre os vocabulários das respostas com os vocabulários de bons artigos do Wikipedia. Todos esses exemplos de uso real da combinação de múltiplas fontes de dados são abordados nesta etapa mais prática do tutorial.
5. **Conclusão e Trabalhos Futuros.** Finalmente, além de abordar os principais conceitos, desafios e exemplos relacionados à coleta, integração e pré-processamento de dados de múltiplas fontes, discutimos problemas em aberto e possíveis trabalhos futuros a fim de incentivar pesquisas relacionadas a este tutorial.

Referências

Batista, N. A., Brandão, M. A., Alves, G. B., Silva, A. P. C. da, and Moro, M. M. (2017) "Collaboration strength metrics and analyses on GitHub." In *WI*, pp. 170-178.

Brandão, M. A., Diniz, M. A., Sousa, G. A., Moro, M. M. (2017) "Visualizing Co-Authorship Social Networks and Collaboration Recommendations With CNARe." *Graph Theoretic Approaches for Analyzing Large-Scale Social Networks*: 173-188.

Dalip, D. H., Gonçalves, M. A., Cristo, M., and Calado, P. (2013) "Exploiting user feedback to learn to rank answers in q&a forums: a case study with stack overflow." In *SIGIR*, pp. 543-552.

Farnadi, G., Tang, J., Cock, M. D., and Moens, M-F. (2018) "User Profiling through Deep Multimodal Fusion." In *WSDM*, pp. 171-179.

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2010) "Walking in facebook: A case study of unbiased sampling of osns." In *IEEE Infocom*, pp. 1-9.