LinkGen: Multipurpose Linked Data Generator

Amit Krishna Joshi, Pascal Hitzler e Guozhu Dong

Data Semantics Lab, Wright State University, Dayton, OH, USA

LinkGen: Gerador de Dados Ligados Multipropósito

Amit Krishna Joshi, Pascal Hitzler e Guozhu Dong

Data Semantics Lab, Wright State University, Dayton, OH, USA

Contexto, Desafio & problema

Contexto

- U = {www}
- Meio de comunicação = web semântica
- Padrão de comunicação = W3C
- Agente: LinkGen

Desafio

- Apresentar uma aplicação que crie um grande dataset sintético para testes em vocabulários quaisquer, ligados semanticamente à web.

Problema

- Há geradores de datasets sintéticos que servem apenas para um # de vocabulários limitado.

LinkGen

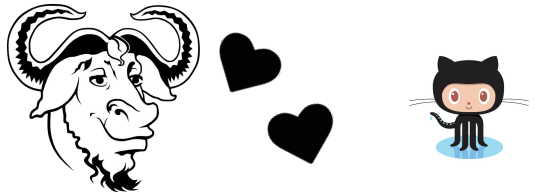

Figura 1: Licença GNU e Github para um gerador de dados ligados de acesso livre.

O que é (What)?

- Gerador automático de dados sintéticos que pode criar uma grande quantidade de dados em RDF
- Trabalha com vasta possibilidade de vocabulários
- Primeiro trabalho (que os autores têm conhecimento) que gera um dataset ligado para qualquer vocabulário e que pode simular um dataset do mundo real com características como distribuição estatística e ruídos.

LinkGen, para quê (For What)?

- Para testar novos vocabulários
- Consultar datasets
- Diagnosticar inconsistências
- Avaliar a performance de datasets
- Testar agregadores de dados ligados
- Avaliar metodologias
- Enfim...fazer pesquisa!

Por quê (Why)?

Contexto corporativo:

- rápida absorção de tecnologias semânticas
- estratégia de gestão de conhecimento

Figura 2: Organizações que têm adotado tecnologias semânticas.

Como (How)?

- Baseia-se em certas distribuições estatísticas
- Pode ser gerado baseado em distribuições de leis de potências para simulação de datasets do mundo real
- Pode-se adicionar ruídos
- Pode ser gerado tanto em modo streaming quanto on-disk
- Instâncias sintéticas podem ser ligadas a entidades do mundo real se o dicionário das entidades do mundo real estiver disponível

Como (How does it work)?

- Para criar diferentes conjuntos de output, LinkGen criam dados aleatórios baseado na “semente” dada pelo usuário.

Figura 3: Como em “João e o Pé de Feijão”, o conto de fadas de origem inglesa, uma semente de dados pode gerar um produto colossal. No caso do LinkGen: um enorme dataset de dados ligados.

- LinkGen usa 2 técnicas estatísticas:
  - distribuição gaussiana

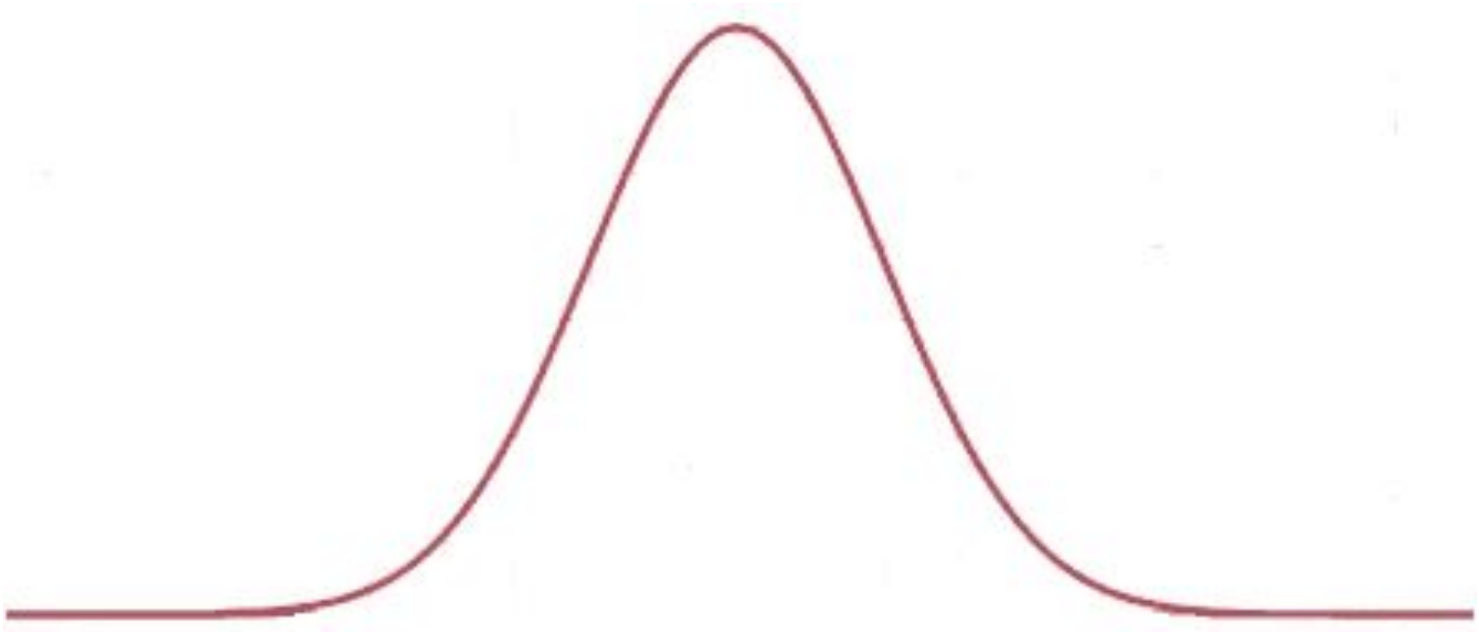

Figura 4: Distribuição gaussiana, também conhecida como a Integral de Euler-Poisson.

- distribuição da lei de potência de Zipf's

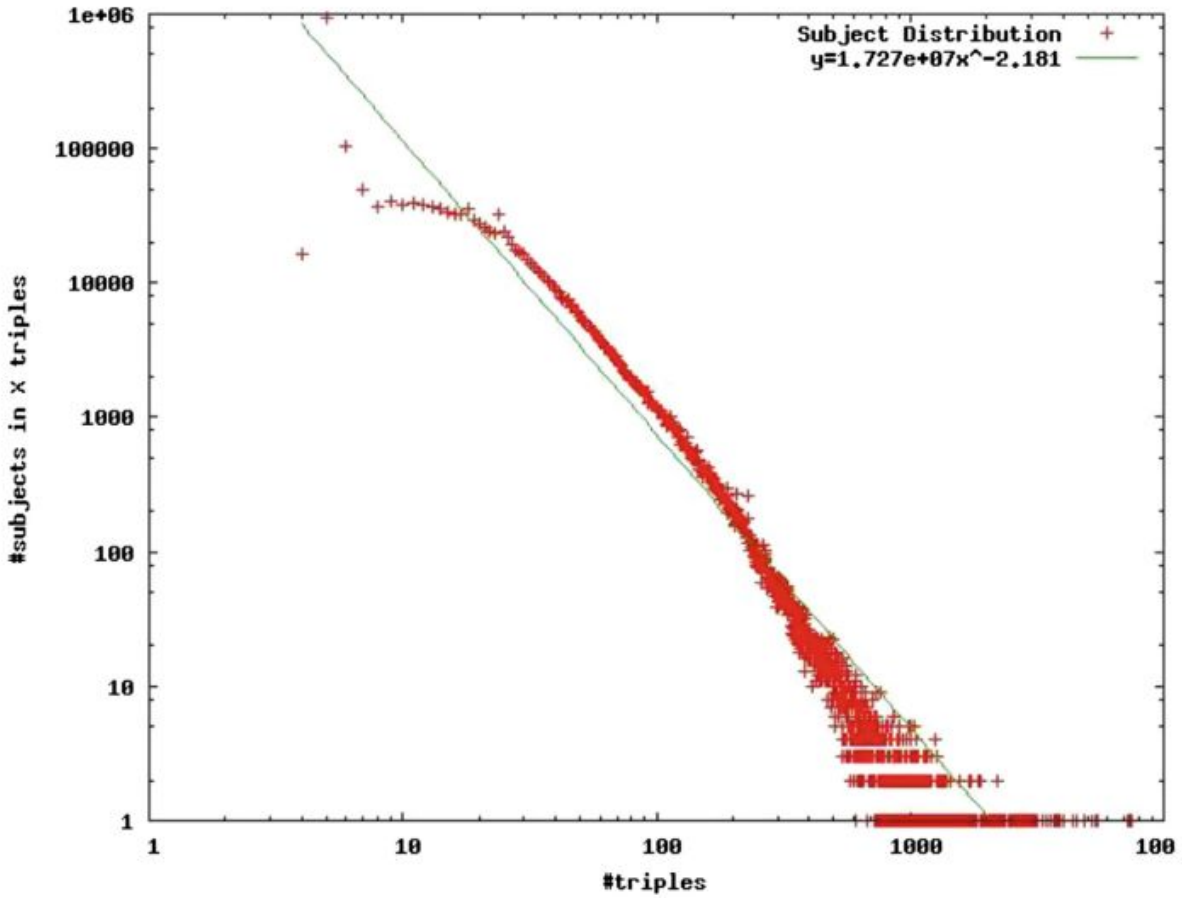

Figura 5: Lei de potência dos sujeitos na wikipedia.

- Ruídos:
  - adicionando dados inconsistentes
  - adicionando triplas com erros sintáticos
  - adicionando declarações erradas por atribuição de domínios inválidos
  - criando instâncias sem informação de tipo
  - combinando parâmetros para a geração de dados ruidosos

Geração dos dados

Passo 1)

- carregar a ontologia e recolher estatísticas sobre todos os componentes da ontologia que não estão definidos (# classes, dataptype, propriedades, propriedades do objeto e propriedades de domínio e alcance).
- guardar a conectividade de cada classe e ordenar as classes com base em frequência.

OBS: classes mais conectadas gerarão um maior # de entidades correspondentes.

Passo 2)

- Usar distribuição estatística para um grande # de entidades e associar pesos para cada uma delas
- Escolha da Lei de Potências que será usada

Passo 3)

- Geração de triplas sintéticas em cada uma das classes, por associação de propriedades e levando em consideração os pesos
- Para cada entidade, pelo menos 2 triplas são adicionadas para indicar seu tipo

Resultados

- Pode gerar datasets sintéticos para qualquer vocabulário expresso em RDFS ou OWL
- Não implementa todas as descrições de classes e restrições de propriedades especificadas na ontologia OWL
- O suporte para blank nodes também não é provido

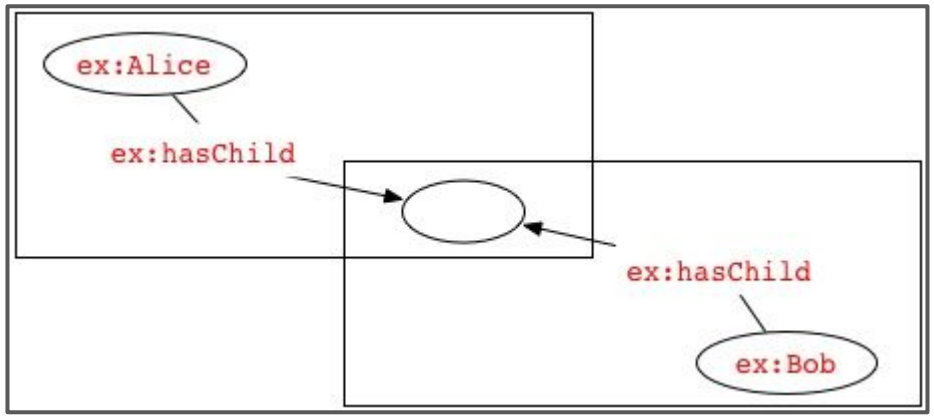

Figura 6: Exemplo de blank node em RDF. Um b-node também pode ser chamado de recurso anônimo. É um recurso onde a URI não é dada.

Tabela 1: características dos datasets usados para a avaliação.

	DBpedia	Schema.org
Number of distinct classes	147	158
Number of distinct properties	2891	1002
Number of distinct object properties	1734	463
Number of distinct data properties	1100	490
distinct properties without domain and/or range specification	685	11

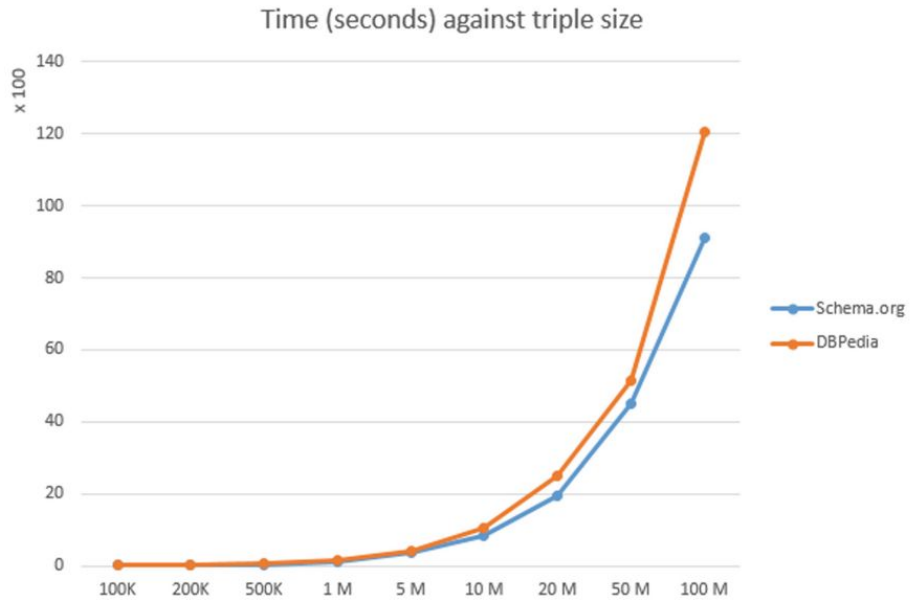

Figura 7: Tempo para geração de datasets de tamanhos variados.

Onde (Where)?

LinkGen

- Github

Trabalhos Relacionados

- Lehigh University Benchmark
- Berlin SPARQL Benchmark
- SP^2Bench
- The Social Intelligence Benchmark
- TontoGen
- WatDiv
- Sygenia

Referências

LinkGen: Multipurpose Linked Data generator. Amit Krishna Joshi, Pascal Hitzler & Guoshu Dong. Wright State University, Dayton, OH, USA. 15th International Semantic Web Conference. Kobe, Japan, October 17-21, 2016. Proceedings, Part II.

Portal da W3C. <https://www.w3.org/>. Acesso em 27/05/2017.

Figura 1: <https://goo.gl/qhRtKx> e <https://goo.gl/XXaD0I> . Disponíveis em 27/05/2017.

Figura 2: <https://goo.gl/Oh1OmY>,
<https://goo.gl/FcX8my>, <https://goo.gl/jNmPta>, <https://goo.gl/CyXfQx>. Disponíveis em 27/05/2017.

Figura 3: <https://goo.gl/qhRtKx>. Disponível em 27/05/2017.

Figura 4: <http://www.gandalf.it/htlws/img/gauss1.gif>. Disponíveis em 27/05/2017.

Figura 5: LinkGen: Multipurpose Linked Data generator. Amit Krishna Joshi, Pascal Hitzler & Guoshu Dong. Wright State University, Dayton, OH, USA. 15th International Semantic Web Conference. Kobe, Japan, October 17-21, 2016. Proceedings, Part II.

Figura 6: <https://goo.gl/es5DNM>. Disponível em 27/05/2017.

Figura 7: LinkGen: Multipurpose Linked Data generator. Amit Krishna Joshi, Pascal Hitzler & Guoshu Dong. Wright State University, Dayton, OH, USA. 15th International Semantic Web Conference. Kobe, Japan, October 17-21, 2016. Proceedings, Part II.

Tabela 1: LinkGen: Multipurpose Linked Data generator. Amit Krishna Joshi, Pascal Hitzler & Guoshu Dong. Wright State University, Dayton, OH, USA. 15th International Semantic Web Conference. Kobe, Japan, October 17-21, 2016. Proceedings, Part II.