SATO: Uma Ferramenta para Geração de Anotações Semânticas

Rodrigo E. Bela, Marilde T. P. Santos, Mauro Biajiz

[1]Departamento de Computação – Universidade Federal de São Carlos (UFSCar)
Caixa Postal 676 – 13.565-905 – São Carlos – SP – Brasil

{rodrigo_bela,marilde,mauro}@dc.ufscar.br

Resumo. *Este artigo apresenta a ferramenta SATO para geração automática de anotações semânticas por meio da utilização de uma ontologia de domínio e de regras de mapeamento. As anotações semânticas geradas têm o objetivo de servir de metadados adicionais que auxiliem na recuperação e compreensão das informações implícitas armazenadas pelo crescente número de bases de dados brutos.*

1. Introdução

Diante do constante crescimento de dados armazenados, várias formas de recuperação de informação que sejam eficientes vêm sendo desenvolvidas. Desde estruturas de indexação até à criação de metadados adicionais que possibilitem a busca e recuperação por meio do significado do conteúdo desejado. Em especial, o uso de metadados adicionais para a explicitação do significado de um conjunto de dados tem sido empregado em bibliotecas digitais, na Web Semântica e em sistemas de bioinformática.

Diante deste contexto, este artigo apresenta a ferramenta SATO (*Semantic Annotation TOol*), desenvolvida para possibilitar a criação automática de anotações semânticas para conjuntos de dados brutos, proporcionando metadados adicionais que podem ser utilizados para resumir e melhorar a compreensão dos dados aos quais estão relacionados. O restante deste artigo está organizado da seguinte forma: a seção 2 descreve o referencial teórico que impulsionou o desenvolvimento da abordagem de anotações semânticas adotada na SATO; a seção 3 descreve as principais funcionalidades da ferramenta SATO; a seção 4 apresenta um estudo de caso de utilização da ferramenta e a seção 5 apresenta as conclusões e trabalhos futuros.

2. Referencial Teórico

Durante a realização de qualquer trabalho é frequente a necessidade de ressaltar a expressividade e o significado de alguns conjuntos de dados como meio de possibilitar a recuperação e compreensão eficiente dos mesmos. Esta necessidade pode ser suprida com o uso de anotações que, segundo [Agosti et al. 2007], são todo o conjunto de marcas, notas, breves comentários e explicações adicionais que são realizadas sobre algum tipo de conteúdo com a função de garantir que o mesmo possa ser facilmente buscado, lido e compreendido. Tais anotações podem assumir diferentes papéis dependendo da maneira com que são aplicados, por exemplo, servindo como metadados, quando são dados adicionais sobre um determinado conteúdo existente ou como conteúdos que explicam ou enriquecem um conteúdo prévio, mantendo uma conexão com tal conteúdo prévio, mas existindo de forma independente e autônoma [Agosti et al. 2007].

Várias ferramentas têm sido desenvolvidas ou dotadas de mecanismos de suporte à geração de anotações. Em bibliotecas digitais elas geralmente são aplicadas como conteúdo adicional relacionado a outros conteúdos [Agosti e Ferro 2006]; em ambientes de gerenciamento e suporte à manipulação de dados genômicos e proteômicos elas são utilizadas como metadados adicionais que descrevem funções biológicas de seqüências de DNA [Teixeira 2008] e na *Web* Semântica elas são utilizadas como meio de explicitar a semântica de um conteúdo, tornando essa semântica computável [Reeve e Han 2005]. As anotações podem ser consideradas semânticas quando representam instâncias de conceitos expressos em ontologias de domínio, pois uma ontologia define um modelo formal para a conceitualização compartilhada entre especialistas de um determinado domínio [Gruber 1993].

No entanto, há um conjunto de aplicações que, de forma semelhante às analises genômicas e proteômicas, somente coletam e armazenam dados em sua forma mais bruta, mas que poderiam se beneficiar do uso de anotações semânticas para facilitar a recuperação e compreensão dos dados armazenados. São exemplos: valores verificados em exames de sangue e armazenados em sistemas médicos, valores capturados por sensores como termômetros, pluviômetros e barômetros em aplicações geográficas e dados sobre interação de usuários em alguns sistemas de ensino. É para estes conjuntos de dados que a ferramenta SATO, descrita na próxima seção, foi desenvolvida.

3. Semantic Annotation TOol - SATO

A Ferramenta SATO foi desenvolvida em linguagem JAVA, usando o *framework* JENA [Carroll et al. 2004] para manipulação de ontologias expressas em OWL e o SGBDR MySQL para persistência dos dados. O objetivo da SATO é suprir a necessidade de gerar metadados adicionais para conjuntos de dados considerados brutos, realizando esta tarefas de forma automática, por meio de regras de mapeamento e utilizando um vocabulário controlado e bem definido em uma ontologia de domínio.

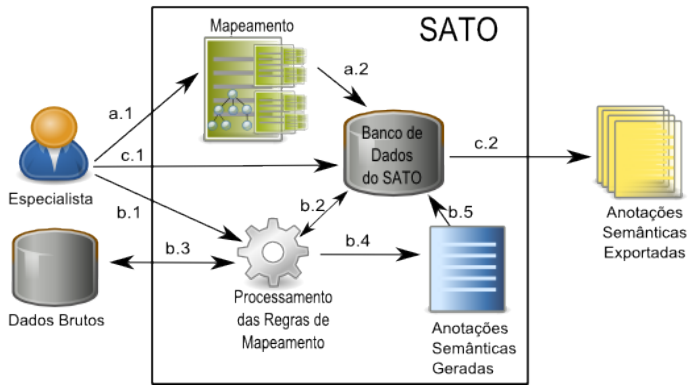

Figura 1. Representação do fluxo das principais atividades realizadas no SATO

A figura 1 apresenta como o SATO pode ser utilizado em três momentos distintos. Num primeiro momento (a1), o especialista do domínio utiliza a ferramenta para informar dados sobre o domínio para o qual deseja gerar anotações semânticas. Nesse momento são informados: um nome, uma descrição e a ontologia do domínio a ser utilizada, sendo que tal ontologia deve estar expressa em OWL. Nesse momento também

devem ser informados: o endereço, nome do banco, nome e senha de acesso do banco de dados onde se encontram os dados brutos a serem anotados. A figura 2(A) apresenta a tela de cadastro de domínio. Após o cadastro do domínio, o usuário deverá informar as regras de mapeamento para aquele domínio. Uma regra de mapeamento descreve como um conceito da ontologia de domínio desejado pode ser encontrado em um conjunto de dados, descrevendo quais dados devem ser analisados e como devem ser analisados. A abordagem implementada pela SATO considera dois tipos de regras de mapeamento:

1) Realizam-se consultas, que podem conter operações ou funções, como cálculo de média, nos dados brutos. O resultado obtido é avaliado frente a um valor estabelecido no domínio como sendo um valor de referência esperado. Esse procedimento permite determinar se o conceito da ontologia está presente ou não nos dados. Por exemplo, considerando uma ontologia do domínio de ensino, que contém o conceito APROVADO para exprimir que um estudante obteve média geral maior que seis em uma disciplina. Para avaliar se esse conceito se aplica ou não aos estudantes de uma determinada disciplina, devem ser obtidas todas as notas referentes àquela disciplina para cada estudante, aplicar uma função nessas notas (média aritmética, por exemplo) e verificar se o resultado obtido, no caso a média, está compatível com o valor de referência estabelecido, ou seja, se a média é maior que seis. Se essa condição for verdadeira então o conceito APROVADO se aplica ao estudante em questão.

2) Verifica-se a presença de um conceito avaliando-se expressões lógicas que envolvem conceitos definidos na ontologia e já anotados nos dados brutos. Tais conceitos são associados nas expressões por operadores lógicos (E, OU). Seguindo no exemplo anterior, considere que a ontologia define que o conceito COLAÇÃO DE GRAU aplica-se a todo estudante que tiver sido APROVADO (conceito já anotado) em um conjunto de disciplinas obrigatórias e obtido determinado NÚMERO DE CRÉDITO TOTALIZADO (outro conceito já anotado) entre disciplinas obrigatórias e optativas. Dessa forma é possível a verificação de conceitos mais complexos que resultam da combinação lógica de vários outros termos presentes no domínio.

A figura 2 apresenta as telas de criação de regras de mapeamento baseadas em consulta (B) e baseadas em regras lógicas (C). Na SATO as regras de mapeamento são organizadas convenientemente de acordo com o objetivo em comum que tais regras possuem. Por exemplo: um conjunto de regras de mapeamento pode ser definido para realizar anotações semânticas num conjunto de dados referentes a exames médicos realizados por pacientes, mas algumas regras deverão ser aplicadas a exames de sangue e outras serão aplicadas a exames de urina, logo, tais regras podem ser agrupadas em dois conjuntos. Para cada conjunto de regras de mapeamento agrupadas é dado o nome de *perspectiva*. Após as definições de regras de mapeamento e perspectivas, as mesmas são persistidas na base de dados da ferramenta (figura 1.a2).

Em outro momento, o especialista informa que deseja gerar anotações semânticas para um determinado domínio (figura 1.b1), a ferramenta recupera então as perspectivas existentes para o domínio desejado. O especialista deve escolher entre as perspectivas recuperadas e, em seguida, informar o conjunto de valores identificadores das transações que deverão ser anotadas, por exemplo, usando o cenário descrito anteriormente, de geração de anotações para um domínio de educação, seriam informados os identificadores dos alunos para os quais se deseja gerar as anotações. Tais identificadores são os parâmetros

esperados pelas regras de mapeamento baseadas em consulta descritas anteriormente. A ferramenta recupera então as regras de mapeamento para aquele domínio e perspectiva (figura 1.b2) e processa, para cada um dos identificadores, as regras definidas, e quando necessário, acessa a base de dados que contém os dados brutos (figura 1.b3). Como resultado do processamento de cada regra de mapeamento, uma lista de anotações semânticas é gerada (figura 1.b4) sendo que a mesma é persistida na base de dados da ferramenta SATO (figura 1.b5) ao fim do processo. As anotações geradas e persistidas na ferramenta SATO podem ser, em um terceiro momento, recuperadas pelo especialista. Estas anotações podem ser visualizadas ou exportadas (figura 1.c1) para diversas finalidades, como, por exemplo, a integração de tais anotações junto à base de dados brutos ou a utilização de tais dados em tarefas de mineração de dados (figura 1.c2).

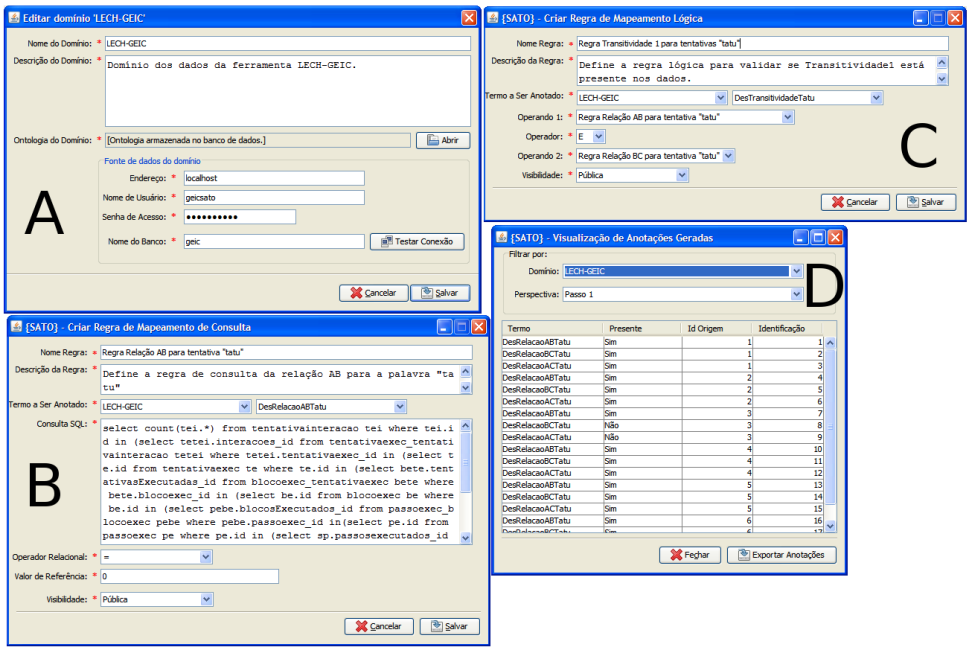

Figura 2. Principais tela da SATO: A)Definição de um domínio; B) Definição de regra de mapeamento baseada em consulta; C) Definição de regra de mapeamento lógica; D) Visualização de anotações semânticas geradas

Além disso, a ferramenta SATO mantém um controle sobre quais foram as atividades realizadas por cada especialista, possibilitando que os mesmos compartilhem os domínios, perspectivas, regras ou anotações semânticas geradas. Por exemplo, nas telas de criação de regras (figura 2 B e C), o especialista pode definir a visibilidade das regras criadas. Entre as visibilidades possíveis têm-se: a) Pública: qualquer especialista que utilize a SATO poderá ver e usar aquela regra; b) Protegida: inicialmente a regra somente pode ser visualizada ou utilizada pelo especialista que criou a mesma. Outros especialistas só terão acesso a tal regra caso o especialista que a criou atribua essa regra a outros especialistas ; c) Privada: somente o especialista que criou a regra pode ver e utilizar a mesma. A visibilidade de uma regra só pode ser alterada pelo especialista que criou a mesma. As anotações semânticas geradas também possuem o controle de visibilidade. Durante o processo de geração as anotações herdam automaticamente a visibilidade atribuída às regras das quais se originam. As visibilidades das anotações podem ser modificadas de forma individual sem alterar a visibilidade da regra das quais se originam ou das anotações rela-

cionadas, ou seja, das anotações realizadas para a mesma transação. Esse gerenciamento de visibilidades e compartilhamento de domínios, perspectivas, regras de mapeamento e anotações semânticas geradas, possibilita que especialistas de um determinado domínio utilizem a ferramenta SATO de forma colaborativa. Na próxima seção será apresentado um estudo de caso de utilização da ferramenta SATO.

4. Utilização da Ferramenta: estudo de caso

A ferramenta SATO foi utilizada para gerar anotações semânticas para os dados provenientes de uma ferramenta de ensino denominada LECH-GEIC, ferramenta desenvolvida no contexto do projeto TIDIA-Ae (Tecnologia da Informação no Desenvolvimento da Internet Avançada - Aprendizado Eletrônico), financiado pela FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) e que é baseada no paradigma de equivalência de estímulos. Entende-se por equivalência de estímulos o surgimento regular e previsível de relações condicionais não ensinadas entre estímulos, derivadas de outras relações ensinadas [de Souza et al. 2009].

Os dados manipulados pelo LECH-GEIC podem ser divididos em três visões diferentes. A primeira armazena as especificações dos especialistas para o ensino de um determinado conteúdo, ou seja, as definições de um programa de ensino (PE). A segunda visão contém as relações entre as crianças que utilizam o sistema e os PEs, controlando quais crianças estão relacionadas a quais PEs e qual o progresso das mesmas em cada um deles. Esse progresso é avaliado por meio do conjunto dos dados obtidos nas interações dos alunos com os PEs. Este conjunto constitui a terceira visão dos dados, sendo o foco do estudo de caso, uma vez que a mesma é utilizada pelos especialistas do domínio para analisar e compreender os comportamentos dos alunos frente ao PE. Os dados do LECH-GEIC possuem características ideais para o uso da ferramenta SATO, uma vez que: a) os dados das interações dos alunos são armazenados em sua forma bruta; b) os dados são analisados manualmente, observando as respostas dadas pelos alunos em um conjunto de tentativa para verificar a presença ou ausência de determinados comportamentos (conceitos); c) o domínio de equivalência de estímulos possui um vocabulário bem definido para a descrever os principais conceitos da área.

Assim, primeiramente foi construída uma ontologia do domínio de equivalência de estímulos, com base em reuniões com especialistas do domínio e no glossário de termos apresentado em [Catania 2006]. Em seguida, foram definidos o domínio e as regras de mapeamento para a verificação dos conceitos *DesRelacaoAB* - desempenho relação AB, *DesRelacaoBC* - desempenho relação BC e *DesRelacaoAC* - desempenho relação AC para tentativas relativas à palavra "tatu"presentes tanto em blocos de treino como em blocos de teste de um determinado passo de ensino. Estes termos foram selecionados porque, por meio deles, é possível inferir se a criança atingiu a transitividade entre três estímulos que diferem na forma, mas que representam o mesmo significado. Assim, uma tentativa que treina ou testa uma relação AB apresenta a som de alguém dizendo a palavra "tatu"e o aluno deve selecionar a imagem de um tatu, uma tentativa que treina ou testa uma relação BC apresenta a imagem de um tatu e o aluno deve selecionar a palavra "tatu"e uma tentativa que treina ou testa uma relação AC apresenta o som de alguém dizendo a palavra "tatu"e o aluno deve selecionar a palavra "tatu".

Cada conceito é considerado presente sempre que todas as execuções das ten-

tativas que o apresentam foram respondidas de forma correta, por exemplo, o conceito *DesRelacaoAB* só é mapeado como presente nos blocos de treino para uma determinada sessão de ensino se o aluno respondeu corretamente todas as tentativas que apresentam tal relação. Algumas das regras geradas podem ser vistas na figura 2(D). Ao todo, foram geradas anotações semânticas para um conjunto de seis sessões de ensino, para validação do protótipo e demonstração das utilidades de tais anotações, que, no caso, foram utilizadas como conjunto de entrada em um processo de mineração de dados usando a tarefa de extração de regras de associação.

5. Conclusão e Trabalhos Futuros

Este artigo apresenta a ferramenta SATO, que gera anotações semânticas sobre dados brutos por meio da utilização de uma ontologia de domínio e de regras de mapeamento. Tais anotações semânticas constituem metadados adicionais que podem ser utilizados para auxiliar na recuperação, visualização e compreensão de um conjunto de dados brutos. Em especial, espera-se que a utilização das anotações semânticas em tarefas de mineração de dados possa possibilitar um ganho na qualidade dos resultados obtidos ao final de tais tarefas. Entre os trabalhos futuros, pretende-se realizar testes envolvendo outros domínios de dados para comprovar a flexibilidade na aplicação da ferramenta para a geração de anotações semânticas para os mais diversos tipos de dados. Além disso, sessões de testes com potenciais usuários estão sendo realizadas para identificar melhorias na usabilidade da ferramenta.

Referências

Agosti, M., Bonfiglio-Dosio, G., e Ferro, N. (2007). A historical and contemporary study on annotations to derive key features for systems design. *International Journal on Digital Libraries*, 8(1):1 – 19.

Agosti, M. e Ferro, N. (2006). A formal model of annotations of digital content. *ACM Transactions on Information Systems*, 1:1–57.

Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., e Wilkinson, K. (2004). Jena: Implementing the semantic web recommendations. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*.

Catania, C. A. (2006). *Learning*. Sloan Publishing, 4 edition.

de Souza, D. d. G., de Rose, J. C., Faleiros, T. C., Bortoloti, R., e Hanna, Elenice Seixas McIlvane, W. J. (2009). Teaching generative reading via recombination of minimal textual units: A legacy of verbal behavior to children in brazil. *International Journal of Psycology and Psychological Therapy*, 9:19–44.

Gruber, T. R. (1993). Toward principles for the design of ontologies used for knowledge sharing. *Formal Ontology in Conceptual Analysis and Knowledge Representation*, 43.

Reeve, L. e Han, H. (2005). Survey of semantic annotation platforms. In *Proceedings of the 2005 ACM symposium on Applied computing*, volume 1, pages 1634–1638.

Teixeira, M. V. C. (2008). Gerenciamento de anotações de biosseqüências utilizando associação entre ontologias e esquemas xml. Mestrado, Universidade Federal de São Carlos - UFSCar.