IntegraWeb: uma arquitetura baseada em mapeamentos semânticos

Felipe L. Pierin[1], **Jaime S. Sichman**[2,1]

[1]Programa de Pós-Graduação em Ciência da Computação
Instituto de Matemática e Estatística (IME) – Universidade de São Paulo (USP)

[2]Laboratório de Técnicas Inteligentes (LTI)
Escola Politécnica (EP) – Universidade de São Paulo (USP)

fpierin@ime.usp.br, jaime.sichman@usp.br

Abstract. *While a large amount of content is produced and published on the Internet by different sources and formats, relevant information about the same domain is spread across the Web in the various portals, which hinders a broad, objective and centralized view of this information. The integration of this data spread in the network allows for smarter queries, with richer results of meaning and closer to the user's interest. However it tends to be costly since there are few reusable and easily integrable models. In this work, we propose an ontology-based architecture for the integration of Internet data and we illustrate its application in real cases on the Internet.*

Resumo. *Dado que grande quantidade de conteúdo é produzida e publicada na Internet por diferentes fontes e formatos, a informação relevante sobre um mesmo domínio acaba espalhada pela Web nos diversos portais, o que dificulta uma visão ampla, centralizada e objetiva sobre esta informação. A integração desses dados espalhados na rede permite consultas mais inteligentes, com resultados mais ricos de significado e mais próximos do interesse do usuário. No entanto, tal integração tende a ser custosa, visto que são poucos os modelos reaproveitáveis e facilmente integráveis entre si. Neste trabalho, propõe-se uma arquitetura baseada em ontologias para a integração de dados da Internet e ilustra-se sua aplicação em casos reais na Internet.*

1. Introdução

A capacidade de armazenar, correlacionar e produzir informação é um tema cada vez mais relevante [Gray et al. 2014]. Nesse contexto, a maioria dos dados que formam a Internet é composta por conteúdo gerado dinamicamente, sem estrutura bem definida, que, de maneira geral, só podem ser compreendidas por humanos mas que só podem ser processados eficazmente por computadores [Stumme et al. 2006]. Além disso as informações não são centralizadas e ficam concentradas em silos de informação como Wikipédia, Facebook e Google Maps, o que pode gerar situações indesejáveis tais como duplicação dos dados, informação incompleta ou excessivamente distribuída.

Um exemplo de duplicação é o cadastro do perfil de um individuo no Facebook e no LinkedIn. Nesse caso, como diferentes organizações não compartilham esse conteúdo, há o custo do armazenamento dos dados que não é compartilhado e o retrabalho do usuário

que preenche o próprio perfil repetidas vezes. Além disso, as informações publicadas na Internet muitas vezes carecem de mecanismos que as inter-relacionem automaticamente o que pode a tornar é incompleta; tal situação poderia ser mitigada pela união dos dados armazenados em diferentes fontes. Considere-se os portais de divulgação de eventos no Brasil: é comum encontrar cenários em que um determinado portal possui a informação da existência e local de uma palestra mas não informa o horário; um segundo portal indica o local da execução da mesma palestra, o horário e o palestrante. Deste modo, não é possível identificar as palestras de um determinado individuo senão pela busca em ambos os sites e composição da informação. Embora trabalhosa, tal pesquisa ainda é possível se limitada a uma palestra especifica de um individuo em especial; no entanto, ao estender a todas as palestras, contidas em diferentes sites da Internet, a mesma pesquisa passa a ser inviável uma vez que são muitos os portais e informações que precisam ser avaliados. Ao combinar as informações de todos os portais acerca desse mesmo domínio, torna-se possível entender e pesquisar melhor as informações sobre o assunto e automatizar processos como, por exemplo, montar a grade de apresentações de um determinado palestrante sem o ônus de pesquisar em diferentes portais. Por exemplo, atualmente um indivíduo que gosta de eventos culturais e que tenha o interesse em decidir entre ir a uma palestra dentro de uma faculdade ou a um evento artístico que acontece em um parque precisa necessariamente navegar por diferentes portais para entender a localização, o horário e então decidir entre uma, outra, ou ambas as atividades. Nesse caso, ao menos um portal de uma faculdade e um portal de eventos artísticos acaba sendo visitado na Internet, já que na realidade do Brasil e na de outros países do mundo o portal que concentra dados sobre cerimônias dentro de uma organização é muitas vezes mantido na própria organização. Nesse sentido, uma nova abordagem para captura e pesquisa da informação distribuída na Internet é necessária.

A integração entre fontes com domínios distintos é outro ponto relevante de atenção. Levando em consideração o exemplo dos eventos distribuídos por diferentes portais na Internet, podemos tornar essa consulta ainda mais rica unindo a essa base de conhecimento os dados sobre outros domínios como, por exemplo, a informação sobre transporte público. Em grandes metrópoles como São Paulo é cada vez mais frequente a adoção de transportes públicos como ônibus, metrô ou táxi para se deslocar pela cidade. No entanto, para pessoas que dependem exclusivamente desses meios de transporte, muitas vezes a escolha de um passeio, restaurante ou estabelecimento em geral pode depender da proximidade, por exemplo, de uma estação de metrô. Indo além, podemos querer saber os restaurantes abertos localizados perto de uma determinada palestra que desejamos assistir. Atualmente os portais de divulgação de bares e restaurantes não dispõem de inteligência para definir o significado de "perto"ou "longe"e por isso não são capazes de trazer esses dados com precisão. Desse modo, ao combinar informações como estações de metrô com os diferentes eventos na cidade e adicionar significado a essa informação de maneira a permitir estabelecer questões como proximidade entre diferentes pontos pode tornar a pesquisa de um individuo ainda mais rica e relevante.

2. Panorama tecnológico

A busca da informação na Internet pode ser melhorada a partir da integração e correlacionamento das informações publicadas na Internet. A definição e uso de ontologias como às

do projeto Schema.org[1] são um passo nesse sentido e o fazem atribuindo significado à informação, marcando o conteúdo a fim de permitir que o computador passe a compreender conceitos mais abstratos como Teatro ou Cinema. Apesar disso, não são suficientes para alcançar a integração dos dados, pois dependem de mecanismos capazes de recuperar as informações contidas nos diferentes portais de dados para marcar a informarção e então convertê-las para a terminologia homogênea pré-definida. Informações sobre uma mesma peça de teatro, uma sessão de cinema ou um evento cultural qualquer podem muitas vezes serem encontradas dentro de diferentes portais na Internet. Desse modo, a recuperação dos dados que estão espalhados na rede deve levar em consideração essa condição. Um sistema capaz de centralizar esses dados dentro da ótica de ontologias deve ser capaz de identificar, tratar e mesclar os conteúdos encontrados propiciando informações mais completas.

A Web Semântica tem o potencial de promover auxilio a tomada de decisão sobre um assunto compartilhado. A proposição de uma arquitetura capaz de alcançar a integração do conteúdo de diferentes portais na Internet e de proporcionar consultas mais próximas do interesse do usuário a partir do uso de ontologias, é portanto, tema muito relevante. Ao recuperar a informação relevante dos portais da Internet e aplicar anotação semântica com o uso de ontologias, torna-se possível alcançar uma condição em que os dados podem ser centralizados, correlacionados, enriquecidos e publicados para novas consultas agora com semântica agregada. Tal condição possibilita responder perguntas que envolvem buscas complexas que dependem da informação que está inicialmente distribuída por entre diferentes portais como, por exemplo, quais restaurantes de comida italiana estão mais próximos a uma exposição que ocorre em São Paulo, quais eventos acontecem próximo ao metrô Butantã, entre outras. Esse é objetivo deste trabalho.

Na Internet os dados são publicados a todo momento mas ficam restritos a grandes silos de informação o que dificulta uma visão homogênea sobre um determinado domínio de interesse[Civili et al. 2013]. No entanto, a necessidade de gerenciar informações provenientes de fontes distintas promove a pesquisa acerca de maneiras mais inteligentes, capazes de lidar com as divergências entre documentos, duplicações ou ruídos, para realizar a integração de dados sobre um mesmo domínio [Vettor et al. 2014]. Esses mecanismos, por sua vez, podem ser descritos dentro de duas abordagens distintas e conhecidas como Global As View (GAV) ou Local As View (LAV) [Abdellaoui and Nader 2015, Wang et al. 2017, Putra and Khalil 2017]. A estratégia GAV é tradicionalmente utilizada para aplicações em que há consultas federadas nas quais uma única consulta dispara pesquisas em múltiplas fontes de dados e unifica a informação recuperada por meio de múltiplas camadas de abstrações. Já o método LAV realiza a materialização desses dados em um banco de dados único. Neste trabalho optamos por aplicar a estratégia LAV que funciona melhor para o contexto da Internet em que existem situações nas quais existem fontes de dados incompletas, que podem estar inacessíveis em determinado momento [Putra and Khalil 2017].

3. Trabalhos relacionados

Em geral os estudos que buscam a integração dos dados na Internet atribuem as máquinas um papel relevante e vão desde o uso de ontologias para mapear um domínio comum

[1] `http://schema.org`

visando a solução do problema da integração de dados heterogêneos [Ahmed 2008], a integração baseada em Sistemas Multi-Agentes [Sui et al. 2009] ou o acesso a informação baseado em ontologias [Civili et al. 2013, Kharlamov et al. 2013].

Levando em consideração que a maior parte dos documentos existentes na Web está definida valendo-se de formatos semi-estruturados, e.g. XML, é de se esperar que a integração de dados seja feita por meio de anotações semânticas. Iniciativas como o SIOC [Bojars et al. 2008] buscam uma proposta valendo-se do apontamento ontológico em RDF para interligar redes sociais como Flickr e Facebook através das APIs disponibilizadas por estes sites, outras estudos como o Deep Annotation [Handschuh et al. 2003] propõem a construção de uma ferramenta capaz de facilitar a anotação semântica dos dados já expostos na Web de anotar os dados. Já trabalhos como o Bottari [Balduini et al. 2012] aliam a interpretação da diversidade de conteúdo produzido das publicações de pessoas no Twitter, seguida de mapeamento semântico desses dados em uma ontologia padronizada para sugerir pontos de interesse. Trabalhos como o SBWS[2] e o ASSAM [Heßand Kushmerick 2003, Heßet al. 2004] buscam realizar mapeamento semântico sobre uma descrição de serviços WSDL[3] que funcionam sobre o procolo SOAP[4].

Outra linha de estudo é o acesso a informação de banco de dados relacionais pré-existentes. Trabalhos como o VirtuosoRDF[5], D2RQ[6], Ontop [Rodríguez-Muro et al. 2013, Calvanese et al. 2016] e o MastroStudio [Civili et al. 2013], são capazes de gerar representações RDF que derivam diretamente de acordos implícitos e explícitos dos bancos de dados (BD) relacionais permitindo assim o acesso à informação baseado em ontologias (ODBA). Neste trabalho busca-se a integração de dados expostos na Internet, por meio da interpretação dos documentos expostos na Web e exposição da informação anotada semanticamente a fim de possibililtar o acesso a dados baseado em Ontologias para um domínio específico.

4. Proposta de arquitetura

Uma arquitetura de integração da informação na Internet depende da definição de uma estrutura que represente com clareza um determinado domínio de conhecimento [Ahmed 2008] e de mecanismos capazes de interpretar, identificar e converter para uma semântica comum os dados relevantes sobre os documentos publicados. O escopo dos domínios de dados deve ser delimitado bem como as fontes de informação da qual serão extraídos. Além disso deve-se considerar também o acesso à informação, a forma pela qual a informação recuperada e anotada poderá ser utilizada. Propomos uma arquitetura formada por uma camada de recuperação de dados, uma camada de persistência e centralização da informação e uma camada de apresentação para o acesso à informação, conforme ilustrado pela Figura 1.

4.1. Domínio de escopo

Nesta proposta o escopo foi delimitado para eventos e restaurantes uma vez que são diversas as fontes de divulgação a respeito de exposições, peças de teatros, palestras, aulas,

[2]http://asio.bbn.com/sbws.html
[3]https://www.w3.org/TR/wsdl
[4]https://www.w3.org/TR/soap/
[5]http://virtuoso.openlinksw.com/
[6]http://d2rq.org/

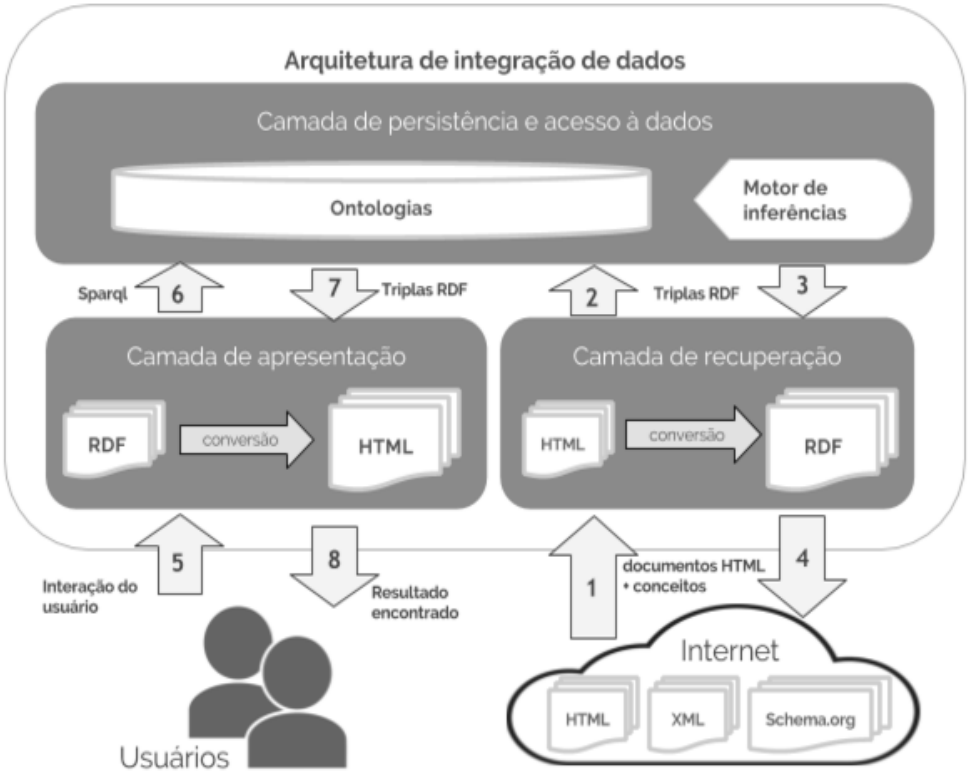

Figura 1. Proposta de arquitetura de sistema de integração de dados da Web

entre outros a fim de tornar possível uma visão ampla sobre os eventos que acontecem em determinada data, horário ou região tomando como base a geolocalização. Quanto às fontes de dados utilizadas, neste trabalho adotaram-se os portais Guia Da Semana[7], Guia da Folha[8] e o portal de Eventos da USP[9]. Os portais foram escolhidos dado a relevância deles na divulgação dos eventos aos quais se dedicam.

4.2. Recuperação e anotação

Delimitado escopo e portais de conteúdo, define-se o processo de recuperação e anotação da informação. Em primeiro lugar, os dados são recuperados dos portais de eventos escolhidos a partir de técnicas de recuperação de informação usando expressões regulares. O conteúdo selecionado é anotado com uma ontologia obtida do portal Schema.org, gerando assim uma informação com semantica agregada. A informação é então armazenada em um repositório de dados para acesso baseado em ontologias. Como resultado deste processo, obtém-se um repositório de dados semanticamente anotado que pode ser consultado de maneira centralizada. Neste trabalho utilizamos os conceitos "Estabelecimento de alimentos" (FoodEstablishment), "Evento" (Event) e suas respectivas derivações.

O processo de identificar e extrair dados nos portais escolhidos acontece na camada de recuperação de dados. Ela contém toda a inteligência da recuperação da informação relevante a partir da mediação da requisição para uma fonte de dados na Internet e a

[7] `https://www.guiadasemana.com.br/`
[8] `http://guia.folha.uol.com.br/`
[9] `http://www.eventos.usp.br/`

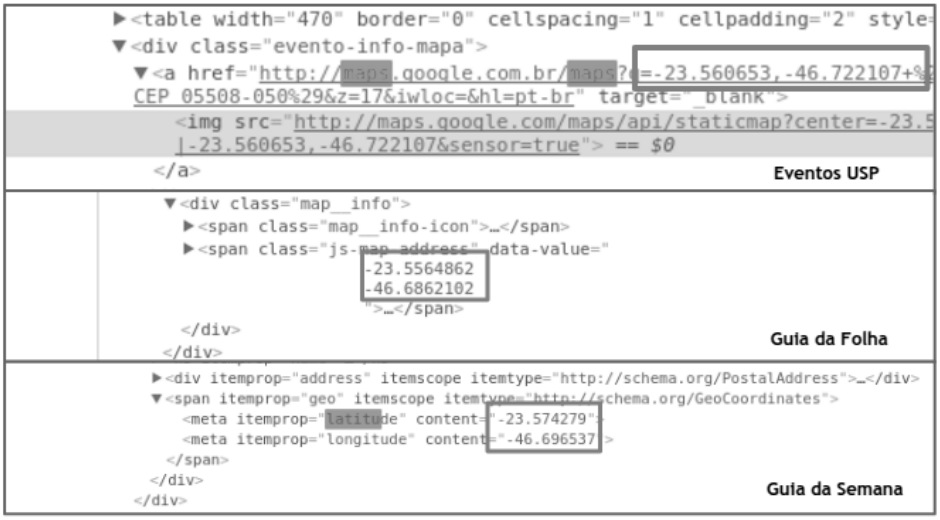

Figura 2. Padrões de repetição em portais de conteúdo

consequente transformação daquele conteúdo escolhido para um documento RDF válido conforme representado pela setas 1 na Figura 1. Essa transformação decorre da criação de um conjunto de regras baseadas em expressões regulares criadas especificamente para cada portal escolhido a partir da identificação de padrões de repetição de dados contidas em cada um dos portais escolhidos. A Figura 2 ilustra a identificação do padrão de repetição para a informação sobre latitude de longitude de eventos em diferentes portais. Desse modo, a partir da junção de diferentes expressões regulares trabalhando conjuntamente sobre o conteúdo dos diferentes documentos dos portais escolhidos é que emerge a inteligência da interpretação do conteúdo relevante nesses portais. Finalmente, toda a informação recuperada é então transformada em triplas RDF que podem ser armazenadas em uma base de dados para consultas posteriores.

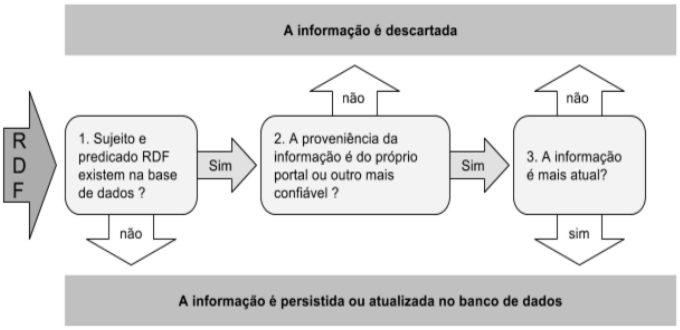

Figura 3. Processo de resolução de conflitos

4.3. Persistência

A camada de persistência e centralização possui a responsabilidade de persistir de forma centralizada as informações recuperadas na camada anterior, o que está indicado pelo fluxo de informação nas setas 2 e 3 da arquitetura. Também é responsável por resolver possíveis conflitos de informação, inferir sobre os dados obtidos e disponibilizar acesso

a eles. A centralização dos dados é importante dada a natureza descentralizada da Internet; uma vez que a informação está espalhada por servidores na Web, estes podem apresentar barreiras para a integração de dados em tempo real como, por exemplo, a velocidade de transferência de dados, quando os servidores estão distantes fisicamente, ou por indisponibilidade de dados, quando o servidor não funciona corretamente. Nesses casos, uma simples busca realizada em um modelo descentralizado pode tornar-se excessivamente lenta, uma vez que toda a informação contida nos diferentes portais escolhidos precisará ser obtida, armazenada e inferida a cada consulta. Já a atualização dos dados acontece por meio de um processo diário de interpretação da informação publicada nos portais escolhidos que ocorre de forma paralela ao processo de inferência sobre a informação. Finalmente o acesso aos dados, representado pelas setas 6 e 7, acontece por meio da exposição de serviços para consulta SPARQL.

Dados sobre eventos e restaurantes podem ser enriquecidos e contextualizados através de suas respectivas geolocalizações. Na proposta de arquitetura deste trabalho a camada de persistência com o uso de ontologias admite a definição de conceitos específicos como a definição do que é perto (ou do que é longe) em relação a outro ponto que possui latitude e longitudes definidos. É possível então entender se determinada palestra ocorre perto de um metrô ou de um determinado restaurante. O conceito "perto" foi definido neste trabalho com a distância de quinhentos metros para atingir tal finalidade. Assim, caso dois pontos "a" e "b" que possuem respectivamente as latitudes e longitudes (x1,y1) e (x2,y2) estiverem distantes em um raio de até quinhentos metros, então uma nova tripla "?a iweb:near ?b" é adicionada na base de dados sugerindo que "a" está perto de "b".

4.3.1. Resolução de conflitos

A resolução de duplicações e conflitos é importante quando informações sobre um mesmo domínio são recuperadas de diferentes portais. Nesses casos, diferentes fontes podem, por exemplo, descrever não somente um mesmo evento mas também informações distintas sobre ele como datas diferentes de uma apresentação de uma peça de teatro ou endereços divergentes sobre um show tornando assim necessário escolher qual das informações sobre as diferentes propriedades será mantida. Neste trabalho a resolução de conflitos acontece por meio de um processo de avaliação baseado em regras pré-definidas, criadas a partir do domínio dos dados escolhido na qual se decide se um tripla RDF será armazenada ou descartada como ilustra a Figura 3. A procedência da informação é também considerada nesse processo. Isso significa que alguns dados podem ser escolhidos em detrimento de outros, baseado em uma ordem de confiabilidade das fontes de informação. Desse modo, a decisão entre duas informações contidas em triplas que possuem objetos diferentes sendo uma extraída de um site A e de outro site B dependem da ordem de precedência pré-estabelecida. Supondo que se estabeleça que a informação do portal B é mais relevante que a do portal A pelo pelo fato deste último ter conteúdo mais preciso, maior abrangência territorial ou qualquer outro, então a informação de A será sempre substituída pela informação de B. Por fim, os dados são substituídos quando a ordem de preferência não é estabelecida.

4.4. Apresentação

A camada de apresentação simplifica e contextualiza a busca de informações sobre o domínio. Representada pelas setas 5 e 8 ela recebe consultas que são traduzidas para SPARQL sem exigência de conhecimento prévio nessa linguagem. Por outro lado, não impede consultas mais elaboradas por usuários mais avançados nessa linguagem. O resultado da consulta é desenhado em cima de uma mapa, que contém informações sobre ruas e estabelecimentos na região dos eventos e estabelecimentos encontrados, como mostrado na Figura 4.

5. Resultados

O primeiro aspecto avaliado foi a capacidade de integrar informações de diferentes portais. A Figura 4 mostra a distribuição geográfica da informação sobre os eventos publicados nas fontes escolhidas. Os eventos recuperados do portal da USP estão rotulados como item A e estão concentrados em regiões próximas a campus da USP, como na região do Butantã, São Carlos e Ribeirão Preto. Os eventos obtidos do portal Guia Da Semana, rotulados como B, estão aglomerados na região central da cidade de São Paulo e abrangem, em maioria, peças de teatro, exposições e shows. Já o Guia da Folha, rotulado como C, possui conteúdo mais diversificado em toda a cidade de São Paulo, abrangendo desde restaurantes a exposições. O item D, por sua vez ilustra uma consulta realizada sobre uma implementação da arquitetura proposta neste trabalho, apresentando todos os resultados que estão próximos do metrô Sé, utilizando o conceito que define a proximidade e agregando os valores dos diferentes portais.

Como mencionado anteriormente, a resolução de conflitos é uma tarefa essencial quando estamos lidando com informações provenientes de diferentes fontes de dados. Um exemplo desta situação se refere à peça de teatro “A Era do Rock”, publicada tanto no portal Guia da Semana quanto no portal Guia da Folha. Enquanto a primeira fonte indica uma localização claramente incorreta fora do país, a segunda a informação marca corretamente o Teatro Porto Seguro, conforme ilustra a Figura 5. Quando a informação é recuperada do portal Guia da Folha, ela é admitida durante o processo de resolução de conflito e como tal portal Guia da Folha possui preferência sobre a informação do site Guia da Semana, a geolocalização incorreta é então substituída e a informação torna-se mais confiável.

A Figura 6 mostra o conflito já resolvido na implementação da arquitetura. Na parte superior da imagem, está o exemplo de quando os dados do evento são capturados de uma fonte de dados com inconsistência da informação e adicionados à base de conhecimento, e abaixo dela os dados já ajustados após o processo de resolução de conflitos por fonte mais confiável. Além disso, outra vantagem observável na proposta apresentada neste trabalho é a capacidade de combinar as informações provenientes das diferentes fontes escolhidas. No exemplo anterior, além de atualizar a informação sobre latitude e longitude do evento, a informação também foi complementada com uma descrição mais específica com o título "overview"o que agrega mais detalhes e consequentemente oferece uma informação mais abrangente ao usuário final.

6. Conclusões

A Web Semântica tem o potencial de correlacionar dados espalhados entre fontes de informação diversas na Web, contidos em diferentes portais e representados em formatos

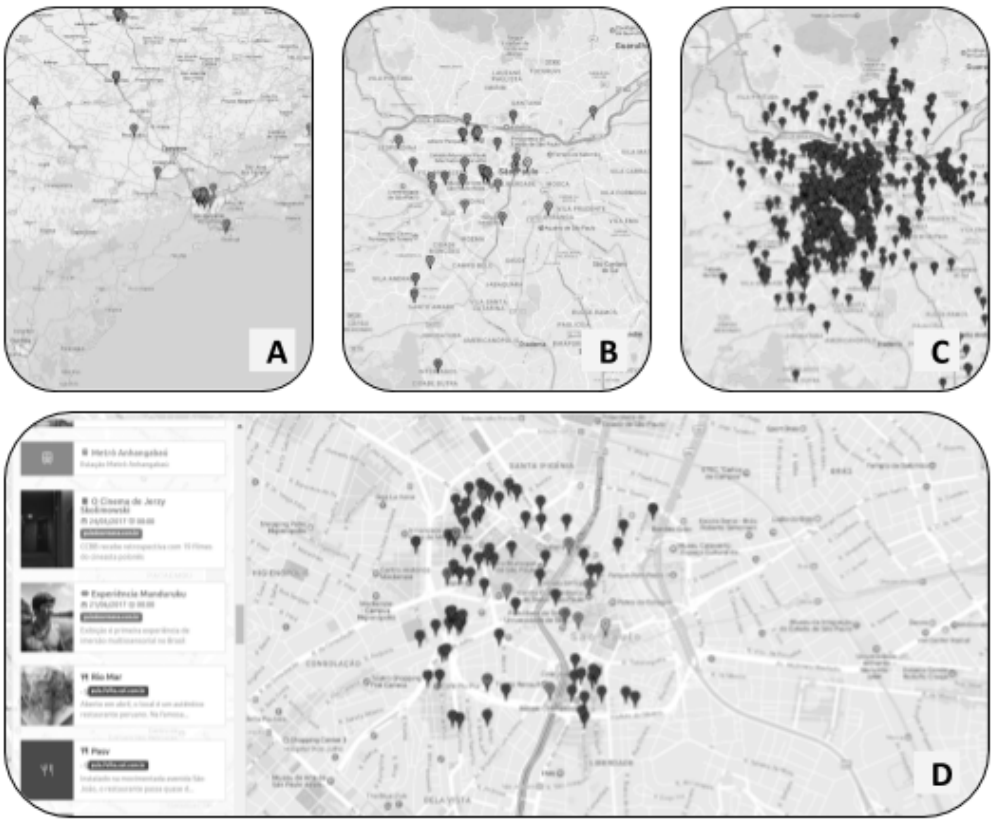

Figura 4. Distribuição e integração de fontes de dados

variados. A aplicação de ontologias auxilia a elevar a qualidade da informação, atribuindo significado aos dados publicados e propiciando consultas mais ricas e mais próximas dos interesses finais dos indivíduos, auxiliando-os na tomada de decisões do dia-a-dia. A premissa deste trabalho é que ao combinar a tecnologia de Web Semântica com mecanismos de recuperação de dados na Web, permite-se que o conteúdo relevante dos portais espalhados pela Internet possa ser extraído de maneira automatizada para oferecer um resultado muito mais expressivo ao usuário final. Nesse contexto, apresentamos uma proposta de arquitetura capaz de permitir a integração da informação contida em portais heterogêneos, a partir do uso de representações bem estabelecidas do portal Schema.org, e a centralização da informação para consulta aos dados com o uso de ontologias.

Há grandes obstáculos a serem superados no que diz respeito à extração de dados. Como a maioria dos portais não anota semanticamente o seu conteúdo, que é gerado dinamicamente, torna-se necessária a aplicação de artifícios para a recuperação de conteúdo a partir da estrutura sintática desses documentos. Propostas como a apresentada neste trabalho são vulneráveis à mudança da maneira pela qual a informação é exposta para os usuários. Em outras palavras, se um determinado portal muda sua forma de apresentação para os seus usuários, o processo de extração de dados deve ser atualizado. Além disso, quanto maior a frequência de modificação da estrutura desses documentos, maior é a quantidade de manutenção na infra-estrutura de recuperação do conteúdo do portal.

A própria natureza da Internet é um obstáculo para a recuperação de dados dada a necessidade de se percorrer variados portais da Web para a construção de um resultado

Figura 5. Conflito de informações em diferentes portais

abrangente. Isso porque nesse tipo de solução há variáveis a se considerar que englobam velocidade do servidor, disponibilidade da informação, estrutura de rede, entre outros. Além disso, é inviável a recuperação de todo o conteúdo de grandes portais a cada consulta, uma vez que essas organizações possuem bases gigantescas de dados. A centralização de conteúdo em uma base de dados semântica diminui os problemas intrínsecos à rede e permite que o conhecimento possa ser construído e atualizado de maneira gradual, o que tende a oferecer mais informações a qualquer momento e independe do gargalo ocasionado pela consulta federada em várias fontes de informação de maneira simultânea. Desse modo a inferência sobre dados também não precisa acontecer após cada consulta realizada, mas sim de maneira assíncrona, o que implica em resultados mais rápidos.

A arquitetura foi implementada e testada utilizando-se diferentes conceitos, tais como exposições, peças de teatro e restaurantes. Além disso, os dados sobre um determinado domínio podem divergir de acordo com a fonte pela qual esta foi extraída, o que pode ser superado por meio de mecanismos de resolução de conflitos como a priorização de fontes. Por fim, informações que dizem respeito a um mesmo conteúdo podem ser agregadas, tornando a informação mais completa e relevante.

Diferente de trabalhos como o VirtuosoRDF e o D2RQ que se valem das definições estruturais de tabelas, colunas entre outras características bem definidas no banco de dados, neste trabalho utilizamos a informação contida na própria Internet para promover a junção das fontes de informação através do mapeamento semântico. Espera-se que este trabalho possa ser mais um incentivo para o reuso de informação exposta na Internet e para o avanço da Web Semântica. Trata-se de uma amostra de que a nova proposta da Web na qual todos os dados estão interconectados não é uma utopia e pode estar mais próxima do que imaginamos.

Figura 6. Resolução de conflito e mesclagem de dados

A conclusão deste trabalho abre portas para novas pesquisas voltadas para a melhoria da integração da informação na Web. Entre as sugestões futuras, encontram-se: (i) a pesquisa de mecanismos mais eficientes de reconhecimento e recuperação de informação relevantes, (ii) a auto-detecção e mapeamento em tempo real dos dados contidos em documentos Web para documentos semanticamente anotados, (iii) a construção de motores capazes de converter consultas SPARQL em HTML, assim como hoje algumas consultas SPARQL podem ser traduzidas diretamente para SQL e (iv) mecanismos capazes de reconhecer a mudança da estrutura sintática das páginas HTML e que tenham a habilidade de se auto ajustar a elas.

Referências

Abdellaoui, S. and Nader, F. (2015). Semantic Data Warehouse at the heart of Competitive Intelligence Systems: design approach. In *2015 6Th International Conference on Information Systems and Economic Intelligence (Siie)*, pages 141–145. IEEE.

Ahmed, E. (2008). Resource capability discovery and description management system for bioinformatics data and service integration - An experiment with gene regulatory networks. In *Proceedings of 11th International Conference on Computer and Information Technology, ICCIT 2008*, pages 56–61. IEEE.

Balduini, M., Celino, I., Dell'Aglio, D., Della Valle, E., Huang, Y., Lee, T., Kim, S. H., and Tresp, V. (2012). BOTTARI: An augmented reality mobile application to deliver personalized and location-based recommendations by continuous analysis of social media streams. *Journal of Web Semantics*, 16:33–41.

Bojars, U., Breslin, J. G., Finn, a., and Decker, S. (2008). Using the Semantic Web for linking and reusing data across Web 2.0 communities. *Web Semantics*, 6(1):21–28.

Calvanese, D., Cogrel, B., and Komla-Ebri, S. (2016). Ontop: Answering SPARQL queries over relational databases. *Semantic Web*, 0(0).

Civili, C., Ruzzi, M., Santarelli, V., Savo, D. F., Console, M., De Giacomo, G., Lembo, D., Lenzerini, M., Lepore, L., Mancini, R., Poggi, A., and Rosati, R. (2013). Mastro Studio: Managing Ontology-based Data Access Applications. *Proceedings of the VLDB Endowment*, 6(12):1314–1317.

Gray, A. J., Groth, P., Loizou, A., Askjaer, S., Brenninkmeijer, C., Burger, K., Chichester, C., Evelo, C. T., Goble, C., Harland, L., Pettifer, S., Thompson, M., Waagmeester, A., and Williams, A. J. (2014). Applying linked data approaches to pharmacology: Architectural decisions and implementation. *Semantic Web*, 5(2):101–113.

Handschuh, S., Staab, S., and Volz, R. (2003). On deep annotation. *Proceedings of the twelfth international conference on World Wide Web - WWW '03*, page 431.

Heß, A., Johnston, E., and Kushmerick, N. (2004). Assam: A tool for semi-automatically annotating semantic web services. *3rd International Semantic Web Conference (ISWC 2004)*.

Heß, A. and Kushmerick, N. (2003). Learning to Attach Semantic Metadata to Web Services. *The Semantic Web - ISWC 2003*, 2870:258–273.

Kharlamov, E., Jiménez-Ruiz, E., Zheleznyakov, D., Bilidas, D., Giese, M., Haase, P., Horrocks, I., Kllapi, H., Koubarakis, M., Özçep, Ö., Rodríguez-Muro, M., Rosati, R., Schmidt, M., Schlatte, R., Soylu, A., and Waaler, A. (2013). Optique: Towards OBDA systems for industry. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7955 LNCS, pages 125–140. Springer Berlin Heidelberg.

Putra, S. J. and Khalil, I. (2017). Context for the intelligent search of information. In *2017 5th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–4. IEEE.

Rodríguez-Muro, M., Kontchakov, R., and Zakharyaschev, M. (2013). Ontop at work. In *PROC. OF OWL: EXPERIENCES AND DIRECTIONS WORKSHOP 2013 (OWLED 2013)*. CEUR-WS.

Stumme, G., Hotho, A., and Berendt, B. (2006). Semantic Web Mining: State of the art and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4:124–143.

Sui, X., Wang, S., and Li, Z. (2009). Research on the model of Integration with Semantic Web and Agent Personalized Recommendation System. In *2009 13th International Conference on Computer Supported Cooperative Work in Design*, pages 233–237. IEEE.

Vettor, P. D., Mrissa, M., Benslimane, D., and Berbar, S. (2014). A Service Oriented Architecture for Linked Data Integration. In *2014 IEEE 8th International Symposium on Service Oriented System Engineering*, pages 198–203. IEEE.

Wang, A., Croft, J., and Dragut, E. (2017). Reflections on Data Integration for SDN. In *Proceedings of the ACM International Workshop on Security in Software Defined Networks & Network Function Virtualization - SDN-NFVSec '17*, pages 65–68, New York, New York, USA. ACM Press.