Mineração da Web Semântica

Estado da arte e direções futuras

Gerd Stumme, Andreas Bettina Berendt Hotho

Resumo

O SemanticWebMining visa combinar as duas áreas de pesquisa de rápido desenvolvimento, o SemanticWeb e o WebMining. Esta pesquisa analisa a convergência de tendências de ambas as áreas: um número crescente de pesquisadores está trabalhando para melhorar os resultados do Web Mining, explorando estruturas semânticas na Web, e eles usam técnicas da Web Mining para construir a Web Semântica. Por último, mas não menos importante, essas técnicas podem ser usadas para minerar a própria Web Semântica.

A Web Semântica é a WWW de segunda geração, enriquecida por informações processáveis por máquina que suportam o usuário em suas tarefas. Dado o enorme tamanho da Web de hoje, é impossível enriquecer manualmente todos esses recursos. Portanto, esquemas automatizados para aprender as informações relevantes estão sendo cada vez mais utilizados. O Web Mining visa descobrir informações sobre o significado dos recursos da Web e seu uso. Dada a natureza principalmente sintática dos dados que estão sendo minerados, a descoberta de significado é impossível com base apenas nesses dados. Portanto, formalizações da semântica de sites e comportamento de navegação estão se tornando cada vez mais comuns. Além disso, a mineração da própria Web Semântica é outra aplicação futura. Argumentamos que as duas áreas de Web Mining e Web Semântica precisam uma da outra para cumprir seus objetivos, mas que todo o potencial dessa convergência ainda não foi realizado. Este documento fornece uma visão geral de onde as duas áreas se encontram hoje e esboça maneiras de como uma integração mais estreita pode ser lucrativa.

Palavras-chave

WebMining, Web Semântica, Ontologias, Descoberta do Conhecimento, Engenharia do Conhecimento, Inteligência Artificial, World Wide Web.

Eu NTRODUÇÃO

As duas áreas de pesquisa de rápido desenvolvimento Semantic Web e Web Mining se baseiam no sucesso da World Wide Web (WWW). Eles se complementam bem porque cada um deles aborda uma parte de um novo desafio apresentado pelo grande sucesso da atual WWW: A natureza da maioria dos dados na Web é tão desestruturada que só pode ser entendida por humanos, mas a quantidade de dados é tão grande que eles só podem ser processados com eficiência por máquinas. A Web Semântica aborda a primeira parte desse desafio, tentando tornar os dados (também) compreensíveis à máquina, enquanto a Web Mining aborda a segunda parte (semi) extraindo automaticamente o conhecimento útil oculto nesses dados e disponibilizando-os como agregação de proporções gerenciáveis.

A. Hotho e G. Stumme estão no Grupo de Engenharia de Dados e Conhecimento, Universidade de Kassel, D-34121 Kassel, Alemanha. E-mail: {hotho, stumme}@cs.uni-kassel.de. B. Berendt trabalha no Instituto de Sistemas de Informação da Universidade Humboldt de Berlim, Spandauer Str. 1, D-10178 Berlim, Alemanha. Correio eletrónico : berendt@wiwi.hu-berlin.de .

A Semantic Web Mining visa combinar as duas áreas: Web Semântica e Web Mining. Essa visão segue nossa observação de que as tendências convergem em ambas as áreas: um número crescente de pesquisadores trabalha para melhorar os resultados do Web Mining, explorando (as novas) estruturas semânticas na Web e utiliza as técnicas de Mineração da Web para construir a Web Semântica. Por último, mas não menos importante, essas técnicas podem ser usadas para minerar a própria Web Semântica. A redação

Mineração da Web Semântica enfatiza esse espectro de possível interação entre as duas áreas de pesquisa: pode ser lido tanto *Semântica (Mineração Web)* e como (*Web Semântica) Mineração.*

Nos últimos anos, houve muitas tentativas de "quebrar a barreira da sintaxe" [1 1] Na internet. Vários deles contam com informações semânticas nos corpora de texto que são implicitamente exploradas por métodos estatísticos. Alguns métodos também analisam as características estruturais dos dados; eles lucram com sintaxe padronizada como XML. Neste artigo, nos concentramos em abordagens de marcação e mineração que se referem a uma *conceituação explícita* de entidades no respectivo domínio. Eles relacionam os tokens sintáticos ao conhecimento de fundo representado em um modelo com *semântica formal.* Quando usamos o termo "semântico", temos em mente um modelo lógico formal para representar o conhecimento.

O objetivo deste artigo é fornecer uma visão geral de onde as duas áreas da Web Semântica e Mineração da Web se encontram hoje. Em nossa pesquisa, descreveremos primeiro o estado atual das duas áreas e, em seguida, discutiremos, usando um exemplo, sua combinação, descrevendo, assim, os tópicos de pesquisas futuras. Forneceremos referências a abordagens típicas. A maioria deles não foi desenvolvida explicitamente para fechar a lacuna entre a Web Semântica e a Mineração da Web, mas se encaixa naturalmente nesse esquema.

Nas próximas duas seções, apresentamos breves visões gerais das áreas Semantic Web e Web Mining. Os leitores familiarizados com essas áreas podem pular as seções II ou III, respectivamente. Em seguida, descrevemos como essas duas áreas cooperam hoje e como essa cooperação pode ser melhorada ainda mais. Primeiro, as técnicas de mineração da Web podem ser aplicadas para ajudar a criar a Web Semântica. A espinha dorsal da Web Semântica são as ontologias, que atualmente são feitas à mão. Esta não é uma solução escalável para uma aplicação abrangente de tecnologias da Web Semântica. O desafio é aprender ontologias e / ou instâncias de seus conceitos de maneira (semi-) automática. Uma pesquisa dessas abordagens está contida na Seção IV.

Por outro lado, o conhecimento de base - na forma de ontologias ou de outras formas - pode ser usado para melhorar o processo e os resultados do Web Mining. Desenvolvimentos recentes incluem a mineração de sites que se tornam cada vez mais sites semânticos e o desenvolvimento de técnicas de mineração que podem explorar o poder expressivo da representação do conhecimento da Web semântica. A seção V discute essas várias técnicas.

Na Seção VI, esboçamos como o loop pode ser fechado: da Mineração da Web à Semântica.

[1 1] Este título foi escolhido por S. Chakrabarti para sua palestra convidada na conferência ECML / PKDD 2004.

Web e vice-versa. Concluímos, na Seção VII, que uma forte integração desses aspectos aumentará bastante a compreensibilidade da Web para máquinas e, assim, se tornará a base para gerações futuras de ferramentas inteligentes da Web. Também voltamos às duas noções de “semântica” e delineamos seus pontos fortes, fracos e complementares. Parte dessa pesquisa substancialmente revisada e ampliada foi apresentada na 1ª Conferência Internacional da Web Semântica [16].

II S EMANTIC W EB

A Web Semântica é baseada na visão de Tim Berners-Lee, o inventor da WWW. O grande sucesso da atual WWW leva a um novo desafio: uma enorme quantidade de dados é interpretável apenas por seres humanos; o suporte da máquina é limitado. Berners-Lee sugere enriquecer a Web com informações processáveis por máquina que suportam o usuário em suas tarefas. Por exemplo, os mecanismos de pesquisa atuais já são bastante poderosos, mas ainda assim retornam frequentemente listas de ocorrências excessivamente grandes ou inadequadas. Informações processáveis por máquina podem apontar o mecanismo de busca para as páginas relevantes e, assim, melhorar a precisão e a recuperação.

Por exemplo, hoje é quase impossível recuperar informações com uma pesquisa por palavra-chave quando as informações estão espalhadas por várias páginas. Considere, por exemplo, a consulta de especialistas em mineração da Web em uma intranet da empresa, onde as únicas informações explícitas armazenadas são as relações entre as pessoas e os cursos que eles participaram, por um lado, e entre os cursos e os tópicos que eles abordam, por outro. Nesse caso, o uso de uma regra declarando que as pessoas que participaram de um curso sobre um determinado tópico têm conhecimento sobre esse tópico pode melhorar os resultados.

O processo de construção da Web Semântica é atualmente uma área de alta atividade. Sua estrutura deve ser definida, e essa estrutura deve ser preenchida com vida. Para tornar essa tarefa viável, deve-se começar pelas tarefas mais simples primeiro. As etapas a seguir mostram a direção em que a Web Semântica está indo:

1. Fornecer uma sintaxe comum para instruções compreensíveis por máquina.
2. Estabelecendo vocabulários comuns.
3. Concordando com uma linguagem lógica.
4. Usando o idioma para trocar provas.

Berners-Lee sugeriu uma estrutura de camadas para a Web Semântica. Essa estrutura reflete as etapas listadas acima. Segue-se o entendimento de que cada etapa por si só já fornecerá valor agregado, para que a Web Semântica possa ser realizada de maneira incremental.

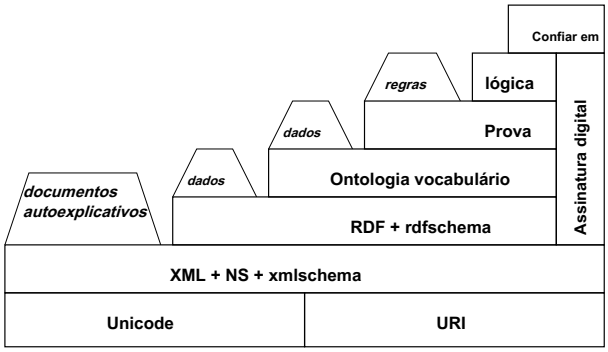

Fig. 1. As camadas da Web Semântica.

A. Camadas da Web Semântica

A Figura 1 mostra as camadas da Web Semântica, conforme sugerido por Berners-Lee.[2] Essa arquitetura é discutida em detalhes, por exemplo, em [138] e [139], que também abordam questões de pesquisa recentes.

Nas duas primeiras camadas, é fornecida uma sintaxe comum. *Identificadores uniformes de recursos (URIs)* fornecer uma maneira padrão de se referir a entidades,[3] enquanto *Unicode* é um padrão para a troca de símbolos. o *Linguagem de marcação extensível (XML)* fixa uma notação para descrever árvores rotuladas e o XML Schema permite a definição de gramáticas para documentos XML válidos. Documentos XML podem se referir a diferentes *namespaces* para tornar explícito o contexto (e, portanto, o significado) de diferentes tags. Atualmente, as formalizações nessas duas camadas são amplamente aceitas e o número de documentos XML está aumentando rapidamente. Embora o XML seja um passo na direção certa, ele apenas formaliza a estrutura de um documento e não o seu conteúdo.

o *Estrutura de descrição de recursos (RDF)* pode ser vista como a primeira camada em que as informações se tornam *compreensível:* De acordo com a recomendação do W3C[4] RDF "é uma base para o processamento de metadados; fornece interoperabilidade entre aplicativos que trocam informações compreensíveis por máquina na Web. "

Os documentos RDF consistem em três tipos de entidades: recursos, propriedades e instruções. Os recursos podem ser páginas da Web, partes ou coleções de páginas da Web ou qualquer objeto (do mundo real) que não faça parte diretamente da WWW. No RDF, os recursos são sempre abordados pelos URIs. Propriedades são atributos, características ou relações específicas que descrevem recursos. Um recurso juntamente com uma propriedade que possui um valor para esse recurso formam uma instrução RDF. Um valor é um literal, um recurso ou outra instrução. Assim, as declarações podem ser consideradas como triplos objeto-atributo-valor.

A parte do meio da Figura 2 mostra um exemplo de instruções RDF. Dois dos autores de

[2] Vejo http://www.w3.org/DesignIssues/Semantic.html

[3] *URL (localizador uniforme de recursos)* refere-se a um URI localizável, por exemplo, um http: // ... endereço. É frequentemente usado como sinônimo, embora os URLs estritamente falando sejam uma subclasse de URIs, consulte http://www.w3.org/Addressing.

[4] http://www.w3.org/TR/REC-rdf-syntax-grammar-20040210/

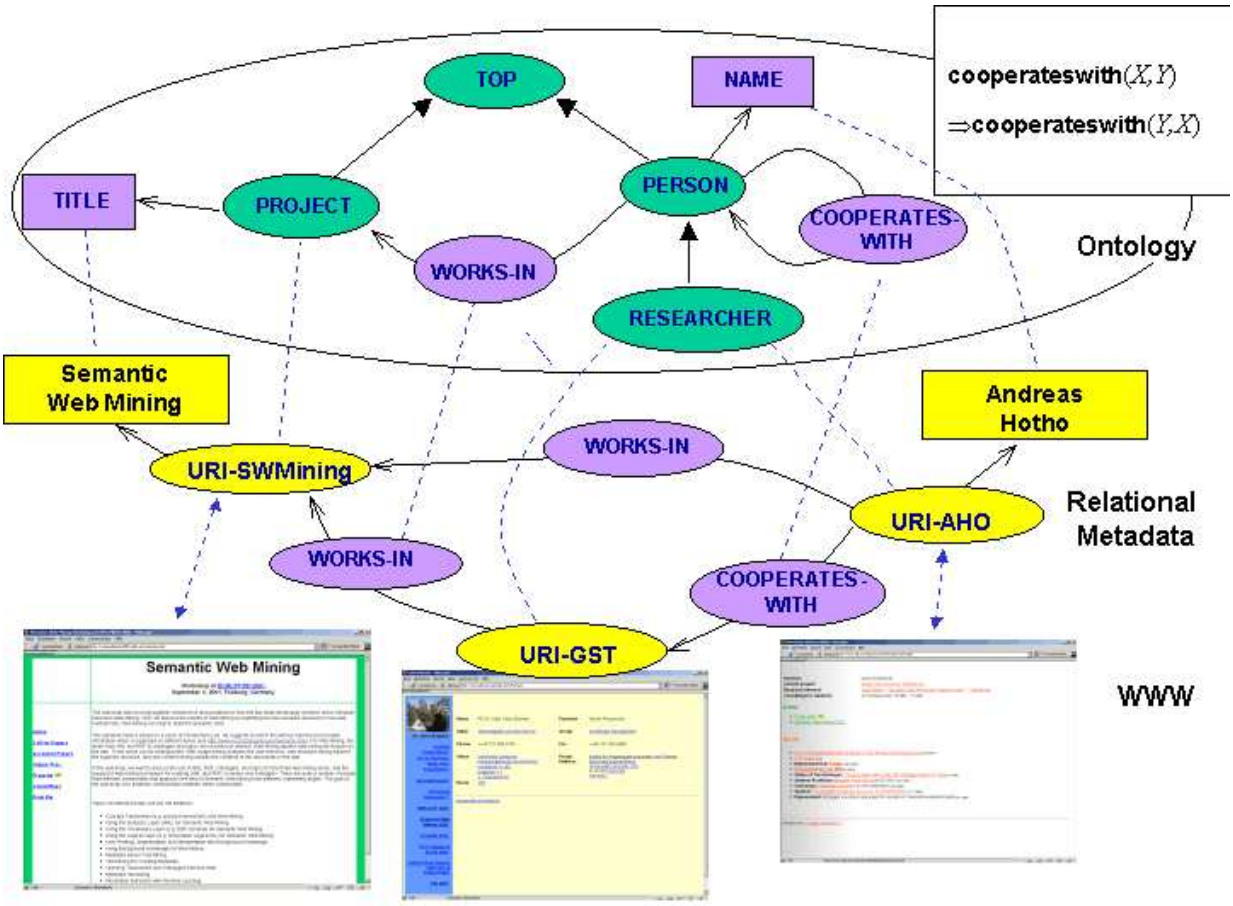

Fig. 2. A relação entre WWW, metadados relacionais e ontologias.

o presente trabalho (ou seja, suas páginas da Web) é representado como recursos 'URI-GST' e 'URIAHO'. A declaração no canto inferior direito consiste no recurso 'URI-AHO' e a propriedade 'coopera com' com o valor 'URI-GST' (que novamente é um recurso). O recurso 'URISWMining' tem como valor para a propriedade 'title' a literal 'Semantic Web Mining'.

O modelo de dados subjacente ao RDF é basicamente um gráfico rotulado direcionado. O esquema RDF define uma linguagem de modelagem simples sobre o RDF, que inclui classes, é um relacionamento entre classes e entre propriedades e restrições de domínio / intervalo para propriedades. RDF e RDF Schema são gravados na sintaxe XML, mas eles não empregam a semântica em árvore do XML.

Os esquemas XML e XML foram projetados para descrever a estrutura dos documentos de texto, como HTML, Word, StarOffice ou L UMA Documentos TEX. É possível definir tags em XML para transportar metadados, mas essas tags não possuem semântica formalmente definida e, portanto, seu significado não será bem definido. Também é difícil converter um documento XML em outro sem nenhuma semântica adicional especi fi cada das tags usadas. O objetivo do XML é agrupar os objetos de conteúdo, mas não descrever o conteúdo. Assim, o XML ajuda a organizar documentos, fornecendo uma sintaxe formal. Isso não é 'semântico' no sentido de nossa pesquisa. Erdmann [53] fornece uma análise detalhada dos recursos do XML, as deficiências do XML em relação à semântica e possíveis soluções.

A próxima camada é a *vocabulário de ontologia.* Seguindo [68], uma ontologia é "uma formalização explícita de um entendimento compartilhado de uma conceituação". Essa definição de alto nível é realizada de maneira diferente por diferentes comunidades de pesquisa. No entanto, a maioria deles tem um certo entendimento em comum, pois a maioria deles inclui um conjunto de *conceitos,* uma hierarquia sobre eles, e *relações* entre conceitos. A maioria deles também inclui axiomas em alguma lógica específica. Discutiremos as abordagens mais importantes em mais detalhes na próxima subseção. Para dar um sabor, apresentamos aqui apenas o núcleo de nossa própria definição [161], [23], como é refletida pelo quadro KAON de Karlsruhe Ontology.[5] Ele é construído de forma modular, para que diferentes necessidades possam ser atendidas pela combinação de peças.

Definição 1: UMA *ontologia central com axiomas* é uma estrutura $O: = (C, \leq_C, R, \sigma \leq_R, A)$ consiste em

- dois conjuntos disjuntos C e R cujos elementos são chamados *identificadores de conceito* e *identificadores de relação,*

resp.,

- uma ordem parcial $\leq_C$ em C, chamado *hierarquia de conceito* ou *taxonomia,*
- uma função $\sigma: R \rightarrow C+$ chamado *assinatura (* Onde $C+$ é o conjunto de todas as tuplas finas de elementos em C)
- uma ordem parcial $\leq_R$ em R, chamado *hierarquia de relações,* Onde $r_1 1 \leq_R r_2$ implica $| \sigma(r_1) = | \sigma(r_2)$ e $\pi_{Eu}(\sigma(r_1)) \leq_C \pi_{Eu}(\sigma(r_2))$, para cada $1 1 \leq Eu \leq |\sigma(r_1)|$, com π_{Eu} sendo a projeção no Eu quinto componente e
- um conjunto UMA de axiomas lógicos em alguma linguagem lógica EU.

Essa definição constitui uma estrutura central que é bastante direta, bem aceita e que pode ser facilmente mapeada na maioria das linguagens de representação de ontologia existentes. Passo a passo, a definição pode ser estendida levando-se em consideração léxicos e bases de conhecimento [161].

Como exemplo, dê uma olhada na parte superior da Figura 2. O conjunto C de conceitos é o conjunto {Topo, Projeto, Pessoa, Pesquisador, Literal} e a hierarquia de conceitos $\leq_C$ é indicado pelas setas com uma ponta de seta cheia. O conjunto R das relações é o conjunto {trabalha, coopera com, nome, título}. A relação 'entrada' tem (Pessoa, Projeto) como assinatura, a relação 'nome' tem (Pessoa, Literal) como assinatura.[6] Neste exemplo, a hierarquia das relações é plana, ou seja, $\leq_R$ é apenas a relação de identidade. (Um exemplo de uma hierarquia não plana de relações será mostrado abaixo na Fig. 3.) Até aqui, o esquema RDF seria suficiente para formalizar a ontologia. Mas muitas vezes ontologias também contêm axiomas lógicos. O da Figura 2 afirma, por exemplo, que a relação 'coopera com' é simétrica. Isso será usado para inferir no nível lógico.

Os objetos do nível de metadados agora podem ser vistos como instâncias dos conceitos de ontologia. Para

[5] http://kaon.semanticweb.org

[6] Por convenção, as relações com o Literal como faixa são traçadas dessa maneira, porque em alguns contextos são consideradas *atributos.*

Por exemplo, 'URI-SWMining' é uma instância do conceito 'Projeto' e, por herança, também do conceito 'Top'.

Lógica é a próxima camada, de acordo com Berners-Lee. Hoje, a maioria das pesquisas trata os níveis de ontologia e lógica de maneira integrada, porque a maioria das ontologias permite axiomas lógicos. Aplicando dedução lógica, pode-se inferir novos conhecimentos a partir das informações que são declaradas implicitamente. Por exemplo, o axioma mencionado acima permite inferir logicamente que a pessoa endereçada por 'URI-AHO' coopera com a pessoa endereçada por 'URI-GST'. O tipo de inferência possível depende muito da lógica escolhida. Discutiremos esse aspecto na próxima subseção em mais detalhes.

Prova e *Confiar em* são as camadas restantes. Eles seguem o entendimento de que é importante poder verificar a validade das declarações feitas na Web (semântica), e que confiam na Web semântica e na maneira como processam as informações aumentará na presença de declarações assim validadas. Portanto, o autor deve fornecer uma prova que deve ser verificada por uma máquina. Nesse nível, não é necessário que a máquina do leitor encontre a prova em si, apenas 'deve' verificar a prova fornecida pelo autor. Essas duas camadas raramente são abordadas na pesquisa de hoje. Portanto, focaremos nosso interesse nas camadas XML, RDF, ontologia e lógica no restante deste artigo.

B. Ontologias: Idiomas e Ferramentas

A priori, qualquer mecanismo de representação do conhecimento [7] pode desempenhar o papel de uma linguagem da Web semântica. *Lógica de quadros (* ou *F-lógica; [*98]) é um candidato, uma vez que fornece uma representação de conhecimento semanticamente fundamentada, com base na metáfora de quadro e slot. Outro formalismo que se encaixa bem na estrutura da RDF são os Gráficos conceituais [148], [42]. Eles também fornecem uma metáfora visual para representar a estrutura conceitual.

Provavelmente, o framework mais popular no momento é o Description Logics (DL). DLs são subconjuntos de lógica de primeira ordem que visam ser o mais expressivos possível enquanto ainda são decidíveis. A lógica da descrição *SHIQ* fornece a base para o DAML + OIL, que, por sua vez, é o resultado da união dos esforços de dois projetos: A Linguagem de Marcação do Agente DARPA, DAML [8] foi criado como parte de um programa de pesquisa iniciado em agosto de 2000 pela DARPA, uma organização governamental de pesquisa dos EUA. OIL (Ontology Inference Layer) é uma iniciativa financiada pelo programa da União Europeia. A versão mais recente do DAML + OIL foi lançada como uma recomendação do W3C sob o nome OWL. [9]

Várias ferramentas estão sendo usadas para a criação e manutenção de ontologias e metadados, bem como

[7] Veja [159] para uma discussão geral.

[8] http://www.daml.org

[9] http://www.w3.org/TR/owl-features/

quanto ao raciocínio dentro deles. *Ontoedit [* 165], [166] é um editor de ontologia conectado a *Ontobroker [* 58], um mecanismo de inferência para a F – Logic. Ele fornece meios para manipulação de consultas baseadas em semântica sobre recursos distribuídos. A F – Logic também influenciou o desenvolvimento do Triple [147], um mecanismo de inferência baseado na lógica Horn, que permite a modelagem de recursos de UML, Mapas de Tópicos ou Esquema RDF. Ele pode interagir com outros mecanismos de inferência, por exemplo, com FaCT ou RACER.

Facto [10] fornece serviços de inferência para o idioma da descrição *SHIQ.* Em [83], o raciocínio dentro *SHIQ* e sua relação com DAML + OIL é discutida. O raciocínio é implementado no mecanismo de inferência do FaCT, que também sustenta o editor de ontologia OilEd [12]. RACER [69] é outro raciocínio para *SHIQ,* com ênfase no raciocínio sobre instâncias.

O Karlsruhe Ontology Framework KAON [23] é uma infraestrutura de gerenciamento e aprendizado de ontologia de código aberto direcionada para aplicativos de negócios. Ele inclui um conjunto abrangente de ferramentas que permite a criação fácil de ontologias, suportada pelo algoritmo e gerenciamento de aprendizado de máquina, além da criação de aplicativos baseados em ontologias. O conjunto de ferramentas também está conectado aos bancos de dados para permitir trabalhar com um grande número de instâncias. Prot´ ege 2000 [132] é independente de plataforma por exemplo ambiente para criação e edição de ontologias e bases de conhecimento. Como o KAON, ele possui uma estrutura de plug-in extensível. O gergelim [91] é uma arquitetura para armazenamento eficiente e consulta expressiva de grandes quantidades de dados RDF (S). Ele fornece suporte para controle de simultaneidade, exportação independente de informações RDF (S) e um mecanismo de consulta para RQL, uma linguagem de consulta para RDF. Uma extensa visão geral das ferramentas de ontologia pode ser encontrada em [64].

C. Áreas de pesquisa e áreas de aplicação relacionadas

Uma das muitas áreas de pesquisa relacionadas à Web Semântica são os bancos de dados. Nos últimos anos, a maioria dos sistemas comerciais de gerenciamento de banco de dados incluiu a possibilidade de armazenar dados XML para acomodar também dados semiestruturados. Como a comunidade de banco de dados trabalha há muito tempo nas técnicas de mineração de dados, pode-se esperar que mais cedo ou mais tarde a 'mineração XML' se torne um tópico de pesquisa ativo. De fato, existem primeiras abordagens nessa direção [111]. Do nosso ponto de vista, isso pode ser visto como um caso especial de Semantic Web Mining.

Mais geral, vários problemas (e soluções) no domínio do banco de dados também são encontrados na engenharia de ontologia, por exemplo, mapeamento de esquema ou integração de fontes de dados distribuídas e heterogêneas. Isso é tratado com mais detalhes na Seção IV-A, onde também discutimos maneiras de derivar ontologias de esquemas de banco de dados.

Outra área de pesquisa relacionada são os Mapas de Tópicos [11] que representam a estrutura das relações entre os sujeitos. A maioria do software para mapas de tópicos usa a sintaxe do XML, assim como o RDF

[10] http://www.cs.man.ac.uk/ ~ horrocks / FaCT

[11] http://www.topicmaps.org/

faz. De fato, Mapas de tópicos e RDF estão intimamente relacionados. Em [8], é fornecida uma estrutura formal para Mapas de Tópicos, que também pode ser aplicada ao RDF. A Mineração da Web Semântica com Mapas de Tópicos foi discutida em [66]. Ferramentas comerciais como "theBrain" [12] fornecem recursos muito semelhantes, como relações nomeadas, mas sem uma semântica formal subjacente.

Diferentes áreas de aplicação se beneficiam da Web Semântica e de uma (re) organização de seus conhecimentos em termos de ontologias. Entre eles estão Web Services [57], [56], [140], [136], [27] e Gerenciamento de conhecimento (consulte [157] para um conjunto de estruturas e ferramentas e [108] para um exemplo de aplicação). No E-Learning, os padrões de metadados têm uma longa tradição (em particular, Dublin Core [13] e LOM, os metadados dos objetos de aprendizagem [14]) Eles são empregados em portais educacionais [15]
e uma mudança geral em direção à notação XML e / ou RDF pode ser observada.

Muitos tipos de sites podem lucrar com uma (re) organização como Sites Semânticos. Portais de conhecimento [16] forneça visualizações sobre informações específicas de domínio na World Wide Web para ajudar seus usuários a encontrar informações relevantes. Sua manutenção pode ser bastante aprimorada usando uma arquitetura de backbone baseada em ontologia e um conjunto de ferramentas, conforme fornecido por SEAL [117] e SEAL-II [85].

Embora os metadados sejam úteis na Web, eles são essenciais para encontrar recursos em redes ponto a ponto. Exemplos incluem EDUTELLA [130] (que transfere o padrão educacional de LOM mencionado acima para uma arquitetura P2P) e POOL [77].

III W EB M INING

Mineração na Web é a aplicação de técnicas de mineração de dados ao conteúdo, estrutura e uso de recursos da Web. É, portanto, “o processo não trivial de identificar padrões válidos, anteriormente desconhecidos e potencialmente úteis” [55] na enorme quantidade desses dados da Web, padrões que os descrevem de forma concisa e em ordens de magnitude gerenciáveis. Como outros aplicativos de mineração de dados, a mineração da Web pode se beneficiar de determinada estrutura de dados (como nas tabelas do banco de dados), mas também pode ser aplicada a dados semiestruturados ou não estruturados, como texto de forma livre. Isso significa que a mineração da Web é uma ajuda inestimável na transformação de conteúdo compreensível por humanos em semântica compreensível por máquina.

Três áreas de mineração na Web são comumente distinguidas: mineração de conteúdo, mineração de estrutura e mineração de uso [177], [106], [155]. Nas três áreas, uma ampla gama de técnicas gerais de mineração de dados, em particular a descoberta de regras de associação, clustering, classificação e mineração de sequência, é empregada e desenvolvida ainda mais para refletir as estruturas específicas de recursos da Web e

[12] http://www.thebrain.com/
[13] http://dublincore.org
[14] Vejo http://ltsc.ieee.org/wg12
[15] por exemplo, http://www.eduserver.de
[16] Um exemplo é http://www.ontoweb.org.

as perguntas específicas colocadas na mineração na Web. Por razões de espaço, apresentaremos a mineração de conteúdo, estrutura e uso da Web apenas brevemente; para uma visão geral aprofundada dos métodos e / ou aplicações, consulte [74], [171], [72], [29], [9].

A. Conteúdo / texto das páginas da Web

Mineração de conteúdo da Web analisa o conteúdo dos recursos da Web. Hoje, é principalmente uma forma de mineração de texto (para visões gerais, consulte [28], [145]). Os recentes avanços na mineração de dados multimídia prometem ampliar o acesso também ao conteúdo de imagem, som, vídeo etc. dos recursos da Web. A mineração de dados multimídia pode produzir anotações semânticas comparáveis às obtidas na mineração de texto; portanto, não consideramos mais esse campo (ver [146], [175] e as referências citadas). Os principais recursos da Web extraídos na mineração de conteúdo da Web são páginas individuais.

A recuperação de informações é uma das áreas de pesquisa que fornece uma variedade de métodos populares e eficazes, principalmente estatísticos, para mineração de conteúdo da Web. Eles podem ser usados para agrupar, categorizar, analisar e recuperar documentos, cf. [149] para uma pesquisa de RI e [106] para uma pesquisa da relação entre RI e mineração de conteúdo da Web. Essas técnicas formam uma excelente base para abordagens mais sofisticadas. Um exemplo principal é a análise semântica latente (LSA) [48]. O LSA e outros métodos de análise fatorial provaram ser valiosos para analisar o conteúdo da Web e também o uso,

por exemplo, [26], [92]. No entanto, LSA refere-se a uma noção mais vaga de "semântica"; é necessário muito esforço para identificar uma conceituação explícita a partir das relações calculadas.

Além das técnicas padrão de mineração de texto, a mineração de conteúdo da Web pode tirar proveito da natureza semiestruturada do texto da página da Web. Tags HTML e marcação XML carregam informações que dizem respeito não apenas ao layout, mas também à estrutura lógica. Levando essa idéia adiante, uma "visão do banco de dados" da mineração de conteúdo da Web [106] tenta inferir a estrutura de um site para transferi-la para um banco de dados que permita melhor gerenciamento e consulta de informações do que uma pura "visão de IR".

A mineração de conteúdo da Web é especi fi camente adaptada às características do texto, como ocorre nos recursos da Web. Portanto, ele se concentra na descoberta de padrões em grandes coleções de documentos e na alteração frequente de coleções de documentos. Uma aplicação é a detecção e rastreamento de tópicos [5]. Isso pode servir para detectar eventos críticos (que são refletidos como um novo tópico no corpus em desenvolvimento de documentos) e tendências que indicam um aumento ou declínio no interesse em determinados tópicos.

Métodos adicionais de mineração de conteúdo que serão usados para o aprendizado de Ontologia, mapeamento e fusão de ontologias e aprendizado de instância são descritos na Seção IV-A. Na seção VI, iremos defini-los ainda mais em relação à Web Semântica.

B. Estrutura entre páginas da Web

Mineração da estrutura da Web geralmente opera na estrutura de hiperlink das páginas da Web (para uma pesquisa,

veja [29]). A mineração se concentra em conjuntos de páginas, variando de um único site à Web como um todo. A mineração da estrutura da Web explora as informações adicionais que estão (geralmente implicitamente) contidas na estrutura do *hiper* texto. Portanto, uma área de aplicação importante é a identificação da relevância relativa de diferentes páginas que parecem igualmente pertinentes quando analisadas em relação ao seu conteúdo isoladamente.

Por exemplo, a pesquisa de tópicos induzida por hiperlink [100] analisa a topologia de hiperlink descobrindo fontes de informações autorizadas para um amplo tópico de pesquisa. Esta informação é encontrada em *autoridade* páginas, que são definidas em relação a *hubs:* Hubs são páginas que apontam para muitas autoridades relacionadas. Da mesma forma, o mecanismo de pesquisa Google [17] deve seu sucesso ao algoritmo PageRank, que afirma que a relevância de uma página aumenta com o número de hiperlinks para ela de outras páginas e, em particular, de outras páginas relevantes [135].

A mineração da estrutura da Web e a mineração de conteúdo da Web são frequentemente realizadas em conjunto, permitindo explorar simultaneamente o conteúdo e a estrutura do hipertexto. De fato, alguns pesquisadores se enquadram na noção de mineração de conteúdo da Web [39].

C. Uso de páginas da Web

No *Mineração de uso da Web,* a mineração se concentra nos registros das solicitações feitas pelos visitantes de um site, geralmente coletadas em um log de servidor da Web [155], [156]. O conteúdo e a estrutura das páginas da Web, em particular as de um site, refletem as intenções dos autores e designers das páginas e da arquitetura de informações subjacente. O comportamento real dos usuários desses recursos pode revelar uma estrutura adicional.

Primeiro, os relacionamentos podem ser induzidos pelo uso, onde nenhuma estrutura específica foi projetada. Por exemplo, em um catálogo on-line de produtos, geralmente não existe uma estrutura inerente (produtos diferentes são simplesmente visualizados como um conjunto) ou uma ou várias estruturas hierárquicas fornecidas por categorias de produtos etc. No entanto, durante a mineração das visitas a esse site, pode-se achar que muitos dos usuários que estavam interessados no produto A também estavam interessados no produto B. O "interesse" pode ser medido por solicitações de páginas de descrição do produto ou pela colocação desse produto no carrinho de compras. Tais correspondências entre o interesse do usuário em vários itens podem ser usadas para *personalização,* por exemplo, recomendando o produto B quando o produto A tiver sido visualizado ("venda cruzada / upselling" no comércio eletrônico) ou tratando um visitante de acordo com o "segmento de cliente" que seu comportamento indica. Exemplos de algoritmos e aplicativos podem ser encontrados em [127], [112], [104] e nas recomendações feitas por livrarias online e outras lojas online.

Segundo, os relacionamentos podem ser induzidos pelo uso em que um relacionamento diferente foi planejado [38]. Por exemplo, a mineração de sequência pode mostrar que muitos dos usuários que saíram da página

[17] http://www.google.com

C para a página D fez isso ao longo de caminhos que indicam uma pesquisa prolongada (visitas frequentes para ajudar e indexar páginas, retorno frequente etc.). Essa relação entre topologia e uso pode indicar

usabilidade problemas: os visitantes desejam acessar D a partir de C, mas precisam procurar porque não há hiperlink direto [96] ou porque é difícil encontrá-lo [19]. Essas informações podem ser usadas para melhorar a arquitetura de informações do site e o design da página.

Terceiro, a mineração de uso pode revelar eventos no mundo mais rapidamente do que a mineração de conteúdo. A detecção e rastreamento de tópicos podem identificar eventos quando eles são refletidos em textos, isto é, no comportamento de escrita de autores da Web. No entanto, a busca de informações geralmente precede a criação e há mais usuários da Web do que autores da Web. Um exemplo é a detecção do início de epidemias (ou o medo de epidemias) no uso de sites de informações médicas [173], [78]. O monitoramento de padrões [10] permite ao analista ir além da análise de séries temporais simples e rastrear evoluções em padrões de acesso mais complexos, como regras ou sequências de associação.

D. Abordagens combinadas

É útil combinar a mineração de uso da Web com a análise de conteúdo e estrutura para "entender" os caminhos frequentes observados e as páginas desses caminhos. Isso pode ser feito usando uma variedade de métodos. As abordagens iniciais basearam-se em taxonomias pré-construídas [176] e / ou em métodos de extração de palavras-chave baseadas em IR [41]. Muitos métodos dependem do mapeamento de páginas em uma ontologia; isso será discutido nas seções VB e VI.

Na seção a seguir, veremos primeiro como as ontologias e suas instâncias podem ser aprendidas. Em seguida, investigaremos como o uso de ontologias e outras maneiras de identificar o significado das páginas podem ajudar a tornar a Web Mining semântica.

IV E XTRACTING S EMANTICA DO W EB

O esforço por trás da Web Semântica é adicionar anotação semântica compreensível por máquina aos documentos da Web para acessar o conhecimento em vez de material não estruturado. O objetivo é permitir que o conhecimento seja gerenciado de maneira automática. O Web Mining pode ajudar a aprender estruturas para a organização do conhecimento (por exemplo, ontologias) e fornecer a população dessas estruturas de conhecimento.

Todas as abordagens discutidas aqui são semi-automáticas. Eles ajudam o engenheiro de conhecimento a extrair a semântica, mas não podem substituí-la completamente. Para obter resultados de alta qualidade, não se pode substituir o humano no circuito, pois há sempre muito conhecimento tácito envolvido no processo de modelagem [24]. Um computador nunca poderá considerar completamente o conhecimento, a experiência ou as convenções sociais de segundo plano. Se fosse esse o caso, a Web Semântica seria supérflua, pois máquinas como mecanismos de busca ou agentes poderiam operar diretamente em sistemas convencionais.

Paginas web. O objetivo geral de nossa pesquisa não é, portanto, substituir o ser humano, mas fornecer-lhe cada vez mais apoio.

A. Semântica criada por Conteúdo e Estrutura

A.1 Aprendizado de ontologia

Extrair uma ontologia da Web é uma tarefa desafiadora. Uma maneira é projetar a ontologia manualmente, mas isso é caro. Em [116], a expressão *Aprendizagem Ontologia* foi cunhado para a extração semiautomática da semântica da Web. Lá, técnicas de aprendizado de máquina foram usadas para melhorar o processo de engenharia de ontologia e reduzir o esforço do engenheiro de conhecimento. Um exemplo é dado na Seção VI.

O aprendizado de ontologia explora muitos recursos existentes, incluindo textos, tesauros, dicionários e bancos de dados (veja [134] como um exemplo do uso do WordNet). Ele se baseia em técnicas de mineração de conteúdo da Web e combina técnicas de aprendizado de máquina com métodos de campos como recuperação de informações [113] e agentes [170], aplicando-os para descobrir a 'semântica' nos dados e torná-los explícitos. As técnicas produzem resultados intermediários que devem finalmente ser integrados em um formato compreensível por máquina, por exemplo, uma ontologia. A mineração pode complementar as taxonomias (Web) existentes com novas categorias (cf. [4] para uma extensão do Yahoo [18]) e pode ajudar a construir novas taxonomias [105].

Um número crescente de sites entrega páginas geradas dinamicamente em uma interação de um banco de dados subjacente, arquitetura de informações e recursos de consulta. Para muitos sites e perguntas de análise, uma ontologia pode ser compilada a partir de fontes internas, como esquemas de banco de dados, opções de consulta e modelos de transação. Essa "engenharia reversa" geralmente envolve uma grande quantidade de trabalho manual, mas pode ser auxiliada por esquemas de aprendizado de ontologia (semi-) automáticos. Por exemplo, muitos sites de varejo e informações possuem catálogos de produtos igualmente estruturados [19], [152]. Assim, um site de turismo pode conter os URLs pesquisar hotel.html, pesquisar iate clube.html, ... que permite deduzir as categorias de produtos hotel, iate clube, etc. [19].

A.2 Mapeando e mesclando ontologias

O crescente uso de ontologias leva a sobreposições entre o conhecimento em um domínio comum. Ontologias específicas do domínio são modeladas por vários autores em várias configurações. Essas ontologias são a base para a construção de novas ontologias específicas de domínio em domínios semelhantes

[18] http://www.yahoo.com

[19] Isso faz parte de um exemplo corrente, a ser usado em todo o artigo, descrevendo um site de turismo de ficção. É baseado no projeto Getess (http://www.getess.de/index en.html), que fornece acesso baseado em ontologia a páginas de turismo na Web para a região alemã Mecklenburg-Vorpommern (http://www.all-in-all.de).

montar e estender várias ontologias a partir de repositórios.

O processo de *fusão de ontologia* recebe como entrada duas (ou mais) ontologias de origem e retorna uma ontologia mesclada. A fusão manual de ontologias usando ferramentas de edição convencionais sem suporte é difícil, trabalhosa e propensa a erros. Portanto, vários sistemas e estruturas para apoiar o engenheiro de conhecimento na tarefa de mesclagem de ontologias foram recentemente propostos [88], [32], [131], [121]. Essas abordagens se baseiam em heurísticas de correspondência sintática e semântica que são derivadas do comportamento dos engenheiros de ontologia confrontados com a tarefa de mesclar ontologias. Outro método é o FCA-M ERGE, que opera de baixo para cima e oferece uma descrição estrutural global do processo [163]. Extrai instâncias de conceitos de ontologia de origem de um determinado conjunto de documentos de texto específicos de domínio, aplicando técnicas de processamento de linguagem natural. Com base nas instâncias extraídas, ele usa o T ITANIC algoritmo [164] para calcular uma estrutura de conceito. A estrutura conceitual fornece um agrupamento conceitual dos conceitos das ontologias de origem. É explorado e transformado interativamente na ontologia mesclada pelo engenheiro de ontologia.

Mapeamento de ontologia é a atribuição dos conceitos de uma ontologia e suas instâncias aos conceitos de outra ontologia. Isso pode ser útil, por exemplo, quando uma das várias ontologias for escolhida como a correta para a tarefa em questão. As instâncias podem ser simplesmente classificadas do zero na ontologia de destino; alternativamente, o conhecimento inerente à ontologia de origem pode ser utilizado com base na heurística de que instâncias de um conceito de fonte provavelmente também serão classificadas em um conceito da ontologia de destino [178].

Uma alternativa para mesclar / mapear ontologias é simplesmente coletá-las em paralelo e selecionar a correta de acordo com a tarefa em questão. Essa visão de um 'corpus de representações' é apresentada em [71], que abre um novo domínio de questões de pesquisa interessantes.

A.3 Aprendizado da instância

Mesmo se ontologias estiverem presentes e os usuários anotarem manualmente novos documentos, ainda haverá documentos antigos contendo material não estruturado. Em geral, a marcação manual de todos os documentos produzidos é impossível. Além disso, alguns usuários podem precisar extrair e usar informações diferentes ou adicionais daquelas fornecidas pelo criador. Para construir a Web Semântica, é essencial produzir métodos automáticos ou semi-automáticos para extrair informações de documentos relacionados à Web como instâncias de conceitos de uma ontologia, tanto para ajudar os autores a anotar novos documentos quanto para extrair informações adicionais de estruturas não estruturadas ou existentes. documentos parcialmente estruturados.

Vários estudos investigam o uso da mineração de conteúdo para enriquecer as conceituações existentes atrás de um site. Por exemplo, em [126], Mladenic usou técnicas de categorização de texto para atribuir páginas HTML a categorias na hierarquia do Yahoo. Isso pode reduzir o esforço manual para

mantendo o índice da Web do Yahoo.

Extração de informações de textos (IE) é uma das áreas mais promissoras das Tecnologias da Linguagem Natural (ver, por exemplo, [43]). O IE é um conjunto de métodos automáticos para localizar fatos importantes em documentos eletrônicos para uso posterior. As técnicas do IE variam desde a extração de palavras-chave do texto das páginas usando o *tf.idf* método conhecido no Information Retrieval, por meio de técnicas que levam em consideração as estruturas sintáticas do HTML ou da linguagem natural, a técnicas que são extraídas com referência a uma estrutura de destino explicitamente modelada, como uma ontologia (para uma pesquisa, consulte [107]).

A Extração de informações é o suporte perfeito para identificação e extração de conhecimento de documentos da Web, pois pode - por exemplo - fornecer suporte na análise de documentos de maneira automática (extração não supervisionada de informações) ou de maneira semi-automática (por exemplo, como suporte para anotadores humanos na localização de fatos relevantes em documentos, por meio de informações destacadas). Um desses sistemas para o IE é o FASTUS [81]. Outro sistema é o GATE. [20] Com o surgimento da Web Semântica, ela foi estendida ao suporte ontológico e, em particular, por exemplo, ao aprendizado [21]. O OntoMat Annotizer [75] foi desenvolvido diretamente para a Web Semântica. Complementa o IE com funcionalidade de criação. A abordagem de Craven et al. [44] é discutido na seção VI. Em [79], [80], técnicas de aprendizado de máquina foram usadas para a anotação semi-automática de serviços da Web.

A.4 Usando conceituações existentes como ontologias e para anotação automática

Para muitos sites, já existe um modelo de domínio explícito para a geração de páginas da Web. Essas formalizações existentes podem ser (re) usadas para marcação e mineração semântica.

Por exemplo, muitos sistemas de gerenciamento de conteúdo geram páginas da Web a partir de um catálogo de produtos, em URLs que refletem o caminho para o produto na hierarquia do catálogo. No exemplo em execução, isso pode levar a URLs como Hotéis / Bem-estarHotéis / BeachHotel.html (URLs semelhantes podem ser encontrados em índices populares da Web). A classificação por hierarquia de produtos é uma técnica comumente usada para mineração de uso da Web, veja, por exemplo, [154], [7], [59] e o conjunto de dados KDDCup 2000 disponível para testar algoritmos. [21] Alternativamente, as páginas podem ser geradas a partir de uma ontologia completa e seu mecanismo de inferência [133], [125]. A adaptação dessa ideia básica a URLs dinâmicos é descrita na Seção VB.1.

Para obter um esquema comum de ontologia e marcação, as páginas podem ser geradas centralmente por um servidor de aplicativos. No caso de autoria distribuída, o uso da ontologia comum pode ser garantido por ferramentas interativas que ajudam autores individuais a marcar suas páginas. Isto provou

[20] http://gate.ac.uk/

[21] http://www.ecn.purdue.edu/KDDCUP

ser uma estratégia bem-sucedida para o desenvolvimento de portais baseados na comunidade. [22]

Outra maneira de usar as informações existentes é descrita em [76]: “Anotação profunda” deriva mapeamentos entre estruturas de informações dos bancos de dados. Esses mapeamentos são usados para consultar informações semânticas armazenadas no banco de dados subjacente ao site. Isso combina recursos de anotação convencional de páginas da Web e geração automática de páginas da Web a partir de bancos de dados.

A.5 Semântica criada por estrutura

Como discutimos na Seção III-B, os resultados da análise da vinculação de páginas da Web pela mineração de uso da Web criam um certo tipo de conhecimento, um ranking de relevância. Outro tipo de conhecimento que pode ser inferido da estrutura é uma semelhança entre as páginas, útil para o popular aplicativo de navegador "Encontrar páginas semelhantes" (para um que foi recuperado pela navegação ou pesquisa): Com base na observação de que as páginas frequentemente citadas juntos de outras páginas provavelmente estão relacionados, Dean e Henzinger [47] propõem dois algoritmos para encontrar páginas semelhantes com base na estrutura do hiperlink. Essas técnicas estruturam o conjunto de páginas, mas não as classificam em uma ontologia.

Por outro lado, a estrutura de hiperlink nas páginas se presta mais diretamente à classificação. Cooley, Mobasher e Srivastava [40], com base em [141], propõem uma ontologia de funções de página, onde a classificação de uma única página com relação a essa ontologia pode ser feita (semi) automaticamente. Por exemplo, as páginas de "navegação" projetadas para orientação contêm muitos links e pouco texto informativo, enquanto as páginas de "conteúdo" contêm um pequeno número de links e são projetadas para serem visitadas pelo seu conteúdo. Isso pode ser usado para comparar o uso pretendido com o uso real [38]. Por exemplo, uma página de conteúdo usada como um ponto de entrada frequente para um site sinaliza um desafio para o design do site: Primeiro, o *pretendido* o ponto de entrada, que provavelmente é a página inicial, deve ser mais conhecido e mais fácil de localizar. Segundo, links adicionais para navegação podem ser fornecidos na página atualmente *real* ponto de entrada. Seu conteúdo pode se tornar um candidato a uma nova categoria de conteúdo de nível superior em várias páginas principais.

A estrutura da marcação dentro da página também pode ajudar na extração do conteúdo da página: concentrar-se nos segmentos da página identificados por referência ao DOM da página (modelo de objeto de documento ou árvore de tags) pode servir para identificar o conteúdo principal de uma página [29, pp. 228ss.] E separá-lo do "ruído", como barras de navegação, anúncios etc. [172].

B. Semântica criada por Usage

A discussão anterior assumiu implicitamente que o conteúdo existe independentemente de seu uso. No entanto, uma grande proporção de conhecimento é construída socialmente. Assim, a navegação não é apenas

[22] Vejo http://www.ontoweb.org e http://www.eduserver.de

impulsionado por relacionamentos formalizados ou pela lógica subjacente dos recursos disponíveis da Web. Pelo contrário, “é uma estratégia de navegação de informações que tira proveito do comportamento de pessoas com idéias semelhantes” ([33, p.18]). Sistemas de recomendação baseados em “filtro colaborativo” têm sido a aplicação mais popular dessa idéia. Nos últimos anos, a idéia foi estendida para considerar não apenas classificações, mas também o uso da Web como base para a identificação de ideias afins (“As pessoas que gostaram / compraram este livro também olharam para ...”; cf. Seção III -C e [94] para uma aplicação clássica).

Extrair essas relações do uso pode ser interpretado como um tipo de aprendizado de ontologia, no qual a relação binária “está relacionada” nas páginas (e, portanto, nos conceitos) é aprendida. Os padrões de uso podem revelar outras relações para ajudar a construir a Web Semântica? Esse campo ainda é bastante novo, portanto apenas descreveremos uma seleção ilustrativa de abordagens de pesquisa.

Ypma e Heskes [174] propõem um método para aprender categorias de conteúdo a partir do uso. Eles modelam a navegação em termos de modelos ocultos de Markov, com os estados ocultos sendo categorias de página e os eventos de solicitação observados sendo instâncias deles. Seu principal objetivo é mostrar que uma categorização significativa da página pode ser aprendida simultaneamente com a rotulagem do usuário e as transições entre categorias; rótulos semânticos (como "páginas de esportes") devem ser atribuídos a um estado manualmente. A taxonomia resultante e a classificação da página podem ser usadas como um modelo conceitual para o site ou para melhorar um modelo conceitual existente.

Chi et al. [35], [34] identificam caminhos frequentes através de um site. Com base nas palavras-chave extraídas das páginas ao longo do caminho, elas calculam o provável "odor de informação" seguido, ou seja, o objetivo pretendido do caminho. O perfume das informações é um conjunto de palavras-chave ponderadas, que podem ser inspecionadas e rotuladas de forma mais concisa usando uma ferramenta interativa. Assim, o uso cria um conjunto de objetivos de informação que os usuários esperam que o site atenda.[23] Esses objetivos podem ser usados para modificar ou estender as categorias de conteúdo mostradas aos usuários, empregadas para estruturar a arquitetura de informações do site ou empregadas no modelo conceitual do site.

Stojanovic, Maedche, Motik e Stojanovic [158] propõem medir o interesse do usuário nos conceitos de um site pela frequência de acessos a páginas que lidam com esses conceitos. Eles usam esses dados para *evolução da ontologia:* estendendo a cobertura do site a conceitos de alto interesse e excluindo conceitos de baixo interesse ou mesclando-os a outros.

A combinação de entrada implícita do usuário (uso) e entrada explícita do usuário (consultas do mecanismo de pesquisa) pode contribuir ainda mais para a estrutura conceitual. A navegação do usuário foi empregada para inferir uma relação tópica, ou seja, a relação de um conjunto de páginas com um tópico, conforme fornecido pelos termos de uma consulta a um mecanismo de pesquisa (“rastreamento colaborativo” [2]). Uma classificação de páginas como “satisfazendo o predicado definido pelo usuário” e “não satisfazendo o predicado” é assim aprendida com o uso, estrutura e

[23] Uma validação empírica mostrou que esse tipo de análise de conteúdo agrupa de fato caminhos que têm o mesmo objetivo de informação [36].

informações de conteúdo. Uma aplicação óbvia é explorar a navegação do usuário para melhorar o ranking dos mecanismos de busca [93], [97].

Muitas abordagens usam uma combinação de conteúdo e mineração de uso para gerar recomendações. Por exemplo, na filtragem colaborativa baseada em conteúdo, a categorização textual de documentos é usada para gerar pseudo-classificações para cada par de documentos usuário [122]. Em [137], ontologias, as técnicas do IE para analisar páginas únicas e o histórico de pesquisa de um usuário juntos servem para gerar recomendações para o aprimoramento de consultas em um mecanismo de pesquisa.

V. U CANTAR S EMANTICA PARA W EB M INING AND M INING THE S EMANTIC W EB

A semântica pode ser utilizada para mineração da Web para diferentes fins. Algumas das abordagens apresentadas nesta seção baseiam-se comparativamente *Ad hoc* formalização da semântica, enquanto outros já podem explorar todo o poder da Web Semântica. A Web Semântica oferece uma boa base para enriquecer a Mineração da Web: Os tipos de (hiper) links são agora descritos explicitamente, permitindo que o engenheiro de conhecimento obtenha insights mais profundos sobre a mineração da estrutura da Web; e o conteúdo das páginas é fornecido com uma semântica formal, permitindo que ela aplique técnicas de mineração que exijam entradas mais estruturadas. Como a distinção entre o uso da semântica para mineração na Web e a mineração da própria Web Semântica é quase nítida, discutiremos os dois de maneira integrada.

A primeira grande área de aplicação é a mineração de conteúdo, ou seja, a codificação explícita da semântica para minerar o conteúdo da Web. Os hiperlinks e âncoras em uma página fazem parte do texto dessa página e, em uma página marcada semântica, são elementos da mesma maneira que o texto. Portanto, o conteúdo e a estrutura estão fortemente interligados (os dois campos são às vezes tratados como um [39]). Na Web Semântica, a distinção entre mineração de conteúdo e estrutura desaparece completamente, pois o conteúdo da página é explicitamente transformado na estrutura da anotação. No entanto, deve-se notar que a distribuição das anotações semânticas em uma página e entre páginas pode fornecer conhecimento implícito adicional.

A. Mineração de Conteúdo e Estrutura

Em [84], ontologias são usadas como conhecimento prévio durante o pré-processamento, com o objetivo de melhorar os resultados do cluster. Nós pré-processamos os dados de entrada (por exemplo, texto) e aplicamos heurísticas baseadas em ontologia para seleção e agregação de recursos. Com base nessas representações, calculamos vários resultados de cluster usando k-Means. Usando a ontologia, podemos selecionar o resultado mais adequado à nossa tarefa em questão. Em [87], demonstramos a melhoria no agrupamento decorrente do uso do WordNet para pré-processar o corpus da Reuters. Um estudo análogo mostrou melhorias na classificação [20].

Outro projeto atual visa facilitar o acesso personalizado ao material de cursos que

é armazenado em uma rede ponto a ponto [24] por meio de agrupamento conceitual. Empregamos técnicas da Análise Formal de Conceito, que foram aplicadas com sucesso no Conceptual Email Manager CEM [37]. O CEM fornece uma hierarquia de conceitos (clusters) de pesquisa baseada em ontologia com vários caminhos de pesquisa. Uma combinação dessa abordagem com o agrupamento de textos e um método de visualização para analisar os resultados são apresentados em [86].

Abordagens ricas em conhecimento na sumarização automática de texto (cf. [118], [119], [89]) visam maximizar as informações em uma quantidade mínima de texto resultante. Eles estão intimamente relacionados à mineração de conteúdo da Web usando semântica, porque na mineração de conteúdo da Web e na sumarização de texto, o texto em linguagem natural precisa ser mapeado em uma representação abstrata. Esse resumo geralmente é representado em alguma lógica e é usado para melhorar os resultados da sumarização de texto. Esperamos que as técnicas de resumo automático de texto tenham um papel importante na Semantic Web Mining.

A mineração da estrutura da Web também pode ser aprimorada levando em consideração o conteúdo. O algoritmo PageRank mencionado na Seção III coopera com um algoritmo de análise de palavras-chave, mas os dois são independentes um do outro. Portanto, o PageRank considerará qualquer página muito citada como 'relevante', independentemente de o conteúdo dessa página refletir a consulta. Ao também levar em consideração o texto âncora do hiperlink e seus arredores, o CLEVER [30] pode avaliar mais especificamente a relevância de uma determinada consulta. O rastreador focado [31] aprimora isso integrando conteúdo tópico ao modelo de gráfico de links e por uma maneira mais flexível de rastrear. O aprendizado Intelligent Crawler [3] estende o Focused Crawler, permitindo predicados que combinam diferentes tipos de consultas tópicas, de palavras-chave ou outras restrições no conteúdo ou nas metainformações da página (por exemplo, domínio da URL).

Um importante grupo de técnicas que podem ser facilmente adaptadas à mineração de conteúdo / estrutura da SemanticWeb são as abordagens discutidas como (*Mineração relacional de dados multi-) (* anteriormente chamado *Programação Lógica Indutiva / ILP) [* 51] A Mineração de Dados Relacional procura padrões que envolvem várias relações em um banco de dados relacional. Ele compreende técnicas para classificação, regressão, agrupamento e análise de associação. Os algoritmos podem ser transformados para lidar com os dados descritos em RDF ou por ontologias. Um ponto de partida para essas transformações é descrito em [67], que analisa diferentes lógicas e desenvolve um novo formato de representação de conhecimento intimamente relacionado à lógica de Horn, uma das lógicas comuns no ILP. Tornar o DataMining relacional acessível à Semantic Web Mining enfrenta dois grandes desafios. O primeiro é o tamanho dos conjuntos de dados a serem processados e o segundo é a distribuição dos dados pela Web Semântica. A escalabilidade para grandes conjuntos de dados sempre foi uma grande preocupação para os algoritmos ILP. Com o crescimento esperado da Web Semântica, esse problema também aumenta. Portanto, o desempenho da mineração

[24] http://edutella.jxta.org

algoritmos precisam ser aprimorados por métodos como amostragem (por exemplo, [144]). Para processar dados distribuídos, é necessário desenvolver algoritmos que executem a mineração de maneira distribuída, de modo que, em vez de conjuntos de dados inteiros, apenas os resultados (intermediários) tenham que ser transmitidos.

B. Mineração de uso

A mineração de uso da Web se beneficia de incluir a semântica no processo de mineração pela simples razão de que o especialista em aplicativos como usuário final dos resultados da mineração está interessado em *eventos no domínio do aplicativo,* em particular o comportamento do usuário, enquanto os dados disponíveis - logs do servidor da Web - são seqüências tecnicamente orientadas de *Solicitações HTTP.* [25] Um objetivo central é, portanto, mapear solicitações HTTP para unidades significativas de eventos de aplicativos.

Nesta seção, apresentaremos primeiro uma estrutura para a modelagem do comportamento do usuário e, em seguida, discutiremos como esse conhecimento básico é usado na mineração. Para ilustrar a estrutura, a usaremos para descrever vários estudos existentes sobre o uso da Web. Vamos nos concentrar nos aspectos semânticos da estrutura. Os estudos que descrevemos usam várias convenções sintáticas diferentes para representar a semântica; esperamos que, no futuro, as notações baseadas em XML (e, portanto, sintaticamente padronizadas) permitirão uma melhor troca e reutilização desses modelos [143], [101].

B.1 Eventos de aplicativos

Eventos de aplicativo são definidos com relação ao domínio do aplicativo e ao site, uma tarefa não trivial que equivale a uma formalização detalhada do modelo de negócios / aplicativo do site (para detalhes, consulte [17]). Por exemplo, eventos relevantes de comércio eletrônico incluem visualizações de produtos e cliques em que um usuário mostra interesse específico em um produto específico, solicitando informações mais detalhadas (por exemplo, do Beach Hotel a uma lista de seus preços nas várias estações do ano). Eventos relacionados incluem cliques em uma categoria de produto (por exemplo, do Beach Hotel, até a categoria Todos os hotéis de bem-estar), cliques em um banner, alterações no carrinho de compras e compras ou lances de produtos.

Esses eventos são exemplos do que chamamos *eventos de aplicação atômica;* eles geralmente correspondem à solicitação de um usuário para uma página (exibição). Eles podem ser caracterizados por seu conteúdo (por exemplo, o Beach Hotel ou, em geral, Todos os hotéis de bem-estar ou Todos os hotéis, consulte a Fig. 3) e o serviço solicitado quando esta página é chamada (por exemplo, a função "pesquisar hotéis por localização") [19] Uma página pode ser mapeada para um ou para um conjunto de eventos do aplicativo. Por exemplo, ele pode ser mapeado para

[25] Observe que esta discussão assume que alguns outros problemas que afetam a qualidade dos dados, por exemplo, a atribuição de solicitações a usuários e / ou sessões, foram resolvidos ou não afetam as inferências baseadas na semântica das páginas da Web solicitadas. Essa é uma idealização, veja [18] para uma investigação do efeito das heurísticas da sessão nos resultados da mineração. O uso de logs do servidor de aplicativos pode ajudar a contornar alguns desses problemas [102]. Na discussão a seguir, também assumimos que outras etapas padrão de pré-processamento foram realizadas [40].

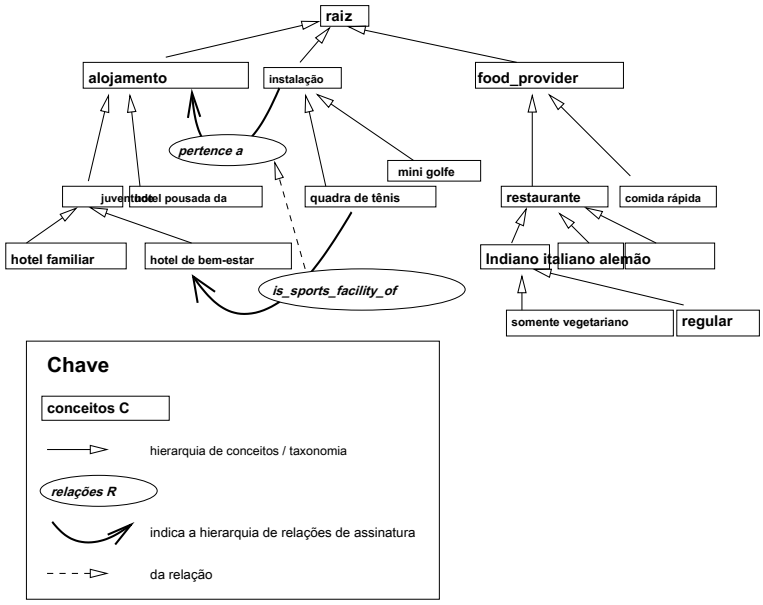

Fig. 3. Partes da ontologia do conteúdo de um site turístico de turismo.

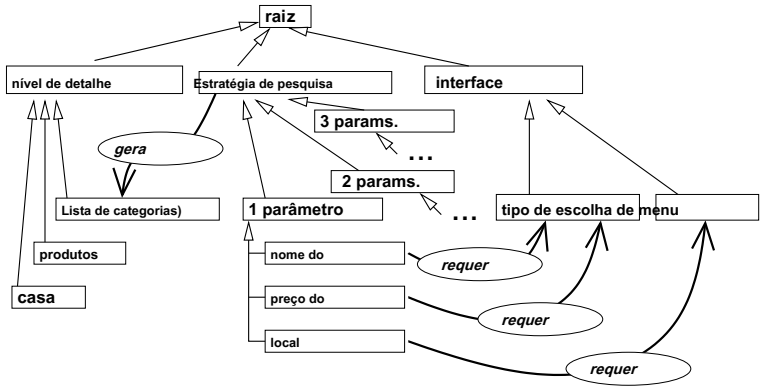

Fig. 4. Partes da ontologia dos serviços do site de exemplo fictício.

o conjunto de todos os conceitos e relações que aparecem em sua sequência de consultas [133]. Como alternativa, palavras-chave do texto da página e das páginas vinculadas a ela podem ser mapeadas para uma ontologia de domínio, com uma ontologia de uso geral como o WordNet servindo como intermediário entre as palavras-chave encontradas no texto e os conceitos da ontologia [52] .

A Figura 4 mostra uma ontologia de serviço para o site de exemplo fictício, modelado após o usado em um exemplo do mundo real em [19]. O site mostra informações relacionadas a acomodações em diferentes níveis de detalhes: como uma página inicial (ou inicial), nas páginas de categorias de produtos (listas de hotéis ou instalações) e nas páginas de produtos pessoais. As estratégias de busca consistem na especificação de um ou mais dos parâmetros localização, preço e nome. Os parâmetros e seus valores são especificados por opção em um menu ou digitando. Em resposta, o servidor gera uma página de categoria com todos os hotéis ou instalações que atendem às especificações especificadas.

Eventos de aplicativos atômicos geralmente fazem parte de unidades significativas de atividades maiores no site, que chamamos de *eventos de aplicativos complexos*. Os exemplos incluem (a) “eventos de compra direcionada” nos quais um usuário entra em uma loja de comércio eletrônico, procura um produto, visualiza esse produto e o coloca em

o carrinho de compras e o compra e, em seguida, sai ou (b) “eventos de construção de conhecimento”, nos quais um usuário navega e pesquisa repetidamente por categorias e visualizações de produtos e geralmente sai sem comprar (mas pode usar o conhecimento criado para retornar e adquirir algo mais tarde) [153]. Eventos de aplicativos complexos são geralmente descritos por expressões regulares cujo alfabeto consiste em eventos de aplicativos atômicos [153], ou por uma estrutura de ordem em eventos de aplicativos atômicos [14].

B.2 Como o conhecimento sobre eventos de aplicativos é usado na mineração?

Depois que as solicitações são mapeadas para os conceitos, os dados transformados estão prontos para mineração. Investigaremos o tratamento de eventos atômicos e de aplicativos complexos, por sua vez.

Em muitas aplicações (veja os exemplos nas Figuras 3 e 4), os conceitos participam de múltiplas taxonomias. A mineração usando múltiplas taxonomias está relacionada às técnicas de cubo de dados OLAP: os objetos (neste caso, solicitações ou URLs solicitadas) são descritos em várias dimensões e hierarquias de conceito ou treliças são formuladas ao longo de cada dimensão para permitir visualizações mais abstratas (cf. [176], [99], [90], [160]).

A abstração taxonômica geralmente é essencial para gerar resultados significativos: primeiro, em um site com páginas geradas dinamicamente, cada página individual é solicitada tão raramente que nenhuma regularidade pode ser encontrada na análise do comportamento da navegação. Em vez disso, as regularidades podem existir em um nível mais abstrato, levando a regras como “as pessoas que ficam nos hotéis Wellness também tendem a comer em restaurantes”. Segundo, padrões extraídos de dados anteriores não são úteis para aplicações como sistemas de recomendação quando novos itens são introduzidos no catálogo de produtos e / ou na estrutura do site: O novo Pier Hotel não pode ser recomendado simplesmente porque não estava no site de turismo até ontem e, portanto, poderia não co-ocorra com nenhum outro item, seja recomendado por outro usuário etc.

Após as etapas de pré-processamento nas quais os dados de acesso foram mapeados em taxonomias, as técnicas de mineração subsequentes podem usar essas taxonomias de maneira estática ou dinâmica. Em abordagens estáticas, a mineração opera com conceitos em um nível escolhido de abstração; cada solicitação é mapeada para exatamente um conceito ou exatamente um conjunto de conceitos (veja os exemplos acima). Essa abordagem geralmente é combinada com o controle interativo do software, para que o analista possa reajustar o nível de abstração escolhido após visualizar os resultados (por exemplo, no minerador WUM; consulte [19] para um estudo de caso). Quando os eventos de aplicativos complexos investigados têm uma estrutura seqüencial, a mineração de sequência é necessária. Geralmente é o caso em investigações de estratégias de busca, compras etc., como mostram os exemplos acima.

Nas abordagens dinâmicas, os algoritmos identificam o nível mais específico de relacionamentos, escolhendo conceitos dinamicamente. Isso pode levar a regras como "As pessoas que ficam em hotéis de bem-estar tendem a

comer em restaurantes indianos apenas vegetarianos ”- vinculando o comportamento de escolha de hotel em um nível comparativamente alto de abstração com o comportamento de escolha de restaurante em um nível de descrição comparativamente detalhado.

Por exemplo, Srikant e Agrawal [154] procuram associações em determinadas taxonomias, usando limites de suporte e confiança para orientar a escolha do nível de abstração. A hierarquia de subsunção de uma ontologia existente também é usada para a descrição simultânea dos interesses do usuário em diferentes níveis de abstração, e essa descrição é usada para guiar a regra de associação e os algoritmos de agrupamento em métodos que vinculam páginas da Web a uma ontologia subjacente de uma forma mais fi nalizada. e maneira flexível [133], [52], [125]. Quando falta uma taxonomia explícita, a mineração pode fornecer agregações para conceitos mais gerais [45].

A Mineração de Uso da Web Semântica para eventos complexos de aplicativos envolve duas etapas de mapeamento de solicitações para eventos. Como discutido na Seção VB.1 acima, eventos complexos de aplicativos são geralmente definidos por expressões regulares em eventos atômicos de aplicativos (em um determinado nível de abstração em suas respectivas hierarquias). Portanto, em uma primeira etapa, os URLs são mapeados para eventos de aplicativos atômicos no nível de abstração exigido. Em uma segunda etapa, um minerador de sequência pode ser usado para descobrir padrões sequenciais nos dados transformados. As formas dos padrões seqüenciais procurados e a ferramenta de mineração usada determinam quanto conhecimento prévio pode ser usado para restringir os padrões identificados. Eles variam de cadeias de Markov de primeira ordem praticamente sem restrições ou de ordem k-ésima ordem [22],

Exemplos do uso de expressões regulares que descrevem cursos de eventos relevantes para aplicativos incluem estratégias de pesquisa [19], uma segmentação de visitantes em clientes e não clientes [152] e uma segmentação de visitantes em diferentes grupos de interesse com base no ciclo de compra do cliente modelo de marketing [153].

Até o momento, existem poucos modelos comuns de comportamento da Web Semântica. A natureza ainda amplamente exploratória do campo implica que ferramentas de mineração e preparação de dados altamente interativas são de suma importância: elas fornecem o melhor suporte para especialistas em domínio que trabalham com analistas para contribuir com seus conhecimentos básicos em um ciclo de mineração iterativo. Um elemento central das ferramentas interativas para exploração é a visualização. Na ferramenta STRATDYN [14], [13], propomos uma visualização semântica do uso da Web que permite ao analista detectar padrões visuais que podem ser interpretados em termos de comportamento no domínio do aplicativo.

Com a crescente padronização de muitos aplicativos da Web e a crescente influência da pesquisa de mineração com a pesquisa no domínio de aplicativos (por exemplo, marketing), é provável que o número de cursos padrão de eventos aumente. Exemplos são os esquemas preditivos de sites de comércio eletrônico (veja o exemplo de [124] mencionado na seção VB.1 acima) e a descrição da navegação

estratégias dadas por [128].

O poder representacional dos modelos que capturam o comportamento do usuário apenas em termos de uma sequência de estados identificados por solicitações de páginas é limitado. No futuro, esperamos mais explorações do significado do tempo de visualização (por exemplo, [60], [11]) e das transições entre estados [14].

Na análise e avaliação do comportamento do usuário, deve-se ter em mente que diferentes partes interessadas têm perspectivas diferentes sobre o uso de um site, o que os leva a investigar processos diferentes (eventos complexos de aplicativos) e também os leva a considerar as ações corretas de diferentes usuários. 'ou' valioso '. Recentemente, foram propostas estruturas para capturar diferentes processos [110], [168], [6] e perspectivas [123].

Em resumo, um desafio central para futuras pesquisas em Semantic Web Usage Mining reside no desenvolvimento, fornecimento e teste de ontologias de eventos de aplicativos.

VI C PERDENDO O eu OOP

Nas duas seções anteriores, analisamos como estabelecer dados da Web Semântica por mineração de dados, como explorar a semântica formal para Mineração da Web e como explorar a Web Semântica. Nesta seção, esboçamos uma das muitas combinações possíveis dessas abordagens. O exemplo mostra como diferentes combinações de Semantic Web e Web Mining podem ser organizadas em um loop de feedback.

Nosso objetivo é obter um conjunto de páginas da Web de um site e aprimorá-las para usuários humanos e de máquinas: (a) gerar metadados que refletem um modelo semântico subjacente ao site, (b) identificar padrões nas páginas ' texto e em seu uso e, com base nessas informações, para aprimorar a arquitetura da informação e o design da página. Para atingir esses objetivos, prosseguiremos com várias etapas nas quais

- empregar métodos de mineração em recursos da Web para gerar estrutura semântica (etapas 1 e 2: aprendendo e preenchendo a ontologia),
- empregar métodos de mineração nos recursos da Web semanticamente estruturados resultantes para gerar mais estrutura (etapas 3 e 4),
- no final de cada etapa, alimente esses resultados com o conteúdo e o design das próprias páginas da Web (visíveis para usuários humanos) e / ou de seus metadados e da ontologia subjacente (visível para os usuários da máquina).

Faremos apenas um esboço para ilustrar nossas idéias, usando o exemplo atual do site de turismo artístico usado ao longo deste artigo.

Pode-se dividir o primeiro passo, *aprendizagem de ontologia,* em duas sub-etapas. Primeiro, uma hierarquia de conceitos é estabelecida usando a metodologia OTK para modelagem de ontologias [167]. Pode ser suportado pelo método formal de modelagem de ontologia O NT E X (Exploração de Ontologia, [62]) que se baseia no

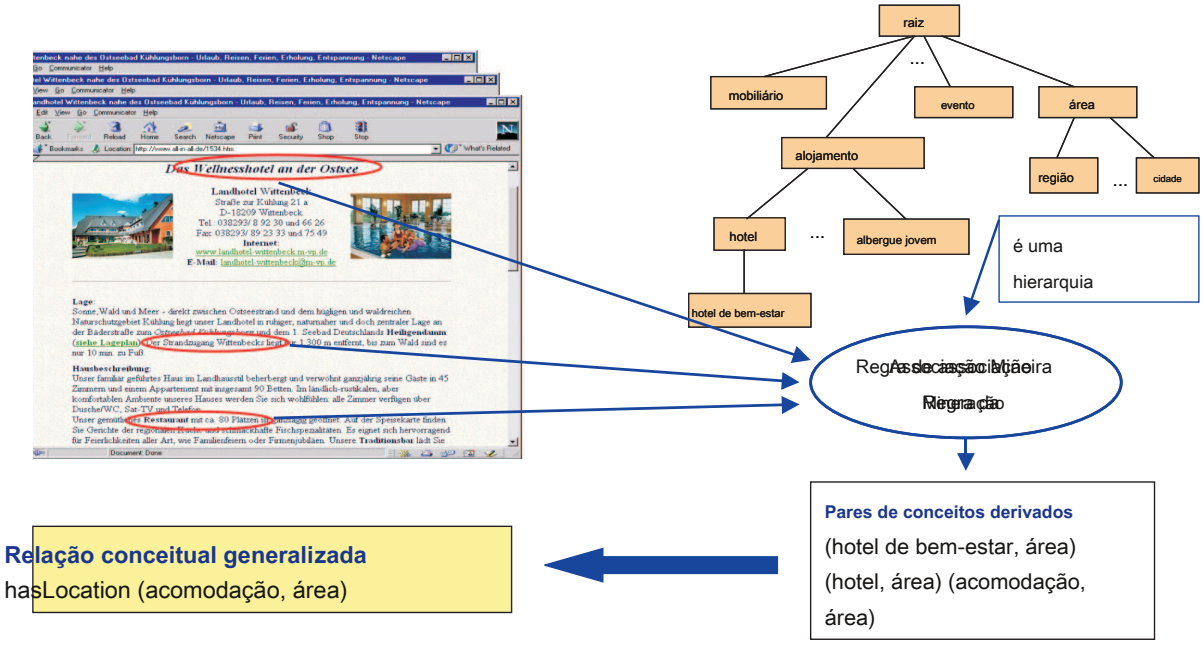

Fig. 5. Etapa 1: Minerando a Web para aprender ontologias.

técnica de aquisição de conhecimento do Attribute Exploration [61], conforme desenvolvido no arcabouço matemático da Análise Formal de Conceito [63]; e garante que o engenheiro de conhecimento considere todas as combinações relevantes de conceitos enquanto estabelece a hierarquia de subsunção. O NT E X toma como entrada um conjunto de conceitos e fornece como saída uma hierarquia sobre eles. Essa saída é então a entrada para o segundo subpasso, juntamente com um conjunto de páginas da Web. Maedche e Staab [115] descrevem como as regras de associação são extraídas dessa entrada, o que leva à geração de relações entre os conceitos de ontologia (ver Fig. 5). As regras de associação são usadas para descobrir combinações de conceitos que freqüentemente ocorrem juntos. Essas combinações sugerem a existência de relações conceituais. Eles são sugeridos ao analista. Como o sistema não pode gerar nomes automaticamente para as relações, o analista é solicitado a fornecê-las.

No exemplo mostrado na figura, a análise automática mostrou que três conceitos frequentemente co-ocorrem com o conceito “área”. Como a ontologia carrega a informação de que o conceito “hotel de bem-estar” é um subconceito do conceito de “hotel”, que por sua vez é um subconceito de “acomodação”, o mecanismo de inferência pode derivar que apenas uma relação conceitual precisa ser inferida com base em essas co-ocorrências: aquela entre “acomodação” e “área”. A entrada humana é então necessária para especificar um nome significativo como "hasLocation" para a relação conceitual generalizada.

Na segunda etapa, *a ontologia está cheia.* Nesta etapa, as instâncias são extraídas das páginas da Web e as relações da ontologia são estabelecidas entre elas usando as técnicas descritas em [44] (veja a Fig. 6), ou qualquer outra técnica descrita na Seção IV-A.3. Além da ontologia, a abordagem precisa de dados de treinamento marcados como entrada. Dada essa entrada, o sistema aprende a

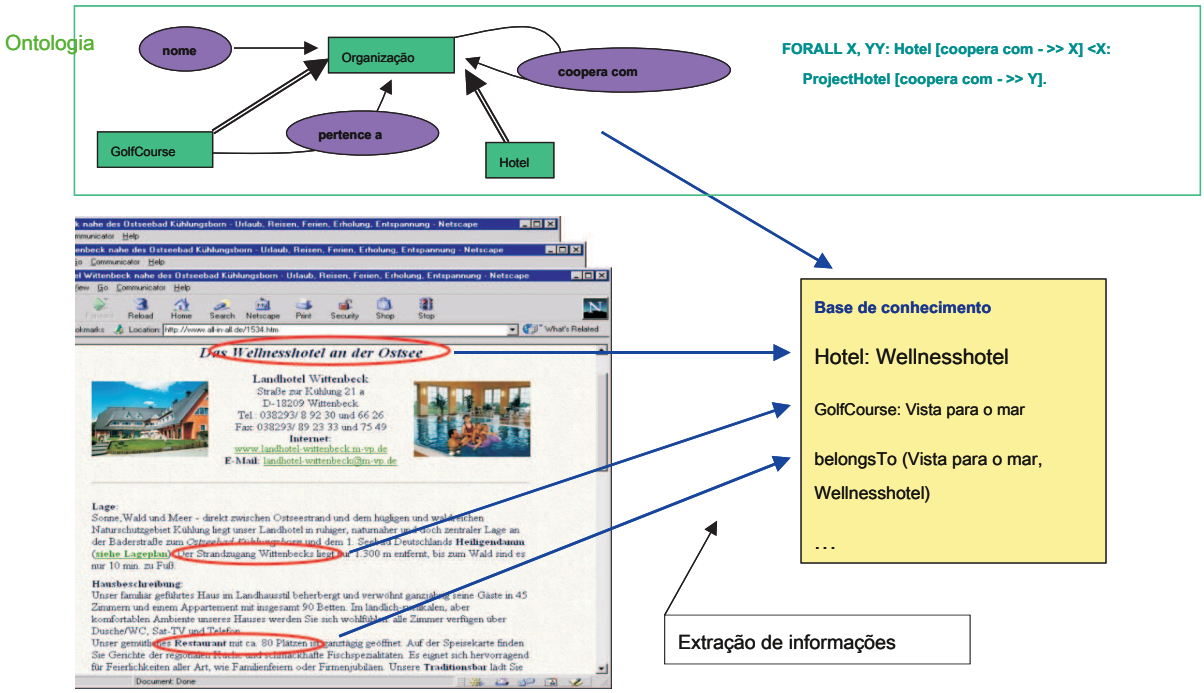

Fig. 6. Etapa 2: Minerando a Web para preencher a ontologia.

extrair instâncias e relações de outras páginas da Web e de hiperlinks.

No exemplo mostrado na figura, a relação "pertence a" entre os conceitos "campo de golfe" e "hotel" é instanciada pelo par (SeaView, Wellnesshotel), ou seja, pelo fato derivado das páginas da Web disponíveis no campo de golfe chamado "SeaView" pertence ao "Wellness Hotel".

Após o segundo passo, temos uma ontologia e uma base de conhecimento, ou seja, instâncias dos conceitos e relações de ontologia entre eles. Esses dados agora são inseridos na terceira etapa, na qual *a base de conhecimento é extraída.* Dependendo da finalidade, diferentes técnicas podem ser aplicadas. Pode-se, por exemplo, calcular regras de associação relacional, como descrito em detalhes em [49] (ver Fig. 7). Outra possibilidade é agrupar conceitualmente as instâncias [164].

No exemplo mostrado na Fig. 7, uma combinação de conhecimento sobre instâncias como o Wellnesshotel e seu campo de golfe SeaView, com outro conhecimento derivado dos textos das páginas da Web, produz a regra de que hotéis com campos de golfe geralmente têm cinco estrelas. Mais precisamente, isso é válido para 89% dos hotéis com campos de golfe e 0,4% de todos os hotéis da base de conhecimento são cinco estrelas que possuem um campo de golfe. Os dois valores são a confiança e o suporte da regra, medidas padrão para regras de associação de mineração.

A estrutura semântica resultante agora pode ser usada para entender melhor os padrões de uso. No nosso exemplo, o agrupamento de sessões do usuário pode identificar um agrupamento de usuários que visitam e examinam atentamente as páginas do “Wellnesshotel”, do “Schlosshotel” e do “Hotel Mecklenburg”. Embora essas informações, por si só, sejam suficientes para gerar uma recomendação dinâmica “Você pode

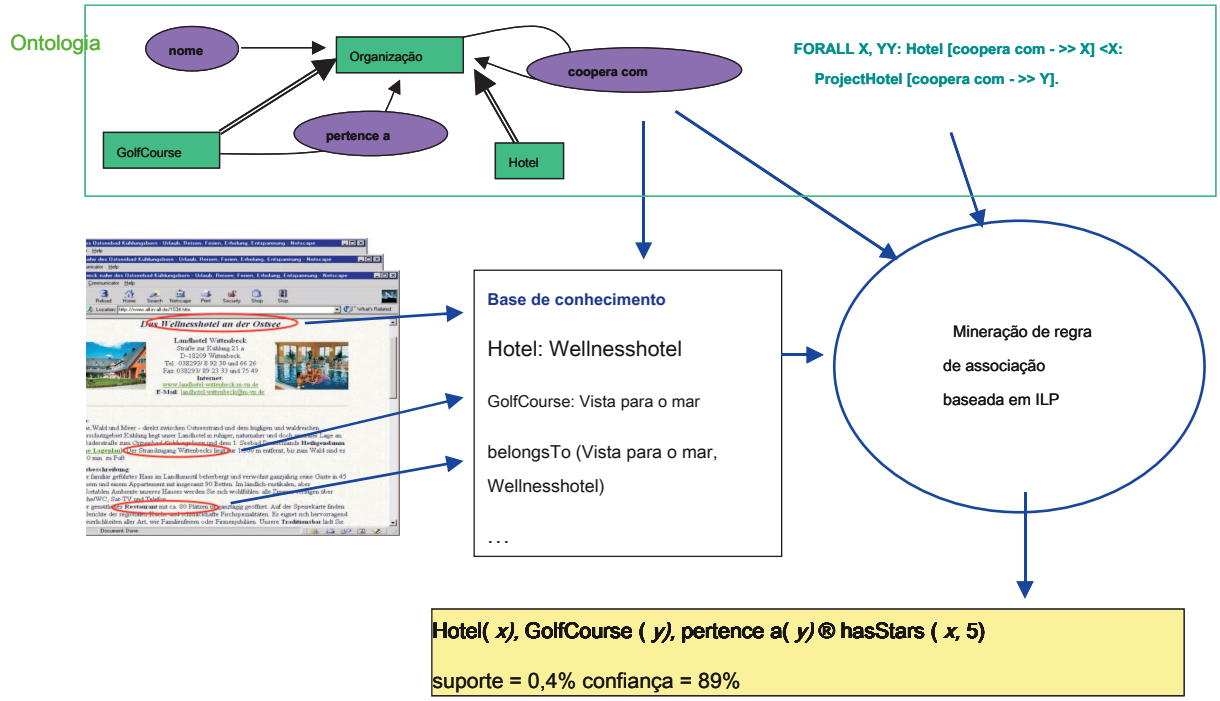

Fig. 7. Etapa 3: usando a ontologia para minerar novamente.

quer também olhar para o Castle Hotel at the Lake ”para novos usuários que visitam o“ Wellnesshotel ”e o“ Hotel Mecklenburg ”, ainda não está claro *porque* esse cluster de hotéis pode ser interessante para um grupo considerável de usuários. Esse problema pode ser resolvido usando nossa ontologia para calcular *perfis de uso em nível de domínio [*45]: Concluímos que todos esses hotéis são caracterizados por possuir um campo de golfe.

Esse entendimento dos padrões de uso pode ajudar-nos a alcançar nosso objetivo inicial (b), a geração de recomendações para o novo design do site. Propomos a introdução de uma nova categoria “hotéis de golfe” na ontologia do site, na arquitetura da informação e no design da página. O aprendizado de instância para esta categoria é simples: todos os "hotéis" para os quais existe um "campo de golfe" que "pertence" ao hotel, e somente esses, se tornam instâncias da nova categoria. O design do site e da página pode, por exemplo, ser modificado adicionando um novo valor "hotel de golfe" ao critério de pesquisa "instalações do hotel" na barra de navegação de pesquisa / navegação do site. Além disso, quando novos hotéis com campos de golfe são inseridos na base de conhecimentos, estes podem ser recomendados dinamicamente aos visitantes das páginas do “Wellnesshotel”, “Schlosshotel” e “Hotel Mecklenburg”.

Nosso objetivo inicial (a), a geração de um modelo semântico e metadados que refletem esse modelo, também foi atingido. Entre outros benefícios, isso permite que uma página do site que descreva o “Fischerhotel” na cidade de “Zingst” seja devolvida em resposta a uma consulta de mecanismo de pesquisa para “acomodação em Ahrenshoop”, porque “hotéis” é conhecido por ser um Sabe-se que a subclasse de "acomodação" e as cidades "Ahrenshoop" e "Zingst" estão localizadas na península "Fischland-Darß". O conhecimento anterior é retirado de nossa ontologia (veja a Fig. 5); o último é

recuperado pelo mecanismo de pesquisa de uma ontologia geográfica de uso geral disponível em outro site semanticamente enriquecido.

Como vimos, os resultados das etapas 3 e 4 podem levar a novas modificações da ontologia e / ou base de conhecimento. Quando novas informações são obtidas, elas podem ser usadas como entrada para as primeiras etapas na próxima etapa do ciclo de vida do site e da ontologia.

Obviamente, o controle final da qualidade, a decisão de manter ou descartar os conceitos da ontologia e a transformação das idéias obtidas na interpretação dos padrões de uso em mudanças no design do site continuam sendo uma responsabilidade humana. No entanto, na consecução de nossos objetivos iniciais, a combinação de métodos de mineração da Web Semântica e da Web economizou uma quantidade considerável de esforço manual necessário quando os dois objetivos são trabalhados isoladamente: o trabalho de criar e instanciar uma ontologia do turismo recursos de um grande número de modelos de páginas dinâmicas, esquemas de banco de dados e HTML bruto, bem como o trabalho de interpretação de padrões de URLs co-ocorrentes encontrados nas sessões do usuário.

Mineração da Web Semântica e outros loops de feedback. O ciclo de feedback descrito nesta seção compartilha vários recursos com abordagens para descobrir o conhecimento da Web que se baseia em uma noção mais vaga de semântica. Um excelente exemplo deste último é o K proposto recentemente AGORA Eu T UMA LL sistema [54]. Ele se baseia em uma abordagem de bootstrapping semelhante à descrita neste artigo: Instâncias de conceitos e relações são extraídas da Web, e a confiabilidade dessas instâncias é então julgada pela quantidade de suporte que essas asserções recebem da Web. . Devido ao tamanho da Web, abordagens totalmente automatizadas como esta parecem ser a principal via para obter acesso instantâneo ao conhecimento implícito em toda a Web, em particular suas partes ad-hoc em rápida mudança.

No entanto, esses sistemas baseados em sintaxe dependem da redundância maciça da Web e, portanto, podem obter acesso apenas a informações que podem ser encontradas em um grande número de páginas da Web (e que podem ser identificadas pelos modelos de linguagem natural necessariamente limitados usados para extração de informações). Nos anos 90, Voorhees [169] afirmou que, por uma questão de princípio, a Inteligência Artificial (e, em particular, a PNL) é incapaz de fornecer resultados de RI significativamente melhores do que essas abordagens sintáticas puras. Desde então, porém, houve progresso. Por exemplo, [129] venceu com distância significativa o concurso TREC na "tarefa de resposta a perguntas de domínio aberto" em 2002, combinando a PNL com representação e raciocínio lógicos do conhecimento.

Portanto, esperamos uma preferência pelas soluções de mineração na Web semântica quando o conhecimento buscado deve abranger o máximo de (ou todos) itens de informação disponíveis e não pode contar com a redundância e os “votos majoritários” implícitos em esquemas de mineração como K AGORA Eu T UMA LL ou PageRank. Devido ao trabalho extra exigido pelo menos pelos autores, os participantes em áreas de aplicação adequadas devem

dedicado à qualidade da informação, uma dedicação induzida por alta motivação intrínseca ou extrínseca. Contextos de aplicativos com uma tolerância acima do normal para serem observados pela mineração de uso se beneficiarão das vantagens adicionais da mineração de uso da Web Semântica. Exemplos proeminentes de áreas de aplicação que exibem essa combinação de recursos são a ciência (onde listas de literatura exaustivas são importantes), comunidades voluntárias unidas por interesses comuns e negócios (onde os custos de transação precisam ser minimizados). Atualmente, [26])

Além da cobertura e da qualidade, a forma da semântica descrita neste artigo tem mais duas vantagens que a tornam adequada para domínios de alto comprometimento. Ambas as vantagens derivam das diferenças de opacidade entre abordagens de processamento de informações sintáticas e semânticas. Primeiro, o processamento de informações de métodos estatísticos que operam exclusivamente em tokens sintáticos permanece opaco para a maioria dos usuários humanos, em particular quando algoritmos proprietários são empregados. Geralmente, não há como explicar, em termos compreensíveis ao usuário, por que um algoritmo chegou a um resultado específico. Por outro lado, uma conceituação explícita permite que pessoas e programas expliquem, raciocinem e discutam sobre significado e, assim, racionalizem sua confiança ou falta de confiança em um sistema. Segundo, sua relativa opacidade força métodos puramente estatísticos-sintáticos a confiar nas habilidades de fazer sentido do usuário individual. A experiência mostra que os usuários compreendem os resultados, mas geralmente de maneira ad hoc que não incentiva a reflexão ou a externalização. Por outro lado, o Semantic Web Mining apóia o desenvolvimento de loops de feedback de princípios que consolida o conhecimento extraído pela mineração em informações disponíveis para a Web em geral.

VII C ONCLUSÃO E O UTLOOK

Neste artigo, estudamos a combinação das duas áreas de pesquisa de rápido desenvolvimento Semantic Web e Web Mining. Discutimos como a Semantic Web Mining pode melhorar os resultados da Web Mining, explorando as novas estruturas semânticas na Web; e como a construção da Web Semântica pode fazer uso das técnicas de Mineração da Web. O exemplo fornecido na última seção mostra os possíveis benefícios de mais pesquisas nessa tentativa de integração.

Uma investigação mais aprofundada dessa interação dará origem a novas questões de pesquisa e estimulará mais pesquisas na Web Semântica e na Mineração da Web - em direção ao objetivo final da Mineração da Web Semântica: "uma Web melhor" para todos os seus usuários, uma "Web melhor utilizável" ”. Um foco importante é

[26] EDI (Electronic Data Interchange) é um formato padrão para troca de dados comerciais, internacionalmente padronizado em ISO

para permitir que os mecanismos de pesquisa e outros programas entendam melhor o conteúdo de páginas e sites da Web. Isso se reflete na riqueza de esforços de pesquisa que modelam as páginas em termos de uma ontologia do conteúdo, os objetos descritos nessas páginas.

Esperamos que, no futuro, os métodos de mineração da Web tratem cada vez mais o conteúdo, a estrutura e o uso de maneira integrada em ciclos iterados de *extração* e *utilizando* semântica, para poder entender e (re) moldar a Web. Entre os ciclos iterados, esperamos ver uma complementaridade produtiva entre aqueles que dependem da semântica no sentido da Web Semântica e aqueles que se baseiam em uma noção mais vaga da semântica.

R EFERÊNCIAS

[1] S. Acharyya e J. Ghosh. Modelagem sensível ao contexto do comportamento de navegação na Web usando árvores de conceito. No *Proc. do Workshop WebKDD sobre mineração na Web e análise de uso da Web,* páginas 1–8, 2003. [2]

CC Aggarwal. Rastreamento colaborativo: experiências de usuário de mineração para descoberta de recursos tópicos. No [*73],* páginas 423-428, 2002.

[3] CC Aggarwal, F. Al-Garawi e PS Yu. Rastreamento inteligente na rede mundial de computadores com predicados arbitrários. No *Anais da Conferência da WWW,* 2001.

[4] CC Aggarwal, SC Gates e PS Yu. Sobre os méritos da construção de sistemas de categorização por cluster supervisionado. No *KDD'1999 - Anais da Quinta Conferência Internacional ACM SIGKDD sobre Descoberta de Conhecimento e Mineração de Dados,* páginas 352–356, 1999. [5]

J. Allan, editor. *Detecção e rastreamento de tópicos: organização da informação baseada em eventos.* Kluwer Academic Publishers, Norwell, MA, 2002. [6]

SS Anand, M. Mulvenna e K. Chevalier. Na implantação da mineração de uso da web. No [*15]* páginas 23-42. 2004. [7]

CR Anderson, P. Domingos e DS Weld. Modelos relacionais de Markov e sua aplicação à navegação na Web adaptável. No [*73],* páginas 143-152, 2002. [8]

Pascal Auillans, Patrice Ossona de Mendez, Pierre Rosenstiehl e Bernard Vatant. Um modelo formal para mapas de tópicos. No *[82]* páginas 69–83, 2002. [9]

P. Baldi, P. Frasconi e P. Smyth, editores. *Modelando a Internet e a Web. Métodos Probabilísticos e Algoritmos.*

[10] S. Baron e M. Spiliopoulou. Monitorando a evolução dos padrões de uso da web. No [*15]* páginas 181–200. 2004. [11] M. Baumgarten, AG B¨ uchner, SS Anand, MD Mulvenna e JG Hughes. Descoberta de padrões de navegação orientada pelo usuário de dados da internet. No [*151],* páginas 74–91. 2000.

[12] Sean Bechhofer, Ian Horrocks, Carole Goble e Robert Stevens. OilEd: Um editor de ontologia razoável para a semântica Rede. *LNCS,* 2174: 396ff, 2001.

[13] B. Berendt. Detalhe e contexto na mineração de uso da Web: Sequências de visualização e visualização mais grossa. No [*104],* páginas 1–24. 2002. [14] B. Berendt. Usando a semântica do site para analisar, visualizar e dar suporte à navegação. *Mineração de dados e descoberta de conhecimento,* 6 (1): 37–59, 2002.

[15] B. Berendt, A. Hotho, D. Mladenic, M. van Someren, M. Spiliopoulou e G. Stumme, editores. *Mineração da Web: da Web Web Semântica. Primeiro Fórum Europeu de Mineração da Web, EWMF 2003. Artigos revisados convidados e selecionados,* volume 3209 de *LNAI.* Springer, Berlim, 2004.

[16] B. Berendt, A. Hotho e G. Stumme. Em direção à mineração semântica da web. No [*82],* páginas 264–278, 2002. [17] B. Berendt, A. Hotho e G. Stumme. Mineração de uso para e na web semântica. No [*95]* 2003. [18] B. Berendt, B. Mobasher, M. Nakagawa e M. Spiliopoulou. O impacto da estrutura do site e do ambiente do usuário na sessão reconstrução na análise de uso da web. No [*120],* páginas 115–129, 2002.

[19] B. Berendt e M. Spiliopoulou. Análise do comportamento da navegação em sites que integram vários sistemas de informação. *O Jornal VLDB,* 9 (1): 56–75, 2000.

[20] Stephan Bloehdorn e Andreas Hotho. Classificação de texto, estimulando alunos fracos com base em termos e conceitos. No

Anais da Quarta Conferência Internacional do IEEE sobre Mineração de Dados. IEEE Computer Society Press, 2004. [21] K. Bontcheva, V. Tablan, D. Maynard e H. Cunningham. Porta em evolução para enfrentar novos desafios na engenharia de idiomas.

engenharia de linguagem natural. *Engenharia de Linguagem Natural,* 10 (3/4): 349-373, 2004. [22] JL Borges e M. Levene. Mineração de dados de padrões de navegação do usuário. No [*151],* páginas 92-111. 2000. [23] E. Bozsak, M. Ehrig, S. Handschuh, A. Hotho, A. Maedche, B. Motik, D. Oberle, C. Schmitz, S. Staab, L. Stojanovic,

N. Stojanovic, R. Studer, G. Stumme, Y. Claro, J. Tane, R. Volz e V. Zacharias. Kaon - em direção a uma web semântica em grande escala. Em K. Bauknecht, A. Min Tjoa e G. Quirchmayr, editores, *Comércio Eletrônico e Tecnologias da Web, Terceira Conferência Internacional, EC-Web 2002, Proceedings,* volume 2455 de *LNCS,* páginas 304–313, Berlim, 2002. Springer. [24] B. Buchanan. Descoberta informada do conhecimento: Usando conhecimento prévio em programas de descoberta. No *KDD 2000 - Processos*

da Sexta Conferência Internacional ACM SIGKDD sobre Descoberta de Conhecimento e Mineração de Dados, Boston, MA, de 20 a 23 de agosto de 2000, página 3, Nova York, 2000. ACM. [25] P. Buitelaar, J. Franke, M. Grobelnik, G. Paaß e V. Sv. ´ atek, editores. *Anais do Workshop sobre Conhecimento*

Descoberta e Ontologias na ECML / PKDD 2004, 2004.

[26] W. Buntine, S. Perttu e V. Tuulos. Usando PCA discreto em páginas da web. No [*65],* páginas 99–110, 2004. [27] Mark H. Burstein, Jerry R. Hobbs, Ora Lassila, David Martin, Drew V. McDermott, Sheila A. McIlraith, Srini Narayanan,

Massimo Paolucci, Terry R. Payne e Katia P. Sycara. Daml-s: descrição do serviço da Web para a web semântica. No [*82],*

páginas 348–363, 2002.

[28] S. Chakrabarti. Mineração de dados para hipertexto: uma pesquisa tutorial. *Explorações SIGKDD,* 1 (2): 1–11, 2000. [29] S. Chakrabarti. *minerando a Web.* Morgan Kaufmann, São Francisco, CA, 2003.

[30] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan e S. Rajagopalan. Compilação automática de recursos por

analisando a estrutura do hiperlink e o texto associado. No *Anais da 7ª conferência na World Wide Web (WWW7), 30 (1-7),*

páginas 65-74, 1998.

[31] S. Chakrabarti, M. van den Berg e B. Dom. Rastreamento focado: uma nova abordagem para a descoberta de recursos da Web específicos de tópicos.

Redes de computadores, 31: 1623-1640, 1999.

[32] Hans Chalupsky. Ontomorph: Um sistema de tradução para conhecimento simbólico. No *Princípios de representação do conhecimento*

e Raciocínio: Anais da Sétima Conferência Internacional (KR2000), páginas 471–482, 2000. [33] C. Chen. *Visualização de informações e ambientes virtuais.* Springer, Londres, 1999.

[34] EH Chi, P. Pirolli, K. Chen e J. Pitkow. Usando o perfume da informação para modelar as necessidades e ações de informações do usuário

a teia. No *Anais da Conferência ACM CHI 2001 sobre Fatores Humanos em Sistemas Computacionais,* páginas 490–497, Amsterdã: ACM Press, 2001.

[35] EH Chi, P. Pirolli e J. Pitkow. O perfume de um site: um sistema para analisar e prever o cheiro, o uso e as informações das informações

usabilidade de um site. No *Anais da Conferência ACM CHI 2000 sobre Fatores Humanos em Sistemas Computacionais,* páginas 161–168, Amsterdã: ACM Press., 2000.

[36] EH Chi, A. Rosien e J. Heer. Descoberta e análise inteligentes da composição do tráfego de usuários da web. No [*120],* páginas 1–15,

2002.

[37] R. Cole e G. Stumme. Cem - um gerente de e-mail conceitual. Em B. Ganter e GW Mineau, editores, *Proc. ICCS 2000,*

volume 1867 de *LNAI,* páginas 438–452. Springer, 2000. [38] R. Cooley. *Mineração de uso da Web: descoberta e aplicação de padrões interessantes a partir de dados da Web.* Tese de Doutorado, Universidade

de Minnesota, maio de 2000.

[39] R. Cooley, B. Mobasher e J. Srivastava. Mineração na Web: descoberta de informações e padrões na Internet. No

Anais da Nona Conferência Internacional do IEEE sobre Ferramentas com Inteligência Artificial (ICTAI'97). IEEE Computer Society, novembro de 1997.

[40] R. Cooley, B. Mobasher e J. Srivastava. Preparação de dados para mineração de padrões de navegação na Internet. *Diário de*

Sistemas de Conhecimento e Informação, 1 (1): 5–32, 1999.

[41] R. Cooley, P.-N. Tang e J. Srivastava. Descoberta de padrões de uso interessantes a partir de dados da web. No [*151],* páginas 163–182.

2000.

[42] O. Corby, R. Dieng e C. H´ ebert. Um modelo conceitual de gráfico para a estrutura de descrição de recursos do w3c. Em B. Ganter

e GW Mineau, editores, *Estruturas conceituais: questões lógicas, linguísticas e computacionais, 8ª Conferência Internacional sobre Estruturas Conceituais, ICCS 2000, Darmstadt, Alemanha, 14 a 18 de agosto de 2000, Anais,* volume 1867 de *LNCS,* páginas 468–482. Springer, 2000.

[43] J. Cowie e Y. Wilks. Manual de processamento de linguagem natural. capítulo Extração de informações. Marcel Dekker, Novo York, 2000.

[44] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam e S. Slattery. Aprendendo a construir conhecimento bases da rede mundial de computadores. *Inteligência Artificial,* 118 (1-2): 69-113, 2000. [45] H. Dai e B. Mobasher. Usando ontologias para descobrir os perfis de uso da web no nível do domínio. No *Procedimentos do o Segundo Seminário de Mineração da Web Semântica na PKDD 2001,* km.aifb.uni-karlsruhe.de/semwebmine2002/papers/full/bamshad.pdf, agosto de 2002. [46] J. Davies, D. Fensel e F. van Harmelen, editores. *Conhecimento direto: Web semântica ativada para gerenciamento de conhecimento.* J. Wiley e Sons, 2002.

[47] J. Dean e MR Henzinger. Localizando páginas relacionadas na Internet. No *Anais da Oitava Internacional Conferência da World Wide Web WWW-1999,* Toronto, maio de 1999.

[48] SC Deerwester, ST Dumais, TK Landauer, GW Furnas e RA Harshman. Indexação por análise semântica latente. *Jornal da Sociedade Americana de Ciência da Informação,* 41 (6): 391-407, 1990. [49] L. Dehaspe e H. Toivonen. Descoberta de padrões freqüentes de registro de dados. *Mineração de dados e descoberta de conhecimento,* 3 (1): 7–36, 1999.

[50] P. Domingos, C. Faloutsos, T. Senador, H. Kargupta e L. Getoor, editores. *KDD'2003 - Anais da Nona ACM Conferência Internacional SIGKDD sobre Descoberta de Conhecimento e Mineração de Dados,* Nova York, 2003. ACM. [51] Saso Dzeroski e Nada Lavrac, editores. *Mineração relacional de dados.* Springer, 2001.

[52] M. Eirinaki, M. Vazirgiannis e I. Varlamis. Sewep: Usando a semântica do site e uma taxonomia para aprimorar a personalização da Web processo de organização. No [*50]* páginas 99-108, 2003. [53] Michael Erdmann. *Ontologien zur konzeptuellen Modellierung der Semantik from XML.* Isbn: 3831126356, Universidade de Karlsruhe, 10 2001.

[54] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, DS Weld e A. Yates. Métodos para extração de informações independentes de domínio da web: uma comparação experimental. No *Proc. da XIX Conferência Nacional de Inteligência Artificial (AAAI-04),* páginas 391–398, Menlo Park, CA, 2004. AAAI / MIT Press. [55] UM Fayyad, G. Piatetsky-Shapiro e P. Smyth. Da mineração de dados à descoberta de conhecimento. Na UM Fayyad, G. Piatetsky-Shapiro, P. Smyth e R. Uthurusamy, editores, *Avanços na descoberta de conhecimento e mineração de dados,* páginas 1–34. AAAI / MIT Press, Cambridge, MA, 1996.

[56] Alexander Felfernig, Gerhard Friedrich, Dietmar Jannach e Markus Zanker. Serviços da Web de configuração semântica no projeto cawicoms. No [*82],* páginas 192–205, 2002.

[57] D. Fensel, C. Bussler e A. Maedche. Serviços Web habilitados para Web Semântica. No [*82],* páginas 1–2, 2002. [58] D. Fensel, S. Decker, M. Erdmann e R. Studer. Ontobroker em poucas palavras. No *Conferência Europeia sobre Bibliotecas Digitais,* páginas 663–664, 1998.

[59] M. Fern´ andez, D. Fiorescu, A. Levi e D. Sucin. Especificação declarativa de sites com strudel. *O Jornal VLDB,* 9: 38–55, 2000.

[60] J. Forsyth, T. McGuire e J. Lavoie. *Todos os visitantes não são criados iguais.* Prática de Marketing da McKinsey, McKinsey & Companhia, abril de 2000, 2000.

[61] B. Ganter. Exploração de atributos com conhecimento prévio. *TCS,* 217 (2): 215-233, 1999. [62] B. Ganter e G. Stumme. Criação e fusão de níveis superiores de ontologia. No (*em preparação),* 2002. [63] B. Ganter e R. Wille. *Análise Formal de Conceitos: Fundamentos Matemáticos.* Springer, Berlim - Heidelberg, 1999.

[64] Asun Gomez-Perez, Juergen Angele, Mariano Fernandez-Lopez, V. Christophides, Athur Stutt e York Sure. Um questionário em ferramentas de ontologia. OntoWeb deliverable 1.3, Universidad Politecnia de Madrid, 2002. [65] M. Gori, M. Ceci e M. Nanni, editores. *Anais do Workshop sobre Abordagens Estatísticas para Mineração na Web em ECML / PKDD 2004,* 2004.

[66] B. Le Grand e M. Soto. Mapas de tópicos em XML e mineração semântica na Web. No [*162],* páginas 67-83, 2001.

[67] Benjamin Grosof, Ian Horrocks, Raphael Volz e Stefan Decker. Descrição Logic Programs: Combining Logic Programas com lógica de descrição. No *Proc. da WWW-2003,* Budapeste, Hungria, 05 de 2003.

[68] TR Gruber. Em direção a princípios para o desenho de ontologias usadas para compartilhamento de conhecimento. Em N. Guarino e R. Poli, editores, *Ontologia formal em análise conceitual e representação do conhecimento,* Deventer, Holanda, 1993. Kluwer. [69] Volker Haarslev e Ralf Moller. Descrição do sistema RACER e suas aplicações. Em DL McGuinness et al, editor, *Anais do Workshop Internacional de 2001 sobre Lógicas da Descrição (DL-2001).* Procedimentos da Oficina CEUR, 2001. [70] P. Haase, M. Ehrig, A. Hotho e B. Schnizler. Acesso a informações personalizadas em um sistema bibliográfico ponto a ponto. No *Proc. do Workshop de Personalização da Web Semântica na AAAI'2004,* páginas 1–12, 2004.

[71] Alon Y. Halevy e Jayant Madhavan. Representação do conhecimento baseado em corpus. Em Georg Gottlob e Toby Walsh, os tors, *IJCAI-03, Anais da Décima Oitava Conferência Internacional Conjunta sobre Inteligência Artificial, Acapulco, México, 9 a 15 de agosto de 2003,* páginas 1567-1572. Morgan Kaufmann, 2003. [72] Han e Kamber. *Mineração de dados. Conceitos e Técnicas.* Morgan Kaufmann, San Francisco, LA, 2001. [73] D. Hand, D. Keim e R. Ng, editores. *KDD - 2002 - Anais da Oitava Conferência Internacional da ACM SIGKDD sobre descoberta de conhecimento e mineração de dados,* Nova York, 2002. ACM. [74] D. Hand, H. Mannila e P. Smyth. *Princípios de mineração de dados.* Cambridge, MA: MIT Press, 2001. [75] Siegfried Handschuh e Steffen Staab. Criação e anotação de páginas da web em creme. No *Anais da Décima Primeira Conferência Internacional da World Wide Web, WWW2002,* páginas 462-473. ACM, 2002. [76] Siegfried Handschuh, Steffen Staab e Raphael Volz. Em anotações profundas. No *Proc. da WWW-2003,* Budapeste, Hungria, 05 2003.

[77] Marek Hatala e Griff Richards. Padrões de metadados globais x comunitários: capacitando usuários para troca de conhecimento. No [*82],* páginas 292–306, 2002.

[78] J. Heino e H. Toivonen. Detecção automatizada de epidemias a partir dos registros de uso do banco de dados de referência de médicos. No *[109]* páginas 180–191, 2003.

[79] Andreas Hess, Eddie Johnston e Nicholas Kushmerick. ASSAM: Uma ferramenta para anotar serviços da Web de maneira semi-automática com metadados semânticos. Em Sheila A. McIlraith, Dimitris Plexousakis e Frank van Harmelen, editores, *A Web Semântica - ISWC 2004: Terceira Conferência Internacional da Web Semântica,* volume 3298 de *Notas de aula em Ciência da Computação,* páginas 320–334. Springer, 2004.

[80] Andreas Hess e Nicholas Kushmerick. Aprendendo a anexar metadados semânticos aos serviços da web. No *A Web Semântica - Proc. Intl. Conferência da Web Semântica (ISWC 2003),* páginas 258–273. Springer, 2003.

[81] Jerry Hobbs, Douglas Appelt, John Bear, David Israel, Megumi Kameyama, Mark Stickel e Mabry Tyson. Fastus: A transdutor de estado finito em cascata para extrair informações de texto em idioma natural. Em E. Roche e Y. Schabes, editores, *Dispositivos de estado finito para processamento de linguagem natural.* MIT Press, Cambridge, MA, 1996. [82] I. Horrocks e JA Hendler, editores. *The Semantic Web - ISWC 2002, Primeira Conferência Internacional da Web Semântica, Anais,* volume 2342 de *LNCS.* Springer, 2002. [83] I. Horrocks e S. Tessaris. Consultando a Web: Uma abordagem formal. No [*82],* páginas 177–191, 2002. [84] A. Hotho, A. Maedche e S. Staab. Cluster de texto baseado em ontologia. No *Anais do Workshop IJCAI-2001 "Texto Learning: Beyond Supervision ", agosto, Seattle, EUA,* 2001.

[85] A. Hotho, A. Maedche, S. Staab e R. Studer. SEAL-II - o ponto fraco entre ricamente estruturado e não estruturado conhecimento. *Journal of Universal Computer Science,* 7 (7): 566-590, 2001.

[86] A. Hotho, S. Staab e G. Stumme. Explicando os resultados do cluster de texto usando estruturas semânticas. No [*109],* páginas 217–228, 2003.

[87] A. Hotho, S. Staab e G. Stumme. Ontologias melhoram o agrupamento de documentos de texto. No *Proc. do ICDM 03, de 2003 Conferência Internacional IEEE sobre Mineração de Dados,* páginas 541–544, 2003.

[88] EH Hovy. Combinação e padronização de ontologias práticas em larga escala para tradução automática e outros usos. No *Proc. 1st Intl. Conf. sobre Recursos e Avaliação de Idiomas (LREC),* Granada, 1998. [89] M. Hu e B. Liu. Mineração e resumo de avaliações de clientes. No [*103],* páginas 695-700, 2004. [90] JZ Huang, M. Ng, W.-K. Ching, J. Ng e D. Cheung. Um modelo de cubo e análise de cluster para sessões de acesso à web. No *[104]* páginas 48-67. 2002.

[91] Frank van Harmelen Jeen Broekstra, Arjohn Kampman. Gergelim: Uma arquitetura genérica para armazenar e consultar rdf e esquema rdf. No [*82],* páginas 54–68, 2002.

[92] X. Jin, Y. Zhou e B. Mobasher. Mineração de uso da Web com base em análise semântica latente probabilística. No [*103],* Páginas 197-205, 2004.

[93] T. Joachims. Otimizando mecanismos de pesquisa usando dados de clique. No [*73],* páginas 133-142, 2002. [94] T. Joachims, D. Freitag e T. Mitchell. Webwatcher: um guia turístico da Internet. No *Procedimentos da Inter-Conferência Conjunta Nacional sobre Inteligência Artificial (IJCAI),* páginas 770-777, San Francisco, CA, 1997. Morgan Kaufmann. [95] H. Kargupta, A. Joshi, K. Sivakumar e Y. Yesha, editores. *Data Mining: desafios da próxima geração e futuro Instruções.* Imprensa AAAI / MIT, Menlo Park, CA, 2004.

[96] H. Kato, T. Nakayama e Y. Yamane. Ferramenta de análise de navegação baseada na correlação entre distribuição de conteúdos e padrões de acesso. No *Notas de trabalho do Workshop sobre mineração na Web para comércio eletrônico - desafios e oportunidades (WebKDD 2000) no KDD 2000,* páginas 95-104, Boston, MA, 2000. [97] C. Kemp e K. Ramamohanarao. Aprendizado de longo prazo para mecanismos de pesquisa na web. No *Anais do 6º Europeu Conferência sobre Princípios de Mineração de Dados e Descoberta de Conhecimento (PKDD 2002),* páginas 263–274, Berlim, 2002. Springer. [98] M. Kifer, G. Lausen e J. Wu. Fundamentos lógicos de linguagens orientadas a objetos e baseadas em quadros. *Jornal da ACM,* 42 (4): 741–843, 1995. [99] R. Kimball e R. Merx. *O Data Webhouse Toolkit - Construindo um Data Warehouse Ativado pela Web.* Wiley Computer Publishing, Nova York, 2000.

[100] Jon M. Kleinberg. Fontes autorizadas em um ambiente com hiperlink. *Jornal da ACM,* 46 (5): 604-632, 1999. [101] N. Koeppen, K. Polkehn e H. Wandke. Um conjunto de ferramentas para dar suporte aos exames de arquivos de log. No Noldus IT AG, editor, *Medindo Behavior 2002, 4th Cofnerence International on Methods and Techniques in Behavioral Research, 27-30 agosto 2002, Amsterdam,* 2002.

R. Kohavi. Dados de comércio eletrônico de mineração: os bons, os ruins e os feios. No *KDD 2001 - Anais da Sétima ACM Conferência Internacional SIGKDD sobre Descoberta de Conhecimento e Mineração de Dados, San Francisco, CA, 26 a 29 de agosto de 2002,* páginas 8–13, Nova York, 2001. ACM.

[103] R. Kohavi, J. Gehrke, W. DuMouchel e J. Ghosh, editores. *KDD'2004 - Anais do Décimo ACM SIGKDD Conferência Internacional sobre Descoberta de Conhecimento e Mineração de Dados,* Nova York, 2004. ACM. [104] R. Kohavi, BM Masand, M. Spiliopoulou e J. Srivastava, editores. *WEBKDD 2001 - Minerando dados de log da Web em todos Pontos de contato do cliente,* volume 2356 de *LNAI.* Springer, Berlin / Heidelberg, 2002. [105] D. Koller e M. Sahami. Classificando documentos hierarquicamente usando muito poucas palavras. No *Anais da 14ª Inter-Conferência Nacional de Aprendizado de Máquina (ML), Nashville, Tennessee, julho de 1997,* páginas 170-178, 1997. [106] R. Kosala e H. Blockeel. Pesquisa de mineração na Web: uma pesquisa. *Explorações SIGKDD,* 2 (1), 2000. [107] AHF Laender, BA Ribeiro-Neto, AS da Silva e JS Teixeira. Uma breve pesquisa sobre ferramentas de extração de dados da web. *SIGMOD Registro,* 31 (2): 84-93, 2002.

[108] T. Lau e Y. Claro. Introdução ao gerenciamento de habilidades baseadas em ontologia em uma grande companhia de seguros. No *Anais da Modellierung 2002,* Tutzing, Alemanha, março de 2002. [109] N. Lavra ˇ c, D. Gamberger, L. Todorovski e H. Blockeel, editores. *Anais da 7ª Conferência Europeia sobre Princípios e práticas de descoberta de conhecimento em bancos de dados: PKDD 2003,* volume 2838 de *LNAI,* Berlin Heidelberg, 2003. Springer.

[110] J. Lee, M. Podlaseck, E. Schonberg e R. Hoch. Visualização e análise de dados de fluxo de cliques de lojas online para entender o merchandising na web. *Mineração de dados e descoberta de conhecimento,* 5 (1/2): 59-84, 2001. [111] Jung-Won Lee, Kiho Lee e Won Kim. Preparativos para mineração xml baseada em semântica. Em Nick Cercone, Tsau Young Lin e Xindong Wu, editores, *Anais da Conferência Internacional IEEE de 2001 sobre mineração de dados, de 29 de novembro a 2 de dezembro de 2001, San Jose, Califórnia, EUA,* páginas 345–352. IEEE Computer Society, 2001. [112] W. Lin, SA Alvarez e C. Ruiz. Mineração eficiente de regras de associação de suporte adaptativo para sistemas de recomendação. *Dados Mineração e descoberta de conhecimento,* 6 (1): 83-105, 2002. [113] A. Maedche. *Aprendizado de Ontologia para a Web Semântica.* Kluwer, 2002.

[114] A. Maedche, M. Ehrig, S. Handschuh, L. Stojanovic e R. Volz. Rastreamento de documentos focado na ontologia e

metadados. No *Anais da décima primeira conferência internacional da World Wide Web WWW-2002,* Havaí, 2002. [115] A. Maedche e S. Staab. Descobrindo relações conceituais do texto. No *ECAI-2000 - Anais do 13º Congresso Europeu*

Conferência sobre Inteligência Artificial, páginas 321-325. IOS Press, Amsterdam, 2000. [116] A. Maedche e S. Staab. Aprendizado de ontologia para a web semântica. *Sistemas Inteligentes IEEE,* 16 (2): 72-79, 2001. [117] A. Maedche, S. Staab, R. Studer, Y. Claro e R. Volz. SEAL - Amarrando a integração de informações e o gerenciamento de sites

por ontologias. *Boletim de Engenharia de Dados do IEEE-CS, Edição Especial sobre Organização e Descoberta da Web Semântica,*

Março de 2002.

[118] I. Mani e M. Maybury. Avanços no resumo automático de texto. páginas 123–136. The MIT Press, 1999. [119] Inderjeet Mani. *Resumo Automático,* volume 3 de *Processamento de linguagem natural.* John Benjamins Publishing Com-

pany, Amsterdã / Filadélfia, 2001.

[120] B. Masand, M. Spiliopoulou, J. Srivastava e OR Zaëıane, editores. *Notas do workshop da quarta mineração WEBKDDWeb*

para padrões de uso e perfis de usuário no KDD'2002, Edmonton, Alberta, Canadá, 23 de julho de 2002. ACM. [121] D. McGuinness, R. Fikes, J. Rice e S. Wilder. Um ambiente para mesclar e testar grandes ontologias. No *No*

Anais da Sétima Conferência Internacional sobre Princípios de Representação e Raciocínio do Conhecimento (KR2000),

páginas 483–493, Breckenridge, Colorado, EUA, 2000.

[122] P. Melville, RJ Mooney e R. Nagarajan. Filtragem colaborativa aprimorada por conteúdo. No *Anais do ACM SIGIR*

Workshop sobre Sistemas Recomendadores, Setembro de 2001. [123] E. Menasalvas, S. Mill ´ um, MS P ´ erez, E. Hochsztain e A. Tasistro. Uma abordagem para estimar o valor das sessões do usuário

usando vários pontos de vista e objetivos. No [*15]* páginas 164–180. 2004. [124] DA

Menasc´ e V. Almeida, R. Fonseca e MA Mendes. Uma metodologia para caracterização da carga de trabalho do comércio eletrônico

sites. No *Anais da Conferência da ACM sobre Comércio Eletrônico,* Nova York, 1999. ACM. [125] Rosa Meo, Pier Luca Lanzi, Maristella Matera e Roberto Esposito. Integrando modelagem conceitual da web e uso da web

mineração. No *Proc. do Workshop WebKDD sobre mineração na Web e análise de uso da Web,* páginas 105-115, 2004. [126] Dunja Mladenic. Transformando o yahoo em classificador automático de páginas da web. No *Conferência Europeia sobre Inteligência Artificial,* Páginas

473-474, 1998.

[127] B. Mobasher, R. Cooley e J. Srivastava. Personalização automática com base na mineração de uso da web. *Comunicações do*

ACM, 43 (8): 142-151, 2000.

W. Moe. Compra, pesquisa ou navegação: diferenciando compradores on-line usando o fluxo de cliques de navegação na loja.

Journal of Consumer Psychology, 13 (1 e 2), 2002.

[129] D. Moldovan, S. Harabagiu, R. Girju, P. Morarescu, F. Lacatusu, A. Novischi, A. Badulescu e O. Bolohan. Ferramentas Lcc para

pergunta respondendo. Em Voorhees e Buckland, editores, *Anais da 11ª Conferência de Recuperação de Texto (TREC-2002),*

NIST, Gaithersburg.

[130] Wolfgang Nejdl, Wolf Siberski, Bernd Simon e Julien Tane. Em direção a uma linguagem de troca de modi

repositórios rdf. No [*82],* páginas 236-249, 2002.

[131] N. Noy e M. Musen. Prompt: Algoritmo e ferramenta para mesclagem e alinhamento automatizados de ontologias. No *Anais da*

Décima Sétima Conferência Nacional de Inteligência Artificial (AAAI-2000), páginas 450–455, Austin, Texas, 2000. [132] NF Noy, M. Sintek, S. Decker, M. Crubezy, RW Fergerson e MA Musen. Criando conteúdo da web semântico com

prot´ por e-2000. *Sistemas Inteligentes IEEE,* 16 (2): 60–71, 2001.

[133] D. Oberle, B. Berendt, A. Hotho e J. Gonzalez. Rastreamento conceitual do usuário. No *Web Intelligence, Primeira Internacional*

Conferência Atlantic Web Intelligence, AWIC 2003, Madri, Espanha, 5-6 de maio de 2003, Anais, volume 2663 de *LNCS,*

páginas 155-164, Berlin, 2003. Springer.

[134] G. Paaß, J. Kindermann e E. Leopold. Ontologias de protótipo de aprendizagem por análise semântica hierárquica latente. No [*25]*

páginas 49-60, 2004.

[135] L. Page, S. Brin, R. Motwani e T. Winograd. A classificação de citação PageRank: Trazendo ordem para a web. No *Procedimentos*

da 7ª Conferência Internacional da World Wide Web, páginas 161-172, Brisbane, Austrália, 1998. [136] Massimo Paolucci, Takahiro Kawamura, Terry R. Payne e Katia P. Sycara. Correspondência semântica de recursos de serviços da Web

ities. No [*82],* páginas 333-347, 2002.

[137] S. Parent, B. Mobasher e S. Lytinen. Um agente adaptável para exploração na Web com base em hierarquias de conceitos. No *Procedimentos da 9ª Conferência Internacional sobre Interação Humano-Computador,* Nova Orleans, LA, 2001. [138] P. Patel-Schneider e D. Fensel. Camadas na web semântica: problemas e orientações. No [*82],* páginas 16–29, 2002. [139] P. Patel-Schneider e J. Sim´ eon. Construindo a web semântica em xml. No [*82],* páginas 147-161, 2002.

[140] Joachim Peer. Reunindo web semântica e serviços da web. No [*82],* páginas 279–291, 2002. [141] Ramana Rao Peter Pirolli, James Pitkow. Seda da orelha de uma porca: extraindo estruturas utilizáveis da web. No *Proc. ACM Conf. Fatores humanos em sistemas de computação, CHI,* páginas 118–125, Nova York, NY, 1996. ACM Press. [142] D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis e M. Dikaiakos. Diretórios da comunidade da Web: um novo abordagem à personalização da web. No [*15]* páginas 113-129. 2004.

[143] JR Punin, MS Krishnamoorthy e MJ Zaki. Logml: linguagem de marcação de log para mineração de uso da web. No [*104],* Páginas 88-112. 2002.

[144] Tobias Scheffer e Stefan Wrobel. Um algoritmo de amostragem seqüencial para uma classe geral de critérios de utilidade. No *Conhecimento Descoberta e mineração de dados,* páginas 330-334, 2000. [145] F. Sebastiani. Aprendizado de máquina na categorização automatizada de texto. *Pesquisas de computação da ACM,* 34 (1): 1-47, 2002. [146] SJ Simoff. Variações na mineração de dados multimídia. Em SJ Simoff e OR Zaëıane, editores, *Continuar-do Workshop MDKM / KDD2000 sobre Mineração de Dados Multimídia,* páginas 104–109, www.cs.ualberta.ca/ zaiane / mdm kdd2000 / mdm00-15.pdf, 2000. _

[147] M. Sintek e S. Decker. Triplo - uma linguagem de consulta, inferência e transformação para a web semântica. No [*82],* Páginas 364-378, 2002. [148] JF Sowa. *Estruturas conceituais: processamento de informações na mente e na máquina.* Editora Addison-Wesley, Reading, MA, 1984.

[149] K. Sparck-Jones e P. Willett, editores. *Leituras em Recuperação de Informação.* Morgan Kaufmann, 1997. [150] M. Spiliopoulou. A maneira trabalhosa da mineração de dados à mineração na web. *International Journal of Computer Systems, Science, & Engenharia,* 14: 113-126, 1999. [151] M. Spiliopoulou e BM Masand, editores. *Avanços na análise de uso da Web e perfil de usuário,* volume 1836 de *LNAI.* Springer, Berlim / Heidelberg, 2000.

[152] M. Spiliopoulou e C. Pohle. Mineração de dados para medir e melhorar o sucesso de sites. *Mineração de dados e Descoberta do conhecimento,* 5: 85-14, 2001.

[153] M. Spiliopoulou, C. Pohle e M. Teltzrow. Estratégias de uso de sites de modelagem e mineração. No *Anais da Multi-Konferenz Wirtschaftsinformatik,* Sep 2002.

[154] R. Srikant e R. Agrawal. Regras de associação generalizadas de mineração. No *Anais da 21ª Conferência Internacional sobre Bancos de dados muito grandes,* páginas 407-419, setembro de 1995.

[155] J. Srivastava, R. Cooley, M. Deshpande e P.-N. Bronzeado. Mineração de uso da Web: descoberta e aplicação de padrões de uso de dados da web. *Explorações SIGKDD,* 1 (2): 12–23, 2000.

[156] J. Srivastava, P. Desikan e V. Kumar. Mineração na Web - conceitos, aplicativos e instruções de pesquisa. No [*95]* 2003. [157] S. Staab, H.-P. Schnurr, R. Studer e Y. Claro. Processos e ontologias do conhecimento. *Sistemas Inteligentes IEEE, Especiais Questão sobre Gestão do Conhecimento,* 16 (1), janeiro / fevereiro de 2001.

[158] N. Stojanovic, A. Maedche, B. Motik e N. Stojanovic. Gerenciamento de evolução de ontologia orientada pelo usuário. No *Procedimentos de os 13◦ Conferência Europeia sobre Engenharia do Conhecimento e Gestão do Conhecimento EKAW'02,* 2002.

[159] R. Studer, VR Benjamins e D. Fensel. Engenharia do conhecimento: princípios e métodos. *Engenharia de conhecimento de dados,* 25 (1–2): 161–197, 1998.

G. Stumme. Processamento analítico on-line conceitual. Em K. Tanaka, S. Ghandeharizadeh e Y. Kambayashi, editores, *Organização da informação e bancos de dados,* capítulo 14, páginas 191–203. Kluwer, Boston, MA, 2002. [161] G. Stumme. Utilizando ontologias e análise formal de conceitos para organizar o conhecimento dos negócios. Em J. Becker e R. Knack-stedt, editores, *Wissensmanagement mit Referenzmodellen - Konzepte f ¨ ur die Anwendungssystem- und Organisationsgestal-tung,* páginas 163-174, Heidelberg, 2002. Physica. [162] G. Stumme, A. Hotho e B. Berendt, editores. *Mineração da Web Semântica,* Freiburg, 3 de setembro de 2001. 12ª Europ. Conf.

sobre Machine Learning (ECML'01) / 5ª Europ. Conf. sobre princípios e práticas de descoberta de conhecimento em bancos de dados (PKDD'01).

[163] G. Stumme e A. Maedche. FCA – Merge: fusão de baixo para cima de ontologias. No *IJCAI-2001 - Anais do dia 17 Conferência Conjunta Internacional sobre Inteligência Artificial, Seattle, EUA, de 1 a 6 de agosto de 2001,* páginas 225–234, São Francisco, 2001. Morgen Kaufmann.

[164] G. Stumme, R. Taouil, Y. Bastide, N. Pasqier e L. Lakhal. Estrutura de computação do conceito do iceberg com titânico. *J. em Engenharia de Conhecimento e Dados,* 42 (2): 189–222, 2002.

[165] Y. Claro, M. Erdmann, J. Angele, S. Staab, R. Studer e D. Wenke. OntoEdit: Desenvolvimento de ontologia colaborativa para a web semântica. No [*82],* páginas 221-235, 2002.

[166] Y. Claro, S. Staab e J. Angele. OntoEdit: Orientando o desenvolvimento de ontologias por metodologia e inferências. No *Continuar-Conferência Internacional de Ontologias, Bancos de Dados e Aplicações do SEmantics ODBASE 2002,* Universidade da Califórnia, Irvine, EUA, 2002.

[167] Y. Claro e R. Studer. Metodologia On-To-Knowledge. No [*46]* capítulo 3, páginas 33–46. 2002. [168] M. Teltzrow e B. Berendt. Métricas de sucesso baseadas na utilização da Web para empresas multicanais. No *Proc. do WebKDD Workshop sobre mineração na Web e análise de uso da Web,* páginas 17–27, 2003.

[169] E. Voorhees. Processamento de linguagem natural e recuperação de informações. Em MT Pazienza, editor, *Extração de informações: Em direção a sistemas escaláveis e adaptáveis,* páginas 32–48. Springer, Berlim etc., 1999.

[170] AB Williams e C Tsatsoulis. Uma abordagem baseada em instâncias para identificar relações de ontologias candidatas dentro de um sistema de agente. No *Anais do Primeiro Workshop sobre Aprendizagem de Ontologia OL'2000,* Berlim, Alemanha, 2000. Décima quarta Conferência Européia sobre Inteligência Artificial. [171] IH Witten e E. Frank. *Mineração de dados. Ferramentas e técnicas práticas de aprendizado de máquina com implementações Java.* Morgan Kaufmann, São Francisco, CA, 2000.

[172] L. Yi, B. Liu e X. Li. Eliminando informações ruidosas em páginas da web para mineração de dados. No [*50]* páginas 296-305, 2003. [173] G. Yihune. *Avaliação eines medizinischen Informationssystems na World Wide Web.* Tese de doutorado, Medizinische Fakultë às der Ruprecht-Karls-Universitë em Heidelberg, Alemanha, 2003.

[174] A. Ypma e T. Heskes. Categorização de páginas da web e agrupamento de usuários com misturas de modelos de markov ocultos. No *[120]* páginas 31–43, 2002.

[175] OR Zaëıane e SJ Simoff. Mdm / kdd: mineração de dados multimídia pela segunda vez. *Explorações SIGKDD,* 3 (2), 2003. [176] OR Zaëıane, M. Xin e J. Han. Descobrindo padrões e tendências de acesso à web aplicando a tecnologia olap e de mineração de dados em logs da web. No *Conferência de Anais de Avanços nas Bibliotecas Digitais (ADL'98),* páginas 19–29, abril de 1998. [177] Osmar R. Zaëıane. Da descoberta de recursos à descoberta de conhecimento na Internet. Relatório Técnico TR 1998-13, Simon Universidade de Fraser, 1998.

[178] D. Zhang e WS Leee. Aprendendo a integrar taxonomias da web. *Journal of Web Semântica,* 2 (2): 131–151, 2004.