Recursos tecnológicos aplicáveis a bases de dados geográficos para extração de informações relevantes na área de Turismo[1]

Margarethe Born STEINBERGER-ELIAS [2]
Thiery OKUYAMA[3]
Universidade Federal do ABC (UFABC), Santo André, SP

Resumo
O ponto de partida deste trabalho é a convicção de que há no Brasil um amplo capital turístico ainda inexplorado por falta de informação. O texto trata da extração de informação relevante sobre turismo local e regional em bases de dados geográficos brasileiros (IBGE) e latino-americanos (Cepal). Os recursos tecnológicos para recuperação de informação baseiam-se na indexação prévia de destinos turísticos reconhecidos. O problema a ser discutido aqui é, em primeiro lugar, o conceito de "destino turístico" e suas condições de aplicação. E, em segundo lugar, a possibilidade de emergência de novos destinos turísticos a partir de uma busca inteligente em bases de dados geográficos. Esse tipo de busca extrai informação implícita através do cruzamento de dados.
Palavras-chave
Bases de dados geográficos; recuperação de informação; relevância; turismo regional

1. Condições de produção de lugares turísticos

O município de Santo André/ SP não é considerado um destino turístico. Se inserirmos o nome "Santo André" no site da Embratur, nenhum resultado será encontrado. Já na base do IBGE, dentre as informações sobre o município, consta um histórico que inclui o distrito de Paranapiacaba, localizado na Serra do Mar e dotado de grande atrativo turístico. Como antiga estação de trem, Paranapiacaba guarda o aposentado trem Maria Fumaça e tem um museu ferroviário. Além disso, é um lugar de grande beleza e recursos naturais, chamando turistas da Grande São Paulo e outras localidades para o ecoturismo. Na base da Embratur, não há correlação entre Paranapiacaba e Santo André, ou seja, numa busca por Santo André, não vai aparecer Paranapiacaba. Nos sites da Cepal e do IBGE também não consta nada que identifique a região como turística. Só o site da prefeitura de Santo André[4] traz Paranapiacaba na página inicial como sede do 9º Festival de Inverno e suas atrações. O link Cultura, Esporte, Lazer e Turismo traz mais informação sobre o Festival. Há também o link

[1] Trabalho apresentado no Grupo de Pesquisa, Geografias da Comunicação (DT7), evento componente do X Congresso de Ciências da Comunicação, Blumenau, 2009 .
[2] Orientadora do trabalho. Professora Doutora do Programa de Pós-Graduação de Engenharia da Informação da Universidade Federal do ABC (UFABC), pesquisadora de sistemas de inteligência social através de redes sócio-comunicativas. mborn@ufabc.edu.br
[3] Thiery Oçkuyama Silva Murakami é mestranda do curso de Engenharia da Informação da UFABC, thieryyama@hotmail.com
[4] http://www.santoandre.sp.gov.br/secretaria/default.asp?categ=sec_cultura site consultado 10-07-09

Gestão de Recursos Naturais de Paranapiacaba e Parque Andreense, com mais informação sobre a vila e seus serviços[5].

Em artigo anterior (Steinberger & Okuyama, 2008) apresentamos o turismo como um modo de produção e apropriação de lugares. Naquele trabalho, mostramos que a divulgação jornalística ajuda a consolidar esse modo de produção e a criar a idéia de destinos turísticos naturalizados. De fato, a expressão "lugares turísticos" é usada como se eles fossem assim em sua essência, como se tivessem nascido como tais, como se a mão do homem não tivesse ido lá e apontado o que via através de suas lentes de rentabilidade econômica. Agora, neste trabalho, já partimos do pressuposto de que os destinos turísticos resultam de uma produção social dos lugares e que a chamada "informação turística" organizada em guias e roteiros não só é uma expressão desse processo, como contribui para consolidá-lo.

A identidade dos lugares tem sido um tema recorrente em nossas pesquisas (cf.Steinberger, 2005). Tomamos como matriz teórica o trabalho de Castoriadis (1975), para quem o imaginário social que produz os lugares organiza-se a partir de "lógicas identitárias" (estabelece identidades a partir de relações). Há dois modos básicos de criar identidades: pelo Dizer e pelo Fazer. Exemplos do primeiro modo são a identificação "Caribe" como destino turístico ao invés de usar o nome de cada país da região; ou a identificação de "América" para referir à América anglo-saxônica apenas. A linguagem organiza identitariamente os lugares como destinos turísticos. O segundo modo de estabelecer identidades é o do fazer. A sociedade "fabrica" lugares ao dotá-los de *valor turístico*. Há lugares que se tornam destinos turísticos por empenho da comunidade local."Explorar" o turismo já designa uma ação instrumental.

Instituir um destino turístico é também instituir um mundo de significações. Os atrativos de um lugar não existem por si, eles valem enquanto *signos*. São Paulo era a terra da garoa, depois tornou-se a metrópole que não podia parar, e hoje é a maior cidade da América Latina. Criam-se estereótipos correlacionados aos lugares, como parte da fabricação sócio-semiótica do modo de ser desses lugares e assim eles se tornam destinos turísticos. O imaginário, depois de estabelecido, ganha autonomia em

[5] http://www.google.com.br/search?sourceid=navclient&aq=0h&oq=pre&hl=pt-BR&ie=UTF-8&rlz=1T4GPTB_pt-BRBR289BR290&q=prefeitura+santo+andre Site consultado em 10.07.09

relação à vida social e gera "conseqüências próprias que vão além de seus motivos funcionais e mesmo às vezes os contrariam", podendo se perpetuar para além das "circunstâncias que o fizeram nascer" (idem p.145). A escolha e delimitação dos lugares a serem nomeados e incluídos num registro turístico fazem parte de um processo social de "discretização da experiência", ou seja, de conversão de uma experiência contínua em uma experiência discreta, que pode ser mensurada e manipulada através de ferramentas simbólicas. Não é uma escolha individual, é socialmente motivada.

Problematizada assim a identificação de lugares turísticos, como então reconhecê-los em bases de informação? As seções seguintes vão descrever recursos tecnológicos que podem contribuir para uma identificação automática de destinos turísticos. A organização das informações no site da Embratur e em duas bases de dados geográficos brasileiros (IBGE) e latino-americanos (Cepal) serão descritas na seção 2; os recursos tecnológicos disponíveis para extração de informação relevante e recuperação de informação em bases de dados serão apresentados nas seções 3 e 4; a web semântica como tecnologia inteligente de busca, na seção 5; e, na seção 6, os primeiros passos para encaminhar uma aplicação.

2. Busca de informação em bases de dados geográficos: IBGE, Cepal e Embratur

Na maioria dos países da América Latina, as amplas bases de dados que disponibilizam informação para o desenvolvimento de políticas públicas não prevêem correlações relevantes para o campo do Turismo. Há uma notória precariedade dos sistemas de informação turística na América Latina, apesar de configurar-se como atividade econômica cada vez mais rentável. O site da Organização Mundial do Turismo nas Nações Unidas (Untwo/OMT)[6] informa que a receita do turismo internacional em 2007 foi de US$ 856 bilhões, com aumento em termos reais de 5,6% em comparação a 2006. Entre janeiro e agosto de 2008, os desembarques turísticos internacionais cresceram 3,7%, comparado ao mesmo período de 2007. E há uma previsão de 1,6 milhões de chegadas turísticas internacionais até 2020.

[6] Agência internacional especializada em turismo, é também fórum mundial sobre políticas de desenvolvimento do setor. Publica a UNWTO Barómetro Mundial do Turismo desde junho de 2003, com informações sobre transporte aéreo e dados econômicos sobre a evolução do setor. http://www.unwto.org/facts/eng/barometer.htm

A Internet já é a principal fonte de informação turística. Perto de 95% dos usuários vão à Internet buscar informação sobre viagens. Em 2001 foram 46,7 milhões de reservas em hotéis de todo o mundo, segundo a rede de distribuição eletrônica de associação , o que gerou US$ 12,9 bilhões em receitas. A natureza da atividade turística tem especificidades econômicas para o desenvolvimento regional e para impulsionar as tecnologias de informação. A Internet cresce no âmbito da busca de informações turísticas, o que torna instigante o estudo de aplicações que extraiam o conhecimento de documentos na web (Staab & Werthener, 2002). Torna-se cada vez mais freqüente o uso de ferramentas automatizadas para buscar, extrair, filtrar e avaliar as informações e os recursos desejados. Com a transformação da Web no principal meio de comércio eletrônico, organizações e companhias investem milhões em tecnologias voltadas a modelos e padrões de acesso. Cresce a procura por serviços e sistemas inteligentes que sejam capazes de minerar conhecimento relacionado ao setor do turismo (Cooley,Mobasher & Srivastava, 1997). Sistemas de informação turística (tourism information systems -TIS) são hoje um novo tipo de negócio. A informação serve de suporte para o *e-tourism* (turismo eletrônico) e abrange serviços de companhias aéreas, redes hoteleiras, locadoras de automóveis e agências de viagens. Os TIS estão diretamente relacionados à criação de produtos e serviços e abarcam uma ampla rede de informações. Para organizar tais informações, criam-se sistemas automatizados de viagens, capazes, por exemplo, de estabelecer comparações entre serviços e seus preços.

A Empresa Brasileira de Turismo (Embratur) foi criada em 1966 no Rio de Janeiro como empresa estatal. Seu objetivo era fomentar a atividade turística no Brasil. Com a criação do Ministério do Turismo, ficou subordinada a ele com o novo nome de Instituto Brasileiro de Turismo e passou a concentrar-se na promoção e marketing de serviços e destinos turísticos. Na base de dados da Embratur há um grande volume de informação sobre destinos turísticos no Brasil, classificados em cinco categorias: “Sol e Praia”, “Cultura” (Arqueologia, Cidades e Patrimônio, Étnico, Intercâmbio e Festas Populares) “Ecoturismo” (Caminhadas, Flutuação, Observação de Pássaros, Observação de Fauna e Turismo em Cavernas), “Esportes”, “Negócios e Eventos”.

O usuário pode buscar informação sobre destinos por palavras-chave. Outra opção é clicando em “Destinos e Roteiros” para escolher um “tour”. Por exemplo, clicando em Cultura, o usuário pode selecionar a região, o estado da federação e o nome

da cidade. Se busca contato com a natureza, mas não sabe ao certo o que quer, encontrará uma descrição do que é possível sentir, vivenciar ou experimentar em cada categoria[7]. Há uma diferença significativa entre uma base de dados voltada a promover o turismo no Brasil e as outras duas bases, do IBGE e da Cepal, que vamos tratar a seguir. Nem sempre a informação já está organizada dentro dos parâmetros de relevância de quem faz a procura. Técnicas de busca mais avançadas, como a mineração de dados, são capazes de mostrar, por exemplo, que um município com sítios naturais de grande beleza é um destino turístico em potencial, se também tiver boas estradas de acesso e infra-estrutura de hospedagem. Informações sobre estradas e transportes estão disponíveis no site do IBGE e poderiam ser cruzadas com outros indicadores favoráveis à atividade turística. As bases de dados do IBGE[8] e da Cepal[9] são exemplos de repositórios que podem oferecer informação relevante para o domínio do Turismo, embora não o façam explicitamente.

O Instituto Brasileiro de Geografia e Estatística (IBGE) é o principal provedor de dados e informações do País. Seus objetivos são identificar e analisar o território, apresentar dados estatísticos da população e mostrar a evolução da economia através das atividades produtivas[10]. Tem status de instituição da administração pública federal, subordinado ao Ministério do Planejamento, Orçamento e Gestão[11].

A Comissão Econômica para a América Latina e o Caribe (Cepal) tem por objetivo realizar estudos, pesquisas e promover o desenvolvimento econômico e social mediante a cooperação e a integração regional e sub-regional. Foi criada pelo Conselho Econômico e Social das Nações Unidas (Ecosoc) em 25 de fevereiros de 1948, com sede em Santiago do Chile. É uma das cinco comissões econômicas das Nações Unidas (ONU) e foi criada para desenvolver e assessorar ações de promoção dos países da área,

[7] No tour "Sol e Praia", por exemplo: "*Se você sonha com uma terra onde o sol brilha o ano inteiro, as águas são mornas, há muita sobra de coqueiros e uma brisa fresca vinda do mar, venha ao Brasil. Aqui, esses elementos combinam-se harmoniosamente nas mais belas praias e esperam por você*".
[8] O Instituto Brasileiro de Geografia e Estatística (IBGE). Consultado em 09.07.2009. http://www.ibge.gov.br/home/
[9] Secretaria da Comissão Econômica para a América Latina e o Caribe (CEPAL). Consultado em 09.07.09. http://www.eclac.org/
[10] A página inicial do IBGE tem um índice composto pela missão do IBGE, Metas e Ações, Diretorias, Princípios Fundamentais das Estatísticas Oficiais, Informações Sociais e Econômicas, Informações Geográficas, Disseminação, Estatuto, Regimento Interno e Obrigatoriedade de prestação de informações estatísticas. Nos itens acima na página inicial temos Indicadores, População, Economia, Geociências, Canais, Download e pesquisa. O primeiro, Indicadores trata de fazer um levantamento sobre a agropecuária, indústria e pesquisa sobre o comércio. No item População temos as opções que tratam sobre indicadores sociais, censos demográficos, contagem da população, estatísticas de registro civil etc. No item Economia temos informações sobre a Indústria, Serviços, Assistência Social Privada sem fins lucrativos entre outros itens que fornece diretrizes da economia no país.. No item Geociências há um mapeamento da área territorial oficial, a geografia e os recursos naturais do País. O item Canais dá acesso a um banco de dados com séries estatísticas de cidades,estados, países etc. No item Canais Temáticos é possível ter acesso a por exemplo, Brasil em Síntese, ou IBGE de 7 a 12 anos, ainda é possível ter acesso a fóruns e apresentações.
[11] http://www.ibge.gov.br/home/disseminacao/eventos/missao/instituicao.shtm consultado em 13-07-09

reforçar suas relações econômicas e promover o desenvolvimento social sustentável. Todos os países da América Latina e do Caribe são membros, asssim como alguns países da América do Norte e Europa. No total são 44 Estados-membros[12] e 8 membros associados.[13] A Cepal tem duas sedes sub-regionais, uma para a América Central, localizada na cidade do México, e outra para o Caribe, situada em Porto Espanha, capital de Trinidad e Tobago. Possui cinco escritórios nacionais: Brasília, Bogotá, Buenos Aires, Montevidéu e Washington.[14]

O portal da Cepal (cf. Fig.1) apresenta um menu com o histórico da instituição, estados-membros, calendário e organização das atividades, programa de trabalho, perguntas freqüentes, oportunidades de emprego. No link do centro de imprensa, estão indexadas as publicações, estudos, entrevistas, manifestações, discursos, etc., relacionados a suas atividades. No link "Análise e investigação" há documentos digitalizados e classificados pela biblioteca do órgão[15]. "Informação Estatística" é outro link da página de abertura, fornece cifras e dados para análise da conjuntura sócio-econômica e ambiental do continente. Os dados são de organismos oficiais dos países e de agências internacionais. Neste link o usuário tem acesso aos itens: Apresentação, Base de Dados, Publicações Estatísticas, Métodos e Classificação. O item Base de Dados dá acesso, por exemplo, a estatísticas sociais como taxas e gráficos de migração populacional. Essa complexa estrutura de hiperlinks permite ao usuário buscar qualquer

[12] Estados-membros: Alemanha, Antigua e Barbuda, Argentina, Bahamas, Barbados, Belize, Bolívia, Brasil, Canadá, Chile, Colômbia, Costa Rica, Cuba, Dominica, Equador, El Salvador, Espanha, Estados Unidos da América, França, Granada, Guatemala, Guiana, Haiti, Honduras, Itália, Jamaica, Japão, México, Nicarágua, Países Baixos, Panamá, Paraguai, Peru, Portugal, Reino Unido da Grã-Bretanha e Irlanda do Norte, República Dominicana, República da Coréia, Santa Lúcia, São Cristóvão e Neves, São Vicente e Granadinas, Suriname, Trinidad e Tobago, Uruguai e Venezuela.

[13] Membros associados: Anguilla, Antilhas Holandesas, Aruba, Ilhas Virgens Britânicas, Ilhas Virgens dos Estados Unidos, Montserrat, Porto Rico, Ilhas Turcas e Caicos.

[14] O programa de trabalho é realizado através das seguintes divisões, unidades e serviços: Divisão de Desenvolvimento Econômico, Divisão de Desenvolvimento Social, Divisão de Desenvolvimento Produtivo e Empresarial, Divisão de Desenvolvimento Sustentável e Assentamentos Humanos, Divisão de Recursos Naturais e Infra-Estrutura, Divisão de Estatística e Projeções Econômicas, Divisão de População e Desenvolvimento, Divisão de Comercio Internacional e Integração, Divisão de Planejamento Econômico e Social (ILPES), Unidade da Mulher e Desenvolvimento, Unidade de Estudos Especiais, Unidade de Recursos Naturais e Energia, Unidade de Transporte, Unidade de Serviços de Informação, Biblioteca, Sedes Sub-regionais e Escritórios Nacionais.

[15] São 18 categorias: Questões Políticas e Jurídicas;Desenvolvimento Econômico e Financiamento do Desenvolvimento;Recursos Naturais e Meio Ambiemte;Agricultura Ciências Florestais e Pesca;Indústria; Transporte e Comunicação;Comércio Internacional; População;Assentamentos Humanos;Saúde;Educação;Emprego;Assistência Humanitária;Questões Sociais;Cultura;Ciência e Tecnologia; Descritores Geográficos;Questões de Organização. Por exemplo o item (Cultura) subdivide-se em seis categorias de acesso: Artes, Literatura e Música;Comunicação e Meios de Comunicação de Massa; Desenvolvimento Cultural;Documentação, Bibliotecnologia, Ciência da Informação,Material de Referência;Filosofia e Religião;Proteção da propriedade intelectual e dos bens culturais. Por exemplo, o item (Filosofia e Religião) dá acesso a "Ética e Filosofia", e se clicarmos em filosofia, teremos acesso a dois outros temas: "A Memória Coletiva e os Restos do Feminismo" e "A Memória Coletiva e os Desafios do Feminismo".

tema relativo à América Latina. O nome "América Latina" é uma ontologia, isto é, um nó de cruzamento de classes informacionais.

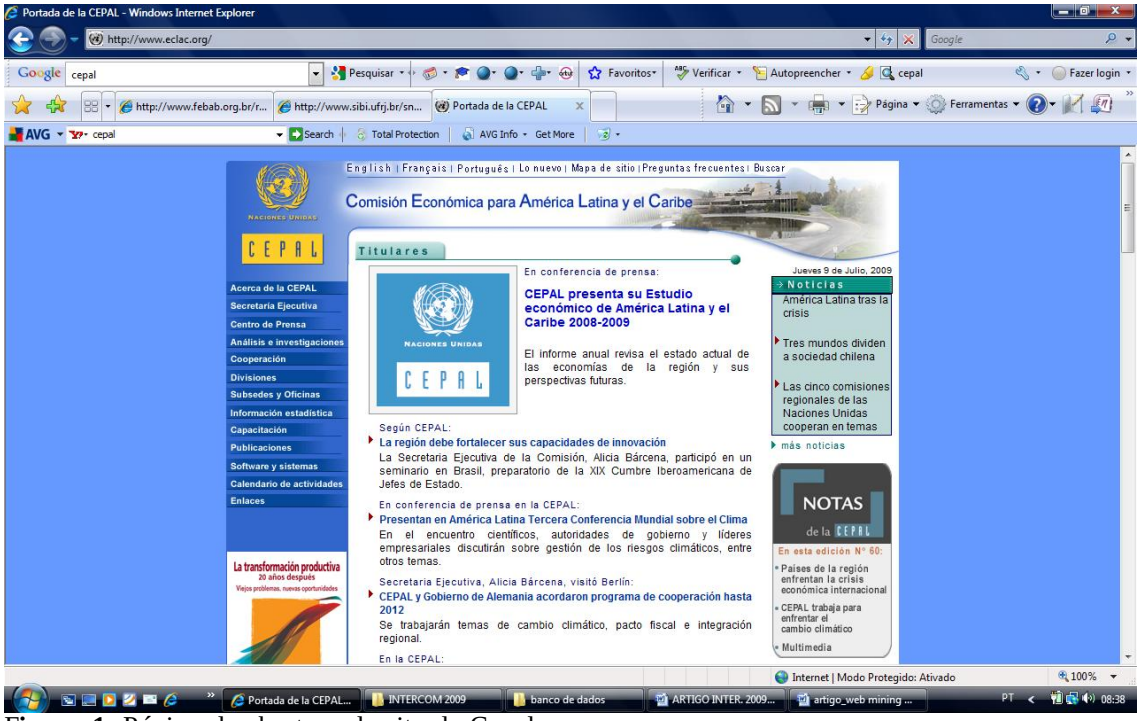

Figura 1. Página de abertura do site da Cepal

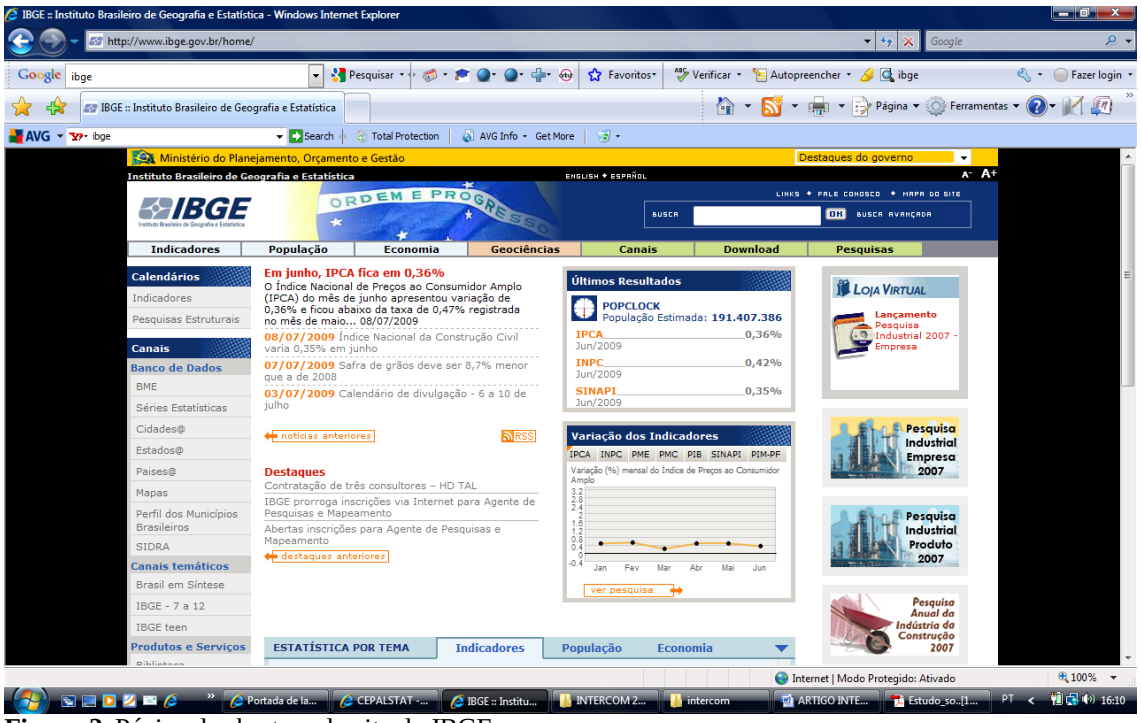

Figura 2. Página de abertura do site do IBGE

3. A mineração de dados e o problema da extração de informação relevante

O conceito da Web foi originalmente proposto por Tim Berners-Lee para o armazenamento de informações sobre as experiências dos aceleradores no CERN[16] . As informações foram organizadas em forma de grafos, ou seja, estruturas onde os nós eram documentos sujeitos a descrição (tais como artigos, departamentos ou pessoas), e as ligações eram relações como "depende de", "refere-se a". A proposta pareceu ideal para acessar grandes quantidades de texto, tais como as que existiam numa grande organização como a CERN . Uma evolução desta idéia foi idealizada para distribuir documentos em uma rede de computadores ao invés de armazená-los em um só computador ou em uma base de dados. Assim nasceu a web. Atualmente estima-se que ela tenha mais de 4 bilhões de páginas e mais de 1 milhão são adicionadas a cada dia (Markov & Larose 2007:143).

A Web Mining pode ser definida como uma ferramenta geral para descoberta e análise de informações acessadas através da World Wide Web. Esta definição ampla abarca técnicas de recuperação automática da informação, desde os recursos disponíveis em milhões de sites e em bases de dados on-line, até a descoberta e análise dos padrões de acesso dos usuários de um ou mais servidores Web ou serviços on-line (Cooley, & Mobasher, 1997). Essas técnicas incluem referência cruzada de sites ou de páginas de um mesmo site, e correlações baseadas em conhecimento implícito sobre o conteúdo do documento (Rezende, 2005:311).

O processo de mineração de dados na Web (Web Mining) pode ser definido como o uso de técnicas de mineração de dados (Data Mining) para descobrir e extrair automaticamente informações relevantes de documentos e serviços ligados a Internet, além de obter conhecimento, encontrar padrões e relações não conhecidos em bases de dados on-line. Web Mining é freqüentemente associada com "recuperação de informação", entretanto trata-se de um processo mais amplo e interdisciplinar, envolvendo técnicas de recuperação de informação, estatística, inteligência artificial e mineração de dados. As técnicas de Web Mining referem-se a aplicações de metodologias de mineração de dados e podem ser decompostas em quatro subtarefas (Kosala & Blockeel, 2000):

- **Busca de documentos:** recupera documentos por palavras-chaves, consiste no processo de extrair dados de páginas web tais como conteúdos de texto em documentos HTML. Envolve portanto técnicas de Recuperação de Informação.

[16] Laboratório de pesquisa em Física de Partículas situado na fronteira entre França e Suiça; reúne cientistas de diferentes países com a missão de construir o maior acelerador de partículas do mundo e reproduzir em experimentação controlada o processo de formação de buracos negros.

- **Seleção de informação e pré-processamento:** seleciona e pré-processa automaticamente as informações, num procedimento que envolve podagem de texto, transformação da representação da informação e adoção de outros formalismos.
- **Generalização:** detecta automaticamente padrões internos nos sites ou padrões entre múltiplos sites da web.
- **Análise:** valida e interpreta os padrões de dados minerados.

A recuperação de informação no processo de mineração de dados pode ser mais eficaz quando aplicada a dados organizados, estruturados por metadados e identificados semanticamente. O uso de metadados permite que os dados sejam filtrados gradualmente em conteúdos mais específicos de modo que se possa saber do que estão tratando (Rezende, 2005:312). A Web Semântica pode ser considerada como um grande conjunto de dados estruturados contidos na Web. Tecnologias semânticas aplicadas a tais dados fazem com que as etapas da mineração, como o pré-processamento e a extração de conhecimento, tornem-se cada vez mais simples e eficientes. A Web atual já possui uma considerável quantidade de problemas pelo fato de armazenar um enorme volume de dados que não são bem estruturados. A proposta da Web Semântica é associar semântica a estes dados e conseqüentemente facilitar as tarefas de Mineração de Dados para atingir seu principal objetivo:a recuperação de informação relevante (Stumme, Berendt & Hotho, 2002).

4. Tipos de bases de dados e opções de recuperação de informação

As técnicas de mineração lidam com bases de dados estruturados, semi-estruturados, não-estruturados. Dados semi-estruturados relacionam documentos da web a bancos de dados. Documentos são dados não estruturados. Bancos de dados são estruturados, o que significa que são organizados em estruturas bem definidas (esquemas). Como exemplo, temos os bancos de dados relacionais, esquemas compostos por nomes de tabela e atributos. Esta estrutura rígida evolui para dados semi-estruturados, ou seja, documentos XML que tem como objetivo permitir a representação menos complexa do mundo real (Kosola & Blockeel,2000)

Segundo Stumme, Berend & Hotho (2002), webmining é uma aplicação de técnicas de mineração de dados que utiliza abordagens de conteúdo, estrutura e uso dos recursos da Web. Desta forma é possível descobrir modelos ou combinações de forma global ou em estruturas locais entre as páginas da Web. Na Web é possível fazer três tipos de mineração: de conteúdo, de estrutura e de uso da internet. A Mineração do

Conteúdo da Web abrange as ferramentas que realizam a recuperação inteligente de informações do conteúdo dos documentos. A Mineração de Estruturas está interessada na informação que está implícita, sendo o seu principal foco as ligações de hipertextos que unem os documentos. Já a Mineração de Uso pode ser definida como a descoberta automática de padrões de acesso dos usuários aos servidores que disponibilizam informações na rede.

4.1. A Webmining de Conteúdo

A Webmining de conteúdo é um processo automático que extrai padrões de informações on-line, como arquivos HTML, imagens ou e-mails, que vão desde extrações de palavras-chave até a simples soma estatística de palavras ou frases em documentos. Segundo Rezende (2005:315), esse procedimento tem a capacidade de reestruturar o conteúdo dos documentos em uma representação apropriada para a manipulação por programas que podem extrair o conhecimento dos documentos ou dos resultados de busca por informação.

4.2. A Webmining de Estrutura

Webmining de estrutura baseia-se na análise da estrutura de links na web, uma de suas propostas é identificar os documentos prioritários. Em tese, a existência de um hyperlink do documento A para o documento B implica que A contém informações correlacionadas às de B. Para Rezende (2005:316), a internet possui mais informações que somente o conteúdo de suas páginas . A referência cruzada entre sites ou páginas de um mesmo site pode revelar um conhecimento implícito sobre os documentos. Essa técnica parte de referências presentes na web e pode descobrir links indiretos pertinentes a uma determinada área de conhecimento (Stumme, Berendt & Hotho, 2002). Ela opera na estrutura hiperlink das páginas da web e explora informações adicionais em hipertexto, portanto, é uma importante aplicação para identificar páginas relevantes.

4.3. A Web Mining de Uso

Servidores Web arquivam e acumulam dados sobre as interações de uso, não importa que tipos de requisições de recursos são recebidos. Analisar os logs de acesso de diferentes websites pode auxiliar a entender o comportamento dos usuários e a estrutura da web, bem como o escopo desta colossal coleção de recursos. Para Resende (2005:316), esse tipo de dado pode, por exemplo, determinar a vida útil de um produto, a estratégia associada a outro ou até mesmo alavancar um terceiro por meio de campanhas específicas.

5. A Web Semântica como tecnologia inteligente de busca

Web semântica é uma recente iniciativa proposta pela web consortium (w3c.org).O objetivo é introduzir através de técnicas uma representação formal do conhecimento na Web. O problema é que os formatos das páginas da Web são de difícil compreensão pelos computadores e até agora as máquinas não conseguiram vencer este desafio. A principal idéia por trás da web semântica é acrescentar uma descrição formal de cada página, que seja invisível para as pessoas, mas torne o conteúdo compreensível pelos computadores. A Web é a maior base de conhecimento do mundo. Com ajuda de ferramentas avançadas de raciocínio, poderia não só fornecer documentos classificados a partir de uma consulta por palavras-chave, mas também seria capaz de responder perguntas e dar explicações. (Markov & Larose, 2007:143)

A Web Semântica atual é enriquecida pela semântica formal baseada em ontologias[17], que capta o significado das páginas e links em linguagens compreensíveis para formulários de máquinas. Segundo Stumme, Berendt & Hotho, (2002), a idéia principal da Web Semântica é enriquecer a Web atual através de uma semântica baseada em ferramentas com maior usabilidade humana. Trata-se de adicionar anotações semânticas aos documentos da Web, a fim de aceder ao conhecimento de forma automática. Há uma expectativa de que as técnicas de mineração da web permitam que a máquina "aprenda" através de definições e crie uma organização estrutural do conhecimento com base, por exemplo, em ontologias. A aplicação de técnicas de mineração combinadas a abordagens de conteúdo, estrutura e uso permitiria a criação automática de uma semântica.

Na Web, como em outros lugares, o conhecimento é socialmente construído e resulta de um comportamento coletivo, ou seja, a navegação não é motivada apenas pelas relações formais de uma lógica subjacente à Web. A navegação em informações é um tipo de comportamento que pode ser aproveitado comercialmente, reflete o gosto, o desejo, o interesse dos usuários. Um sistema de recomendação de mercadoria baseado em filtragem colaborativa permite identificar o gosto pessoal de um eventual cliente. Na *Amazon*, a mensagem "as pessoas que gostam deste livro também compram este livro" revela uma afinidade entre produtos diferentes a partir da sua co-ocorrência nas buscas

[17] Representações em redes semânticas usadas para modelar objetos, propriedades e relações em um domínio de conhecimento.

de usuários, servindo, portanto, para uma aproximação de venda.

No contexto da mineração de uso da Web, que abarca a recuperação e análise do comportamento dos usuários, alguns pesquisadores associam o uso de informações semânticas a informações de uso. Stumme, Berendt & Hotho (2002), por exemplo, combinam pesquisas em web semântica e webmining, analisando convergências. Outros estudiosos tentam melhorar os resultados da mineração "minerando" novas estruturas semânticas na Web e usam para isso as próprias técnicas de webmining (Mobasher & Daí 2005) e (Koutri, Avouris & Hotho 2004).

6. O recurso a bases de dados geográficos para definir potenciais destinos turísticos

Como já assinalado, na maioria dos países da América Latina, bases de dados como as do IBGE e da Cepal, que disponibilizam informação para o desenvolvimento de políticas públicas, não prevêem correlações relevantes para o campo do Turismo. Como usar a informação disponível em tais bases de modo a favorecer a atividade turística no Brasil? No caso de destinos turísticos já reconhecidos como tais, a manipulação das informações das bases de dados poderia contribuir para o planejamento de um turismo sustentável[18]. No caso de destinos potenciais, que é o objeto especifico do presente estudo, o recurso ao cruzamento das informações disponíveis nas bases de dados será capaz de revelá-los e, com auxílio das ferramentas apresentadas, poderá fazê-lo de modo automático.

A possibilidade de extração de informação, a partir de tais bases pode ser informalmente avaliada através de alguns caminhos de busca[19]. Relatamos a seguir um exemplo desses caminhos, tomando o município de Santo André, na região do ABC paulista, como referência. Não serão apresentados resultados, já que nosso objetivo aqui é apenas descrever as ferramentas tecnológicas disponíveis, deixando a testagem para um próxima etapa da pesquisa. A consulta às bases de dados sobre o município de Santo André revelou que, no IBGE, é possível fazer uma busca inserindo o nome do

[18] O planejamento turístico é o processo que tem como "finalidade ordenar ações humanas sobre uma localidade turística" (Ruschmann & Widner, 2001:66). Segundo os autores, um turismo equilibrado poderia, ao mesmo tempo, manter a atratividade dos destinos turísticos naturais e culturais e preservá-los de degradação e danos ambientais.

[19] A escolha das duas bases, do IBGE e da Cepal, sugere a intenção de uma complementaridade entre a visão "macro" da Cepal sobre dados da estrutura econômica, política, territorial e social da América Latina, e a visão "micro" do IBGE com ênfase em dados locais e mais específicos sobre o espaço brasileiro. Não foi possível, no estágio atual da pesquisa, encontrar elementos que sustentassem essa idéia. A extração de informação com auxílio de ferramentas computacionais poderá revelar se há ou não uma concentração de nós "macro" na base da Cepal e de nós "micro" na do IBGE com relação ao domínio turístico.

município (cf.Fig.5). O sistema devolve informações sobre histórico, estimativa da população, área da unidade territorial. São oferecidas informações estatísticas sobre finanças públicas, serviços de saúde, posição no mapa de pobreza e desigualdade dos municípios, produto interno bruto, entre outras. A base do IBGE contém também estudos sobre turismo, como, por exemplo, o documento "Economia do turismo: uma perspectiva macroeconômica 2000 a 2005".

A aplicação de uma lógica identitária à ontologia "Santo André" baseia-se no Dizer (cf.seção 1), isto é, em anotações semânticas como: "localização" (região metropolitana da cidade de São Paulo, SP); "contexto regional" (região do Grande ABC paulista); dados demográficos (população estimada pelo senso de 2007 em 667.891 habitantes); "área territorial" (175 km²); "peculiaridades locais" (sede da Universidade Federal do ABC (UFABC) criada em 2005).

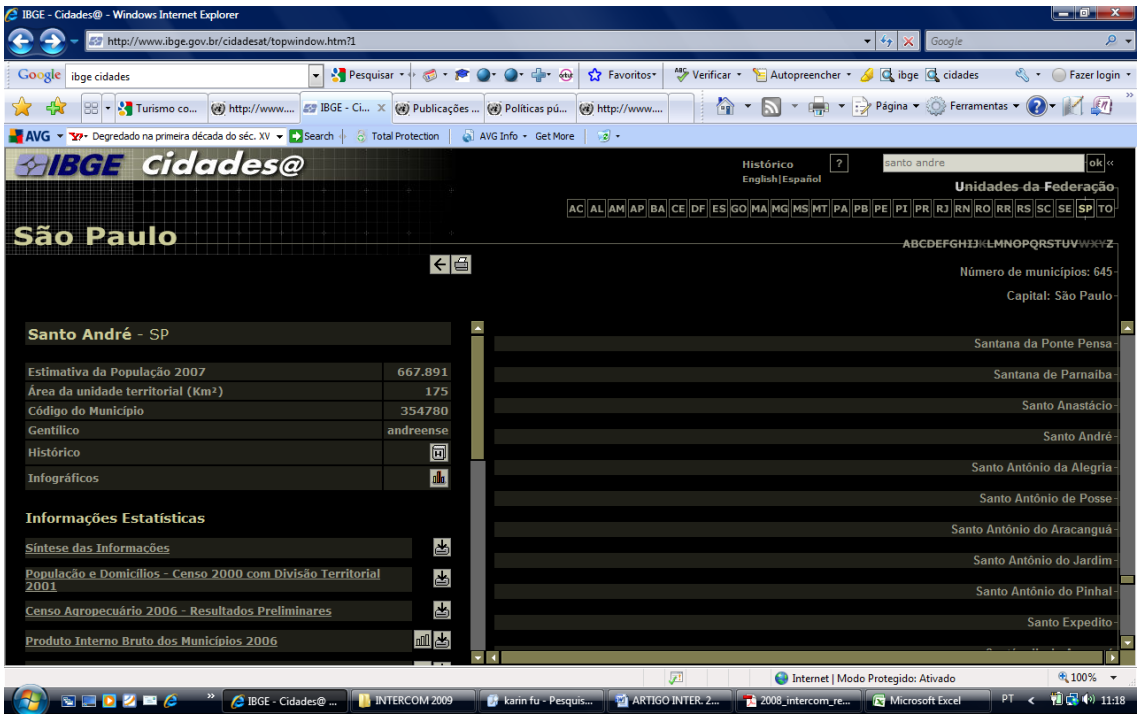

Figura 5. Página do IBGE com dados sobre o município de Santo André, SP.

Na base de dados da Cepal, localizamos o dossiê "O Sistema Municipal e o combate à pobreza no Brasil", 56 páginas, datado de 2004 e assinado pelas consultoras do órgão Beatriz Azevedo e Tereza Lobo. O dossiê trata das políticas municipais de combate à pobreza no Brasil. O município de Santo Andre não é diretamente citado, mas sua designação como "município" estabelece uma correlação automática com o documento, que toma o conjunto de municípios brasileiros como referência. Por exemplo, outros índices de correlação possíveis de "Santo André" ao documento podem

ser estabelecidos em "políticas urbanas", "bolsões de miséria", "capacidade de intervenção", "área social", "parcerias", "organizações". A cada índice serão atribuídos valores em uma escala. Por exemplo, numa escala de 1 a 5, um município que tenha valor 5 em "bolsões de miséria" não tem muita probabilidade de ser definido como destino turístico potencial[20].

"Santo André" pode ser tratado como um *frame* (Steinberger & Okuyama, 2005), isto é, como uma grande rede semântica implícita a cujos nós sejam atribuídas probabilidades de maior ou menor centralidade. Assim, a probabilidade de que o predicado "destino turístico" se aplique a "Santo André" e que "turismo" seja um nó central na rede pode ser mais alta ou mais baixa, dependendo dos demais nós associados à rede e aos valores a eles correlacionados. A organização de um mapeamento semântico dos municípios de uma região poderá indicar o grau de probabilidade de cada um deles ser um destino turístico e permite criar um ranking.

Para um tratamento automático com menor granularidade semântica podem-se adotar parâmetros gerais associando feixes de nós. Em pesquisas anteriores (Steinberger & Okuyama, 2005; 2008), propusemos, com base na análise de cadernos de jornalismo turístico, algumas categorias para classificar destinos turísticos na América Latina. Tais categorias inspiram agora a proposta de quatro parâmetros para avaliação de lugares como potenciais destinos turísticos: Histórico/Cultural, Geográfico, Econômico e Político. O parâmetro Histórico/Cultural correlaciona dados sobre a evolução histórico/ cultural da região, a identidade local, atrativos e interesses que a diferenciam. O parâmetro Geográfico correlaciona dados sobre clima, relevo, hidrografia, vegetação, infra-estrutura, população etc. Numa rede semântica, a aplicação com valor positivo ou negativo a predicados geográficos como "praia", "montanha", "lençóis de águas sulfurosas" "areia monazítica" atribuídos a sítios locais pode estar associada pela lógica identitária do Fazer (cf.seção 1) com os nós "surfe", "alpinismo", "natureza", "curas e tratamentos", etc. O parâmetro Econômico correlaciona dados sobre infra-estrutura, transportes, serviços, comércio, etc. E o parâmetro Político correlaciona dados sobre políticas, incentivos, fomento, isenção fiscal, etc. Com ferramentas de indexação, podemos converter documentos em dados semânticos estruturados e correlacionar suas informações a parâmetros de avaliação para identificar potenciais destinos turísticos.

[20] A correlação entre turismo e pobreza só é produtiva numa modalidade de "turismo social" em que os turistas são convidados a visitar favelas e são ciceroneados através de becos e ruelas, pagando para conhecer de perto a "cor local".

Referências bibliográficas

COOLEY, R, MOBASHER B & SRIVASTAVA J.(1997) "Web mining: information and pattern discovery on the World Wide Web". Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence.

KOSALA R. & BLOCKEEL, H. (2000) "Web Mining Research: A Survey In: Knowledge Discovery and Data Mining community SIGKDD Explorations Vol.2, Issue 1, NY, USA

GOLDSCHMIDT.R. & PASSOS E.(2005). **Data Mining, um guia prático**. Editora Campus,Rio de Janeiro:Elsevier.

KOUTRI, M.,AVOURIS, N., DASKALAKI, S.(2004) "A survey on web usage mining techniques for web-based adaptive hypermedia systems". In: Adaptable and Adaptative Hypermedia Systems. Idea Publishing Inc. Hershey.

MARKOV, Z. & LAROSE, D. T.(2007) "Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage". Wiley-Interscience.

MOBASCHER, B. & DAÍ, H.(2005) "Integrating Semantic Knowledge with Web Usage Mining for Personalization". In Web Mining: Applications and Techniques. Anthony Scime (ed.),Idea Group Publishing.

OKUYAMA, Thiery (2004) "O Imaginário jornalístico brasileiro sobre a América Latina nos cadernos de turismo". Monografia defendida na PUC-SP sob orientação da Profa.Dra.Margarethe Born Steinberger-Elias, Programa de Pós-Graduação em Comunicação Jornalística.

REATEGUI, B Eliseu & CAZELLA C. Silvio. (2005) "Sistemas de Recomendação"In: XXV Congresso da Sociedade Brasileira de Computação, ENIA. São Leopoldo RS, pag.306-348

REZENDE S. (2005) **Sistemas Inteligentes, Fundamentos e Aplicações**. Barueri, SP:Ed. Manole.

RUSCHMANN, V. M. Doris & WIDNER M. Gloria (2001) in ANSARAH, Marilia Gomes dos Reis. **Turismo: como aprender, como ensinar Vol 2.** São Paulo: Ed. Senac, páginas 65- 86.

STAAB, S. & WERTHNER H. (2002) "Intelligent Systems for Tourism" In: Trends e Controversies, IEEE Intelligent Systems

STEINBERGER, Margarethe Born (2005) **Discursos geopolíticos da mídia: jornalismo e imaginário internacional na América Latina.** São Paulo: Fapesp, Educ e Cortez.

_____ (2004) "Jornalismo e imaginário internacional sobre o Mercosul". Revista Estudos de Jornalismo e Mídia, Vol II Nr. 2 , 2º semestre de 2005, Florianópolis, Universidade Federal de Santa Catarina.(UFSC)

STEINBERGER, M. & OKUYAMA, T. (2005) " O Imaginário jornalístico brasileiro sobre a América Latina nos cadernos de turismo" Trabalho apresentado ao NP de Comunicação, Turismo e Hospitalidade, XXVIII Congresso Brasileiro de Ciências da Comunicação (Intercom), UERJ, Rio de Janeiro.

STEINBERGER, M. & OKUYAMA, T. (2008) "Estudo sobre condições de criação de um sistema informacional no campo turístico". Trabalho apresentado na NP Comunicação Turismo e Hospitalidade, XXXI Congresso Brasileiro de Ciências da Comunicação, UFRN, Natal.

STUMME, G., BERENDT, B., HOTHO, A.(2002): "Usage Mining for and on the Semantic Web". Next Generation Data Mining. Proc. NSF Workshop, Baltimore, pp. 77-86.

Site IBGE- http://www.ibge.gov.br/home/default.php, perfil dos municípios base2008[1].zip - Arquivo ZIP, tamanho descomprimido 26.443.776 bytes; Acesso 12-12-08 e 10-07-09

Site CEPAL http://www.eclac.org/ Acesso 10-07-09