

INTEGRAÇÃO DE DADOS NA RECUPERAÇÃO DE INFORMAÇÃO NA WEB E EM REDES SOCIAIS

Modesto, L. R.¹

RESUMO

O presente trabalho tem como objetivo fazer um levantamento bibliográfico dos métodos e técnicas capazes de realizar a recuperação de informação na WEB e em Redes Sociais. A integração de dados possibilita ao usuário o acesso a várias fontes de dados heterogêneos por meio de um esquema global. Esta integração leva em consideração dois níveis: integração de esquemas (consiste em consolidar todos os esquemas locais em um esquema global que será usado como base para a formulação das consultas) e integração a nível de instâncias ou objetos que são popularmente chamados de dados (refere-se à população global do sistema, seja virtual ou de forma materializada). O objetivo de um sistema de recuperação de informação é encontrar documentos relevantes, portanto, a relevância pode ser considerada como um conceito central para a recuperação de informação. Este trabalho mostra que as atividades de integração de dados e recuperação de informação são complementares e que para o usuário obter resultados relevantes de suas buscas é necessário que essas duas atividades se complementem. São abordados assuntos relacionados à integração de dados, arquitetura para integração de dados e as abordagens para integração de dados. Além disso são apresentados os conceitos de recuperação de dados e recuperação de dados na WEB. E por fim são abordados os conceitos relativos a mecanismos de busca na WEB: buscadores e metabuscadores e a conclusão deste material.

Palavras-Chave: Integração de Dados, Recuperação de Informação Web e Redes Sociais, Buscadores e Metabuscadores.

ABSTRACT

The present work aims to conduct a bibliographic survey of methods and techniques capable of performing information retrieval on the web and in social networks. Data integration enables users to access various sources of heterogeneous data through a global schema. This integration takes into account two levels: schema integration (involving consolidating all local schemas into a global schema that will be used as a basis for query formulation) and instance-level integration or objects commonly referred to as data (referring to the overall population of the system, whether virtual or materialized). The objective of an information retrieval system is to find relevant documents, thus relevance can be considered a central concept for information retrieval. This work demonstrates that data integration and information retrieval activities are complementary and that, for users to obtain relevant results from their searches, these two activities need to complement each other. Topics related to data integration, data integration architecture, and approaches to data integration are addressed. Additionally, concepts of data retrieval and web data retrieval are presented. Finally, concepts related to web search mechanisms are discussed: search engines and metasearch engines, concluding this material.

Keywords: Data Integration, Web Information Retrieval and Social Networks, Search Engines and Metasearch Engines.

¹ Dr. Lisandro Rogério Modesto. Coordenador da Pesquisa. Docente do Curso de Bacharelado em Ciência da Computação da Universidade Estadual do Paraná – UNESPAR. Apucarana- Pr. 2023. Contato: lrmomento@gmail.com

INTRODUÇÃO

A integração de dados se tornou um requisito essencial para a sociedade da informação desde os primeiros anos de desenvolvimento da ciência da computação. A idéia do compartilhamento de informações entre aplicações diferentes foi perseguida inicialmente buscando integrar dados distribuídos em sistemas de bancos de dados centralizados, proprietários e incompatíveis (Busse, Leser e Weber, 1999).

A integração de dados possibilita ao usuário o acesso a várias fontes de dados heterogêneos por meio de um esquema global. Esta integração leva em consideração dois níveis: integração de esquemas (consiste em consolidar todos os esquemas locais em um esquema global que será usado como base para a formulação das consultas) e integração a nível de instâncias ou objetos que são popularmente chamados de dados (refere-se à população global do sistema, seja virtual ou de forma materializada).

Quando se fala em um Sistema de Integração de dados deve se levar em consideração seu principal objetivo que é oferecer ao usuário uma interface de acesso à diferentes fontes de dados, de forma que os mesmos definam as suas consultas especificando o que se deseja encontrar e o sistema determine onde o resultado pode ser encontrado, apresentando assim as respostas para as solicitações do usuário (Salgado e Lóscio, 2001).

Os Sistemas de Integração de dados trabalham com fontes de dados que podem ser autônomas e dinâmicas, ou seja, as fontes continuam a suportar aplicações locais podendo atualizar tanto os dados quanto o esquema que disponibilizam para o Sistema de Integração de Dados. Como consequência dessa natureza dinâmica das fontes de dados é necessário que o sistema de integração seja extensível, ou seja, o sistema deve ser capaz de se adaptar a visão oferecida aos usuários, bem como gerenciar os novos tipos de informação oferecidos pelas fontes de dados (Salgado e Lóscio, 2001).

Um sistema de integração de dados também precisa lidar com a heterogeneidade das fontes de dados, as quais, na maioria das vezes, foram projetadas independentemente utilizando diferentes modelos de dados e diferentes representações para os mesmos conceitos do mundo real. Desta forma, um sistema de integração de dados deve ser capaz de sobrepor a heterogeneidade de forma a oferecer uma visão integrada dos dados. É importante notar que um sistema de

integração não precisa oferecer necessariamente uma única visão integrada, ou seja, o sistema de integração de dados pode oferecer diferentes visões integradas que são definidas de acordo com as necessidades dos usuários (Salgado e Lóscio, 2001).

Para que o sistema de integração de dados permita que os dados integrados mantenham o mesmo significado que tinham em suas fontes originais, alguns problemas precisam ser resolvidos, entre eles, formato de datas, unidade de medida, fator de escala e representação de notas (Amaral e Salgado, 2005).

ARQUITETURAS PARA INTEGRAÇÃO DE DADOS

A integração de dados ainda possui alguns problemas a serem resolvidos, entre eles podemos citar a arquitetura dos dados. Para a resolução do problema de arquitetura dos dados já existem algumas propostas. Entre elas destacam-se: Arquitetura Clássica para Integração de Dados, Arquitetura para Integração de Dados na Web, Arquitetura de mediadores e Arquitetura de Data Warehouse. Sendo que as que merecem maior destaque atualmente por sua aplicabilidade são as Arquiteturas de Mediadores e Data Warehouse.

Na Arquitetura Clássica para Integração de Dados as soluções são baseadas na construção de um esquema global a partir da integração dos esquemas das fontes de dados locais, de forma que as diversas fontes de dados distribuídas e heterogêneas podem ser acessadas de maneira uniforme e transparente através deste esquema global (Salgado e Lóscio, 2001).

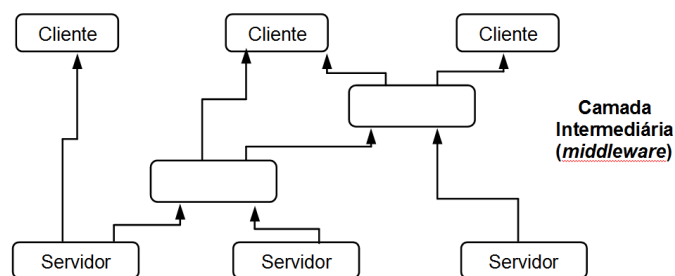


Figura 1 - Arquitetura Multi-camada (Abiteboul e Buneman, 2000).

Arquiteturas para Integração de Dados na Web é uma forma alternativa aos tradicionais sistemas de gerenciamento de banco de dados, sendo que o contexto da Web é baseado em uma abordagem “multitier” ou multi-camadas, onde a camada de mais baixo nível consiste nas fontes de dados, também chamadas de servidores, que podem ser servidores de bancos de dados convencionais, sistemas legados,

servidores de arquivos ou qualquer aplicação que produz dados, conforme mostrado na figura 1.

A camada mais alta é a camada do cliente e consiste em interfaces de usuários ou aplicações que consomem dados. Entre estas duas camadas existe uma coleção de camadas intermediárias que oferecem o “middleware” - software para transformar e integrar dados (Salgado e Lóscio, 2001).

Na arquitetura de mediadores (figura 2) a consulta é submetida ao mediador que a decompõe em sub-consultas e as submete às fontes de dados, enquanto na arquitetura de data warehouse as consultas são feitas diretamente no warehouse não havendo necessidade de acessar diretamente as fontes de dados no momento da consulta, porém manter os dados do warehouse consistentes com os dados das fontes de dados é um problema nesta arquitetura (Trindade e Netto, 2002).

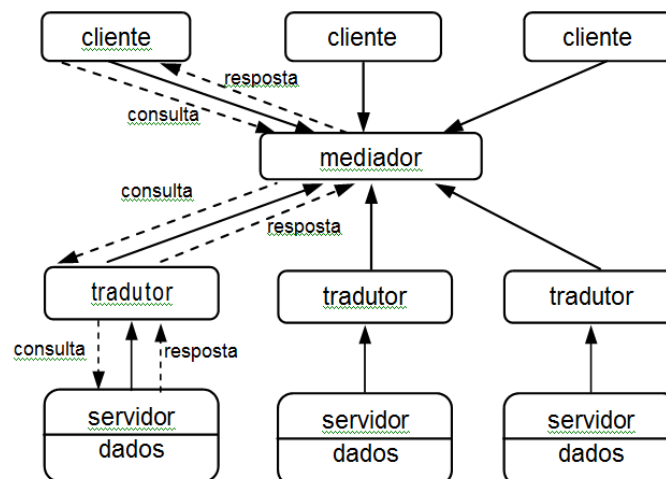


Figura 2 - Arquitetura de Mediadores (Abiteboul e Buneman, 2000).

Inicialmente a arquitetura de mediadores foi criada para integração de dados distribuídos em múltiplos bancos de dados, porém atualmente seu uso está sendo feito também em sistemas de integração de dados na Web.

ABORDAGENS PARA INTEGRAÇÃO DE DADOS

Sistemas de integração de dados servem como ferramentas para oferecer acesso uniforme às fontes de dados distribuídas e heterogêneas no ambiente Web. Isso pode ser feito através da resolução das heterogeneidades nos dados e fornecendo uma visão integrada das fontes de dados, onde também são refletidos requisitos de usuários. Os usuários do sistema de integração submetem consultas

sobre a visão integrada sem precisar perder tempo com longas buscas em páginas Web isoladas (Abiteboul e Buneman, 2000).

Os sistemas de integração de dados podem ser divididos em duas abordagens: virtual e materializada. O que difere essas duas abordagens é a forma com que os dados e suas fontes são tratados.

A abordagem virtual tem as informações são extraídas das fontes de dados no momento de uma requisição, ou seja, a consulta é submetida ao mediador que a decompõe em sub-consultas e as submete às fontes de dados. Esta abordagem mantém os dados sempre atualizados, mas a recuperação destes dados pode ser lenta.

Na abordagem materializada as informações são extraídas das fontes de dados, integradas e armazenadas em um repositório responsável por atender às requisições, ou seja, não há a necessidade de acessar diretamente as fontes de dados no momento da consulta. Nesta abordagem os dados podem estar desatualizados ou inconsistentes, porém a recuperação dos dados é rápida.

Abordagem virtual	Abordagem materializada
Passo 1: o sistema de integração recebe uma consulta, determina o conjunto de fontes de dados necessárias para respondê-la gera as subconsultas apropriadas para cada fonte de dados.	Passo 1: as informações relevantes de cada fonte de dados são extraídas, traduzidas e filtradas, em seguida estas informações são integradas com as informações relevantes de outras fontes de dados e armazenadas em um repositório central.
Passo 2: o sistema coleta os resultados das sub-consultas de cada fonte de dados, executa as traduções apropriadas, filtra e integra as informações, e em seguida, retorna o resultado final para o usuário.	Passo 2: quando uma consulta é requisitada ao sistema de integração, a consulta é avaliada diretamente no repositório, sem a necessidade de acessar as fontes de dados originais.

Tabela 1 - Processo de extração de informações das abordagens virtual e materializada (Salgado e Lóscio, 2001).

Sendo assim, o uso adequado de uma dessas abordagens depende o que se espera rapidez ou garantia de dados atualizados. Para exemplificar melhor a ajudar

na definição de qual abordagem utilizar, abaixo é apresentada uma tabela que aborda o processo de extração de informação em cada abordagem.

Em aplicações mais complexas e em grande escala, as duas abordagens são necessárias, pois para manipular tais aplicações, um sistema de integração de informação ideal é aquele no qual algumas informações são recuperadas, processadas e integradas previamente e, em seguida, armazenadas em um repositório do sistema, enquanto que outras informações são recuperadas e processadas somente quando consultas são requisitadas (Salgado e Lóscio, 2001).

Em resumo, o ambiente de integração de dados processa três tipos de consultas: consultas virtuais a dados que serão obtidos sob demanda, ou seja, acessados diretamente das fontes de dados; consultas a dados materializados que serão obtidos a partir de um data warehouse e consultas cujos resultados previamente armazenados serão recuperados em uma cache local ao sistema de integração (Batista e Salgado, 2003).

RECUPERAÇÃO DE INFORMAÇÃO NA WEB E EM REDES SOCIAIS

O objetivo de um sistema de recuperação de informação é encontrar documentos relevantes, portanto, a relevância pode ser considerada como um conceito central para a recuperação de informação. Não obstante à sua importância e as enormes quantidades de pesquisas sobre este tema no passado, a relevância ainda não é um conceito bem entendido por causa de uma terminologia inconsistente usada.

Quando se fala em busca na Web, os motores de busca oferecem a possibilidade de o usuário indicar um documento no conjunto de resultados como sendo exemplar do ponto de vista de satisfação de suas necessidades informacionais.

A Web é uma vasta coleção de documentos heterogêneos. Possui natureza dinâmica e milhões de páginas surgem e desaparecem todos os dias. Por isso sente-se um anseio para que a Web realmente alcance todo o seu potencial e se torne uma ferramenta mais utilizável, eficaz e compreensível (Marinho e Girardi, 2003).

Devido ao crescimento exponencial do volume de dados disponível na Web e nas redes sociais, a recuperação e filtragem de documentos e informações relevantes torna-se uma tarefa extremamente árdua. Neste contexto, surgem as máquinas de busca e recuperação de informação. Os principais problemas associados à filtragem

de informações na Web decorrem do fato de estas tarefas terem sido, inicialmente, concebidas para serem realizadas sob o controle direto de pessoas. A Web coloca à disposição dos usuários uma grande quantidade de documentos e informações, no entanto, de modo geral, estes documentos contêm texto não estruturado ou semi-estruturado e cabe às pessoas filtrá-los usando máquinas de busca, que normalmente são baseadas em palavras-chave (Trucolo e Coello, 2008).

O processo de recuperação de informação na Web, geralmente, é feita por meio de sites de busca, metabuscadores, navegação ou em diretórios Web. Cada uma dessas formas possui particularidades que determinam seu uso. Nas consultas (querying) feitas em sites de busca, o usuário estrutura uma solicitação da informação desejada e recebe como resposta um conjunto de páginas, sendo que esse conjunto de páginas é apresentado de forma indexada de acordo com a regra definida pelo buscador. Os metabuscadores são sistemas que realizam buscas em outros sistemas de busca em paralelo, apresentando os resultados. Estes sistemas não mantêm banco de dados próprio. No caso da navegação (browsing) o usuário visita várias páginas acessando seus hyperlinks e documentos referenciados, esse processo pode ser complementar às consultas, uma vez que após o buscador apresentar os resultados para a pesquisa é necessário acessar as páginas apresentadas, porém, a navegação pode acontecer diretamente informando o endereço (URL) da página Web.

É possível acessar também os diretórios web, são listas de assuntos organizadas em categorias, geralmente com uma estrutura hierárquica (árvore), esses diretórios apresentam informações e documentos disponíveis para download, esses diretórios são acessados via endereço (URL) e um exemplo são os serviços de transferência de arquivo FTP (File Transfer Protocol).

Encontrar a informação depende, principalmente, do uso eficiente das ferramentas de busca. Para explorar todo o potencial dos buscadores, o usuário precisa conhecer: como é coletada e estruturada a informação em diferentes bancos de dados; suas características e limitações; todas as possíveis formas de interação e suas linguagens de busca. Conceitua-se estrutura da informação como a sua organização lógica para posterior recuperação e linguagem de busca como os comandos que permitem a recuperação da informação através de palavras contidas nos títulos, resumos ou outros campos de dados (Branski, 2003).

A seção abaixo apresenta os mecanismos de busca na web, sendo dividida

em buscadores e metabuscadores.

MECANISMOS DE BUSCA NA WEB

Mecanismos de busca, sites de busca ou buscadores são softwares específicos para a recuperação de informações na Web, os mesmos recuperam e armazenam informações sobre os conteúdos dos sites em bancos de dados que ficam a disposição do usuário para pesquisa. Essas informações armazenadas estão disponíveis em banco de dados, que são estruturas flexíveis e fáceis de utilizar e manter. Os mecanismos de busca são divididos em buscadores e metabuscadores.

Ao utilizar os mecanismos de busca, o usuário descobre a localização do material encontrado podendo acessá-lo, tanto a busca quanto o acesso a esses materiais encontrados pode ser feito por meio de Navegadores Web (Browser). Um navegador é um programa de computador que habilita seus usuários a interagirem com documentos virtuais e páginas da Internet, entre outros navegadores podemos citar Microsoft Edge, Mozilla Firefox e Google Chrome, conforme pode ser verificado na figura 3.

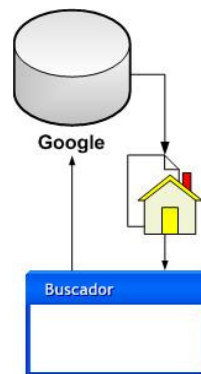


Figura 3 - Estrutura de recuperação de informação de um buscador.

Por meio dos navegadores o usuário acessa a página de um buscador digita algum termo, geralmente palavras ou frases no campo de busca, solicita a pesquisa e os spiders fazem a varredura. Os buscadores procuram o termo de busca em seus bancos de dados, sendo assim toda busca realizada é feita no banco de dados dos buscadores e não diretamente na web.

As ferramentas de busca oferecem como resposta ao usuário, páginas onde estão relacionados todos os sites armazenados em seu banco de dados onde foram verificadas a ocorrência da linguagem de busca. Os resultados são apresentados na forma de links de hipertextos, isto é, clicando com o mouse sobre uma das frases

realçadas (links), o próprio site, que está fora do banco de dados do buscador, é trazido para o computador do usuário. As páginas que os mecanismos de busca oferecem como resposta não são armazenadas no servidor. Elas simplesmente desaparecem após a consulta. Estas páginas são transitórias porque é mais barato e fácil reconstruí-las novamente do que armazená-las com todas as possíveis opções existentes de consulta (Branski, 2003).

O mecanismo de busca Google inovou no critério utilizado para apresentação dos documentos recuperados, alcançando resultados bastante satisfatórios. Este buscador define seus resultados de acordo com o número de links apontando para cada um dos documentos armazenados em sua base de dados. Isto é, na relação dos documentos recuperados pelo buscador, ocuparão os primeiros lugares os sites que tiverem sido mais citados por outros sites. Esta forma de estruturar a informação tem como premissa a idéia de que os sites mais populares oferecem informações de melhor qualidade.

METABUSCADORES

Metabuscadores são sistemas de pesquisa WEB que permitem aos usuários realizar consultas simultâneas em várias fontes de informações, como mecanismos de busca, bancos de dados, sites de comércio eletrônico, redes sociais e muito mais (Branski, 2003).

Nos buscadores tradicionais submete-se os termos de busca a um único banco de dados e recebe-se uma relação dos documentos onde constam os termos pesquisados. Utilizando um metabuscador o usuário está buscando, simultaneamente, em vários buscadores, conforme pode ser visto na figura 4.

Os metabuscadores não possuem banco de dados próprio e funcionam como um agente intermediário que repassa a pesquisa, obtém as respostas dos buscadores individualmente e, então, apresenta um resultado unificado, extraído das diversas fontes. Em poucos segundos os metabuscadores compilam e apresentam os resultados obtidos em diversos mecanismos de busca. Embora o seu uso possa significar economia de tempo, já que a pesquisa é feita em um único site, a qualidade dos resultados varia muito de acordo com a ferramenta escolhida (Branski, 2003).

Para recuperação de informação e conteúdos na Web, são utilizados programas específicos conhecidos como spiders que trabalham como robôs, e são

r

alcançar a Internet e recuperar dados pode ser chamado bot. Um spider é um bot especializado e construído para vasculhar sites baseados em conteúdo encontrados em sites já conhecidos. Um spider trabalha iniciando em uma página web simples ou, algumas vezes, em várias páginas. Essa página web é varrida na busca por referências a outras. O spider então visita tais páginas e repete esse processo, continuando indefinidamente. Essa aplicação constitui a estratégia inicial dos mecanismos de busca na web. Em função do dinamismo da Internet, essas aplicações precisam ser automatizadas, ou seja, precisam desempenhar seu papel sem haver a freqüente intervenção humana.

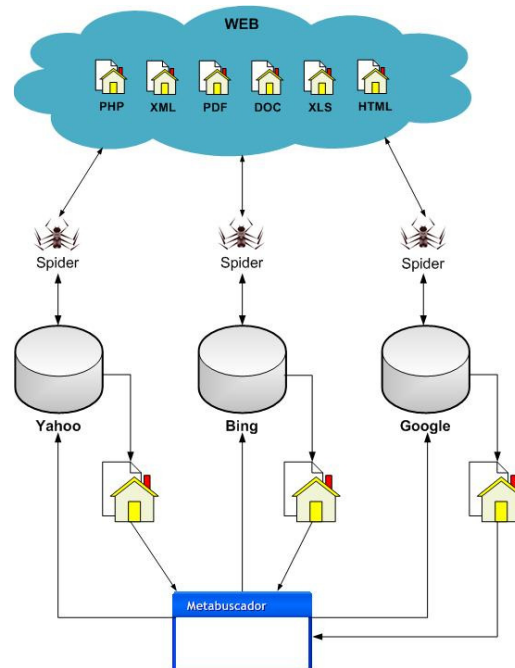


Figura 4 - Estrutura de recuperação de informação de um metabuscador.

INTEGRAÇÃO DE DADOS NA RECUPERAÇÃO DE INFORMAÇÃO NA WEB E NAS REDES SOCIAIS

Com a aplicação do conceito de integração de dados, o usuário pode acessar várias fontes de dados heterogêneas por meio de esquemas globais. Essa integração tornou-se imprescindível para a atual sociedade da informação uma vez que o compartilhamento de conteúdos e informações entre aplicações diferentes é de extrema importância possibilitando um acesso pleno e eficaz. O real objetivo de um sistema de integração de dados é oferecer ao usuário uma interface de acesso a

diferentes fontes de dados, bastando para isso, definir suas consultas especificando o que se deseja encontrar.

Um sistema de recuperação de informação tem como objetivo central encontrar conteúdos e informações relevantes para o usuário, esse processo pode ser dividido em duas partes: a primeira realizada pelo usuário e a segunda realizada pelo sistema de recuperação, ou seja, existe um processo híbrido entre humanos e máquinas. O sucesso da recuperação de informação depende da qualidade e exatidão com que um termo é descrito em um motor de busca (buscadores). Diante disso é correto afirmar que um sistema de informação terá sucesso dependendo do modelo que o mesmo adota, podendo ser usados modelos quantitativos e modelos dinâmicos.

É fácil perceber que o sucesso do processo de busca da Web depende diretamente de dois conceitos: integração de dados e recuperação de informação. Quando se executa uma busca utilizando um buscador, este processo a abordagem materializada uma vez que o resultado apresentado está na base de dados do buscador em questão. Porém a busca é executada em metabuscador a abordagem utilizada é a materializada, pois o resultado apresentado é referente à informações armazenadas em várias bases de dados de vários buscadores.

O grande desafio é fazer com os servidores e páginas Web estejam com as informações integradas de forma efetiva para que os sistemas de recuperação de informação possam executar um trabalho eficaz retornando conteúdos e informações relevantes para o usuário.

Com esse trabalho foi possível verificar que as atividades e integração de dados e recuperação de informação são complementares e que para o usuário obter resultados relevantes de suas buscas é necessário que essas duas atividades se complementem e sejam cada vez mais automatizadas, pois a web cresce a cada dia e cada página web pode ser considerada uma base de dados e além dos servidores as mesmas também precisam estar integradas para que possam ser localizadas.

METODOLOGIA

A metodologia adotada para este trabalho foi revisão bibliográfica, consistindo em um estudo das formas de recuperação de informação na WEB e em redes sociais, para tanto foi necessária uma análise exploratória e aprofundada em páginas WEB e em plataformas de redes sociais a fim de verificar o formato com que os dados e as

informações estão disponibilizados para desenvolver métodos e técnicas capazes de realizar a recuperação de informação em tais plataformas. Este trabalho foi desenvolvido tendo como base um recorte de algumas plataformas de redes sociais, como Facebook, Instagram e LinkedIn; além disso páginas Web aleatórias de instituições, empresas, organizações não governamentais, entre outras.

Para o desenvolvimento deste trabalho de pesquisa foram estudados padrões de recuperação de dados e informação possibilitando a descrição das formas de pesquisa nas páginas WEB e nas plataformas de redes sociais transformando dados desestruturados em dados estruturados que possam ser utilizados de forma efetiva por quem busca obter respostas para questões que definem o perfil dos usuários, descrevendo os caminhos relacionados aos recursos digitais.

RESULTADOS E DISCUSSÃO

Como resultados e discussão do tema proposto, foram apresentados métodos e técnicas que possibilitam o usuário acessar várias fontes de dados heterogêneas compartilhando conteúdos entre aplicações diferentes e possibilitando um acesso pleno e eficaz às informações.

Para tanto, o primeiro passo foi a apresentação dos conceitos relativos à integração de dados, da recuperação de informação na web e em redes sociais e dos mecanismos de busca na web, para só então abordar a integração de dados na recuperação de informação na web e nas redes sociais.,

Sendo assim é possível notar a oferta, para o usuário, de uma interface de acesso a diferentes fontes de dados, bastando para isso, definir suas consultas especificando o que se deseja encontrar.

CONSIDERAÇÕES FINAIS

A melhoria nos serviços de disponibilização de informações por meio das páginas WEB e das plataformas de redes sociais é uma busca que se tornou uma necessidade social e comercial, e ajudar a promover esta melhoria é obrigação de todo cidadão principalmente dos pesquisadores ligados à área de Ciência da Computação e Ciência da Informação.

As redes sociais tornaram-se uma grande base de dados onde pessoas e

organizações disponibilizam e compartilham dados e informações que guardam características importantes que podem definir perfis passíveis de interação e manipulação que ajudam nas tomadas de decisões mais assertivas, o que justifica o desenvolvimento de um trabalho desta natureza, que busca meios para descrever de forma efetiva o caminho que leva às informações disponibilizadas nas páginas WEB e nas plataformas de redes sociais apresentando métodos e técnicas para a integração de dados na recuperação de informações.

Atualmente, a cada nova interação feita em uma página Web ou plataforma de rede social, novas decisões podem ser tomadas por parte das pessoas e organizações com base no perfil que pode ser definido e compartilhado a fim de propor melhorias nas interações sociais.

Com isso foi possível concluir que um sistema de integração de dados oferece ao usuário uma interface de acesso a diferentes fontes de dados, bastando para definir suas consultas especificando o que se deseja encontrar.

REFERÊNCIAS

- ABITEBOUL, S.; BUNEMAN, P. Data on the Web, 1st. Ed. Morgan Kaufmann Publishers, 2000.
- AMARAL, Ana Carolina do; SALGADO, Ana Carolina. Resolvendo alguns conflitos semânticos em Sistemas de Integração de Dados. Centro de Informática da Universidade Federal de Pernambuco. 2005.
- BATISTA, Maria da Conceição Moraes, SALGADO, Ana Carolina. Centro de Informática da Universidade Federal de Pernambuco. 2003.
- BRANSKI, Regina Meyer. Recuperação de informações na Web. 2004.
- BUSSE, S., KUTSCHE, R., LESER, U., WEBER, H. (1999), Federated Information Systems: Concepts, Terminology and Architectures. Technische Universität Berlin, April.
- HEATON, J. Programming Spiders, Bots, and Aggregators in Java. California - USA: Sybex, 2002. 2–3 p. (1).
- MARINHO, Leandro Balby; GIRARDI, Rosário. Mineração na Web. Departamento de Informática da Universidade Federal do Maranhão.
- SALGADO, Ana Carolina; LÓSCIO, Bernadette Farias. Integração de Dados na Web. Centro de Informática da Universidade Federal de Pernambuco. 2001

TRUCOLO, Felipe Fernandez; COELLO, Juan Manuel Adán. Recuperação de Informação na Web Semântica. 2008.