Sistemas de informação distribuídos: uma breve análise do contexto atual

Carlos Eduardo Benevides Bezerra
Universidade Federal do Rio Grande do Sul
Bento Gonçalves, 9500, Porto Alegre, RS, Brasil
E-mail: carlos.bezerra@inf.ufrgs.br

I. Introdução

Com o surgimento do computador – máquina que processa e armazena dados – em meados do século XX, tornou-se possível guardar uma grande quantidade de informações em espaços cada vez menores. Com o surgimento das redes de computadores (o que levou a sistemas distribuídos [TODO:ref couloris e lamport]) e, posteriormente, das interconexões destas redes com, por fim, o surgimento e popularização da Internet, houve uma revolução na maneira como as pessoas têm acesso a informações. A quantidade de dados que são transmitidos entre pontos distintos do globo, assim como a rapidez com que isso acontece, ajudaram a definir a atual era como Era da Informação.

Contudo, devido justamente à liberdade com que é criado conteúdo – e disponibilizado na Internet, por exemplo –, assim como ao crescente número de indivíduos, grupos e organizações que disseminam informações, surgem alguns desafios no que se refere ao tratamento e filtragem dessas informações. Por um lado, tem-se acesso a dados a respeito de praticamente qualquer coisa que se imagine. Devido à enorme quantidade desses dados, é necessário prover alguma ferramenta para buscar as informações onde elas estejam. Além disso, devido à liberdade com que se publica conteúdo na Internet, praticamente não há um padrão para a exibição da informação. Por último, mas não menos importante, após localizar, extrair e normalizar os dados, é mandatório classificar aquelas informações de acordo com algum padrão de qualidade, já que praticamente não há controle sobre o que se publica na Internet.

Este trabalho tem por objetivos: dar uma visão geral sobre as metodologias de coleta, tratamento e classificação das informações extraídas, geralmente, de sistemas distribuídos – como a Internet – e fazer uma análise de alguns problemas que merecem atenção dos pesquisadores.

II. Muitas informações, em muitos locais

Como foi dito, existe uma enorme quantidade de conteúdo disponível hoje em dia. Para se fazer o melhor uso possível desta grande base de dados, é necessário, primeiramente, localizar e extrair os dados distribuídos em diversos repositórios. Após a extração desses dados, é preciso normalizá-los, de maneira a torná-los adequados com o indivíduo que os está visualizando. Por exemplo, alguém poderia querer informações a respeito de uma cidade, com dados como temperatura e distâncias entre pontos turísticos sendo apresentados em unidades que lhe sejam conhecidas, ou que as informações lhe sejam apresentadas em uma linguagem compatível com sua compreensão e sem detalhamento excessivo. Por último, uma informação só deve ser apresentada, ou recomendada, a um usuário se for de qualidade, o que dependerá de critérios, que por sua vez são verificados através de métricas. Nas seções a seguir, serão dados alguns exemplos e brevemente explicadas essas etapas.

A. Localização

Para que as informações possam ser localizadas, podem ser utilizados alguns métodos, como indexação. Ao invés de consultar cada base de dados em busca de um casamento com a chave utilizada para busca, é mantida uma lista de tags que serão utilizadas para busca, formando um índice. Por exemplo, o Google Scholar [TODO: ref googlescholar] tem agentes autônomos – ou crawlers – que têm acesso permitido às enormes bases de dados das maiores e mais bem conhecidas editoras de material científico (como IEEE, ACM, Springer e outras). Esses agentes vasculham essas bases de dados e indexam seu conteúdo, baseado em informações relevantes, como nomes de autores, títulos dos trabalhos científicos, abstract etc., o que será enviado aos grandes servidores do Google para serem utilizados nas buscas feitas pelos usuários.

Outra maneira de buscar conteúdo é através de redes P2P, descentralizadas [TODO: ref P2P]. Quando um dos participantes deseja determinado arquivo, por exemplo, ele envia a requisição a seus pares, que lhe respondem ou encaminha a requisição a outros pares. No entanto, algumas dessas redes tendem a saturar rapidamente a banda dos pares, por basearem as buscas em inundação de mensagens. Existem alternativas comprovadamente mais eficientes, tanto na teoria quanto na prática, que utilizam DHTs (tabelas hash distribuídas), mas em que o pedido é feito com um identificador único (o hash do arquivo que está sendo procurado, por exemplo), e não em palavras-chave. Exemplos dessas redes baseadas em DHT são: Chord[TODO:ref], Pastry[TODO:ref] e Can[TODO:ref].

B. Extração

Uma vez que os dados são localizados, seja baseado em índices com tags ou em busca por identificador único, como um hash, eles devem ser extraídos e armazenados em um formato padrão, de maneira que conteúdos oriundos de

diferentes fontes possam ser agregados ou comparados. O RoadRunner [TODO: ref roadrunner], por exemplo, tem por objetivo extrair porções de informação de páginas em HTML, baseado em casamento de padrões nas suas marcações. Para isso, é definida uma linguagem regular, baseada num exemplo genérico. Casando o padrão dessa linguagem com o código de cada página, são encontradas as estruturas dos dados presentes naquela página.

Outras ferramentas para extração de informações são apresentadas em [TODO: ref brief survey]. Algumas delas são: STALKER [TODO: ref no brief], RAPIER [TODO: rnb] e Web-OQL [TODO:].

Um detalhe importante é que o XML [TODO:ref xml] vem como um facilitador dessa extração de dados. Sendo uma linguagem de marcação, permite que cada bloco de informação seja apresentado com atributos e sua relação com outros pedaços de informação seja representado por um grafo. Um arquivo em XML, ao mesmo tempo que apresenta uma hierarquia e possibilita uma representação rica dos dados e metadados, permite que essa mesma representação não siga uma estrutura rígida, o que é adequado à falta de uniformidade da apresentação dos dados nos diferentes repositórios.

C. Normalização

Após os dados serem extraídos de suas fontes, é desejável que sejam apresentados de acordo com determinada norma. Isso é mais claro de se entender quando se trata de unidades de medida, como distância e temperatura. Quais unidades de medida serão utilizadas depende, em última análise, de quem as está visualizando. Um brasileiro provavelmente prefira saber a distância entre duas cidades em quilômetros, enquanto um americano deverá querer a distância em milhas.

No entanto, a normalização também serviria para adequar o conteúdo apresentado, a linguagem e o seu detalhamento ao nível de compreensão e/ou interesse de quem o estivesse visualizando. As informações em uma bula de remédio seriam apresentadas de maneira completamente diferente para um médico e para um paciente com pouca escolaridade, por exemplo.

D. Qualidade

Por fim, é necessário classificar as informações de acordo com sua qualidade. Para determinar a qualidade de cada informação, podem ser utilizadas diversas métricas [TODO:ref arg], atribuindo um ou mais índices de qualidade para cada informação.

Como exemplo, pode ser citado o OrtoQualis [TODO:corrigir nome e por ref do palazzo], que utilizou um conjunto de critérios definidos pela CAPES para avaliar conferências, obtendo uma classificação semelhante àquela realizada pessoalmente pelos membros do comitê de Qualis da instituição.

III. Problemas

IV. Conclusão

References