Sistemas de informação distribuídos: uma breve análise do contexto atual

Carlos Eduardo Benevides Bezerra
CMP112 - Sistemas de Informação Distribuídos
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 [1], [2]) 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 de sistemas distribuídos – na Internet, geralmente – 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 ao 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 se 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 [3] 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, resumo 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 [4]. Quando um dos participantes deseja determinado arquivo, por exemplo, ele envia a requisição a seus pares, que lhe respondem ou encaminham 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 [5], Pastry [6] e CAN [7].

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 [8], 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 [9]. Algumas delas são: STALKER [10], XWRAP [11] e Web-OQL [12].

Um detalhe importante é que o XML [13] 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 em 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 preferirá saber 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 uma farmacêutica 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 [14], atribuindo um ou mais índices de qualidade para cada informação.

Como exemplo, pode ser citado o OntoQualis [15], 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 e desafios

Foi visto que se busca fazer a recuperação de informações relevantes em uma grande base de dados distribuída, com estes dados armazenados sem seguir um determinado padrão de formato ou estrutura, adequá-los a quem os está visualizando e classificá-los de acordo com determinados critérios de qualidade. Embora tenha sido feita vasta pesquisa nesta área, o que pode ser confirmado pela coletânea de referências bibliográficas apresentadas neste trabalho e em trabalhos citados aqui, ainda persistem algumas questões.

Uma das questões mais críticas se refere ao tempo que se leva para que se possa recuperar informações, contado a partir do momento em que são disponibilizadas. Varrer o conteúdo de repositórios – como *websites*, bancos de dados e bibliotecas digitais – cada vez que se faz uma busca, ainda que seja mantido um cache, é impraticável, devido ao tempo que se levaria para casar o dado ou padrão procurado com as informações armazenadas nesses repositórios. O que se faz, em geral, é deixar que agentes percorram as bases de dados e as indexem, para posterior uso em uma ferramenta de busca, ou apenas fazer uma busca com base em um identificador único, com tempo de resposta constante ou logarítmico. No caso dos agentes indexadores, podem-se levar dias para que uma determinada informação seja adicionada ao índice, dificultando um dos objetivos desejáveis da recuperação de informações, que é a atualidade das mesmas.

Outro desafio atual é quanto às metodologias para recuperar, normalizar e classificar informações. Para cada área de interesse, é criado todo um conjunto de critérios para realizar as buscas e ordenar os resultados de acordo com a qualidade de cada um. Porém, o ideal seria que isso fosse automatizado. Com uma única ferramenta, deveria ser possível encontrar de maneira automática esses critérios de busca e classificação, independente de qual área de interesse estivesse sendo pesquisada, ou do usuário. Obviamente, isto implicaria a implementação de algum tipo de agente inteligente que pudesse ser treinado para encontrar padrões em conjuntos de dados considerados bons, assim como determinar como esses dados devem ser apresentados a cada usuário diferente, provavelmente baseado em algum tipo de histórico.

IV. Conclusão

Neste trabalho, foi dada uma visão geral de algumas das etapas atualmente utilizadas para a busca e filtragem de informações disponíveis em sistemas de informação distribuídos, sendo algumas: localização, extração, normalização e classificação de acordo com qualidade dessas informações, sendo que foram citados trabalhos existentes e/ou exemplos de como cada uma delas funciona. Por fim, foram apresentadas algumas idéias do que deve ou pode estar por vir na área. Tendo em vista o que foi apresentado neste texto e em suas referências bibliográficas, é possível perceber o quanto a atual era se encaminha para um futuro onde as informações estarão disponíveis em qualquer lugar, de maneira transparente e – se forem resolvidas as questões referente à velocidade com que as informações são disponibilizadas em ferramentas de busca – tão cedo quanto elas forem publicadas.

References

[1] G. Bücher Couloris, J. Dollimore, and T. Kindberg, “Distributed Systems,” 2000.

[2] L. Lamport, "Time, clocks, and the ordering of events in a distributed system," 1978.
[3] P. Jacsó, "Google Scholar: the pros and the cons," *Online Information Review*, vol. 29, no. 2, p. 208, 2005.
[4] H. Balakrishnan, M. F. Kaashoek, D. Karger, R. Morris, and I. Stoica, "Looking up data in p2p systems," *Commun. ACM*, vol. 46, no. 2, pp. 43–48, 2003.
[5] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup service for internet applications," in *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*. ACM New York, NY, USA, 2001, pp. 149–160.
[6] A. Rowstron and P. Druschel, "Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems," in *IFIP/ACM International Conference on Distributed Systems Platforms (Middleware)*, vol. 11. Heidelberg, 2001, pp. 329–350.
[7] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Schenker, "A scalable content-addressable network," in *Proceedings of the 2001 SIGCOMM conference*, vol. 31, no. 4. ACM New York, NY, USA, 2001, pp. 161–172.
[8] V. Crescenzi, G. Mecca, P. Merialdo *et al.*, "Roadrunner: Towards automatic data extraction from large web sites," in *Proceedings of the International Conference on Very Large Data Bases*, 2001, pp. 109–118.
[9] A. Laender, B. Ribeiro-Neto, A. da Silva, and J. Teixeira, "A brief survey of web data extraction tools," *ACM Sigmod Record*, vol. 31, no. 2, pp. 84–93, 2002.
[10] I. Muslea, S. Minton, and C. Knoblock, "Hierarchical wrapper induction for semistructured information sources," *Autonomous Agents and Multi-Agent Systems*, vol. 4, no. 1, pp. 93–114, 2001.
[11] L. Liu, C. Pu, and W. Han, "XWRAP: an XML-enabled wrapper construction system for Webinformation sources," in *Data Engineering, 2000. Proceedings. 16th International Conference on*, 2000, pp. 611–621.
[12] G. Arocena and A. Mendelzon, "WebOQL: Restructuring documents, databases, and webs," *Theory and Practice of Object Systems*, vol. 5, no. 3, pp. 127–141, 1999.
[13] C. M. Sperberg-McQueen, "Xml and semi-structured data," *Queue*, vol. 3, no. 8, pp. 34–41, 2005.
[14] S. D. Uvarow, G. Llambías, and F. Toledo, "Extracción automática de valores de calidad de datos," 2007.
[15] M. Souto, M. Warpechowski, and J. de Oliveira, "An Ontological Approach for the Quality Assessment of Computer Science Conferences," *LECTURE NOTES IN COMPUTER SCIENCE*, vol. 4802, p. 202, 2007.