From the Editor in Chief...

Deep Web Structure

Munindar P. Singh • *North Carolina State University* • *singh@ncsu.edu*

Google's popularity has turned link analysis into a major approach for organizing and retrieving information on the Web. The simple underlying idea is to model a Web page's authoritativeness by the number of links to it from other pages, while recursively factoring in the referring pages' authoritativeness.[1] Links among pages induce a structure on the Web that has two especially interesting properties: The number of incoming links follows a power-law distribution wherein few pages get most of the links, and the pages are clustered such that a core of well-connected pages is separated by paths of fewer than 20 hops.[2]

This structure, which has come about because of links made by people, enables current techniques for locating information on the Web, albeit in a coarse manner that doesn't respect individual needs. However, as the Web acquires various Web services, which typically access information from different databases and repositories, information on the Web will increasingly be interpreted by programs. These changes will likely affect the Web's structure, and they will almost certainly affect the way we locate information (or services) on the Web.

Authoritativeness and Relevance

Traditional document retrieval, which is geared toward closed collections, considers only the contents of documents (that is, the words in a given document relative to the words in others in the same collection) to determine how relevant a given document is to a given query. Even neglecting well-known problems with identifying concepts by words alone, a remaining problem arises on the Web when determining the relevance of a page (viewed as a document) to a query: Wishing to appear more authoritative than they really are, page authors might include arbitrary words to fool search engines into treating their pages as relevant to a particular set of queries. Commercial Web sites often live and die by the number of visitors they can attract, for example, and that number depends on how authoritative the site appears to potential visitors. (Sites can also attempt to imply authoritativeness through paid placement, but I'll put that aside for now.)

In an open environment such as the Web, we must therefore introduce an external evaluation mechanism for estimating a page's authoritativeness. By looking at information in pages other than those that are under direct consideration, link analysis provides a ready, if crude, basis for external evaluation. That is, a page can probably be viewed as valuable if other pages consider it worthy of linking to. Factoring in recursive authoritativeness makes it difficult for Web sites to deceive the system and fake their authoritativeness.

Link-based approaches make a fundamental assumption, however, that the existence of a link from one page to another indicates a recommendation by (the author of) the referring page. The success of search engines like Google indicates that this assumption is often valid in the aggregate for locating information that is widely accessible and well-known (if not to the seeker). It is easy to construct pathological scenarios, however. If several authoritative sites list a site that contains spurious information as one to avoid, for example, link analysis could consider the site authoritative because of the many links to it. Moreover, this approach to determining authoritativeness leaves much to be desired when the user has specialized information needs.

Deep Web

Our current understanding of Web structure is based on large graphs created by centralized crawlers and indexers. They obtain data almost exclusively from the so-called *surface* Web, which consists, loosely speaking, of interlinked HTML pages. The *deep* Web, by contrast, is information

 1089-7801/02/$17.00

that is reachable over the Web, but that resides in databases; it is dynamically available in response to queries, not placed on static pages ahead of time. Recent estimates indicate that the deep Web has hundreds of times more data than the surface Web (www.press.umich.edu/jep/07-01/bergman.html).

The deep Web gives us reason to rethink much of the current doctrine of broad-based link analysis. Instead of looking up pages and finding links on them, Web crawlers would have to produce queries to generate relevant pages. Creating appropriate queries ahead of time is nontrivial without understanding the content of the queried sites. The deep Web's scale would also make it much harder to cache results than to merely index static pages.

Whereas a static page presents its links for all to see, a deep Web site can decide whose queries to process and how well. It can, for example, authenticate the querying party before giving it any truly valuable information and links. It can build an understanding of the querying party's context in order to give proper responses, and it can engage in dialogues and negotiate for the information it reveals. The Web site can thus prevent its information from being used by unknown parties. What's more, the querying party can ensure that the information is meant for it.

Autonomy and Interaction

Individual querying parties and explorations would heavily affect the linkage structures that emerge from deep Web queries and responses. Instead of the current centralized notion of authority — the same for all comers — we could consider authoritativeness as personalized to each user. In other words, there would be not one Web, but several. While phenomena such as the power-law and short path-length cores might still hold, they would apply within the Webs observed by different users. The structure of the individual Webs would affect how different users locate authoritative information and how it may be shared with others.

Beyond the data it provides, the deep Web forces us to recognize that the Web is active and its denizens — human and computational — autonomous. Consequently, the way they interact depends on their relationships with each other. That leads us to the question of trust, which I will pick up next time.

References

1. M.R. Henzinger, "Hyperlink Analysis for the Web," *IEEE Internet Computing*, vol. 5, no. 1, Jan. 2001, pp. 45-50.
2. J. Kleinberg and S. Lawrence, "The Structure of the Web," *Science*, vol. 294, Nov. 2001, pp. 1849-1850.

How to Reach IC

Articles: We welcome submissions about Internet application technologies. For detailed instructions and information on peer review, *IEEE Internet Computing*'s author guidelines are available online at computer.org/internet/author.htm.

Letters to the Editor: Please send letters, including reference to articles in question, via e-mail to swoods@computer.org.

IEEE INTERNET COMPUTING

IEEE Computer Society Publications Office
10662 Los Vaqueros Circle, PO Box 3014
Los Alamitos, CA 90720-1314