Dark Matter on the Web

Peter Bailey, Nick Craswell
Dept. Computer Science, FEIT,
The Australian National University
Canberra ACT Australia
{peterb,nick}@cs.anu.edu.au

David Hawking
Mathematical and Information Sciences,
CSIRO
Canberra ACT Australia
David.Hawking@cmis.csiro.au

1. Obtaining Web Information

When searching the Web, users of search engines may be under the erroneous impression that their search will cover ``all of the Web.'' We introduce the concept of *dark matter* - information on the Web that is not or cannot be discovered by an individual or a search engine. The term ``dark matter'' was originally coined by astronomers to describe observations by Fritz Zwicky over 50 years ago that demonstrated a substantial amount of the matter of the universe is not visible. Dark matter seems an equally appropriate term for an analogous situation in Web space. Its main contribution as a concept for the Web is in providing a language to describe how and why information is not discoverable. Our definition characterises why obtaining pages is *dependent on the characteristics of the observer*. Observers are either people who download pages from the Web using some browser technology, or crawler agents of certain kinds of Web search systems. People typically find pages by: learning of the existence of a URL, following links from pages which they download, forging URLs based on cultural knowledge, or using search systems which present them with pages of links. Web crawlers find pages by following links from some seed set of URLs.

2. Dark Matter

We understand the Web to be all pages (both their URL and their information) which can be accessed through the HTTP protocol by some person or computer somewhere. We can ``see'' parts of the Web using search engines, which collect and organise information for us. But there are parts of the Web which are not visible from any of these search engines. Lawrence and Giles's estimate of the size of the Web [**Lawrence and Giles 1999**] was that in February 1999 the Web contained over 800 million ``publicly indexable'' pages. They also found that the biggest search engine at that time, Northern Light, indexed 16%, and combined coverage of all the search engines examined was 42% of the available Web. There are many reasons why search engines and human browsers are unable to find material that exists on the Web. The reasons why material is not discoverable give rise to a classification taxonomy.

2.1 Rejected Dark Matter

The first kind of dark matter arises because a page is *rejected* for some reason and thus did not load. Human observers frequently reject pages, mainly due to a lack of interest in the contents indicated by a link. They may also reject it if the page requires a new kind of browser plugin to be viewable. Web crawlers frequently reject pages for policy reasons. Most current search engines' Web crawlers have resource limits to the extent any one server's Web data will be crawled. Web crawlers may also choose to exclude themselves from parts of a site by following the `robots.txt` convention. The page loading policy of a Web crawler may exclude files with particular extensions, such as `.tar.gz` or `.gif` and many others. Dynamically generated Web matter, usually indicated by URLs containing `cgi-bin` or ?'s, are also mostly rejected.

2.2 Restricted Dark Matter

Second, there is material which is publicly linked to, but is *restricted* to observers with the appropriate permissions. User/password combinations are a common mechanism used by Web sites to restrict access to specified users. Without the combination, Web crawlers cannot access the information, even if its creation is free. Similarly, a Web server administrator can limit access to particular content to users in nominated internet domains.

2.3 Undiscovered Dark Matter

The third kind of dark matter encompasses pages which remain *undiscovered* by an observer. The reason for this is straightforward: the necessary links which locate the material are never found. Very often there is material on

the Web which has public links to it, but the linking pages are themselves dark matter to the observer. An interesting subcategory of undiscovered dark matter is material which is not publicly linked to at all. We refer to this scenario as *private dark matter*. Access to the material relies on knowing or guessing the URL. For example, information can be placed on a Web server and the URL sent to someone via email, who can then download it.

2.4 Removed Dark Matter

The Web changes constantly: new servers go online, old servers are turned off, new pages are created, old pages are changed, or removed. The fourth kind of dark matter is that which is available at some point in time, but later on has been *removed* and is no longer available. There are two subcategories of removed dark matter, which are differentiated by whether the URL is still valid, or whether it too has vanished. Removed *ephemeral dark matter* exists when information which was once available is removed, and new information takes its place. For example, many media organisations replace the entire contents of their news pages on a daily or faster basis. The other subcategory of interest is *dead dark matter*: information which has been successfully loaded at some previous time, but is no longer available when an attempt is made to observe it now. In general, all Web matter is likely to become dead dark matter given a long enough time.

3. Conclusions

As the importance of the Web as the predominant information medium continues to rise, and search engines continue to be a prime information discovery mechanism, the existence of information that is dark to them assumes greater significance. Our categorisation of the different forms of Web matter indicate that not only is universal coverage unlikely to be feasible, but that it is impossible without huge resources and omnipotent security privileges. We believe that expanded use of meta-search systems combined with increasing local search facilities on Web document servers is the most likely means by which the quantity of dark matter, particularly undiscovered dark matter, may decline on the Web.

Acknowledgements This work was carried out within the CRC for Advanced Computational Systems established under the Australian Government's Cooperative Research Centres Program.

Bibliography

Lawrence and Giles 1999 Accessiblity and Distribution of Information on the Web, *Nature* 400, 107-109.