Final Year Project Interim Report
Content-based Prediction of Web 2.0 Page Updates

U096883L Shawn Tan

1 Introduction

With the increasing number of Web 2.0 sites, sites with forums, or similar thread-based discussion features are extremely common. Data found on discussions like these provide useful feedback for content providers. As more users are involved with the content-generation process of these sites, maintaining an updated database of crawled content becomes increasingly difficult.

Web crawlers which maintain the 'freshness' of a database of crawled content are known as incremental crawlers. Two tradeoffs these crawlers face cited by Yang et. al. 2009 [8] are *completeness* and *timeliness*. *Completeness* refers to the extent which the crawler fetches all the pages, without missing any pages. *Timeliness* refers to the efficiency with which the crawler discovers and downloads newly created content.

Let us define all such thread-based discussion styled sites as forums. Ideally, an incremental crawler of such user-generated content should be able to maintain a fresh and complete database of content of the forum that it is monitoring. A naive way to approach this would be to aggressively download these pages at a frequent rate. This, however, would (1) incur excessive costs when downloading un-updated pages, and (2) raise the possibility of the web master blocking the requester's IP address.

Thus, we need a strategy of revisiting pages that will reduce the cost of downloading unchanged pages, while at the same time downloading them as soon as possible after it's update.

2 Related work

In order to devise such a strategy, we need to predict how often any user may update the a page. Some work has been done to try to predict how often page content is updated by the page owner.

Many such works have used the Poisson distribution to model page updates. Coffman et. al. [5] analysed the theoretical aspects of doing this, while Cho and Garcia-Molina trace the change history of 720,000 web pages collected over 4 months, and compared the result against what the Poisson process model predicts [2], and then proposed different revisiting or refresh policies [3, 4] that attempt to maintain the 'freshness' of the database. The Poisson distribution were also used in Tan et. al. [6] and Wolf et. al. [7]. However, the Poisson distribution is memoryless, and in experimental results due to Brewington and Cybenko [1], the behaviour of site updates are not.

	T	FB L	FB S	G +1	Rating	L	DL	C	PV	Follows
`http://www.lifehacker.com`	1	1		1				1	1	
`http://digg.com/`	1	1				1	1	1	1	
`http://9gag.com/`	1	1	1	1		1		1	1	
`http://www.flickr.com/`						1		1	1	
`http://news.ycombinator.com/`						1		1		
`http://stackoverflow.com/`						1		1	1	
`http://www.youtube.com/`						1	1	1	1	
`http://www.reddit.com/`						1	1	1		
`http://www.stumbleupon.com/`						1		1	1	
`http://delicious.com/`	1	1						1	1	1

Table 1: Features of popular Web 2.0 sites

Yang et. al. [8], attempted to resolve this by using the list structure of forum sites to infer a sitemap. With this, they reconstruct the full thread, and then use a linear-regression model to predict when the next update to the thread will arrive.

These methods of estimating page updates either rely on previously gathered information about the page updates through repeated polling of the page, or through timestamps gathered from the individual posts. They have two shortfalls:

Lack/Improperly formatted Timestamp Information While most comment threads or forum sites tend to have timestamps, they often try to optimise readability. For example, timestamps of comments that were posted 8 months ago may be displayed as "more than 4 months ago".

Requires previous time series data If the individual threads are treated independently of each other, a new thread (1 or 2 posts) would not have sufficient data to fit a Poisson model. Yang et. al. [8] accounts for this by factoring into their regression model other threads with a similar recent history

The lack of these pieces of information may result in a poorer estimate, or no estimate at all. We argue, that the content within the posts of the thread should be important in predicting the thread updates. While there is little existing work using content to predict page updates, we will review some existing work related to analysing thread-based pages which we think will aid us in our efforts to do content-based prediction.

3 Possible Approaches

3.1 Hidden Markov Models

3.2 Lexical Chaining

References

[1] Brewington, B. E., and Cybenko, G. Keeping up with the changing web. *Computer* (2000), 52–58.

[2] Cho, J. The evolution of the web and implications for an incremental crawler. *Science* (1999), 1–18.

[3] Cho, J., and Garcia-Molina, H. Effective page refresh policies for Web crawlers. *ACM Transactions on Database Systems 28*, 4 (Dec. 2003), 390–426.

[4] Cho, J., and Garcia-molina, H. Estimating Frequency of Change.

[5] Coffman, E., and Liu, Z. Optimal robot scheduling for web search engines. *Sophia* (1997).

[6] Tan, Q., Zhuang, Z., and Mitra, P. Designing efficient sampling techniques to detect webpage updates. *Proceedings of the 16th 1*, 3 (2007).

[7] Wolf, J., Squillante, M., and Yu, P. Optimal crawling strategies for web search engines. *on World Wide Web* (2002).

[8] Yang, J., Cai, R., Wang, C., and Huang, H. Incorporating site-level knowledge for incremental crawling of web forums: A list-wise strategy. *on Knowledge* (2009), 1375–1383.