Intelligent Crawling On Open Web for Business Prospects

Bharat Bhushan[1], Narender Kumar[2],

1Department of Computer Science & Applications, Guru Nanak Khalsa College, Yamuna Nagar (Haryana), India
2Narender Kumar , Research Scholar, Enrollment No. 1050103206 ,Singhania University , V.P.O. - Pacheri Bari, Dist. Jhunjhunu, Rajasthan - [INDIA]

Summary
Dynamic nature of web based systems requires continuous system updating. Information retrieval depends upon crawlers that crawl the web exhaustively, but business corporates expect from their crawlers to retrieve the specific information as per their applications. Crawlers help to download the required information using hyperlinks that occur in Web pages but the information is usually partial & fails to fulfill user's aspirations. To retrieve updated information from one single link/url is very simple but if many urls give the same information, it becomes difficult to analyze which url/link is giving desired, sufficient, updated & up to date information. Moreover, it becomes difficult how to remove duplicate stories from same link domain. In the present paper attempt has been made to discuss the issues related to intelligent crawling by proposing various techniques to assist the scenario concerned with web mining for business prospects.

Key words:
Web crawler, latency, ethics, reliability, longevity

Introduction

Web provides a vast resource for business intelligence. Web crawling is an important method for collecting data and keeping up to date with the rapidly expanding Internet. However, large size of the Web, its exponential growth and dynamic nature makes the task of retrieving appropriate information quite challenging. A large number of web pages are added every day and the quantum and nature of information gets changed. A web crawler is a tool for the search engines and other information seekers to enable them to keep their databases up to date.

For a particular one-time project web crawlers may operate one time only, If its purpose is long term crawling, as is the case with search engines, they may be programmed to comb through the internet periodically to ascertain whether there have been any significant changes. If a site is experiencing heavy traffic or technical difficulties, the spider may be programmed to take a note of it and revisit it after a certain time gap, hopefully, when the technical issues have subsided.

Generally web crawling doesn't fulfill the need of business prospects as far as Timely, Sufficient, Updated information is concerned and, at the same time, such crawling is not sensitive in eliminating duplicate Data.

Manuscript received June 5, 2012
Manuscript revised June 20, 2012

These issues have negative impact on the corporate business. Various factors need to be reconsidered to meet the needs so that the business corporate may get intelligent, up to date & reliable information for business growth well in time. Following are the factors taken into consideration for the same:

Latency: Latency is a measure of time delay experienced in a system. Latency in crawl schedule means the difference in crawling start time and under process time. It is because of web standard errors codes like 702, 3-x, 404 etc., same domain crawls, crawling on close times and lack of maintenance on crawl schedules the latency results in documents not being crawled, documents being crawled late and increase in the number of un crawled topics. Since the document doesn't get crawled on time therefore, after a certain time gap it become old. Due to this, a lot of time goes wasted and inconsistency causes hampering the quality of the product in terms of business perspective.

Crawler Issues: Crawling issues may be due to updating of information at a very small interval e.g. once in an hour/day. If a site gets updated every hour whereas crawling frequency is once in a day; then there is a possibility of crawling old content from that particular site/source.

Related Work

Gautam Pant and Filippo Menczer [1] investigated the use of topical crawlers in creating a small document collection that helps locate relevant business entities. The problem of locating business entities is encountered when an organization looks for competitors, partners or acquisitions. Their results underscore the importance of identifying good hubs and exploiting link contexts based on tag trees for accelerating the crawl and improving the overall results. They formalized the problem, created a test bed, introduced metrics to measure the performance of crawlers, and compared the results of four different crawlers.

Soumen Chakrabarti, et. al. [2] described a new hypertext resource discovery system called a Focused Crawler i.e. goal-directed crawling. The goal of a focused crawler is to selectively seek out pages that are relevant to a pre-

defined set of topics. The topics are specified not using keywords, but using exemplary documents. Rather than collecting and indexing all accessible web documents to be able to answer all possible ad-hoc queries, a focused crawler analyzes its crawl boundary to find the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the web. This leads to significant savings in hardware and network resources, and helps keep the crawl more up-to-date because focused crawling acquires relevant pages steadily while standard crawling quickly loses its way, even though they are started from the same root set

Junghoo Cho, Hector Garcia-Molina [3] discussed the case for estimating the change frequency of data to improve web crawlers, web caches and to help data mining. They first identify various scenarios, where different applications have different requirements on the accuracy of the estimated frequency. Then they developed several frequency estimators for the identified scenarios, showing analytically and experimentally how precise they are. In many cases, they proposed estimators predict change frequencies much more accurately and improve the effectiveness of applications.

Christopher Olston and Sandeep Pande [4] characterized the longevity of information found on the web, via both empirical measurements and a generative model that coincides with these measurements. They developed new re-crawl scheduling policies that take longevity into account. They experimentally showed that their policies obtain better freshness at lower cost, compared with previous approaches over real web data.

Chengjie Liu and Pie Cao [5] proposed a strong cache consistency algorithm to improve access latency and reduce network and server load for scaling the internet to meet increasing demands of the users. Compared with previous strong consistency, their proposed algorithm is more efficient.

C. Lee Giles, Yang Sun, m Isaac G. Councill [6] proposed quantitative models to measure the web crawler ethics such as spam and service attacks based on their behaviors on web servers. They investigated and defined rules to measure crawler ethics, referring to the extent to which web crawlers respect the regulations set forth in robots.txt configuration files. They proposed a vector space model to represent crawler behavior and measure the ethics of web crawlers based on the behavior vectors. The results showed that ethicality scores vary significantly among crawlers. Most commercial web crawlers' behaviors were found to be ethical. However, many commercial crawlers still consistently violate or misinterpret certain robots.txt rules. They also measured the ethics of big search engine crawlers in terms of return on investment. The results showed that Google has a higher score than other search engines for a US website but has a lower score than Baidu for Chinese websites.

Viv Cothey [7] investigated the web crawling reliability, in the social science sense, of collecting informetric data about the World Wide Web by web crawling. The investigation included a critical examination of the practice of web crawling and contrasts the results of content crawling with the results of link crawling. It was shown that Web crawling by search engines is intentionally biased and selective. They also illustrated experimental simulation of web crawling to study the effects of different crawling policies on data collection and it was found that the reliability of web crawling as a data collection technique is improved by complete reporting of relevant crawling policies

Junghoo Cho, Hector Garcia-molina [8] studied how can one maintain local copies of remote data sources fresh, when the source data is updated autonomously and independently during web crawling. In this context, remote data sources do not notify the copies of new changes, so one need to periodically poll the sources to maintain the copies up-to-date. Since polling the sources takes significant time and resources, it is very difficult to keep the copies completely up-to-date. They proposed various refresh policies and studied their effectiveness. They first formalized the notion of freshness of copied data by defining two freshness metrics, and proposed a Poisson process as the change model of data sources. Based on this framework, they examined the effectiveness of the proposed refresh policies analytically and experimentally and found that they improve the freshness of data very significantly.

Proposed Model

An attempt has been made to fetch the important information beneficial to business corporate as per their business plans by resolving the above issues.

The Latency issue can be managed by timely checking the crawler load and do the crawl redistribution on weekly/monthly bases and setting correct crawl frequency of new links by keeping in mind the no. of average documents coming on that link, time of maximum updating on the URL and region (country) of the source.

Crawler issues can be handled by setting the crawling frequency accordingly and changes in original site behavior related to http standard error code links need to be updated accordingly.

Also make a list of key words which produce irrelevant information of business prospects contents on the basis of pattern analysis & stop those links in which these keywords comes in url/link during crawling.

Case Study: A plan of setting the crawl frequency on the bases of crawled content has been employed. The Links

were using 1,70,000 crawls per day. The reduction in latency was achieved by using:

1) Redistribution cycle plan for links

Distribution of Links

Table 1

Crawl Frequency	No of Docs/ Day (including upper limit)	Links Count	Total Crawl
24 Hrly Crawl	< 0.5	32332	32332
12 Hrly Crawl	0.5 - 4	21432	42864
6 Hrly Crawl	4 - 14	7685	30740
3 Hrly Crawl	14 - 20.5	1398	11184
2 Hrly Crawl	20.5 - 35.7	1165	13980
1 Hrly Crawl	>= 35.7	700	16800
	Total	64712	147900

The table 1 shows that the links are using only 1,47 ,900 crawl instead of 1, 70,000 crawls per day.
These Links were further distributed on different crawl timings keeping the following two factors in mind:

a. by evenly distributing the load in each hour.
b. by keeping the same domain links on different timings.

The table 2 shows Daily Crawl, Twelve Hourly Crawl and Six Hourly Crawl. They can be used efficiently to minimize each load from the crawler.
Also timely check the standard error urls code found in visited urls and update them or remove them from the system to effectively and efficiently use the crawler.

Table 2

	24 hrly	12 hrly	6 hrly	3 hrly	2 hrly	1 hr		Total Crawl/Hour
00 PDT	2646	3699		1398	1165	700		9608
23 PDT		8585				700		9285
22 PDT	7639				1165	700		9504
21 PDT			7685	1398		700		9783
20 PDT	7220				1165	700		9085
19 PDT	8526					700		9226
18 PDT	5917			1398	1165	700		9180
17 PDT	385					700		1085
16 PDT					1165	700		1865
15 PDT			7685	1398		700		9783
14 PDT					1165	700		1865
13 PDT		9148				700		9848
12 PDT		3699		1398	1165	700		6962
11 PDT		8585				700		9285
10 PDT					1165	700		1865
09 PDT			7685	1398		700		9783
08 PDT					1165	700		1865
07 PDT						700		700
06 PDT				1398	1165	700		3263
05 PDT						700		700
04 PDT					1165	700		1865
03 PDT			7685	1398		700		9783
02 PDT					1165	700		1865
01 PDT		9148				700		9848
								147900
Crawls	32332	42864	30740	11184	13980	16800	147901	
Links	32332	21432	7685	1398	1165	700	64713	

2) Examine the Crawl frequency & distribute it on different crawlers as per tricks; like:

By finding out the domains/website on the basis of documents crawled in last one month or three month or 6 month & find out how much updated content came and

make the list of domains & distribute it and send in respective crawler.
Separate list of magazines/Journals which are mostly updated once in a month. Separate the list of newspaper sites which are always updated daily.

[a] Domains with heavy crawler activity in last 7 days [Table 3]
[b] Distribution of documents across crawlers for last 2 days [Table 4]
[c]. Details for key crawler metrics (total crawls, size of downloaded data etc.) over the last two weeks for each crawler.

Consider a sample set of 5 heavy domain category in which three different crawlers are involved & following table shows the domains with good content as per business needs.

Table 3: Domains with heavy crawler activity in last 7 days

Domains with most requests in all crawlers		
Total: 5 start points		
Domain	Downloads	Crawls
www.marketwatch.com	13461	1738
rss.prnewswire.com	9724	1960
www.emailwire.com	7215	61
www.forbes.com	5404	1040
www.reuters.com	5128	519

Table 4:Crawler Activity trends based on distribution of for last 2 days

Crawler Identification No.	Crawler Name	Documents Crawled in Last 2 Days
CID1	CQM1	31873
CID2	CQM2	16230
CID3	CQM3	32589

Table 5:Details for key crawler metrics (total crawls, size of downloaded data etc.) over the last two weeks for each crawler.

Domain	Docs All	CQM1	CQM2	CQM3
All Domains	80538	31822	16166	32550
www.prlog.org	1185	948	237	0
www.businessweek.com	861	354	507	0
www.investegate.co.uk	665	0	665	0
in.news.yahoo.com	639	0	64	575
hosted.ap.org	632	150	69	413
pr-usa.net	610	610	0	0
www.bloomberg.com	582	0	582	0
feeds.feedburner.com0	537	314	126	97
www.consultingmag.com	536	536	0	0
biz.yahoo.com	511	14	0	497

Crawler Latency

Table 6:
[a] Listing of most recent abnormal crawls

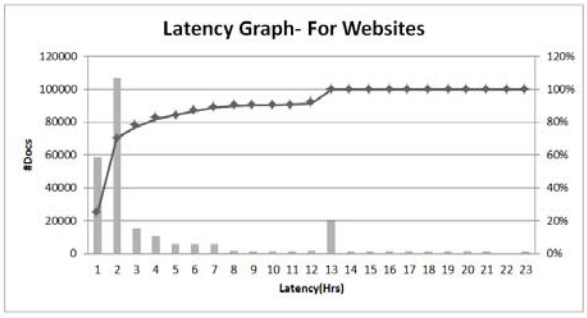
Links Id	Crawler	Exit Code	Down loaded	Bytes	Queue time	Start Latency Minutes	Start time	Crawl Duration Minutes	End time
244001	CQM1	-1	0	14,564	09-04-2012 00:21	186	00:27:54	1	0:28:32
384453	CQM1	-1	0	20,072	09-04-2012 00:20	187	00:27:38	1	0:28:05
192568	CQM1	-1	0	11,883	09-04-2012 00:20	187	00:27:08	0	0:27:51
179584	CQM1	-1	0	1,188	09-04-2012 00:20	181	00:21:12	1	0:22:09
372598	CQM1	-1	0	16,788	09-04-2012 00:18	175	00:13:18	0	0:13:54

Table 7:
[b] Overall breakdown of scheduling latencies for various crawl scheduling intervals

Latency (Hrs)	docs	cumulative %age
1	58444	25%
2	107030	71%
3	15479	77%
4	10937	82%
5	5798	84%
6	5708	87%
7	5826	89%
8	1910	90%
9	553	90%
10	296	90%

11	562	91%
12	1884	91%
13	19985	100%
14	6	100%
15	5	100%
16	16	100%
17	17	100%
18	9	100%

Various histograms of latencies and crawl durations



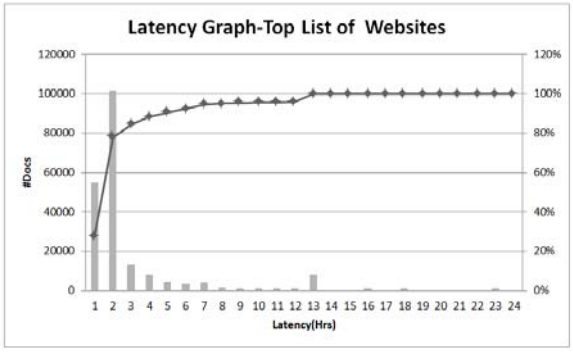


Table 8: Top list of Websites with their cumulative percentage

Latency (Hrs)	docs	cumulative %age
1	55020	27%
2	101711	78%
3	13091	84%
4	8156	88%
5	4432	91%
6	3573	92%
7	4150	94%
8	1409	95%
9	352	95%
10	198	95%
11	439	96%
12	415	96%
13	8303	100%
14	0	100%
15	0	100%
16	16	100%
17	0	100%
18	9	100%
19	0	100%
20	0	100%
21	0	100%
22	0	100%
23	9	100%
24	0	100%

Crawler issues:

Crawler issues can be handled using following techniques

1) Setting the crawling frequency:

a) Analyze the site/source and setting the crawling frequency accordingly and further

b) Update accordingly in original site behavior related to http standard error code links

2) Searching the irrelevant information:

Make a list of key words which produce irrelevant information of business prospects contents on the basis of pattern analysis & stop those links in which these keywords comes in url/link during crawling.
e.g. The keywords might be changed as per requirement of docs as per business prospects Keywords:
#, $, %, .flv, .atom, .hyperlink, .jpg, .mp3, //br., //hk., //id., //kr., /;id=, /?redirect, /accounts, /adclick, /adlinks, /admarket, /adpay, /adperfect, /adproducts, /ads., /ads/, /adserver/, /adv/, /advchart/, /advisory_board, /album, /aph., /buyersGuideForVendors/, /cn., /contact, /copyright, /coupons/, /crossborder, /email_marketing/

Discussion & Conclusion

By using the above proposed solutions, it is found that one can easily resolve the above mentioned latency and crawlers issues to crawl the important information as per the needs of the business prospective. This information will be beneficial to business corporate as per their business plans. By using these methods we also reach that level where one can automate few things which help to reduce manual work and also helpful for further analysis on these data. This paper will prove an asset in improving a web crawler, update policy of a data warehouse, web caching, data mining and irregular information interval from web resources with minimum latency and relevant contents.

REFERENCES

[1] Gautam Pant and FilippoMenczer "Topical Crawling for Business Intelligence" in proc. 7th european conference on research and advanced technology for digital libraries (ecdl 2003}

[2] Soumen Chakrabarti Martin van den Berg Byron Dom "Focused crawling: a new approach to topic-specific Web resource discovery" Journal Computer Networks: The International Journal of Computer and Telecommunications Networking archive Volume 31 Issue 11-16, May 17, 1999 Pages 1623 - 1640 Elsevier North-Holland, Inc. New York, NY, USA

[3] Junghoo Cho, Hector Garcia-Molina "Estimating Frequency of Change" Published in: • Journal ACM Transactions on Internet Technology (TOIT) TOIT Homepage archive Volume 3 Issue 3, August 2003 Pages 256 - 290 ACM New York, NY, USA

[4] Christopher Olston ,SandeepPandey " Recrawl scheduling based on information longevity " Proceeding WWW '08 Proceedings of the 17th international conference on World Wide Web Pages 437-446 ACM New York, NY, USA ©2008

[5] Chengjie Liu and Pie Cao " Maintaining strong cache consistency in the www" IEEE TRANSACTIONS ON COMPUTERS , VOL. 47 , NO. 4 , APRIL 1998

[6] C. Lee Giles,Yang Sun,Isaac G. Councill "Measuring The Web Crawler Ethics" Proceedings of the 19th international conference on World wide web Pages 1101-1102

[7] VivCothey "Web-crawling reliability" Journal of the American Society for Information Science and Technology - Special issue: Webometrics archive Volume 55 Issue 14, December 2004 Pages 1228 - 1238 John Wiley & Sons, Inc. New York, NY, USA

[8] Junghoo cho, hector garcia-molina "Effective Page Refresh Policies for Web Crawlers" Cho, Junghoo and Garcia-Molina, Hector (2003) Effective Page Refresh Policies for Web Crawlers. Technical Report. Stanford InfoLab.

Bharat Bhushan received the M Sc (Physics), from Panjab Univ. Chandigarh and M.Sc. (Comp. Sc.), MCA degrees from Guru Jambeshwar University respectively. Presently working as Associate Professor & Head, Department of Computer Science and Applications, Guru Nanak Khalsa College, Yamuna Nagar (affiliated to Kurukshetra University, Kurukshetra-Haryana, India) and senior most teacher of computer science in Haryana since 1984. He is a member of Board of Studies of Computer Science, Kurukshetra University and member of Advisory Board of educational programme (EDUSAT) launched by Govt. of Haryana to impart online education. His research interest includes Software engineering, Digital electronics, networking and Simulation Experiments. He has published more than 25 research papers in national and international journals of international repute.

Narender Kumar, Research Scholar, Enrollment No. 1050103206, Singhania University, V.P.O. - Pacheri Bari, Dist. Jhunjhunu, Rajasthan - 333 515 [INDIA]. He received B.Sc. (Computer Science & Applications) degree from Guru Nanak Khalsa College, Yamunanagar and MCA from from Ch. Devil Lal Post Graduate Regional Center of Kurukshetra University.