International Journal of Computer Science and Business Informatics

IJCSBI.ORG

Simultaneous Use of CPU and GPU to Real Time Inverted Index Updating in Microblogs

Sajad Bolhasani
Department of CSE,
Lorestan Science and Research College,
Lorestan, Iran

Hasan Naderi
Assistant Professor
Department of CSE,
Iran University of Science and Technology,
Tehran, Iran

ABSTRACT

Nowadays, with attention to developing the different data networks, the wide masses of data are producing and updating continually. Managing the great data enumerate the fundamental challenges in data mining. One of the considered main subjects in this context is how searching among the wide masses of data. Therefore, require to producing the typical powerful, expansible and efficient file of documents and data for using in search motors is necessary. In this study, with surveying the done prior works, implementing the inverted index with the immediate updating capability from the dynamic and little data of microblogs is targeted. With utilization from processing multicore facility, the approach of the graphical processing unit (GPU) is presented that as expansible and without decreasing the attention, the index file is prepared with suitable speed, as the mentioned file is usable in inquiry unit. This method tries to feed the updating unit continually with separating the operation for the system Central Processing Unit (CPU) and suitable utilization of parallel processing capability of CUDA core. Also, in parallel to increasing the quality, one Hint method is presented for employing the vacant cores and compactor function for decreasing the index file mass. The results indicate that the presence of necessary hardware, the presented method in identity to immediate updating slogan, have the upper speed for making the inverted index of microblogs than to available samples.

Keywords

Inverted index, Microblog, GPU, Update.

1. INTRODUCTION

In data mining and contextual managing information inverted index is a main key for each searching process by having this file search engines have the ability to stream a search without any repeated attention to the content of any documents. The structure of inverted index is generally upon the hash table frame and consists of a word dictionary and some values. Creator of inverted index in a process of searching skim the words in document, analyze and stemming, after all adds them to the dictionary. In this Platform each term is a specific key in dictionary of words. Any keyword in dictionary refer to a list of ID these ID's refer to those documents that, containing keywords. While a change applied on a document there is a necessity to update the ID files. So this updating process has some costs. The ultimate goal for each dynamic inverted index is reducing updating speed and near to zero or real-time it [1, 2, 3]. And the way we introduce in this article to reach that goal is dividing and paralleling of making inverted index operation. By using the capability of multi core, multi thread GPU's help us to near our goal [4]. Cuda cores have capability to operate simultaneously multi tasks give us the opportunity to divides the instructions for paralleling in little blocks. Microblogs introduce as data entries for inverted index in this article. In conclusion with the approach introduced above for making inverted index use any possibility from microblog's documents recognized by crawler so files makes with possible lower cost and use with real-time update.

2. BACKGROUND STUDY

Time in updating inverted index is an important characteristic of measurement on search engines. Insert time in multi barrel hierarchy index [5], consist of different index size that will finally contribute each other with this functions.

$$(1)\quad I_1(n) = O\left(\frac{1}{\log(k)}\log(n)\right)\frac{P_s(n)}{n}$$

$$(2)\quad I_2(n) = O\left(\frac{K}{\log(k)}\log(n)\right)\frac{P_s(n)}{n}$$

In function (1) $I_1(n)$ is the average time for insert n new documents with different sizes. $P_s(n)$is also the time for making statistic inverted index with the size of n. And also using $\frac{1}{\log(k)}\log(n)$ verses $log_k(n)$to show how much use of k has positive effect on ability of system. In function (1) with increasing the value of k the average insert time will near to $\frac{P_s(n)}{n}$ that it will be our ideal but it will increase search time. In function (2) by increasing k search speed will improved.

2.1 Inverted Index

Search engines consist of three parts. Crawler that find web pages, indexer that indexes inverted index and crawled web pages and a ranker that answers query by using indexes. Dynamic documents are document that change and update continually. Static inverted index [6] has a dictionary structure. Dictionaries are making from split word in text and find their stems by using algorithm called "porter stemmer" and prepare them for indexing. The reason for saving stems is to reduce memory size and indexing more documents in search result.

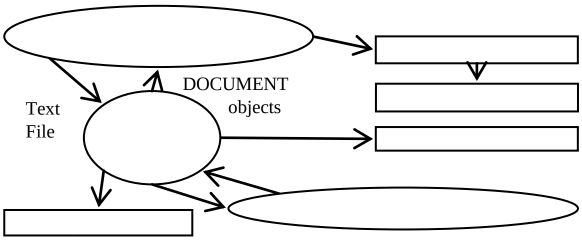


Figure 1. Structure of index file [7]

All of information that gathered by search engine use as entry for inverted index after process of save in barrel.

2.2 GPU structure and capability of paralleling non graphical tasks

The structure of graphic processor consists of many simple processing units called thread. These units are only make for simple calculations like add or subtract. By introducing CUDA form NVIDIA CO. the limitations of non-graphical tasks has taken from graphic units.in a graphic card ,elements have separated memory so designers know them as independent device or even a computer. Regarding to this knowledge each computer has a processor that works with a unique memory.in GPU a shared memory specify to each block; also each stem has two specific memories. Local and stable (fixed). Local memory use for global memory data or shared memory. This memory is similar to computer’s hard drive that use as a kind of main memory in a graphic unit. In this structure commands process in stems (set of threads) simultaneously [8,9,10].

3. RELATED WORKS

In this way microblogs real time identification has discussed [11]. Microblogs update their contents many times. Main core of this structure consist of inverted indexes with the value of$I_1 = I_0 * 2$.new microblogs place in smallest index so these set of indexes gather in larger indexes. This hierarchy makes passive updating. Results obtained in their studies regarding to multi threads capabilities.

QiuyingBai [12] and coworkers apply a way for real time updating in subject oriented search engines. They are designing an inverted index with three parts, primary inverted index, additional inverted index and a list of omitted files. This way of real time updating is useful for subject oriented search engine.

In a different research [13] introduce a way in which by using a graphic processor and multi core processing among web without any distribution and only one computer system the operations of build and update of index file will be done. one of the advantage for this way is that the graphic unit will done all processing and in result enteral processing unit will be free to different task so that is not faced to the problem of decreasing the capability of system.

4. METHODOLOGY

In this study we present a shared method in which by using two different processors the operation of both build and updating of inverted index will done in real time. In this method we take incoming documents as preset frame from microblogs, the reason for this task is innate characteristic of the method plan, where data income to graphic processor and divided in smaller units (block). The sizes of units are very restricted (small) in GPU. Each block run through a thread from a CUDA and process a carrying command or part of it .In this stream of processing structure, cores of CUDA will be like warp and blocks are like woof and regarding to this advantage that each block done its process task in a clock so incoming documents that take from crawler are consider as small graphic unit blocks. First, documents depend on adding to index or deleting take stickers of I or D and places in a lineup (Q) on basic memory of system so inter to the central processing unit. From now on, the cores of CPU divide in two sets: half splitters and half updaters. Documents are inter to splitters and change to term and prepare to send them to next step. After each task of splits, by time lapse documents in separated sets are inter to graphic unit to continue the task in parallel. While documents inter to GPU from microblogs and set into blocks .graphic unit depends on number of both blocks and CUDA core will done the process phase to phase .In each block from graphic processor unit, two different tasks of operation for build a doc index will done. Build in Insertion index barrel and build in deleted index barrel. The kind of operation is shown in header of document.

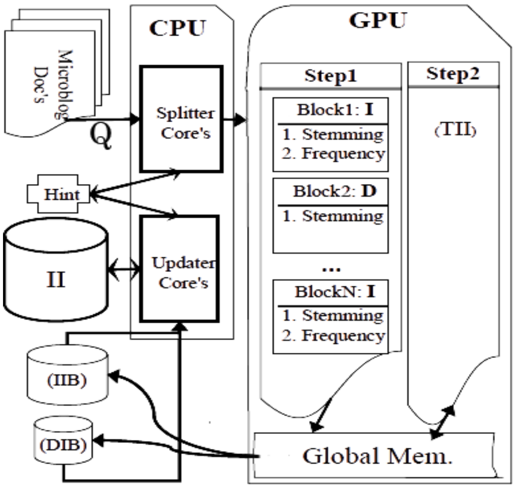
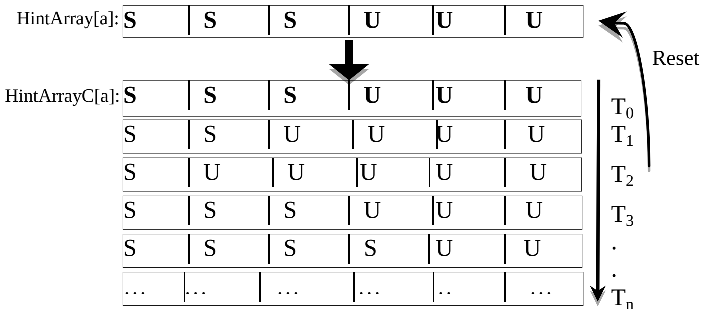


Figure 2. Tasks of real time inverted index by sharing processors.

In first part of GPU processing, inserted blocks both identify how many times a word repeated in each document and the number of deleting blocks and saves in global memory of graphic unit. Soon threshold processing done and threshold inverted index makes. Hint is a notifiable function (mechanism) designed for times the CPU cores do nothing (non-function). This function identifies free times of each core and shares them temporarily to each other so the amount power of processing for those very busy units will increase.



HintArray[a]:	**S**	**S**	**S**	**U**	**U**	**U**
HintArrayC[a]:	**S**	**S**	**S**	**U**	**U**	**U**
	S	S	U	U	U	U
	S	U	U	U	U	U
	S	S	S	U	U	U
	S	S	S	S	U	U
	…	…	…	…	..	…

Figure 3. Change routine in GPU cores by using hint function.

In picture above the cores of main processor of system divided in two sets: splitters (S) and updaters (U). that depends on how much busy an entrance of a core ,the hint function will change real time those free cores to S or U to facilitate busy points.

5. EVALUATION

Consider to ultimate goal of research that is both build inverted index and real time updating of it so measurement range is already “time” for this research. As you can see in below chart two tasks that are use graphic unit or without graphic unit is presented. In both the portion of size with time have direct relation. But execute time is decrease when we use the capability of multi core CUDA graphic processor and it is guide us to reach real time execution.

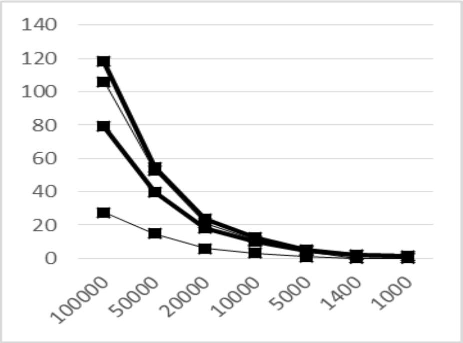


Figure 4. Evaluation algorithm by CPU and CPU&GPU.

In this chart horizontal line is the number of documents and vertical line is the time for building file. Two upper lines related to required time to build index and send for CPU and two lower lines is required time for GPU and CPU.

6. CONCLUSIONS

In this article there is an effort to use graphic units processing power and also use of central processing unit simultaneously for build a real time updating index system from microblogs contents. Also present a way to strongly use of non-function cores. Finally presented the algorithm that has the function to build inverted index from documents in microblogs and updating them in real time.

The future for this research is an end to create a static structure and present a unit called index managing to distribute streams of processing between processor units.

7. ACKNOWLEDGMENTS

We would like to thank Muhammad Saleh Mousavi for their exceptionally useful reviews.

REFERENCES

[1] P. Mudgil, A. K. Sharma, and P. Gupta, An Improved Indexing Mechanism to Index Web Documents, *Computational Intelligence and Communication Networks (CICN)*, 2013 5th International Conference on, 27-29 Sept. 2013, pp. 460 - 464.

[2] R.Konow, G.Navarro, and C. L. A. Clarke, Faster and Smaller Inverted Indices with Treaps, *artially funded by Fondecyt grant 1-110066 , by the Conicyt PhD Scholarship Program*, Chile and by the Emerging Leaders in the Americas Program, Government of Canada ACM, 2013.

[3] S. Brin and L. Page, Reprint of: The anatomy of a large-scale hypertextual web search engine, *The International Journal of Computer and Telecommunications Networking*, 2012, 3825-3833.

[4] Z. Wei and J. JaJa, A fast algorithm for constructing inverted files on heterogeneous platforms, *J. Parallel Distrib*. Comput, 2012.

[5] N. Grimsmo, Dynamic indexes vs. static hierarchies for substring search, *Trondheim*, 2005.

[6] R. A. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman Publishing Co, Inc., 1999.

[7] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Book, ISBN:0521865719 9780521865715 , 2008.

[8] NVIDIA CUDA™, NVIDIA CUDA C Programming Guide, Book, *www.nvidia.com*, 2012

[9] W. Di, Z. Fan, A. Naiyong, W. Fang, L. Jing, and W. Gang, A Batched GPU Algorithm for Set Intersection, *Pervasive Systems, Algorithms, and Networks (ISPAN),*

2009 10th International Symposium on, 978-1-4244-5403-7, 14-16 Dec. 2009, pp. 752 - 756.

[10] Z. Wei and J. JaJa, A fast algorithm for constructing inverted files on heterogeneous platforms, *J. Parallel Distrib*. Comput. 2012.

[11] W. Lingkun, L. Wenqing, X. Xiaokui, and X. Yabo, LSII: An indexing structure for exact real-time search on microblogs, in Data Engineering (ICDE), IEEE 29th International Conference, 2013.

[12] Q. Bai, C. Ma, and X. Chen, A new index model based on inverted index, *Software Engineering and Service Science (ICSESS), 2012 IEEE 3rd International Conference on*, 978-1-4673-2007-8, 22-24 June 2012, pp. 157 - 160.

[13] N. N. Sophoclis, M. Abdeen, E. S. M. El-Horbaty, and M. Yagoub, A novel approach for indexing Arabic documents through GPU computing, *Electrical & Computer Engineering (CCECE)*, 2012 25th IEEE Canadian Conference on, 978-1-4673-1431-2, April 29 2012-May 2 2012, pp. 1- 4.

This paper may be cited as:

Bolhasani, S. and Naderi, H., 2014. Simultaneous Use of CPU and GPU to Real Time Inverted Index Updating in Microblogs. *International Journal of Computer Science and Business Informatics, Vol. 12, No. 1, pp. 25-31*.