Scalable Approaches for Recommendation in Social Networks

Yifan Li

Scalable Approaches for Recommendation in Social Networks

Yifan LI*
LIP6, Université Paris 6
Paris, France
yifan.li@lip6.fr

ABSTRACT

Nowadays, the online social network has become a significant part of our life, and deeply influenced our activities in many aspects than one might imagine before. Hereby, with the benefit from its increasing growing, there are more prediction opportunities emerging for recommendation system deployment. Rather than as only a supplementary of the traditional use of *collaborative filtering* (CF) method, in some cases[7], the social affinity among users can provide more precision in recommendation result(scores) calculation.

In this doctoral thesis, we propose to design a scalable recommendation approach over large social graph, taking into consideration not only graph topological properties but also those user semantic content, e.g. user interest, (hash-)tags, item ratings, to gain a good and fast recommendation evaluation.

Keywords

Social Network, Recommendation System, Graph, Pregel, Personalized PageRank

1. INTRODUCTION

Social network, e.g. Facebook, Linkedin and Weibo, as a new but sharply growing media, has permeated our lives. For instance, by April 2014, the main micro-blogging service in the world, Twitter, has build a social graph consisting of more than 570 million users. Moreover, around one million new accounts are registered in Twitter every week, and 500 million fresh tweets are posted every day. Hence, this rapidly increasing data volume and natural structure dynamism present twofold serious challenges to our existing data management system, 1) *distributed* graph storage and computation and 2) *scalable* algorithm performing.

To deal with the first problem: most of social graphs are too large to fit in memory and disk on a generic machine, several distributed graph computation models, such as Pregel [6], PowerGraph[4], GraphX [8], have been proposed to meet this requirement. The graph partitioning strategy study, however, has *not* been fully conducted in these systems in which it is indeed the basis for distribution. For this reason, our first task is to propose an efficient flexible partitioning approach for processing large social graphs in a distributed way, in stead of a simple *hash function* method with *high communication cost*.

For answering the query for a given node in social graph, we need to evaluate the calculation of *topological similarity* between two nodes in graph. Whereas the volume of the social network and its dramatical growth raise scalability issue. However these existing graph algorithms don't give sufficient weight on either increasing *computation parallelism* or exploring *topological feature* inside graph partition. Thus, in my thesis we plan to propose an segments-based similarity computation algorithms that can be implemented on large scale graph processing systems that offer transparent scalability, high computation parallelism and fully using of graph partitioning feature.

*supervised by Prof. Cedric du Mouza(CNAM, Paris) and Prof. Camelia Constantin(LIP6-UPMC, Paris)

BDA 29 September 2015, Porquerolles, France.

2. SCALABLE GRAPH PROCESSING

2.1 Graph Partitioning

Naturally, how to partition a social network is the foundation of a distributed graph processing system. Given a graph application, the partitioning strategy has a huge impact on its performance. From perspective of querying, a good graph partitioning also can advance the query evaluation in a data/computation-distributed context.

Hence, here we proposed a novel workload-aware *Block-based Partitioning Strategy* to acquire targets: 1) Minimised Communication, 2) Balanced Workload

Block Construction using Seeds. In this thesis, a block can be seen as one, or part of tightly knit cluster in graph, e.g. a community in social network, which is formed by several edges, rather than vertices. We propose to find those K inherent blocks B_1, B_2, ..., B_K from graph G by exploring its full potential topology. Especially, there are K pre-selected seed vertices, of which one is per block as its centre. Then we compute the closenesses, from seeds, to the two nodes of every edge, which will be used to determine the ownership of edge finally.

Closeness between Nodes. Since the block an edge belongs is determined by the closeness(namely reachability) from its seed(center) to that edge's two end-nodes. Firstly we propose to offer a method to measure that closeness, by

using a simpler *inverse P-distance*:

$$D_i(j) \equiv \sum_{p \in \{i \sim j\}} S(p)$$

where the p is a path from node i to node j or vice versa, which is formed by a sequence of edges. *S(p)* is the inverse distance value of path p: v_0, v_1, ..., $v_{(k-1)}$ with length k, and is defined as

$$S(p) = (1-\alpha)^k \cdot \prod_{i=0}^{k-1} \frac{1}{outDeg(v_i)}$$

where $\alpha \in (0, 1)$is the teleporting probability in the random walk, which means the probability to return origin. $outDeg(v_i)$ is the out-degree of vertex v_i.

Edge Allocation. For each edge triplet, we will calculate the two closeness-score vectors together to obtain the final result: the block(seed) id it belongs.

Block Refinement. Due to the limits for workload balancing, or computation balancing further, we need to conduct some kind of necessary refinement operations on blocks, e.g. to combine several small ones or split a large one.

Demostration. Up to now, from our existing experiments, the communication measurement[4], *vertex replication factor*, has been reduced to around *40 percent*, and it has also cut the graph processing time in half, compared with the EdgePartition2D in [8] which has best performance generally. Then, we tend to perform more experiments to observe boundary-crossed communication between partitions, esp. for our segments-based algorithm.

In addition, we will further give a fresh analysis on *computation balancing*, rather than storage balancing, for the first time.

2.2 Scalable Similarity Algorithm

The main target for recommendation on a social network is to find the most *similar* node(s) for a given node. This task can be considered from two aspects: 1) searching the topologically close node(s) and 2) searching the semantic-close node(s).

Among those existing graph-based similarity computation algorithms, e.g. Hitting Time, PageRank, Graph Distance, we made an adapted segments-based Monte Carlo method to approximate Fully Personalized PageRank(PPR) to measure the topological *similarity* between nodes [1, 5].

Here, as the ideas in [3, 2], instead of simulating a complete long random walk over the whole graph, several short walks(of lengths geometrically distributed on teleporting factor) are conducted starting from the common node. Finally the approximate PPR value can be obtained by counting the frequencies of visits to each node in short walks. Especially, still, each short random walk can be divided into shorter segments, thus in this way we can increase the parallelism of graph computation which is implemented in Pregel model. Moreover, the considerable short segments(random walk) starting from the same initial node are expected to stay in a common graph partition.

3. FUTURE WORK

3.1 Seeds Selection Strategy

The quality of seeds has a big effect on graph partitioning result, for this reason, we need to make more effort to seek out a way for adaptive seed selection with various graph topologies.

3.2 Personalized PageRank

We have had a straight-forward segments-based random walk simulation algorithm that can be applied to PPR easily, but nonetheless, more optimisations, w.r.t Pregel computation model, are still worthwhile.

3.3 Graph Evolving

The dynamism is another big challenge in computing large graphs of real world, social network etc. We plan to extend the recommendation model to deal with the change of social network structure and user content, with necessary but acceptable additional cost.

4. REFERENCES

[1] K. Avrachenkov, N. Litvak, D. Nemirovsky, and N. Osipova. Monte carlo methods in pagerank computation: When one iteration is sufficient. *SIAM J. Numer. Anal.*, 45(2):890–904, Feb. 2007.

[2] A. Das Sarma, S. Gollapudi, and R. Panigrahy. Estimating pagerank on graph streams. In *Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '08, pages 69–78, New York, NY, USA, 2008. ACM.

[3] D. Fogaras and B. RÃącz. Towards scaling fully personalized pagerank. In *In Proceedings of the 3rd Workshop on Algorithms and Models for the Web-Graph (WAW*, pages 105–117, 2004.

[4] J. E. Gonzalez, Y. Low, H. Gu, D. Bickson, and C. Guestrin. Powergraph: Distributed graph-parallel computation on natural graphs. In *Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation*, OSDI'12, pages 17–30, Berkeley, CA, USA, 2012. USENIX Association.

[5] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th International Conference on World Wide Web*, WWW '03, pages 271–279, New York, NY, USA, 2003. ACM.

[6] G. Malewicz, M. H. Austern, A. J. C. Bik, J. C. Dehnert, I. Horn, N. Leiser, G. Czajkowski, and G. Inc. Pregel: A system for large-scale graph processing. In *In SIGMOD*, pages 135–146, 2010.

[7] A. Olteanu, A.-M. Kermarrec, and K. Aberer. Comparing the predictive capability of social and interest affinity for recommendations. In B. Benatallah, A. Bestavros, Y. Manolopoulos, A. Vakali, and Y. Zhang, editors, *Web Information Systems Engineering âĂŞ WISE 2014*, volume 8786 of *Lecture Notes in Computer Science*, pages 276–292. Springer International Publishing, 2014.

[8] R. S. Xin, J. E. Gonzalez, M. J. Franklin, and I. Stoica. Graphx: A resilient distributed graph system on spark. In *First International Workshop on Graph Data Management Experiences and Systems*, GRADES '13, pages 2:1–2:6, New York, NY, USA, 2013. ACM.