



Scalable Approaches for Recommendation in Social Networks

Yifan Li

► To cite this version:

Yifan Li. Scalable Approaches for Recommendation in Social Networks. BDA2015, Sep 2015, Toulon, France. hal-01398199

HAL Id: hal-01398199

<https://hal.science/hal-01398199>

Submitted on 18 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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, sev-

eral 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.

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, \dots, 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

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

