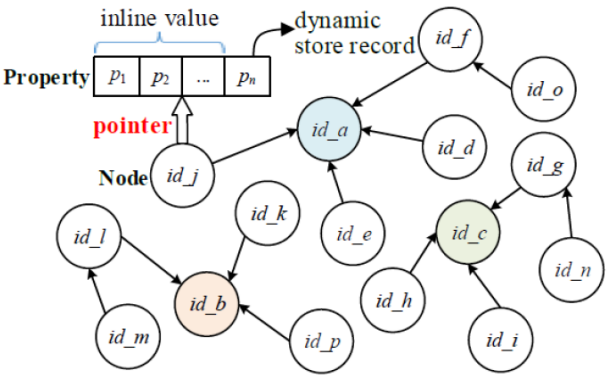
Efficient Graph Data Compression Model based on the Germ Quotient Set Structure

Dawei WANG, Wanqiu CUI

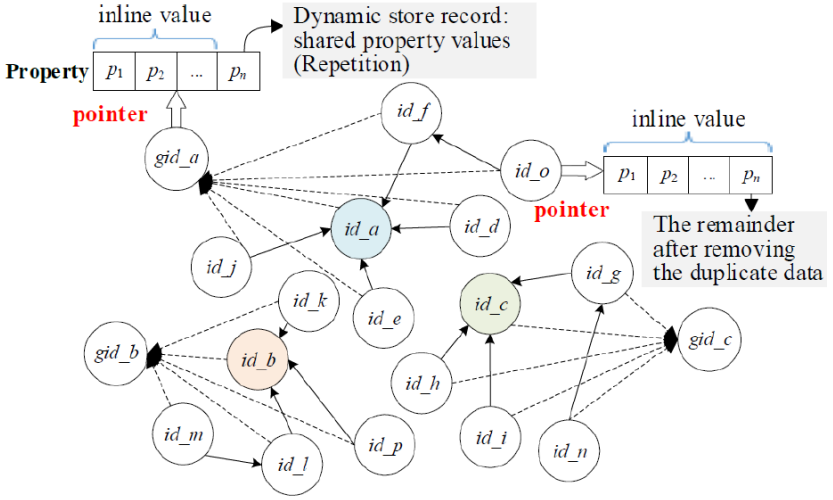
Frontiers of Computer Science, DOI: 10.1007/s11704-022-1489-7

Problems & Ideas

- Problems of existing graph data models:
  - Social networks topic can trigger extension discussion, which leads to a proliferation of highly repetitive data.
  - The property graph model loads all message nodes and its property. It causes a serious waste of space.
- Ideas: the hot data is separated from the original property set and unified storage in germ nodes.



(a)The storage form of the property graph model.



(b)The storage structure of GNode.

The comparison of storage form. Left: *id_a, id_b*, *id_c* are original messages, and other nodes are the retweets. The tweet content of nodes include duplicated data and stored in different dynamic store record p_n; Right: *gid_a*, *gid_b*,and *gid_c* are germ nodes. The duplicate part is stored by the germ node and shared by other associated nodes. The message nodes store the remainder in p_n. Thus, GNode reduces the property storage and realizes the space compression.

Main Contributions

- Contributions:
  - We design a graph data compression model GNode aiming at the intensive graph in a social network. It eliminates data redundancy and achieves compressed storage;
  - We extract representative data from the relationship network and define the germ quotient set structure based on germ node;
  - The evaluate results indicate that GNode can achieve effective compression storage for social networks graph data.

Table 1 Compression ratios of the GNode model.

Dataset	#Tweets	#Compression objects	Size of stored data in the property graph model	Size of stored data in GNode	Compression ratios
Dataset I	70,347	18994 (27%)	7.82M	5.83M	25.4%
Dataset II	138,599	63756 (46%)	15.40M	9.06M	41.2%
Dataset III	215,276	133471 (62%)	23.92M	12.01M	49.8%

Compression ratios of the GNode model. The third column is the number of compression objects and its proportion. The size of the file stored in GNode is significantly smaller than the size of the property graph model by comparing columns 4 and 5. The last column in the Table 1 is the compression ratio of GNode for three datasets.