Description of Datasets

Md. Saiful Islam

27 November 2019

1 Datasets

We use two large real datasets: `OAG` (Open Academic Graph)[1] [1] and `Gowalla`[2] that reflect the real life application scenarios to evaluate our proposed algorithms for keyword-aware influential community search.

`OAG` is generated by linking two large academic graphs: Microsoft Academic Graph (MAG) and AMiner. Here, we represent each author as a vertex and co-authorships between authors as edges. This dataset contains more than 150 million academic articles with metadata like title, abstract, authors, keywords provided by authors, etc. We first choose $1,000$ most frequent author-provided keywords as the set of keywords for the attributed graph. Each vertex of the graph is augmented with the semantically relevant set of keywords for the author provided keywords. The score is modeled as the author's percentile rank (scaled to 1.0), considering the number of citations in publications relevant to a keyword. We skipped non-English articles and the articles with no citation. Finally, there were $10,714,737$ articles in our dataset. In the attributed graph, we considered the first 1 million authors as vertices and $15,677,940$ co-authorship relations among the authors as edges.

In `Gowalla` dataset, users are modeled as vertices, and the location ids are considered as keywords. Edges represent the friendship between two users. Each user is augmented with the locations where she checked in. The influence score of a user at a location is modeled as the user's percentile rank considering the number of check-ins posted by the user at that location. There are $407,533$ vertices, $2,209,169$ edges, and $2,727,464$ keywords in this attributed graph.

2 Query generation

Regarding query generation, to generate a query for `OAG` datasets, first, we choose the number of query terms randomly within [1,3]. Then we randomly choose a keyword from the most frequent $10,000$ author-provided keywords. This keyword is taken as the first query term. We choose the rest of the query terms randomly. Any keyword that appears with the first query term in the

[1] https://aminer.org/open-academic-graph
[2] http://www.yongliu.org/datasets/index.html

author-provided keyword list is a candidate to be chosen as a query term. Additionally, we augment each keyword with its semantically similar keywords using our `semantic keyword similarity model`. For each set of query terms, we take both AND and OR predicates. Note that low-frequency keywords are usually very specific, and when we consider AND predicate among them, the communities may not be meaningful. For example, queries like **"game simulation" AND "soil heat storage"** do not yield meaningful communities. For this reason, we do not choose straightforward random keywords as a query. Parameters r and k_{min} are set to default values unless otherwise specified.

For `Gowalla` dataset, we randomly choose a location as the first keyword, and then choose 0-4 additional random locations within the radius of $5km$ from the first location. We do so because if the locations are very far from each other, there will be very few users who visit all the locations, thereby making the AND queries meaningless. For the entire query setup, we only consider the locations where at least 50 users checked in.

References

[1] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: extraction and mining of academic social networks. In *SIGKDD*, pages 990–998, 2008.