Description of Datasets

Md. Saiful Islam, Mohammed Eunus Ali, Yong-Bin Kang,
Timos Sellis, Farhana M. Choudhury

07 July 2020

1 Datasets

We use three large real datasets: OAG (Open Academic Graph)[1] [1], Twitter[2] and Gowalla[3] 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.

Twitter dataset is a social network graph of 140,371 users with 2,283,875 follow relations among the users. This dataset also contains the tweets published by these users. Due to Twitter's privacy policy, any dataset revealing both the users and their tweets can not be shared publicly. However, we share the processed attributed graph where the users are given untraceable identifiers. We process each tweet to filter out tags, website links, and very short words (1-3 characters). Then, we apply tokenization that converts each tweet into a set of token streams. We apply lemmatization on each token to find the root word, and it's parts-of-speech. Finally, we carefully parse the nouns which are used as keywords in our dataset. We only consider the most frequent $10,000$ keywords for building the attributed graph. The impact of each tweet is evaluated by

[1] https://aminer.org/open-academic-graph

[2] Raw dataset cannot be shared publicly due to Twitter's privacy policy. Please contact the authors if needed.

[3] http://www.yongliu.org/datasets/index.html

the number of retweets. We calculate the total number of retweets a user gets from all the tweets associated with a single keyword. In this way, we can assess how popular the user is in Twitter if we consider only that single keyword. We use the percentile of a user's popularity in a certain keyword (topic) to model his/her influence in that keyword. For example, if a user gets more retweets from "music" related tweets than 98% other users, the user's influence in "music" keyword is modeled as 0.98.

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.

Attributed graphs are publicly available at `https://drive.google.com/drive/folders/1yU32kH6E2xvQow8hxCbCtMoFASDVn4TE?usp=sharing`.

2 Query generation

To generate a query for `OAG` dataset, 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 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.

Queries for `Twitter` dataset are generated similarly as the `OAG` dataset. Keywords in a query are carefully chosen such that irrelevant keywords are not grouped together to ensure that we are searching for meaningful communities. We augment each keyword with its semantically similar keywords as well.

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.