Overview of Keyword Aware Influential Community Search

Md. Saiful Islam, Mohammed Eunus Ali, Yong-Bin Kang,
Timos Sellis, and Farhana Murtaza Chowdhury

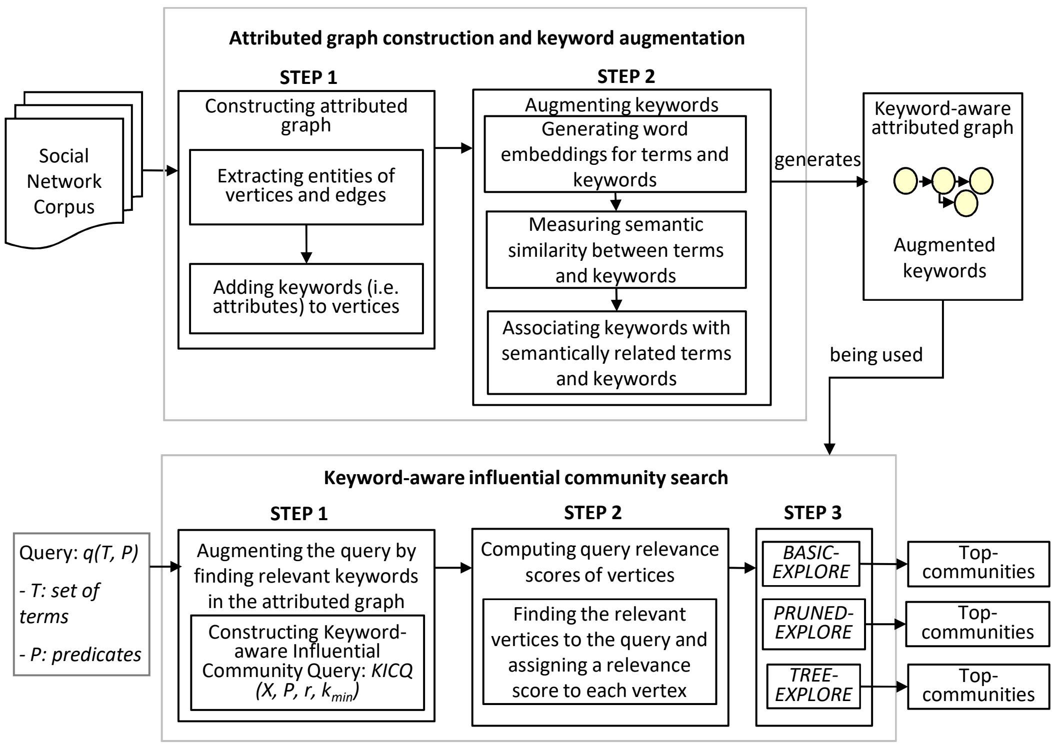

Figure 1: The overview of our system

An overview of our system is presented in Figure 1. The system is mainly divided into two phases. First, we construct a keyword-aware attributed graph from a social network corpus that may consist of a combination of structured and/or unstructured (i.e., text) data. In an academic domain, the corpus can be scientific publications of researchers (e.g., authors, titles, abstracts, author-provided keywords, etc). Second, we focus on searching keyword-aware influential communities using the constructed attributed graph, given a query indicating search conditions in terms of keywords and predicates.

The distinctive features of the first phase can be outlined as follows:

STEP 1 : First, we build an attributed graph from a domain corpus by extracting entities of possible vertices and edges to represent the social network. One example of such a graph can be seen as an academic graph, where vertices represent authors and edges represent co-authorships between two authors. Each vertex is associated with the relevant keywords (i.e. attributes) of the corresponding author. Here, such keywords represent the expertise topics of the authors in the graph.

STEP 2 : To enable keyword-aware influential community search, we augment keywords with their semantically related terms and keywords. These augmented vocabulary will be used for identifying $X \in KICQ$. For this purpose, we build *word embedding vectors* as an external knowledge source for associating keywords in the graph with semantically related terms and/or keywords. These vectors are generated from a large text corpus relevant to the source of the graph using the algorithm [1] that has shown the power of encoding semantics of words. The semantic relatedness is estimated by exploiting the word embedding vectors. These augmented keywords will enhance the capability of our community search algorithms. The output of this step is a graph, called *keyword-aware attributed graph*.

Further, the unique features of the second phase of our system can be briefly highlighted below:

STEP 1 : Initially, a query raised by a user is given in the form of a pair $q(T, P)$ consisting of a set of query terms T, and a predicate P. The terms need to match with the keywords in the keyword-aware attributed graph to find meaningful communities. We acknowledge the difficulty faced by the users to put the exact terms while raising a query. For example, it is highly likely that some of the users will input "song" instead of "music". To help the users to easily raise a query, we propose augmenting each query term with a semantically meaningful set of representative keywords. The output of this step is a KICQ.

STEP 2 : Given a KICQ query, our objective is to find the vertices relevant to the query and then compute their query relevance scores. Relevance score of vertices are used to compute the scores of potential most influential communities. We argue that a measure that rewards both the cohesiveness of the community and high influence of the members, and does not require user input of any internal parameters (e.g., k in k-core) is more preferable than the existing influence measures. We propose a linear weighted summation of the cohesiveness of the community and the total influence of the members of the community to estimate the overall score of a community.

STEP 3 : Given the augmented query and the influential score function, our focus is now to retrieve top-r most influential communities relevant to the query. Since we are the first to propose the keyword-aware influential community search problem, and existing pre-computation based approaches are not suitable to retrieve communities for any given query, we first present a basic solution named `BASIC-EXPLORE` followed by two efficient algorithms: `PRUNED-EXPLORE` and `TREE-EXPLORE`.

References

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th*

International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, pages 3111–3119, USA, 2013. Curran Associates Inc.