A Proposal for Analyzing Mobile Social Networks in the Physical World

Győző Gidófalvi
gyg@geomatic.dk

May 27, 2007

Abstract

Social networks – structures whose nodes represent people or other entities embedded in a social context, and whose edges represent interaction, collaboration, or influence between entities – have been studied by researchers in mathematics and the natural and social sciences since the 1930's [2]. Mainly due to their availability and ease of processing, most recent studies in the area focus on social networks, whose edges are explicitly present in some digital media such as publications linking authors, or project assignments linking employees in companies, etc. However, for the most natural social network, people meeting in the physical world, until today no such links exist. With the advances of mobile technology, such as accurate short-range or satellite–based localization methods, datasets recording the links implicitly are expected to become available in the future. These datasets are expected to (1) be highly dynamic and very large, (2) have links that are annotated with attributes such as space, time and duration providing a context and strength for the interaction. The aim of this paper is to investigate existing analysis methods and their applicability to the future datasets.

1 Introduction

Social networks have been studied in the academic circles since the 1930's. Since the study of social networks is very much related to graph theory, it is not surprising that the academically most well–known social networks is the network of research collaborators of Paul Erős, the most prolific mathematician of the twentieth

century. Professionals in other disciplines also form and maintain social networks, f.ex., the Digital Bibliography & Library Project (DBLP) has been established to organize authors, institutions, publications, proceedings and conferences/journals in the field of computer science. The importance of social networks has also been recently placed in the focus of the public eye through popular books [1]. Hence, concepts like *six degrees of separation*, the *Kevin Bacon game*, and *academic genealogy* became common knowledge.

The public's interest to build and maintain social networks is so fundamental that websites facilitating these activities, such as Facebook, MySpace, Friendster, YouTube, Flickr, Blogger, are amongst the fasters growing and most successful websites [4]. The user–generated content, i.e., the links to the entities of interest or the entities themselves, vary from website to website, but they all establish a social network amongst the users. The benefits of automatic analysis of these social networks are manyfold. Such analysis can give rise to semantic representations (meta data) of real-world entities and complex concepts that can be processed automatically by machines. Such semantic representations can aid the development in machine intelligence, and related issues such as the semantic web.

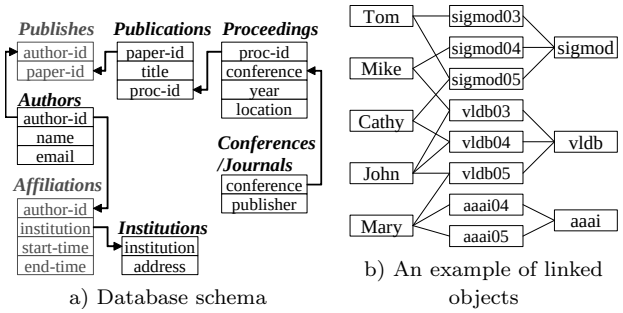
Clustering, the process of partitioning objects into groups according to their similarities with eachother, have been studied extensively in the past. In many real–world applications, the most explicit information about the similarity between objects is in the form of linkages among objects of different types. To this extent, Yin et. al proposed a linkage–based clustering of objects of different types [3], which is reviewed in Section 2.

On the above social networking websites users *explicitly* express their interest by linking to eachother and to the entities of interests. For the most natural social network, people meeting in the physical world, no such explicit links are available. However, with the advances in mobile and positioning technologies, data sets about this physical social network are expected to become available in the future, where *collocation* of entities in the physical space will imply implicit links between the entities. The main entities in this physical social network are `person`s, `place`s, and `social` and `commercial activitie`s. In the following this natural social network will be referred to as Mobile Social Network (MoSoNet). Based on the unique characteristics of MoSoNet and the reviewed linkage–based clustering method, Section 3 proposes a distributed architecture for gathering and analyzing relevant linkage information of MoSoNet.

2 Clustering via heterogenous semantic links

Clustering, the task of discovering natural groupings of objects according to their similarities based on object attributes and / or hierarchies have been studied for decades. In many real–world applications, the most explicit information about the similarity between objects is in the form of linkages between objects of different types. This section gives a high level overview of a recently proposed method that uses these linkages to cluster objects of different types [3].

As an illustrative example, consider the task of clustering entities (authors, institutions, publications, proceedings, and conference/journals) in a publication database depicted in Figure 1(a). As an example of linked objects in the publication database consider Figure 1(b). While clustering objects based on their attributes alone is possible, the linkage information between objects is likely to be more relevant to the clustering task. The links can be used to cluster objects of one type based on their links to objects of another type. However, clustering objects based on *direct* links may not capture the true similarity between objects. For example, because Tom publishes only VLDB papers and James publishes only SIGMOD papers, since the similarity between the two conferences is not considered, Tom and James will have zero similarity, while possibly working on the same topic.



a) Database schema

b) An example of linked objects

Figure 1: A publication database

To overcome this shortcoming, an earlier linkage–based clustering algorithm, *SimRank*, adopts a recursive definition of similarity of objects as follows: the similarity between two objects is the average similarity of objects linked with them. As illustrated in Figure 2, final goal of *SimRank* is to group entities together based on *indirect* links using the this recursive linkage–based similarity definition. Al-

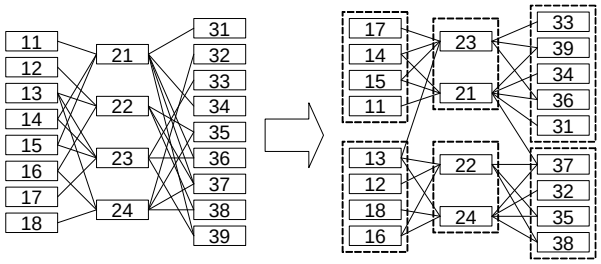
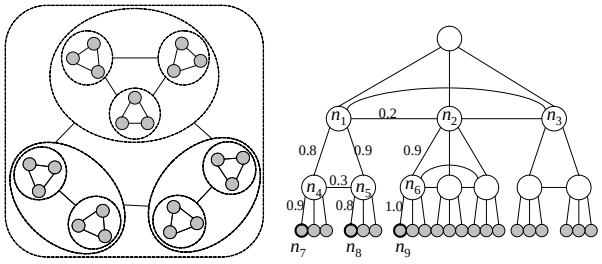


Figure 2: Indirect linkage–based clustering

though objects 12 and 18 do not have common neighbors, they are considered similar, because they both link to objects 22 and 24, which are similar due to their common linkages.

While effective in terms of the clustering task, *SimRank* is computationally not efficient as it computes similarities between every pair of objects in an iterative manner. Based on the observations that 1) object hierarchies naturally exist for many types of objects, and that 2) there exists a power law distribution among linkages in social networks, Yin et. al propose a hierarchical data structure called *SimTree*, to represent similarities between objects in a compact way. Two views of an example *SimTree* are shown in Figure 3. Leaf nodes in the *SimTree* correspond to objects, and non–leaf nodes group similar, lower–level nodes together. Instead of computing and storing the pairwise similarities between objects, the only measures that a *SimTree* stores are 1) the similarities between sibling leaf nodes, 2) the average similarities between sibling non–leaf nodes, and 3) the specificity of each child node compared to its parent, i.e., the ratio between the similarity of the child to all of the objects and the similarity of the parent to all of the objects. Similarities between two non–sibling leaf nodes are estimated as the products of the above measures along the shortest path between the two nodes in the *SimTree*. Note that in the above measures all similarities are linkage–based, and linkages are inherited from leaf nodes along the hierarchy.



a) Structure of a SimTree b) Another view of the SimTree

Figure 3: Two views of an example *SimTree*

Using *SimTree*s, Yin et. al propose an efficient and effective linkage–based clustering method, *LinkClus*. *LinkClus* consists of two phases: an initialization phase, and an iterative improvement phase. In the initialization phase, a *SimTree* is built for every type of object in a bottom–up fashion by greedily grouping nodes (objects) based on the intersection size of their neighbors. Since in the initialization phase inter–object similarities (indirect linkages) are not considered, in the second phase of *LinkClus*, the *SimTree*s are iteratively adjusted / restructured based on the rule that two nodes are similar if they are linked with similar objects.

3 Proposal for analyzing entities in MoSoNet

As it is briefly described in Section 1, advances in mobile technology and localization methods allow the gathering of implicit linkage information (based on collocation of entities) in MoSoNet. The number of interactions between objects (links) in MoSoNet are far larger than in the described publication database. The interactions are also more dynamic. For example a `person` might go to a `place` several times indicating the strength of the interaction. Centrally storing and analyzing the individual interactions between all objects are likely to be infeasible and unnecessary. Instead, interactions can be aggregated in a distributed fashion on the smart clients ,carried by the `person`s, as follows. Using satellite–based localization or short–range communication technology together with sensors, smart clients can associate their location with `place`s. Smart client can store temporal statistics about the `place`s they visit. Such statistics can include the time of day, the duration, the frequency and the number of visits. Since a `person`'s interests change over time the statistics can be `aged` to provide an accurate view of the `person`'s current interests. These client–side, temporal statistics about `place`s can be combined to indicate *strengths* of interactions between the `person` and the `place`s visited. Interactions between a `person` and *significant* `place`s can be uploaded to a central server / database. The server then can perform linkage–based clustering taking into consideration the strengths of interactions to derive *global* groups of similar `person`s and `place`s. The hierarchical groupings of `person`s and `place`s can be used as multi–level descriptors of the entities.

Interactions between `person`s can also be detected in a distributed fashion on the smart clients using short–range communication technology. Clients can broadcast their pseudonym and record other clients' pseudonyms that are within range. The clients can also relate *frequent* pseudonyms, *friends* of the `person` carrying the client, to some *social context* through the multi–level descriptors of the `place`s where the interactions take place. Taking into account the similarities between `place`s, and the strengths of interactions between *friends*, the clients then can calculate linkage–based similarities between *friends* to derive groups of similar *friends* in the social context, f. ex., roommates, labmates, party–friends.

It is important to note, that while the global grouping of `person`s and `place`s is server–side knowledge that can be used for example for commercial purposes, knowledge about the a `person`'s friends is derived purely on the client–side and can be used for example for personalization of client–side applications, such as information sharing, blogging, and other common applications in the social networking arena.

4 Conclusion

The popularity and importance of social networking is apparent from the success of social networking sites on the Internet. The analysis of the linkage structure of such social networks can reveal useful semantic information about the entities in such social networks. To this extent, the paper reviews a recent method for clustering of intra–linked objects of different types.

In light of the recent advances in mobile communication and localization technologies, the paper forecasts the near–future availability and unique characteristics of the datasets about a Mobile Social Network (MoSoNet), i.e., mobile users carrying smart phones and interacting with eachother and other entities in the physical world. The main entities in such a MoSoNet are `person`s, `place`s, and `social` and `commercial activitie`s, and interactions are indicated implicitly by collocation of objects. Furthermore, the paper proposes a distributed architecture for gathering and analyzing linkage information in MoSoNet. The proposed analysis architecture can reveal: 1) *public* knowledge about global similarities and groupings of `person`s and `place`s, and 2) *private* knowledge about similarities and groupings of *friends* in the social context.

References

[1] A.-L. Barabasi. "Linked: How Everything Is Connected to Everything Else and What It Means." Perseus Books Group, 2002, ISBN-10: 0738206679

[2] M. S. Martin and C. Gutierrez. A Database Perspective of Social Network Analysis Data Processing. Presented at *Sunbelt*, 2006.

[3] Y. Xiaoxin, H. Jiawei, and Y. S. Philip LinkClus: Efficient Clustering via Heterogeneous Semantic Links. In *Proc. of VLDB*, pp. 427–438, 2006.

[4] M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *Proc of SIGCHI*, pp. 970–981, 2007.