Potential Collaboration Discovery using Document Clustering and Community Structure Detection

Cristian K. dos Santos
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
+552125627388
c.klen@coc.ufrj.br

Alexandre G. Evsukoff
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
+552125627388
evsukoff@coc.ufrj.br

Beatriz S.L.P. de Lima
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
+552125627388
bia@coc.ufrj.br

Nelson F. F. Ebecken
COPPE/UFRJ
Federal University of Rio de Janeiro
Rio de Janeiro, Brazil
+552125627389
nelson@ntt.ufrj.br

ABSTRACT
Complex network analysis is a growing research area in a wide variety of domains and has recently become closely associated with data, text and web mining. One of the most active areas in the study of complex networks is the detection of community structure, which can be related to the clustering problem in data mining. This paper employs a community structure detection algorithm for document clustering in order to discover potential relationships in a social network. The proposed approach is explored in a case study of potential collaboration discovery among the research staff in the Graduate Civil Engineering Department of the Federal University of Rio de Janeiro, Brazil. The results show that the combined use of both techniques provides useful insights on the relationships, both existent and potential, among individuals in the social network.

Categories and Subject Descriptors
H.3.4 [Systems and Software]: Information networks, H.2.8 [Database Applications]: Data Mining; H.3.3 [Information Search and Retrieval]: Clustering; I.7.5 [Document Capture]: Document analysis.

General Terms
Algorithms, Management, Documentation, Human Factors.

Keywords
Community structure detection, complex networks, text mining, documents clustering, spectral clustering.

1. INTRODUCTION
The Internet has allowed social networking to become a worldwide phenomenon that integrates people who would probably never be connected through their conventional social acquaintances.

Individuals are very willing to express themselves in online social networks. The ways they understand the world are expressed by their opinions and thoughts, which are shared, with more or less agreement, with many others. Most content is released in the form of text in a wide variety of formats, which can be useful for the discovery of potential links in a social network. The discovery of potential relationships in a social network stimulates the integration of people and also enhances the information flow in the network.

One of the most active areas in the study of complex networks is the detection of community structure [1]-[8]. The graph theoretic approach is widely used for community structure detection in complex networks and is also the base formalism for spectral clustering [9]-[11], so it is a natural way to integrate these techniques.

The concept of a good clustering or community partition is very difficult and can be formally defined in many ways, so that many different algorithms can be derived. In the graph theoretic approach, the algorithms are usually formulated as graph partition problems, in which the weight of each edge is the similarity between points that correspond to vertices connected by the edge. The goal of this algorithm is to find the minimum weight cuts in the graph, which is a combinatorial problem. The problem is thus usually addressed through spectral decomposition techniques, as described in recent excellent reviews on the subject [10][11].

The most popular class of methods to detect communities is, perhaps, the maximization of the function known as "modularity," introduced by Newman and Girvan [1]. This measure is by far the most used and best-known function to quantify the "goodness" of possible subdivisions of a given network into communities [1]. The modularity measure is, however, not able to detect very small communities, as it has been recently pointed out [6].

Community detection algorithms have been recently studied for document clustering [12][13], where the Newman algorithm [3] was compared to spectral clustering techniques. The Newman algorithm produced better results.

The main contribution of this work is methodology to integrate document clustering and community detection for the discovery of potential relationships in a social network. The results are explored in a case study of potential collaboration discovery in the Graduate Civil Engineering Department of the Federal University of Rio de Janeiro. The Newman algorithm is used both for community detection in the co-authorship network and for document clustering. The co-authorship network is obtained from the collaboration of Professors in MSc and PhD thesis supervision. The corpus used to identify potential links in this

CNIKM'09, November 6, 2009, Hong Kong, China.

network is the set of abstracts of MSc and PhD theses produced in the department from 2004 to 2008.

The problem of potential link discovery in networks has being studied in the recent literature [14]. In the Relational Topic Model [15], it is also possible to predict links using texts' content. The authors have also developed a method to uncover the relationships encoded in a collection of texts using an approach based on a probabilistic topic model [16], which allows inferring descriptions of the network's elements. Kemp et al. [17] have presented an approach to cluster one or more sets of entities and discover the relationships between clusters that are possible or likely.

In this work, as so as in related approaches [14-17], use not only the link structure, but also the features of the network's elements, in this case textual documents, in order to analyze the network for link prediction or discovery. This seems to be more effective as additional information is included into the analysis instead os using only the link structure to predict links [18].

The paper is organized as follows. The proposed methodology is presented in the next section. The modeling of a document collection as a graph and the formulation of the document clustering as a graph cut problem are presented in section three. In section four the Newman algorithm is introduced, and in section five the case study is discussed. The paper finishes with conclusions and future studies in section six.

2. POTENTIAL LINK DISCOVERY

As is typical in complex network analysis, a co-authoring network definition starts with a bi-partite graph defined over two sets of objects [23]. In this work there are three sets of objects, as shown in Figure 1. The first set is the set of individuals of the social network under study. Each document in the set of documents is related to one or more individuals. The set of terms may be a set of keywords within a controlled dictionary or, more generally, the set of generic terms appearing in the documents.

In the case study presented in section 5, the set of individuals are the permanent staff and external collaborators of the Graduate Civil Engineering Department of the Federal University of Rio de Janeiro. The set of documents are the abstracts of the MSc and PhD theses produced in the department from 2004 to 2008. The relationships between individuals and documents are the collaborations in thesis supervision. Each document can have up to three supervisors. The set of terms are the words (stems) appearing in the documents.

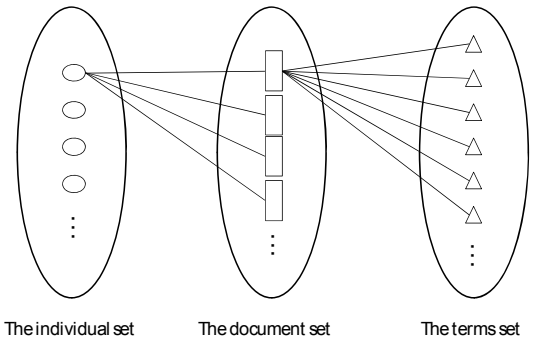

Figure 1. The sets of objects for the definition of the networks in the proposed method

Three kinds of networks can be defined from the object sets shown in Figure 1. The first one is the collaboration network, which is a kind of co-authorship network, in which the nodes are the individuals and a link represents the co-supervision between two individuals in at least one document. This is the base social network used in this study and represents the existing relationships among the individuals in their social interactions. In a more general setting, the existing relationships can also be represented by other kind of networks, such as hyperlinks in a set of blogs, friendship in a social network web service, or emails. The base network may also be weighted or unweighted, directed or undirected. In the case study presented in this work, the base network is un-weighted and undirected.

The second network is the document network, in which the nodes represent the documents and the links represent the similarities among documents, as determined by the terms appearing in the documents. This network is generated artificially using a threshold in the similarity value representing a strong similarity between two vertices. The document network is weighted and undirected.

The third network is a combination of the first two, in which the nodes are the individuals and each link represents the similarity between the content produced by the two individuals. The community structure in this network reveals groups of individuals interested in the same subjects. The comparison of the network structure found for this network with the structure of the base network makes it possible to reveal potential relationships that are not yet present in the base network.

The definition of the document network plays a central role in the methodology described in this work and is discussed in the next section.

3. SPECTRAL CLUSTERING IN DOCUMENT NETWORKS

Unstructured information in document databases presents intrinsic characteristics such that data mining algorithms must be adapted to solve text-mining tasks. The most usual representations for text mining rely on the vector space model of documents, usually in information retrieval [19]. In such a model, the order of words is not considered, and each document in a collection is represented by a vector, in which the components are related to relevant words appearing in the document collection.

In the vector space model, the document collection is represented as the $n \times m$ sparse matrix $\mathbf{X}$, in which the lines are related to the documents and the columns are related to the terms. An element x_{ij} accounts for how the term T_i is related to the document D_j, often computed by the tf-idf frequency [20].

3.1 The document network

A document collection can be viewed as a complex network, in which the nodes are the documents and the edges are weighted according to document similarities.

The document network can be defined from $\mathbf{X}$ as a weighted and undirected proximity graph $G(V,E)$, in which the set of vertices $V = \{v_1,...,v_n\}$ corresponds to the n documents and the set of edges E is defined through the symmetric adjacency matrix $\mathbf{A} \in \Re^{n \times n}$. Each element entry $a_{ij} \in \mathbf{A}$ represents the pair-wise similarity between the documents D_i and D_j, computed as:

$$a_{ij} = \begin{cases} h(\mathbf{x}_i, \mathbf{x}_j) & \text{if } i \neq j \text{ and } h(\mathbf{x}_i, \mathbf{x}_j) \geq \varepsilon, \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The function h measures the local neighborhood relationships between two vertices, which should be greater than or equal to the parameter ε that defines the radius of proximity among the documents. This parameter is very important in the definition of the document network structure since $\varepsilon = 0$ defines a complete network. Different results are obtained with different values of ε; this issue is further exploited in section 5.

The similarity function h can be computed by different functions. In spectral and kernel clustering literature, the Gaussian similarity function is usually employed. In text mining applications, the cosine similarity function is usually employed within the vector space model [20]. The cosine similarity function has shown good results in previous studies of document clustering within a framework of spectral clustering [22]. It is defined as:

$$h(\mathbf{x}_i, \mathbf{x}_j) = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\sqrt{\langle \mathbf{x}_i, \mathbf{x}_i \rangle \langle \mathbf{x}_j, \mathbf{x}_j \rangle}} \quad (2)$$

where $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j$ is the scalar product.

3.2 Spectral clustering

The graph cut problem aims to separate a subset of vertices $S \subset V$ from its complement $V - S$ denoted by $\overline{S}$ [9][21]. The graph cut problem can be formulated in several different ways, depending on the choice of the objective function to be optimized [11][21]. One of the options is the cut function, whose minimization favors partitions containing isolated vertices. It is defined as follows:

$$cut(S, \overline{S}) = \sum_{i \in S, j \in \overline{S}} a_{ij} \quad (3)$$

To overcome the weakness of the cut function and achieve better balance between partitions, it is recommended to use its normalized version, that is, the normalized cut function [21]:

$$Ncut(S, \overline{S}) = cut(S, \overline{S}) \left(\frac{1}{vol(S)} + \frac{1}{vol(\overline{S})} \right) \quad (4)$$

where $vol(S)$ is the volume of S, computed as:

$$vol(S) = \sum_{i \in S} d_i \quad (5)$$

The degree d_i of a vertex $v_i \in V$ is the number of edges incident to the vertex and is defined as:

$$d_i = \sum_{j=1}^{n} a_{ij} \quad (6)$$

The minimization of the function (5) is an NP-hard problem that can be relaxed by introducing the graph Laplacian matrix [11][21].

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \quad (7)$$

where the degree matrix $\mathbf{D}$ is defined as the diagonal matrix of the degrees $d_1, ..., d_n$.

The graph Laplacian is a positive semi-definite matrix, such that its eigenvalues are always positive real-valued. Some spectral clustering approaches are based on the solution of the generalized eigenvalue problem [1]:

$$\mathbf{LU} = \mathbf{\Lambda DU} \quad (8)$$

where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues, which are ordered in ascending order, $0 \leq \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$. The orthogonal matrix $\mathbf{U}$ is the matrix in which the columns are the generalized eigenvectors.

Good clustering algorithms for graph clustering depend on the quality of the objective function being used. Recently, a cost function called the modularity function was proposed by Newman and Girvan, [1] to overcome limitations of the previous measures for measuring community structure, as discussed in the next section.

4. MODULARITY-BASED COMMUNITY DETECTION

4.1 The modularity and the community structure in networks

A community structure in a network $G(V, E)$ is defined as a partition P_K of the set of vertices into K subsets $C_j, j = 1...K$, such that $\bigcap_{j=1...K} C_j = \varnothing$ and $\bigcup_{j=1...K} C_j = V$.

One can think about group structure in graph clustering problems as clusters with high density of edges within them, and a lower density of edges among them.

Newman and Girvan [1] defined a quantitative measure called modularity to evaluate an assignment of nodes into communities. This measure can be used to compare different assignments of nodes into communities. The network modularity Q is defined over a network partition P_K as:

$$Q(P_k) = \sum_{j=1}^{K} \left(\frac{R(C_j, C_j)}{R(V, V)} - \left(\frac{R(C_j, V)}{R(V, V)} \right)^2 \right) \quad (9)$$

where $R(C', C'') = \sum_{i \in C', j \in C''} a_{ij}$ measures the association among the nodes of the subsets C' and C''. Thus, $R(C_j, C_j)$ measures the within-community sum of edge weights; $R(C_j, V)$ measures the sum of weights over all edges attached to nodes in community C_j; and $R(V, V)$ is the normalization term that measures the sum over all edge weights in the entire network.. Considering binary weights, the first term $R(C_j, C_j) / R(V, V)$ is the empirical probability that both vertices of a randomly selected edge fall in subset C_j. The second term $\left(R(C_j, V) / R(V, V) \right)^2$ is the empirical probability that only one of the ends (either one) of a randomly selected edge falls in subset C_j. Thus, the modularity measures the deviation between observed cluster structure and what could be expected under an independent random model. If the number of within-community edges is no better than random, then the value $Q = 0$. A value of $Q = 1$, which is the maximum,