Detecting and Understanding Web communities

Athena Vakali
Aristotle University of Thessaloniki,
Department of Informatics
AUTH Campus Thessaloniki, Greece
avakali@csd.auth.gr

Yiannis Kompatsiaris
Informatics and Telematics Institute
Centre for Research and Technology
Thermi-Thessaloniki, Greece
ikom@iti.gr

Abstract

Web communities comprising of documents and/or users activities have been formed and are continuously expanding and transforming as Web users role shifts from typical navigations to content managing and regulating. Defining, deriving and exploiting communities is not a trivial task since several parameters (large-scale, complexity, evolving information etc) are involved. This paper aims at providing answers for crucial questions raised about communities emerging in the Web and it summarizes different community definitions such that then, the problem of community detection (which is well matured and researched in the past) is understood. The paper emphasizes and discusses the most important methodologies and techniques which deal with large populations of Web documents participating in vast hyperlinked networks, or networks formed from crawling (part of) the web and more recently, networks reflecting the social relations and/or interactions among people. It is important to understand and categorize community identification efforts by taking into account that different levels of granularity and different views are often used for community identification. The emphasis is on the intuition behind all these methodologies and implementations, and on their practical impact for tasks of recommendation, searching, content outsourcing, etc.

Communities definitions & scales

Collective user activities on multiple, often heterogeneous and evolving Web sources contributes in the formation of Web communities which are either derived from Web documents/pages, or by users navigational tasks and more recently by tags and social frameworks. Defining, deriving and exploiting communities is not a trivial task since several parameters (large-scale, complexity, evolving information etc) are involved.

We categorize community identification algorithms based on graph structures under both macroscopic and microscopic perception of Web entities, so communities are identified at the following scales :

- document level communities : to deal with the Web content as individual logical documents with internal organization [1]; *microscopic view - networks of unprecedented size and complexity hard to interpret;*
- Web site level communities : to identify communities in the context of a Web site [2]; *from microscopic to macroscopic view; for reducing user-perceived latency;*
- Web level communities. to deal with the World Wide Web as a whole [3], [4]; m*acroscopic view; computation- expensive and hard to apply in a streaming manner.*

A community is typically defined out of a graph structure and the community is defined as a cluster of information which is relevant and/or (hyper-) linked.

Certainly, members of a community are strongly related whereas at the same time they are loosely related with members of the other communities. Therefore, detecting a community is highly relevant with the efforts to measure "within clusters" density versus "between clusters" sparsity. An intuitive and typical definition of a community is that :

Community definition : Having a graph $G=(V,E)$ where V is the set of vertices and E is the set of the edges, a community c is a vertex subset of V, such that for each of the vertices v which belong to c, there are many edges connecting v (strong connectivity) to the other vertices in c and few edges connecting v (weak connectivity) with the vertices in V-c.

Based on this typical structured-oriented definition, a community is defined as a vertex subset such that for all of its vertices, the number of links connecting a vertex to the cluster is higher than the number of links connecting the vertex to the remaining of the graph.

Detecting Communities : methodologies

A community is typically defined out of a graph structure and the obvious choice is to consider a graph-clustering for revealing a community. Community identification algorithms have been employed on several graphs and networks, including networks deriving from Web data (such as emails, user logs etc), from social interactions, from metabolic and gene networks etc.. Therefore, the problem of communities identification is well studied and a variety of graph-clustering algorithms have been presented in the literature.

On the Web scene communities detection is important since a lot of information can be clustered towards improving Web applications and practices (searching, indexing etc). Apart from having communities of documents/pages earlier efforts have focused on identifying communities of users and communities of tags [5], [6]. Typically, such communities are identified by similarity-based clustering techniques which identify appropriate functions which characterize users and relevant tags closeness.

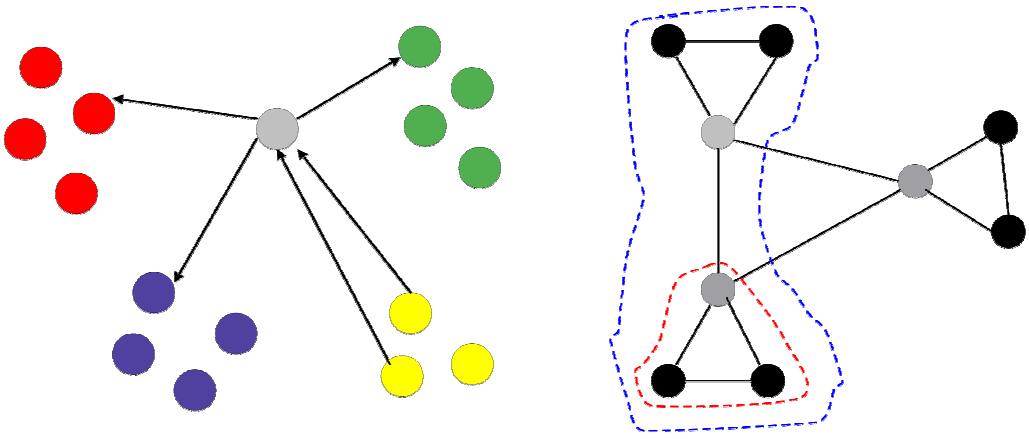
The methodologies used so far in order to tackle the extremely high complexity of providing an exact solution to the community detection problem, *which* has been proven to be NP-hard, primarily focus on :

- *graph-oriented methods* : communities being identified as dense bipartite subgraphs [4] out of a (micro- or macro-scopic) Web graph;
- *divisive-agglomerative methods* : communities revealed by progressive divisive or agglomerative tasks which are guided by certain metrics and criteria which meet community density requirements [7], [8];
- *flow and flooding algorithms* : communities are detected by maximum flow strategies which are used along with iterative crawling processes [3], or by using a flooding process originating from seed/hub nodes [9];
- *hybrid methodologies* : communities are detected by using several of the above methods with emphasis on initiating a community detection by a graph-oriented technique which is adjusted to maximize modularity.

We distinguish the above methodologies into two different approaches in identifying communities based on the regional level on which they work on. More specifically, we highlight the following region scales on which communities are detected :

- *micro-scale methods*, which operate on a graph structure on which they evaluate similarities between pairs of nodes [10]. These methods are often characterized as bibliographic since they evaluate similarities on the basis of the popular to bibliographics metrics namely, on the co-citation and the bibliographic coupling. Such methods cannot scale well to large scale graphs and Web originating datasets since they work on a local/micro level and their span is limited;
- *macro-scale* methods, which operate on a graph as a whole and they consider linkage of the global graph structure which is then divided into sub-graphs based on some nodes which serve to be the starting elements which then attract the other nodes [9]. Such methods are characterized by their spectral focus which enables identifying minimal sets of links which can define a community or minimal cuts which guarantee maximum flows. Such methods are suitable for the Web graph since they work on a large scale view and they may reach optimal solutions especially when additional (metadata, keywords etc) information is given along with the graph structure.

Revisiting Community definition



We raise some issue in terms of the typical community definition which seems to lack ability to handle some non-obvious or non-clearly linked vertices. For example, as authors emphasized in [2] the above definition for community detection fails to include grey vertex (left part of Figure) in one of the four different communities whereas it also fails to identify communities at the right part of the graph.

Therefore, we have proceeded in another perspective for communities definition which follows a more macro-scale generalized approach so we define communities as next.

Global Community definition : Having a graph $G=(V,E)$ where V is the set of vertices and E is the set of the edges, a community c is a vertex subset of V, such that the sum of all edges among the vertices v which belong to c, is greater than the sum of edges which connect the vertices of the community c with the rest of the graph V-c.

This definition identifies communities in the above graph and proceeds with a more global confrontation of graph structures in order to reveal vertices inter and intra connections. Moreover, this definition facilitates proposing an appropriate algorithm for community detection and identification which follows the following steps :

1: Consider a graph $G = (V, L)$
2: Start with each vertex being a community seed $C = \{c_1, c_2, ..., c_{|V|}\}$ with $c_i = \{v_i\}$
3: For each vertex evaluate its linkage and define optimization criterion
4: **while** criterion is not met
   a. Find communities c_i and c_j with greatest intra-linkage
   b. Merge c_i and c_j if they improve global community approach
end while

Web Communities Understanding & Exploiting

Communities understanding and their exploitation is highly relevant with clustering exploitation and as authors have emphasized in [2] and [5], we may propose the following tracks for communities exploitation and utilization :

- *Web users/customers targeted activities* such as market advertisement campaigns, e-commerce attractive notifications targeted to communities of users revealed by Web graph community identification methodologies. The proposed community definition is suitable for addressing users with common activities based on their assignment to communities, and moreover, non-obvious or regular users might be captured by the proposed general-global scope community approach;
- *Recommendation tasks* under specific application frameworks by understanding users' communities trends and profiles. Having communities identified by the proposed global level definition we result in communities with strong linkage and relevance of scope so recommendation acts can be better focused, correlated and user-tailored;
- *Web portals and Web sites management* functionality can be improved since communities of documents may be cached and/or prefetched together and users management can be employed on a community rather than on a single user basis. Such a community-oriented Web portal functionality tuning can improve accessing times and performance whereas at the same time quality of information will be better targeted;
- *Content delivery networking* can be employed in a more advanced manner since detecting communities as proposed here has been proven in [2] to outperform earlier conventional community detection approaches. In such content delivery network frameworks, the detected communities become the core outsourcing units and their appropriate storage improves both performance and infrastructure exploitation rates.

References

[1] N. Eiron and K. S. McCurley. Untangling compound documents on the web. Proceedings of the fourteenth ACM conference on Hypertext and hypermedia, pages 85-94, New York, NY, USA, 2003.

[2] D. Katsaros, G. Pallis, K. Stamos, A. Vakali, A. Sidiropoulos, and Y. Manolopoulos. Cdns content outsourcing via generalized communities. IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 1, Jan. 2009.

[3] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of Web communities. KDD '00: 6th ACM SIGKDD International Conference on Knowledge discovery and data mining, pp. 150-160, 2000.

[4] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. Computer Networks, 31(11-16):1481-1493, 1999.

[5] S. Petridou, V. Koutsonikola, A. Vakali, G. Papadimitriou : Time Aware Web Users Clustering, IEEE Transactions on Data and Knowledge Engineering, Vol. 20, No. 5, pp. 653-667, May 2008.

[6] Eirini Giannakidou, Vassiliki Koutsonikola, Athena Vakali, Ioannis Kompatsiaris, "Co-Clustering Tags and Social Data Sources", Proc. of the 9th Int. Conf. on Web-Age Information Management, IEEE Computer Society Press, 2008.

[7] M. E. J. Newman. Fast algorithm for detecting community structure in networks. Physical Review E, 69:066133, 2004.

[8] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. Physical Review E, 69:026113, 2004.

[9] L. da F. Costa, Hub-based community finding, cond-mat/0405022, http://arxiv.org/PS_cache/cond-mat/pdf/0405/0405022v1.pdf, May 2004.

[10] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the Web's link structure. IEEE Computer, 32(8):60–67, 1999.