J. Parallel Distrib. Comput. 70 (2010) 282–295

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

An adaptive overlay network inspired by social behaviour

Vincenza Carchiolo [a], Michele Malgeri [a], Giuseppe Mangioni [a,*], Vincenzo Nicosia [b]

[a] Dipartimento di Ingegneria Informatica e delle Telecomunicazioni, Facoltà di Ingegneria, Università di Catania, V.le A. Doria 6, 95125 Catania, Italy

[b] Laboratorio sui Sistemi Complessi, Scuola Superiore di Catania, Via S. Nullo 5/i, 95123 Catania, Italy

ARTICLE INFO

Article history:
Received 21 October 2008
Received in revised form
11 February 2009
Accepted 27 May 2009
Available online 10 June 2009

Keywords:
Peer-to-peer systems
Overlay networks
Socially inspired systems

ABSTRACT

Nature is a great source of inspiration for scientists, because natural systems seem to be able to find the best way to solve a given problem by using simple and robust mechanisms. Studying complex natural systems, scientists usually find that simple local dynamics lead to sophisticated macroscopic structures and behaviour. It seems that some kind of local interaction rules naturally allow the system to auto-organize itself as an efficient and robust structure, which can easily solve different tasks. Examples of such complex systems are social networks, where a small set of basic interaction rules leads to a relatively robust and efficient communication structure. In this paper, we present ***PROSA***, a semantic peer-to-peer (P2P) overlay network inspired by social dynamics. The way queries are forwarded and links among peers are established in ***PROSA*** resemble the way people ask other people for collaboration, help or information. Behaving as a social network of peers, ***PROSA*** naturally evolves to a small world, where all peers can be reached in a fast and efficient way. The underlying algorithm used for query forwarding, based only on local choices, is both reliable and effective: peers sharing similar resources are eventually connected with each other, allowing queries to be successfully answered in a really small amount of time. The resulting emergent structure can guarantee fast responses and good query recall.

1. Introduction

In the last decades, sociologists have been focused on studying social networks in order to understand why the collaboration of several people with different behaviour could magically result in an organized community of people. This field is becoming even more interesting for physicists, because many structural similarities between social networks and discrete matter organization have been discovered. Studies performed by Newman, Albert, Barabasi [23–25,45,2], and others underline the fact that almost all networks of cooperating elements, even if cooperation is based on really simple rules, naturally evolves to a small world. The existence of small worlds in social networks was empirically demonstrated by Milgram [22]. Through a famous experiment, he showed that two randomly chosen American people are connected by a very short chain of relationships.

This result is at the same time surprising and astonishing: how it is possible that all of two hundred millions people are connected by just "six degrees of separation"? A first model of the small-world effect was proposed by Watts and Strogatz [44]: they supposed to add random links to an ordered mesh of nodes, and discovered that the average path length among nodes was dramatically lowered by the addition of just a few long-distance links. On the other hand, not only networks of people and collaborations, but also some artificial networks, such as the Internet or the World Wide Web (WWW), are small worlds.

The most appreciable characteristic of a small world is that messages from one node of the network to any other one can be delivered in a few steps, thanks to long-distance links.

Since one of the main issues of many peer-to-peer (P2P) overlay networks is that searching and retrieving documents is slow and inefficient, we propose a novel P2P overlay structure called ***PROSA*** (P2P Resource Organization by Social Acquaintances) [5,6], that tries to mimic the way social links among peers are established and evolve, in order to build an efficient and self-organizing P2P network for resource sharing.

The paper is organized as follows. Section 2 is a brief overview of recent studies in the field of semantic-driven P2P networks; Section 3 explains the basic ideas ***PROSA*** is inspired by; in Section 4 we give a formal description of the algorithms involved; Section 5 describes the simulation framework used to test ***PROSA*** features; in Section 6 topological aspects of the network are discussed, aside with simulation results; Section 7 reports ***PROSA*** performance in resource retrieval; Section 8 reports some results about ***PROSA*** robustness and Section 9 summarizes obtained results.

* Corresponding author.
E-mail addresses: car@diit.unict.it (V. Carchiolo), mm@diit.unict.it (M. Malgeri), gmangioni@diit.unict.it (G. Mangioni), vincenzo.nicosia@ct.infn.it (V. Nicosia).

0743-7315/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jpdc.2009.05.004

2. Related works

P2P networks have gained popularity and interest in the last ten years mainly because cheaper and faster access to the Internet is nowadays available for the public at large. In this new scenario people are relatively free from price and bandwidth constraints, and sharing documents and multimedia files has become more attractive and effective. The problem is that traditional protocols for file transfer and resource publishing (such as File Transfer Protocol and Hyper Text Transfer Protocol) are intrinsically asymmetric: on the one side you have servers, where resources are put and can be accessed; on the other side there are clients, which can just download resources from servers. If a server is out, all resources it provides are unreachable, even if many clients want to download them. But this is a quite unnatural way of sharing resources. Users usually want to send to friends or colleagues resources downloaded from servers, and want their friends to do the same, avoiding being bound to a given "provider".

P2P overlay networks were born to address this problem. In a P2P network, each "node" acts as a client if it is asking for a resource, while it becomes a server whenever it is asked a certain resource. Nodes in a P2P network are functionally equivalent, and can collaborate in order to reach the common target: making their resources available in a fast and reliable way, to a large number of nodes in the network, while being able to retrieve wanted resources themselves, even if some other nodes are down or unreachable.

For this reason, researchers have proposed many models for P2P overlay networks, the most common being "Distributed Hash Tables" (DHTs) [39] and "Unstructured Overlays".

DHT-based P2P networks are able to give a certain internal structure to the network. Nodes are usually assigned unique IDs and are connected to a certain number of nodes that result in being "near to them" according to a given hashing function and to a similarity index. Queries are usually routed using the best available link and the underlying structure allows forwarding queries in a really efficient way, usually in $O(\log n)$ hops, where n is the number of nodes in the network. The main problem with DHTs is that resources have to be queried using their ID, which is not a human readable string.

Unstructured Overlay networks are the simplest P2P networks. Each node is directly connected to a relatively small number of other nodes. One of the early examples of such a network is Gnutella [10]. In this overlay network, queries for resources are based on limited flooding, i.e. the query is forwarded to all connected nodes, except the one that sent it, provided that the same query is not forwarded twice from the same node, and that queries live just for a limited number of "hops". The worst aspect of the Gnutella overlay network is that flooding is not efficient, and limited flooding does not guarantee finding matching results.

To address the problems of the Gnutella searching algorithm, many alternative scheme have been proposed. These include iterative deepening [46], the k-walker random walk [21], modified random Breadth First Search (BFS) [16], the two-level k-walker random walk [12], directed BFS [46], intelligent search [16], local indices based search [46], routing indices based search [8], attenuated bloom filter based search [33], adaptive probabilistic search [42], and dominating set based search [47]. Many of them are variations of the BFS, while others are Depth First Search (DFS) based. The interested reader can refer to [17] for a survey of major searching techniques in P2P networks.

Some recent works [3,50,11,18,49,48,34] focused their interest on introducing a certain amount of semantics in P2P overlays, allowing a query to become more readable and understandable by users, while trying to maintain good performance in terms of resource availability, recall and robustness.

In particular, SETS [3] uses a hybrid P2P network where all nodes are spread into a predefined number of clusters (topic segments), depending on the kind of resources they are sharing, and a network manager (a super-peer) is responsible for periodically reassigning nodes to each cluster in order to maintain consistency. The main issue of SETS is that the reliability and performance of the network is devoted to a single super-peer, which represents a possible point of fault. On the other hand, the number of clusters is an arbitrary value, and it is not clear how it impacts on performance.

A completely different approach is used in GES [50]. No super-peer is in charge of creating clusters of similar nodes and network management is completely distributed. Each peer decides to link to some other peer depending on a similarity index, after a handshake phase. Queries are forwarded using a hybrid algorithm: if a peer does not have links to relevant nodes, it forwards the query along a random path. Otherwise, the most relevant neighbour is selected as the next hop. The problem with GES is that network management requires additional messages to be exchanged among neighbours. As in SETS, in GES resources are also represented by a Vector Space Model, in terms of vectors of TF–IDF[36] coefficients.

A similar approach is used in INGA [11,18]. The selection of the next peer to forward a query is based upon the probability that a node could answer it. INGA uses an ontological approach to model resources and relevance among query and resources. Semantic similarity among nodes is mapped onto topological relationships: peers are usually linked to neighbours that share similar resources. The main issue of INGA is that resources need an ontological description, and this description is usually different for different classes of resources. Moreover, the semantic cluster in INGA does not change as a consequence of query routing, resulting in an almost static structure.

Looking at issues of these approaches, we tried to find a valuable model for a P2P network that could solve the problems introduced by organization needs and by semantics. We found that such a network does exist, in nature, and it is the network of social relationships. In fact, as explained in the next section, social relationships allow us to find good solutions to many problems, from collecting resources to gaining collaborations or finding help. On the other hand, it has been widely observed [27,20,29,37,43] that social networks are usually small worlds, and they also allow queries to be routed in a really fast and effective way in very large networks. From here came the idea of copying natural human behaviour, implementing the mechanisms involved in acquiring, modifying and cutting social links, so that the resulting P2P overlay could evolve into an effective and reliable small-world network.

PROSA tries to face some important issues of P2P systems. First of all, query routing is based on a local evaluation of relevance between nodes and queries themselves: messages are not flooded to a great number of nodes, as in Gnutella, but just to nodes that can probably answer them. Second, the network organization is entirely distributed and unsupervized: nodes naturally link to other nodes that share similar resources, once they "meet" them as a consequence of searching resources. It is not necessary to have super-peers in charge of deciding where to put each node, as in SETS, and all nodes participate in building the structure of the network. Third, no overhead messages are needed in order to build or remove links among nodes: query messages are used to establish new links and to renew them, with no extra messages for network management, as in GES. Finally, since routing and searching algorithms are based on social behaviour, simulations show that ***PROSA*** naturally evolves to a small-world network of peers, where each node can find the required resources just a few hops away. Peers naturally get divided into "semantic groups", i.e. emergent groups of nodes that result in being interested in a topic, that share resources in that topic and, thanks to the underlying link management algorithm, usually try to link to each other. The structure of