Computer Networks 56 (2012) 970–982

Contents lists available at SciVerse ScienceDirect

Computer Networks

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

Talent scouting in P2P networks

N. Koenigstein *, Y. Shavitt

School of Electrical Engineering, Tel Aviv University, Israel

ARTICLE INFO

Article history:
Available online 25 November 2011

Keywords:
Spatial data mining
P2P networks
Information retrieval

ABSTRACT

Record labels would like to identify potential artists as early as possible in their career, before other companies approach the artists with competing contracts. However, there is a huge number of new artists, and the process of identifying the ones with high success potential is labor intensive. This paper demonstrates how data mining in P2P networks can be used together with social marketing theories in order to mechanize most of this detection process.

Using a unique intercepting system over the Gnutella network we captured an unprecedented amount of geographically identified queries, allowing us to investigate the diffusion of music related content in time and space. Our solution is based on the observation that successful artists, start by growing a discernible stronghold of fans in their hometown area, where they are able to perform and market their music. Only then they manage to breakthrough to national fame. In a file sharing network, their initial local success is reflected as a delta function spatial distribution of content queries. Using this observation, we devised a detection algorithm for emerging artists that suggests a short list of artists with breakthrough potential, from which we showed that about 30% translate the potential to national success.

1. Introduction

Record label companies are constantly looking for the "next big thing". Each year, a small number of new artists succeed in breaking out of their anonymity and rise to stardom. The artists and repertoire (A&R) divisions in record labels are responsible for discovering these artists out of the masses of unknown talent on the market. A&R executives rely mostly on the word of mouth of trusted associates, critics and business contacts. Human scouts are expected to understand the current tastes of the market, but in the case of unfamiliar new artists, they have little else to rely on but their gut instinct. They therefore, tend to favor artist coming from their own city, where they can "feel the scene" [1]. Nonetheless, their predictions are usually wrong, and only around 20% of signed artists will make money for their label.[1] Locating artists with high success potential is thus of great importance for the music industry.

We examine a database of query strings to the Gnutella file sharing network, and use our understanding of the Gnutella protocol in order to identify the ones which can be located geographically. Since our aim is to gain advantage in the artists scouting market, we look at fairly rare queries, those that are not even on the top 2000 list, trying to detect emerging artists with higher potential to make a national level breakthrough (in the US). We mathematically model the popularity diffusion patterns of emerging artists in their first steps, and explain why local popularity is an important factor when new, previously unknown artists are considered. This research demonstrates how P2P query strings can be used by the music industry for intelligent decision making.

* Corresponding author. Tel.: +972 522653599.
E-mail addresses: noamk@eng.tau.ac.il (N. Koenigstein), shavitt@eng.tau.ac.il (Y. Shavitt).

[1] Taken from: www.telegraph.co.uk/technology/3778304/Could-software-find-the-X-Factor.html (Accessed: June 2010).

1389-1286/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.comnet.2011.10.021

The rest of this paper is organized as follows: In Section 2, we explain how the database of Gnutella geo-aware query strings is created. Then, in Section 3, we review the small world model for the spatial and temporal diffusion of new innovations. In Section 4 we discuss the diffusion of digital content from emerging artists. Finally, in Section 5 we describe the detection algorithm and test its predictions in Section 6.

2. Data collection

Our scouting algorithm (described in Section 5), can be applied in different P2P networks, as well as music related websites such as MySpace or even YouTube. In this study we used data collected from the Gnutella network which was the most popular file sharing network on the Internet at the time of data collection [2].

Capturing a large quantity of Gnutella queries is achieved by deploying several hundred ultrapeer nodes[2] in the network. This has been done in several previous academic studies [3–6]. However, inferring the origin IP address (essential for the geographical mapping), is not a trivial task. The Gnutella protocol does not maintain an origin address in queries. Instead it keeps an "Out Of Band" (OOB) return IP address, which is the origin IP address in principle. However, for firewalled clients which cannot accept connections from the outside the OOB address belongs to the ultrapeer acting as a proxy on behalf of the query originator. For such clients there are no known methods to determine their IP address. Furthermore, there is no explicit indication in the query whether it was issued from a firewalled client or not. In this study we have used the same data collection system as in [6]. More details about the Gnutella protocol, and the system implementations can be found in their paper. Here we will only explain how we overcome the above problem of IP resolution, as this is crucial for the geographical mapping of the queries used by our algorithm.

To overcome the above difficulties, we wish to use only queries originating from non-firewalled clients where the OOB field carries the leaf's IP address. We therefore, devised a process to distinguish firewalled from non-firewalled queries. Our technique is based on another field carried by the query, the hop count, which is set to 1 by the originator, and incremented by one each time a node forwards the query to its neighbors. To understand this technique, observe the small network in Fig. 1. The figure depicts an intercepting node along with other ultrapeers and leafs. Ultrapeer B is directly connected to the intercepting node. Thus, any query that traversed only a single hop must have come from it. Leaf A, leaf C (firewalled), and ultrapeer D are at a distance of two hops away. All queries coming from A, C and D, will have a hop count of two. Queries from Leaf A and Ultrapeer D will contain their own IP address in the OOB field. However, queries originating at C will contain B's OOB return address as C is firewalled or otherwise unable to accept incoming connections. As we are directly connected to ultrapeer B, we can simply compare the query's OOB address with B's address. If they are not identical, the query must have come from A or D, and the address is guarantied to be the origin's address. If the query contain's B's address but passed two hops, it must be coming from a firewalled client (leaf C) using ultrapeer B as a proxy. In this case C's address is not available, and the query is not recorded. Ultrapeer F and leaves E (firewalled) and G are at a distance of three hops or more. When we intercept their queries we cannot know whether the OOB IP address belongs to them, or perhaps to ultrapeers D or F (acting as a proxy). Thus, any query that traversed three hops or more is discarded. As a result, an intercepting node records traffic originating from its immediate neighborhood only (having a hop count $\leqslant 2$), thus requiring a massive deployment of such nodes.

The described setting eliminates most of the bias against popular queries which travel only short distances before being satisfied. Discarding such queries cancels the advantage of "rare" queries that stay in the network longer. However, this setting does introduce a bias against queries from firewalled clients, as only queries that can receive incoming connections are recorded.

2.1. Dataset statistics

The removal of queries which traveled more than two hops, non-Limrewire clients, firewalled queries and non-OOB enabled queries, amounts for approximately 75% of the intercepted queries. We remained with 25–40 million IP identified queries every day. We then used the commercial IPligence database to resolve the geographical location of the IP addresses bounding country, state, city, latitude and longitude to each query string. This allowed us to pin point the source of each query string to the level of cities and sometimes even smaller areas like the boroughs of NYC. Since we concentrate our study on American artists, we removed all the non-US queries reducing an additional 55–60% of the data records.

Our data-set comprised of query strings collected over a period of nine and a half months from mid October 2006 until July 2007. The activity on the Gnutella networks increases by 20–25% over the weekend [6]. We thus used weekly samples taken on a Saturday or a Sunday of every week of that period. The sample from the 51st week of 2006 and the samples from the 24th and 25th weeks of 2007 were not recorded as a result of technical difficulties. We thus remained with 38 samples instead of 41. The total number of unique geo-aware query strings processed in this study is **310,380,190**, making it the largest P2P queries study thus far.

Using the geo-aware query strings we generated weekly global (national) and local (per city) popularity charts. For each string, its global and local popularity was calculated by aggregating the number of appearances intercepted. The chart rankings were calculated by sorting the queries according to popularity. This means that in principle, more than a single string can be ranked in a given position. However, since we dealt with millions of queries this rarely ever happened among the popular strings. These popular-

[2] Ultrapeer nodes are nodes that were selected to form the backbone of the Gnutella network. As such they route queries and responses for other nodes connected to them.