Querying Similarity in Metric Social Networks

Jan Sedmidubský, Stanislav Bartoň, Vlastislav Dohnal, and Pavel Zezula

Masaryk University
Brno, Czech Republic
[xsedmid,xbarton,dohnal,zezula]@fi.muni.cz

Abstract. In this paper we tackle the issues of exploiting the concepts of social networking in processing similarity queries in the environment of a P2P network. The processed similarity queries are laying the base on which the relationships among peers are created. Consequently, the communities encompassing similar data emerge in the network. The architecture of the presented metric social network is formally defined using the acquaintance and friendship relations. Two version of the navigation algorithm are presented and thoroughly experimentally evaluated. Finally, learning ability of the metric social network is presented and discussed.

1 Introduction

The area of similarity searching is a very hot topic for both research and commercial applications. Current data processing applications use data with considerably less structure and much less precise queries than traditional database systems. Examples are multimedia data like images or videos that offer query-by-example search, product catalogs that provide users with preference-based search, scientific data records from observations or experimental analyses such as biochemical and medical data, or XML documents that come from heterogeneous data sources on the Web or in intranets and thus does not exhibit a global schema. Such data can neither be ordered in a canonical manner nor meaningfully searched by precise database queries that would return exact matches.

This novel situation is what has given rise to similarity searching, also referred to as content-based or similarity retrieval. The most general approach to similarity search, still allowing construction of index structures, is modeled in metric space. Many index structures were developed and surveyed recently [12]. However, the current experience with centralized methods [6] reveals a strong correlation between the dataset size and search costs. Thus, the ability of centralized indexes to maintain a reasonable query response time when the dataset multiplies in size, its *scalability*, is limited. The latest efforts in the area of similarity searching focus on the design of distributed access structures which exploit more computational and storage resources [2,7,4,3]. Current trends are optimizing and tuning the well known distributed structures towards better utilization of the available resources.

Another approach to design the access structure suitable for large scale similarity query processing that is introduced in this paper emerges from the notion

T. Enokido, L. Barolli, and M. Takizawa (Eds.): NBiS 2007, LNCS 4658, pp. 278–287, 2007.

of *social network*. A social network is a term that is used in sociology since the 1950s and refers to a social structure of people, related either directly or indirectly to each other through a common relation or interest [10]. Using this notion, our approach places the peers of the distributed access structure in the role of people in the social network and creates relationships among them according to the similarity of the particular peer's data. The query processing then represents searching for a community of people, i.e., searching for peers related by a common interest, for example, maintaining similar data.

Using this data point of view our designed metric social network is a cognitive knowledge network according to the terminology stated in [9] that is described as *who thinks who knows what* where it is not who you know but it is what who you know knows. This means that the network links are created on the basis of the particular peers' knowledge (stored data) rather than on being acquainted with other peers. As for the navigation, social networks exhibit the *small world network topology* [11] where most pairs of peers are reachable by a short chain of intermediates – usually the average pairwise path length is bound by a polynomial in $\log n$. Therefore it is anticipated that a small amount (around six) of transitions will be needed to find the community of peers holding the answer to a query posed at any of the participating peers in the network.

Unlike the usual access structures that retrieve a total answer to each query, the presented approach focuses on retrieving the *substantial part* of the answer yet with *partial costs* compared to the usual query processing. The concepts of social networking towards the approximative query processing in large scale data have already been introduced in related works [1,8].

The paper is structured in following sections. In Section 2, the architecture of the access structure is detailed. In Section 3, the preliminary experimental results are reviewed to present the nature and behavior of the proposed approach. Finally, the conclusions are drawn in Section 4.

2 Architecture

Our social network comprises of nodes (peers) and relationships between them. The relationships identified always relate to a particular query processed by the network and its retrieved answer. In general, we distinguish relationships of two types: the friendship and a relation of acquaintance.

2.1 Nodes of the Social Network

A node (peer) itself, besides the assigned piece of data, remembers also the history of the queries that it has been asked. To each query the recognized set of friends and acquaintances is also remembered for future optimization of a similar query processing. So, a network peer P is $P = (D, H)$ where $D = \{o_1, \ldots o_l\}$ represents the assigned piece of data and $H = \{h_1, \ldots h_m\}, h_i = (Q, L_P^{Acq}(Q), L_P^{Fri}(Q))$ represents the history of queries with the pair of ordered lists of retrieved acquaintances and friends regarding the particular query Q. For example as a query a usual range query can be considered: $R(q, r)$ where