Adaptive Approximate Similarity Searching through Metric Social Networks

Jan Sedmidubský, Stanislav Bartoň, Vlastislav Dohnal, Pavel Zezula
{xsedmid,xbarton,dohnal,zezula}@fi.muni.cz
Masaryk University, Brno, Czech Republic

Abstract

Exploiting the concepts of social networking represents a novel approach to the approximate similarity query processing. We present a metric social network where relations between peers, giving similar results, are established on per-query basis. Based on the universal law of generalization, a new query forwarding algorithm is proposed. The same principle is used to manage query histories of individual peers with the possibility to tune the tradeoff between the extent of the history and the level of the query-answer approximation. All algorithms are tested on real data and real network of computers.

I. Introduction

Current data processing applications use data with considerably less structure and pose much less precise queries than traditional database systems. Examples are multimedia (images or videos), biochemical and medical data. Such data collections 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 retrieval. The most general approach to similarity search, still allowing construction of index structures, is modeled as a metric space. Many index structures were developed and surveyed recently [1]. Due to the massive growth of volume of digital data, we need efficient tools for searching in such archives. The latest efforts in this area concentrate on: (1) structured peer-to-peer networks [2] and (2) unstructured or self-organized networks. In this paper, we focus on the second approach which emerges from the notion of *social network*. A social network is a term that is used in sociology since the 1950s and refers to a social structure of individuals, related either directly or indirectly to each other through a common relation or interest. Using this notion, our approach places the peers of a network in the role of individuals in a social network and creates relationships among them according to the similarity of the particular peer's data. The query processing then represents the search for a community of individuals, i.e., peers having similar data. This concept is closely related to *semantic overlay networks* [3] which relate peers in the network semantically. One can view semantics as a way of expressing similarity. Semantic overlays are defined over an existing P2P network, so they can exploit properties of the underlying network, such as navigation. Unlike the usual access structures that retrieve a total answer to each query, these approaches try to retrieve the *substantial part* of the answer yet with *partial costs* compared to the usual (complete) query processing.

In this paper, we address the major disadvantages of the metric social network [4], i.e., the poor behavior in larger P2P networks. The contributions of this paper are: (1) extending the concept of *confusability*, (2) adaptive navigation algorithm, and (3) managing the query history.

A. Related Work

P2P networks were traditionally used for file-sharing (Napster, Gnutella, Freenet). Semantic overlays created upon an ordinary P2P network connect semantically similar peers via relationships in order to route queries efficiently or to speed up files downloading. Tribler is a representative of an overlay for file sharing. Indexing documents by terms or keywords is used in PROSA, Jin et al. [5], SempreX, 6S [6] and Routing Index. Resource Document Format (RDF) as a model is used in REMINDIN, INGA and GridVine. Linari et al. [7] uses language models. PARIS defines a schema of peers' data. A concise survey of semantic overlay networks from different perspectives given as a tutorial is available in [3]. These systems relate peers according to the global knowledge inferred from the whole data content in the peer or according to the results returned by a specific query. Most systems are based on

[0] Partially supported by: the EU IST FP6 project 045128 (SAPIR), the national research project 1ET100300419, the Czech Grant Agency project 201/07/P240.

the former approach. The latter alternative allows the ties to be more semantically related, so the navigation can be more effective. REMINDIN, INGA as well as our metric social network exploit this strategy.

We model data as a metric space, so our approach is generic and can be applied to a variety of application domains ranging from pure keyword search to image or video retrieval. In metric spaces, the social (semantic) connections in the network created from semantic closeness of whole peer's data are very rough even if the data collection is clustered and distributed over peers. Thus, navigation algorithms would visit almost every peer by following these connections. Again, in our approach we define connections related to individual queries.

II. Metric Social Network

The Metric Social Network (mSN) is a P2P network where peers are directly linked if they are semantically similar with respect to a certain query. By the query we assume a range query as defined in a metric space $\mathcal{M} = (\mathcal{D}, d)$, i.e., $R(q, r) = \{o | o \in \mathcal{D}, d(o, q) \leq r\}$. In particular, each peer maintains a list of queries it has asked or answered, called *the query history*. The history represents the peer's knowledge about the network and is exploited by a navigation algorithm. Every query in the query history has associated an acquaintance and a list of friends. The acquaintance (week tie) is the peer returning the best answer to the specific query. Whereas the friends (strong ties) are the peers which contributed to the query answer significantly, i.e., maintaining similar data. An example of mSN is provided in Figure 1a. The acquaintance and the friendship ties are depicted as dashed and solid lines, respectively. The ties are created based on sub-answers of peers that were asked to evaluate a query. Assume that the peers P_1, P_2 and P_3 responded to the query Q_5 with nine, one and eight objects, respectively. The peers P_1 and P_3 will become friends, while P_1 will be the acquaintance. The friends will store the query with the acquaintance tie and the friend ties. The query issuing peer stores the query with the acquaintance only unless it has been identified as one of the friends. Figure 1b shows the network status after evaluating Q_5. Detailed specification of the mSN is available in [4].

A. Adaptive Search Algorithm

The query processing using the social network follows the common world concepts for searching. The peer P_{start} initiates a query Q. In its query history, it finds the *template query* Q_t which is the most similar query to Q. The query Q is forwarded to the acquaintance P_{acq} retrieved from Q_t. The contacted peer P_{acq} repeats the same procedure as P_{start} and forwards Q further if the next peer's quality is higher than the quality of P_{acq}. The definition of peer's quality is available in [8]. Notice that in each query-forwarding step a different template query can be used. When the best acquaintance P_{best} is found, it evaluates the query Q and returns its answer to P_{start}. Then P_{best} asks all friends associated with the current template query to answer the query Q as well and it returns the sub-answer back to P_{start}. Eventually, P_{start} analyzes the results and identifies the acquaintance and friends which will store the query Q in their query histories.

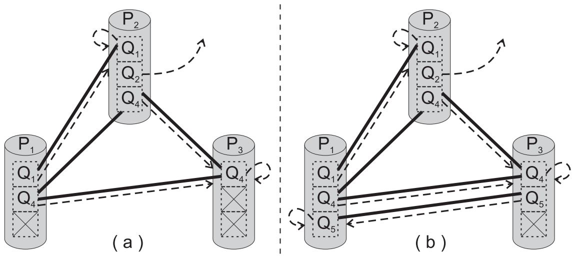

Fig. 1. Metric Social Network: (a) before and (b) after processing the query Q_5.

In this paper, we propose a new search algorithm which allows to use more template queries for query forwarding when the new query is not very similar to any query in the query history. Thus, this algorithm can adapt to unknown queries fast by flooding the network more. The number of template queries used is inversely proportional to the value of *confusability*. On no account, we use more than five template queries at a time. By utilizing the *generalized universal law of generalization* [9], we define the confusability of two queries as a weighted sum of a distance similarity D, an intersection of query regions I and a temporal aspect T: $Confusability(Q, Q_t) = w_D \cdot D(Q, Q_t) + w_I \cdot I(Q, Q_t) + w_T \cdot T(Q, Q_t)$.

B. Managing the Query History

We also introduce the query history management procedure which optimizes the content of query history. When a new query has been inserted into the query history, existing queries are checked whether they are replaceable by the new query. If such queries exist, they get simply deleted. The decision process is based on the confusability again, however, the temporal aspect is used inversely – the older the query is, the higher the probability to be replaced. To distinguish this view from the confusability used in searching, we define *replaceability* of two queries as follows: $Replaceability(Q, Q_t) = w_D \cdot D(Q, Q_t) + w_I \cdot I(Q, Q_t) + w_T \cdot (1 - T(Q, Q_t))$. The algorithm of the query-history management is specified in [8].

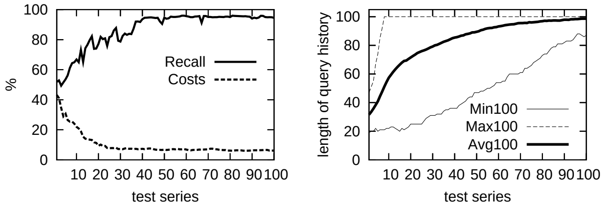

Fig. 2. Values of recall and costs (left) and the lengths of peers' query histories (right).

III. Experimental Results

In this section, we present an experimental evaluation of the proposed adaptive navigation algorithm along with managing the query history. The experiments have been conducted on two real-life datasets containing 200,000 three and forty five dimensional vectors which represent extracted color image features. We present results obtained by experimenting on the 3-D dataset only because the results for the other dataset were very much alike.

We have investigated average values of recall and costs obtained by executing test series (consisting of twenty range queries with randomly picked objects and the radii fixed) for one-hundred times. The costs represent the ratio between the number of accessed peers and the number of all peers in the network. The evolution of the metric social network is ensured by executing a batch of 150 random queries between individual runs of test series.

In Figure 2a, at the beginning, the recall was 50% although the navigating algorithm contacted more than 40% of all peers in the network. After several iterations, the social information was updated and the confusability values increased, so the flooding of the network was decreased automatically. Finally, after all iterations, the recall values reached 95% while the searching costs decreased to 7%. Figure 2b describes the evolution of lengths of peers' query histories. We set the maximum limit of the length of peers' histories to 100. If the length on a peer is exceeded, the oldest query is removed. Curves labeled with $Min100$, $Max100$ and $Avg100$ represent the minimum, maximum and average length of peer's history. In early stages (first twenty iterations), the social information was poor, so the peers' histories grew rapidly – new queries were inserted and none got deleted because of low values of both the confusability and replaceability. Next, the length of query histories became saturated and the replaceability function instructed the query management algorithm to supersede some template queries. Such a hard limit had surprisingly no impact on the recall values in comparison with an experiment without any limit set.

IV. Concluding Remarks and Future Work

We have proposed two algorithms which address two drawbacks of the metric social network – namely the navigation algorithm which limited exploration of the network and the ever-growing query history which contained also obsolete items. The principles of these algorithms are based on the *law of generalization* formed in the *confusability* measure. The presented experiment trails confirm suitability and auspiciousness of such advances. In our system, we have no automatic exploration implemented, no background actions are done by peers automatically, and peers do not exchange any profiles about their data but the query results during querying. A future research challenge is to study the behavior of the metric social network when the data change in terms of adding or deleting data items but also in terms of joining two distinct networks. We also plan to verify the metric social network properties on large-scale networks consisting of thousands of peers.

References

[1] P. Zezula, G. Amato, V. Dohnal, and M. Batko, *Similarity Search: The Metric Space Approach*, ser. Advances in Database Systems. Springer, 2005, vol. 32.

[2] M. Batko, D. Novak, F. Falchi, and P. Zezula, "On scalability of the similarity search in the world of peers," in *Proceedings of First International Conference on Scalable Information Systems (INFOSCALE 2006), Hong Kong, May 30 - June 1.* ACM Press, 2006, pp. 1–12.

[3] K. Aberer and P. Cudré-Mauroux, "Semantic overlay networks." in *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB 2005), Trondheim, Norway, August 30 - September 2, 2005.* ACM Press, 2005, p. 1367.

[4] J. Sedmidubský, S. Bartoň, V. Dohnal, and P. Zezula, "Querying similarity in metric social networks," in *Proceedings of the 1st International Conference on Network-Based Information Systems (NBIS 2007), Regensburg, Germany, September 3-7, 2007*, ser. Lecture Notes in Computer Science, vol. 4658. Springer, 2007, pp. 278–287.

[5] H. Jin, X. Ning, H. Chen, and Z. Yin, "Efficient query routing for information retrieval in semantic overlays," in *Proceedings of the 2006 ACM Symposium on Applied Computing (SAC 2006), Dijon, France, April 23-27, 2006.* ACM Press, 2006, pp. 1669–1673.

[6] R. Akavipat, L.-S. Wu, F. Menczer, and A. Maguitman, "Emerging semantic communities in peer web search," in *Proceedings of the international workshop on Information retrieval in peer-to-peer networks (P2PIR 2006).* New York, NY, USA: ACM Press, 2006, pp. 1–8.

[7] A. Linari and G. Weikum, "Efficient peer-to-peer semantic overlay networks based on statistical language models," in *Proceedings of the international workshop on Information retrieval in peer-to-peer networks (P2PIR 2006).* New York, NY, USA: ACM Press, 2006, pp. 9–16.

[8] J. Sedmidubský, S. Bartoň, V. Dohnal, and P. Zezula, "Adaptive approximate similarity searching through metric social networks," Faculty of Informatics, Masaryk University Brno, http://www.fi.muni.cz/reports/files/2007/FIMU-RS-2007-06.pdf, Tech. Rep. FIMU-RS-2007-06, November 2007.

[9] N. Chater and P. M. Vitanyi, "The generalized universal law of generalization," *Journal of Mathematical Psychology*, vol. 47, no. 3, pp. 346–369, 2003.