Indexing of sequences of sets for efficient exact and similar subsequence matching

Witold Andrzejewski, Tadeusz Morzy, and Mikołaj Morzy

Institute of Computing Science
Poznań University of Technology
Piotrowo 3A, 60-965 Poznań, Poland
{wandrzejewski,tmorzy,mmorzy}@cs.put.poznan.pl

Abstract. Object-relational database management systems allow users to define complex data types, such as objects, collections, and nested tables. Unfortunately, most commercially available database systems do not support either efficient querying or indexing of complex attributes. Different indexing schemes for complex data types have been proposed in the literature so far, most of them being application-oriented proposals. The lack of a single universal indexing technique for attributes containing sets and sequences of values significantly hinders practical usability of these data types in user applications. In this paper we present a novel indexing technique for sequence-valued attributes. Our index permits to index not only sequences of values, but sequences of sets of values as well. Experimental evaluation of the index proves the feasibility and benefit of the index in exact and similar matching of subsequences.

1 Introduction

Through unprecedented development of computer techniques witnessed in recent years, the databases are paving their way to many application areas, such as scientific, banking, industrial, retail, and financial systems. Broad applicability of database systems in diverse domains results in the development of novel data types. Traditional simple data types, such as strings, numbers, and dates, are often insufficient to describe complex structure of real-world objects. Complex data structures, such as sets and sequences, are used to reflect the complexity of the modeled reality. Sequential data are present in numerous different domains, including protein sequences, DNA chains, time series, and Web server logs. Another example of common sequence data are purchases made by customers in stores. Here, elements of a given customer sequence are not atomic, but consist of sets of products ordered by timestamps representing the date of each purchase. Contemporary object-relational database management systems support the definition and storage of complex user-defined data types as collections and nested tables. On the other hand, efficient querying and indexing of such data types is currently not supported by any commercially available database management system.

Several indexing schemes have been proposed so far, most notably for time series and sequences of atomic values. Alas, no proposals are given for indexing of sequences of sets. The original contribution of this paper is the proposal of a new indexing structure capable of efficient retrieval of sequences of sets based on non-contiguous subsequence containment and similarity. We present the physical structure of the index and we develop algorithms for query processing based on subsequence matching and subsequence similarity. In addition, we present a novel algorithm for subsequence matching with tolerance thresholds on subsequence similarity.

The rest of the paper is organized as follows. In Section 2 we introduce basic definitions used throughout the paper. Section 3 contains an overview of the related work. We present our index in Section 4. Experimental evaluation of the index is presented in Section 5. Finally, the paper concludes in Section 6 with a summary and a future work agenda.

2 Basic Definitions

An *element of a sequence* is a pair $S_i = (v(S_i), ts(S_i))$, where $v(S_i)$ denotes the *value* of the element, and $ts(S_i)$ denotes the *timestamp* of occurrence of the element S_i. A *sequence* S is an ordered set of elements S_i arranged according to their timestamps $ts(S_i)$. A *subsequence* S' of the sequence S is a sequence created from the sequence S by removing arbitrary elements. A sequence $S' = \langle (v(S'_1), ts(S'_1)), \ldots, (v(S'_k), ts(S'_k)) \rangle$ is called a *continuous subsequence* of a sequence $S = \langle (v(S_1), ts(S_1)), \ldots, (v(S_n), ts(S_n)) \rangle$ (denoted $S' \subset S$) if

$$\exists w : \forall i = 1, \ldots, k \;\; v(S_{i+w}) = v(S'_i) \wedge ts(S_{i+w}) = ts(S'_i)$$

A sequence Q such that the first element of Q has the timestamp $ts(Q_1) = 0$ is called a *query sequence*. Each query sequence Q has a *tolerance sequence* T associated with it. The tolerance sequence T has the same cardinality as the query sequence Q. The elements of the tolerance sequence T are numbers, and their timestamps are consecutive integers. The elements of the tolerance sequence T form tolerance ranges for corresponding elements of the query sequence Q of the form $(ts(Q_i) - v(T_i), ts(Q_i) + v(T_i))$. In addition, tolerance ranges must not disturb the order of elements, i.e., $ts(Q_i) + v(T_i) < ts(Q_{i+1}) - v(T_{i+1})$.

An *allocation* $A(Q, S')$ is a mapping of every query sequence element to an element of S' such that $\forall i = 1, \ldots, |Q| \;\; ts(S'_i) - ts(S'_1) - ts(Q_i) \in \langle -v(T_i), +v(T_i) \rangle$.

Given a query sequence Q. The *subsequence query* retrieves all sequences S having a continuous subsequence S', such that the following condition is fulfilled

$$l = n \wedge \forall i = 1, \ldots, n \;\; v(Q_i) \subset v(S'_i) \wedge ts(S'_i) - ts(S'_1) - ts(Q_i) \in \langle -v(T_i), +v(T_i) \rangle$$

Let ϵ denote the threshold value of minimum similarity between two sequences. Given an allocation $A(Q, S')$ of the query sequence Q to the sequence S'. The *similarity query* retrieves all sequences S such that $\exists S' \subset S : sim(Q, S') > \epsilon$, where $sim(x, y)$ is any measure of similarity between two sequences.

3 Related Work

Most research on indexing of sequence data focused on three distinct areas: indexing of time series, indexing of strings of symbols, and indexing of text. Most indexes proposed for time series support searching for similar or exact subsequences by exploiting the fact, that the elements of the indexed sequences are numbers. This is reflected both in index structure and in similarity metrics. Most popular similarity metrics include Minkowski distance [3, 15], compression-based metrics [4], and time-dimension deformation metrics [12]. Often, a technique for reduction of the dimensionality of the problem is employed, such as discrete Fourier transform [1, 2]. String indexes usually support searching for subsequences based on identity or similarity to a given query sequence. Most common distance measure for similarity queries is the Leveshtein distance [5], and index structures are built on suffix tree [8, 10, 11, 14] or suffix table [7].

Indexing of sequences of symbols differs significantly from indexing of strings. The main difference is the fact, that symbols in a sequence of symbols are assigned a timestamp that must be taken into consideration when processing a query. Most proposals for indexing of sequences of symbols transform the original problem into the well-researched problem of indexing of sets [9]. The transformation of a sequence into a set first maps all sequence elements into set elements, and then adds additional elements representing the precedence relation between the elements of the original sequence. The main drawback of this technique is the fact, that it ignores the timestamps associated with sequence elements. This leads to an additional verification phase, where sequences returned from the index are verified against the query sequence to prune false hits.

ISO-Depth index [13] is an indexing structure that efficiently supports searching of sequences based on subsequence containment and similarity. ISO-Depth index stores all continuous subsequences of given length in a trie structure. Additionally, trie nodes are numbered in a way permitting to quickly determine the nature of the relationship between the nodes. The order of the nodes in the trie corresponds to the order of symbols represented by those nodes in sequences pointed at in the trie leaves. Diversification of symbols in the trie (symbols differ depending on the distance from preceding symbols in a sequence) allows to answer queries containing timestamp constraints. After creating the trie structure, ISO-Depth lists and position lists are read off the trie to form the ISO-Depth index.

An interesting proposal of SEQ-join index was presented in [13]. This index uses a set of relational tables and a set of $B+$-tree indexes. Each table corresponds to a single symbol appearing in the indexed sequences and contains ordered timestamps of the symbol together with a pointer to an appropriate sequence. Preparing a subsequence query consists in creating a directed graph with nodes representing query sequence elements and edges representing order constraints between sequence elements. Answering a subsequence query consists in performing a join between symbol tables using $B+$-tree index joins. Detailed description of subsequence query algorithms using SEQ-join is presented in [6].

4 Generalized ISO-Depth Index

In this paper we extend the basic ISO-Depth index to support efficient indexing of sequences of sets. The new structure allows to search for similar subsequences and uses a similarity measure that is based on user-defined similarity measure for sets. We make no further assumptions on the similarity measure used to compare sets that are elements of sequences, but we require the measure to (i) increase with the increase of the size of intersection of sets, and (ii) decrease with the increase of the Hamming distance between the sets.

To the best of authors' knowledge, there are no similarity measures for sequences of sets. Therefore we introduce two new measures that can be used when formulating similarity queries on sequences of sets. Given a query sequence Q and a subsequence S' of a sequence S, such that a valid allocation $A(Q, S')$ of Q to S' exists. *Liminal similarity* is defined as the minimum similarity between any pair of sets in the allocation. Formally,

$$sim_L(Q, S') = \min_{i=1,\ldots,|Q|} \{setsim(Q_i, S'_i) : (Q_i, S'_i) \in A(Q, S')\}$$

where $setsim(Q_i, S'_i)$ is the value of user-defined similarity measure for sets that fulfills the above mentioned requirements. *Average similarity* is the average similarity between all pairs of sets in the allocation $A(Q, S')$. This similarity is given in the formula below.

$$sim_A(Q, S') = \frac{1}{|Q|} \sum_{(Q_i, S'_i) \in A(Q, S')} setsim(Q_i, S'_i)$$

It is easy to notice that for any pair of sequences (Q, S') the value of the average similarity is always greater or equal to the value of the liminal similarity between the sequences.

Below we present the algorithm for constructing the Generalized ISO-Depth index. Given a database D consisting of n sequences S^k and the width of a moving window ξ.

1. For every sequence of sets $S^k \in D$ perform the following actions
   (a) Sequence S^k is transformed into a sequence of binary signatures B^k, such that $|\mathcal{S}^k| = |\mathcal{B}^k| \wedge \forall \mathcal{S}^k_i : \mathcal{B}^k_i = (sig(\mathcal{S}^k_i), ts(\mathcal{S}^k_i))$. Timestamp values should be discretized prior to building binary signatures. Query sequences should be transformed analogously.
   (b) A moving window is used to read all continuous subsequences of B^k of the length lesser or equal to ξ. For each such subsequence B'^k, the sequence identifier k is stored along with the position, where B'^k starts within B^k.
   (c) Subsequences B'^k are transformed into symbol sequences of the form x_i, where $x \in sig(S^k) \wedge i \in N \cup \{0\}$ using the function

$$f(B'^k) = \langle x_1, \ldots, x_n \rangle \text{ where: } x_i = \begin{cases} v(B'^k_i)_0 & \text{if } i = 1, \\ v(B'^k_i)_{ts(B'^k_i) - ts(B'^k_{i-1})} & \text{if } i > 1. \end{cases}$$

(d) Symbol sequences created in the previous step are then inserted into a modified trie structure. We modify the original trie structure in the following way: instead of defining an additional terminator symbol we add subsequence identifier to a trie node in which a given subsequence terminates. In general, there can be several subsequences terminating in a given node. Therefore, each node of the trie contains a list of subsequence identifiers.

2. The trie is traversed and all nodes are numbered using the depth-first search order. Additionally, each node is marked with the highest number of the node contained in a sub-trie starting at a given node. Those two numbers determine the range of node numbers contained in a given sub-trie. The distance of a given node from the beginning of the subsequence is simply the sum of indexes of symbols on the path to a given node.
3. The trie is used to extract ISO-Depth lists of the form $(s, (a, b))$, where s is a signature of a set and the range (a, b) is the range of node numbers stored in the node pointed at by the edge representing the signature s. Each ISO-Depth list orders elements according to the value of a, and for all nodes stored in the list the distance of the node from the beginning of the subsequence is the same.
4. After creating ISO-Depth lists the trie is used to generate position lists. Each position list stores information corresponding to sequences that terminate in a given node. A position list is generated for each node where a sequence terminates.
5. ISO-Depth lists and position lists together form the Generalized ISO-Depth index. The trie structure is not used anymore and can be safely discarded.

Algorithms for processing of sequence-oriented queries using the Generalized ISO-Depth index use the following lemma.

Lemma 1. *Ranges of node numbers stored on a ISO-Depth list for a given distance from the beginning of the sequence are disjoint. Given ISO-Depth lists for distances $d_k < d_l$ from the beginning of the sequence. Let the entries on the lists be of the form $(s^k, (a^k, b^k))$ and $(s^l, (a^l, b^l))$, respectively. If $a^k < a^l \leq b^l \leq b^k$, then the database contains a sequence, such that a subsequence exists that contains sets with signatures s^k, s^l, respectively. Moreover, if the timestamp of the first element of this subsequence is subtracted from other timestamps of the subsequence elements, then the timestamps of those sets are d_k, d_l.*

The algorithm for processing of subsequence queries is given below. Let us assume that the query sequence is given as $Q = \langle (v(Q_1), 0), \ldots, (v(Q_n), ts(Q_n)) \rangle$.

1. For each timestamp $ts(Q_i)$ retrieve the ISO-Depth list for the distance equal to the timestamp.
2. Search the lists recursively. For each ISO-Depth list entry $(s^1, (a^1, b^1))$ check, if the signature $sig(Q_1)$ is contained in s^1. If true, search the ISO-Depth list corresponding to the next element of the search sequence looking for an entry $(s^2, (a^2, b^2))$, such that $a^1 < a^2 \leq b^1$ and find signatures s^2 containing

$sig(Q_2)$. For each such s^2 search the list corresponding to the next element of the query sequence retrieving only the entries contained in (a^2, b^2).
3. Continue this procedure until the last element of the query sequence is reached. Signatures retrieved during each recursive call, along with the timestamps corresponding to the subsequent ISO-Depth lists, form the searched subsequence.
4. If a signature s^n is found such that s^n contains $sig(Q_n)$, use position lists to find all pointers to subsequences stored in the nodes with numbers in the range (a^n, b^n). Store those pointers for the sake of future verification. Return to the recursive traversal of ISO-Depth lists.
5. Read the subsequences accessed via stored pointers to verify the actual subsequence containment (this is required due to ambiguity introduced by binary signature generation procedure).

Algorithms for subsequence similarity matching are similar to the algorithm presented above. We design two algorithms, one capable of using tolerance sequences when searching for a similar subsequence, and one used for strict similarity subsequence searches. Both algorithms use the upper bound of approximation of similarity between compared sequences. This approximation is based on the upper bound of the intersection and the lower bound of Hamming distance between sets that are elements of the compared sequences. Using this approximation allows for significant pruning of sequences. The upper bound approximation is used during step (2) of the algorithm, instead of checking for the containment of $sig(Q_i)$ in B^i. For queries allowing tolerance sequences, the algorithm needs to retrieve in step (1) not only ISO-Depth lists for the distance equal to the timestamp $ts(Q_i)$, but all ISO-Depth lists for distances from the range $(ts(Q_i) - v(T_i), ts(Q_i) + v(T_i))$ and merge these lists into a single list.

5 Experimental Results

The efficiency of the index is experimentally evaluated and the results of the conducted experiments are presented below. For each experiment 40 different sequence databases were generated. Elements of sets contained in sequences were generated using homogeneous and Zipf distributions. Table 1 summarizes the parameters used in experiments.

After building indexes the sets of query sequences were generated. For each database 7 different sets of 10 query sequences were prepared. Each set consisted of subsequence queries and similarity queries (with and without tolerance) for similarity thresholds of 70%, 80%, and 90%.

Experiment 1 measures the efficiency of the index with respect to increasing the size of the database. Figure 1 presents the performance of the Generalized ISO-Depth index (using 8 bit and 16 bit signatures) for subsequence queries (*Subseq*), exact similarity queries (*Sim*), and similarity queries with tolerance (*Tol*). Figure 2 presents the results for the same queries without the index. It can be easily seen that the index is 2 to 4 orders of magnitude faster than the naive approach. Query processing time grows linearly with the number of

Table 1. Synthetic data parameters

parameter	**Exp.1**	**Exp.2**	**Exp.3**
size of the domain	150 000	150 000	150 000
minimal distance between sets	1	1	1
maximal distance between sets	100	100	100
minimal set size	1	1	5–100
maximal set size	30	30	15–110
minimal number of sets in sequence	2	5–100	2
maximal number of sets in sequence	20	15–110	2
number of sequences	10 000–100 000	10 000	10 000
signature length	8b,16b	8b,16b	8b,16b
page/node size	4096B	4096B	4096B
window width (ξ)	250	250	250

sequences stored in the database. Indexes using 8 bit signatures are faster for all classes of queries. We attribute this to the fact that shorter signatures induce smaller trie structure, less nodes in the trie, and shorter ISO-Depth lists. Of course, shorter signatures produce more ambiguity and more false hits have to be verified. Nevertheless, our experiments show that the benefit of using shorter signatures surpasses the cost of additional false hit verification.

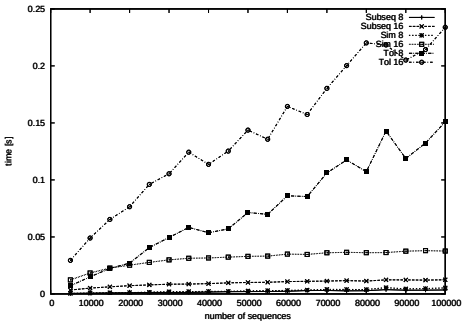

Fig. 1. Number of sequences

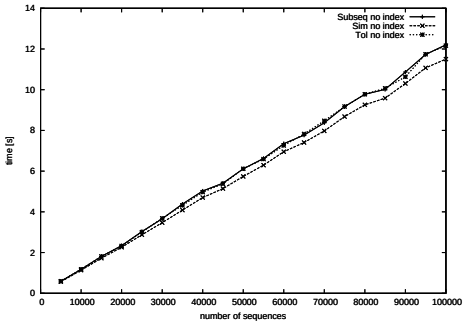

Fig. 2. Number of sequences (no index)

Experiment 2 studies the impact of the average number of sets in indexed sequences on the performance of the Generalized ISO-Depth index. We vary the average number of sets from 10 to 105. Figure 3 shows the performance of our index for three classes of queries. The results for the same queries without the index are depicted in Figure 4. Both figures exhibit the results similar to the results obtained in Experiment 1. This similarity can be easily explained. The number of subsequences inserted into the trie depends both on the number of sequences in the database, and the number of sets in indexed sequences. Conclusions of the Experiment 1 apply equally to the results of Experiment 2.

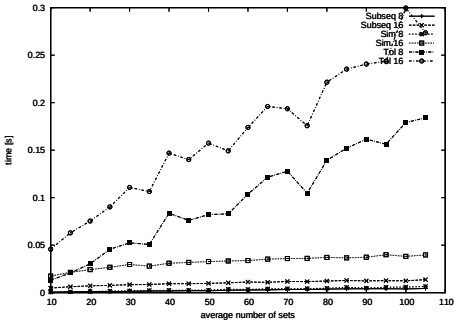

Fig. 3. Average number of sets

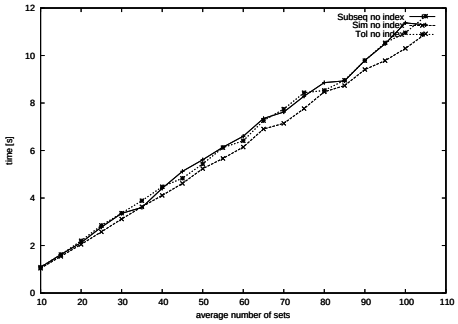

Fig. 4. Average number of sets (no index)

Experiment 3 measures the impact of the average size of sets being elements of the indexed sequences on the performance of the Generalized ISO-Depth index. We vary the average size of sets from 10 to 105. Figure 5 presents the results of three classes of queries when using the index, while Figure 6 shows the results of the same queries when not using an index. The shapes of curves presented in both figures can be easily explained. As the average size of a set grows, the probability that all positions of the signature corresponding to a given set would be set to '1' also increases. In other words, the increase of the average set size causes the saturation of signatures. Therefore, the diversity of signatures diminishes, and the set of all signatures stored in the trie becomes more compact. As the result, the number of nodes in the trie decreases and ISO-Depth lists become shorter. This in turn results in shorter processing times, although increases the number of false hits that need to be pruned. As we have already mentioned, our experiments suggest that this additional verification phase still pays off because of the shortened access time. After reaching a certain threshold, the signatures are fully saturated with bits set to '1' and the processing time stabilizes.

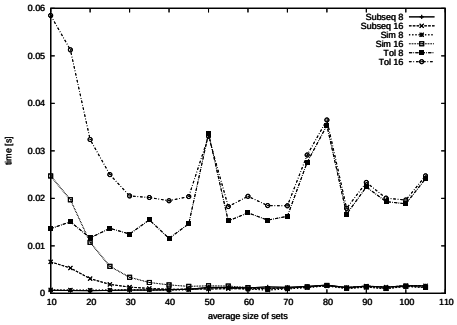

Fig. 5. Average size of sets

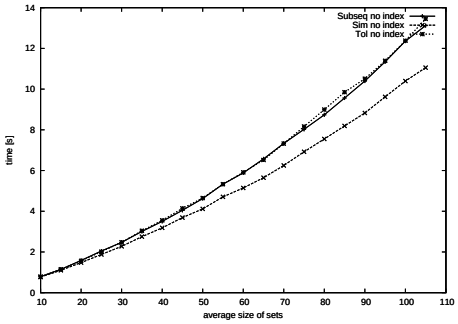

Fig. 6. Average size of sets (no index)

6 Conclusions

To the best of authors' knowledge, Generalized ISO-Depth index presented in this paper is the only index structure for sequences of sets proposed so far. Our index supports different classes of sequence-oriented queries, such as subsequence queries and similarity queries. The experiments show that the ratio of speed-up for those queries is 2 to 4 orders of magnitude when compared to brute-force approach. Possible applications of Generalized ISO-Depth index include, but are not limited to, indexing of customer purchase data, indexing of multimedia databases, or analytical processing systems.

Still, further research is required. Our future work agenda includes optimization of the physical structure of the index and designing efficient algorithms for index maintenance. Inserting and deleting of sequences from the index is not supported yet. Creating of new algorithms for insertion and deletion of sequences is our next goal. We also plan to run excessive experiments on real-world data sets to prove the practical usability of the proposed index.

References

1. R. Agrawal, C. Faloutsos, and A. N. Swami. Efficient similarity search in sequence databases. In *Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms*, pages 69–84. Springer-Verlag, 1993.
2. C. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast subsequence matching in time-series databases. In *Proceedings of the 1994 ACM SIGMOD international conference on Management of data*, pages 419–429. ACM Press, 1994.
3. E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra. Locally adaptive dimensionality reduction for indexing large time series databases. In *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, pages 151–162. ACM Press, 2001.
4. E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 206–215. ACM Press, 2004.
5. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademia Nauk SSSR*, 163(4):845–848, 1965.
6. N. Mamoulis and M. L. Yiu. Non-contiguous sequence pattern queries. In *Proceedings of the 9th International Conference on Extending Database Technology*, 2004.
7. U. Manber and G. Myers. Suffix arrays: a new method for on-line string searches. In *Proceedings of the first annual ACM-SIAM symposium on Discrete algorithms*, pages 319–327. Society for Industrial and Applied Mathematics, 1990.
8. E. M. McCreight. A space-economical suffix tree construction algorithm. *J. ACM*, 23(2):262–272, 1976.
9. A. Nanopoulos, Y. Manolopoulos, M. Zakrzewicz, and T. Morzy. Indexing web access-logs for pattern queries. In *WIDM '02: Proceedings of the 4th international workshop on Web information and data management*, pages 63–68. ACM Press, 2002.
10. E. Ukkonen. Constructing suffix trees on-line in linear time. In J.v.Leeuwen, editor, *Information Processing 92, Proc. IFIP 12th World Computer Congress*, volume 1, pages 484–492. Elsevier Sci. Publ., 1992.

11. E. Ukkonen. On-line construction of suffix trees. *Algorithmica*, 14(3):249–260, 1995.
12. M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh. Indexing multidimensional time-series with support for multiple distance measures. In *ACM KDD*, 2003.
13. H. Wang, C.-S. Perng, W. Fan, S. Park, and P. S. Yu. Indexing weighted-sequences in large databases. In *Proceedings of International Conference on Data Engineering*, 2003.
14. P. Weiner. Linear pattern matching algorithms. In *Proceedings 14th IEEE Annual Symposium on Switching and Automata Theory*, pages 1–11, 1973.
15. B.-K. Yi and C. Faloutsos. Fast time sequence indexing for arbitrary lp norms. In *Proceedings of the 26th International Conference on Very Large Data Bases*, pages 385–394. Morgan Kaufmann Publishers Inc., 2000.