Sequential pattern mining

From Wikipedia, the free encyclopedia

Sequential pattern mining is a topic of data mining concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence.[1] It is usually presumed that the values are discrete, and thus time series mining is closely related, but usually considered a different activity. Sequential pattern mining is a special case of structured data mining.

There are several key traditional computational problems addressed within this field. These include building efficient databases and indexes for sequence information, extracting the frequently occurring patterns, comparing sequences for similarity, and recovering missing sequence members. In general, sequence mining problems can be classified as *string mining* which is typically based on string processing algorithms and *itemset mining* which is typically based on association rule learning. *Local process models* [2] extend sequential pattern mining to more complex patterns that can include (exclusive) choices, loops, and concurrency constructs in addition to the sequential ordering construct.

Contents

String mining

String mining typically deals with a limited alphabet for items that appear in a sequence, but the sequence itself may be typically very long. Examples of an alphabet can be those in the ASCII character set used in natural language text, nucleotide bases 'A', 'G', 'C' and 'T' in DNA sequences, or amino acids for protein sequences. In biology applications analysis of the arrangement of the alphabet in strings can be used to examine gene and protein sequences to determine their properties. Knowing the sequence of letters of a DNA or a protein is not an ultimate goal in itself. Rather, the major task is to understand the sequence, in terms of its structure and biological function. This is typically achieved first by identifying individual regions or structural units within each sequence and then assigning a function to each structural unit. In many cases this requires comparing a given sequence with previously studied ones. The comparison between the strings becomes complicated when insertions, deletions and mutations occur in a string.

A survey and taxonomy of the key algorithms for sequence comparison for bioinformatics is presented by Abouelhoda & Ghanem (2010), which include:[3]

- **Repeat-related problems:** that deal with operations on single sequences and can be based on exact string matching or approximate string matching methods for finding dispersed fixed length and maximal length repeats, finding tandem repeats, and finding unique subsequences and missing (un-spelled) subsequences.
- **Alignment problems:** that deal with comparison between strings by first aligning one or more sequences; examples of popular methods include BLAST for comparing a single sequence with multiple sequences in a database, and ClustalW for multiple alignments. Alignment algorithms can be based on either exact or approximate methods, and can also be classified as global alignments, semi-global alignments and local alignment. See sequence alignment.

Itemset mining

Some problems in sequence mining lend themselves discovering frequent itemsets and the order they appear, for example, one is seeking rules of the form "if a {customer buys a car}, he or she is likely to {buy insurance} within 1 week", or in the context of stock prices, "if {Nokia up and Ericsson up}, it is likely that {Motorola up and Samsung up} within 2 days". Traditionally, itemset mining is used in marketing applications for discovering regularities between frequently co-occurring items in large transactions. For example, by analysing transactions of customer shopping baskets in a supermarket, one can produce a rule which reads "if a customer buys onions and potatoes together, he or she is likely to also buy hamburger meat in the same transaction".

A survey and taxonomy of the key algorithms for item set mining is presented by Han et al. (2007).[4]

The two common techniques that are applied to sequence databases for frequent itemset mining are the influential apriori algorithm and the more-recent FP-growth technique.

Applications

With a great variation of products and user buying behaviors, shelf on which products are being displayed is one of the most important resources in retail environment. Retailers can not only increase their profit but, also decrease cost by proper management of shelf space allocation and products display. To solve this problem, George and Binu (2013) have proposed an approach to mine user buying patterns using PrefixSpan algorithm and place the products on shelves based on the order of mined purchasing patterns.[5]

Algorithms

Commonly used algorithms include:

- GSP algorithm
- Sequential PAttern Discovery using Equivalence classes (SPADE)
- FreeSpan

- PrefixSpan
- MAPres[6]

See also

- Process mining
- Sequence analysis (bioinformatics)
- Sequence clustering
- Sequence labeling

References

1. Mabroukeh, N. R.; Ezeife, C. I. (2010). "A taxonomy of sequential pattern mining algorithms". *ACM Computing Surveys*. **43**: 1–41. doi:10.1145/1824795.1824798 (https://doi.org/10.1145%2F1824795.1824798).
2. Tax, N.; Sidorova, N.; Haakma, R.; van der Aalst, Wil M. P. (2016). "Mining Local Process Models". *Journal of Innovation in Digital Ecosystems*. Elsevier. **3** (2): 183–196. doi:10.1016/j.jides.2016.11.001 (https://doi.org/10.1016%2Fj.jides.2016.11.001).
3. Abouelhoda, M.; Ghanem, M. (2010). "String Mining in Bioinformatics". In Gaber, M. M. *Scientific Data Mining and Knowledge Discovery*. Springer. ISBN 978-3-642-02787-1. doi:10.1007/978-3-642-02788-8_9 (https://doi.org/10.1007%2F978-3-642-02788-8_9).
4. Han, J.; Cheng, H.; Xin, D.; Yan, X. (2007). "Frequent pattern mining: current status and future directions". *Data Mining and Knowledge Discovery*. **15** (1): 55–86. doi:10.1007/s10618-006-0059-1 (https://doi.org/10.1007%2Fs10618-006-0059-1).
5. George, A.; Binu, D. (2013). "An Approach to Products Placement in Supermarkets Using PrefixSpan Algorithm". *Journal of King Saud University-Computer and Information Sciences*. **25** (1): 77–87. doi:10.1016/j.jksuci.2012.07.001 (https://doi.org/10.1016%2Fj.jksuci.2012.07.001).
6. Ahmad, Ishtiaq; Qazi, Wajahat M.; Khurshid, Ahmed; Ahmad, Munir; Hoessli, Daniel C.; Khawaja, Iffat; Choudhary, M. Iqbal; Shakoori, Abdul R.; Nasir-ud-Din, (1 May 2008). "MAPRes: Mining association patterns among preferred amino acid residues in the vicinity of amino acids targeted for post-translational modifications". *Proteomics*. **8** (10): 1954–1958. PMID 18491291 (https://www.ncbi.nlm.nih.gov/pubmed/18491291). doi:10.1002/pmic.200700657 (https://doi.org/10.1002%2Fpmic.200700657).

External links

- SPMF (http://www.philippe-fournier-viger.com/spmf/) includes open-source implementations of GSP, PrefixSpan, SPADE, SPAM and many others.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Sequential_pattern_mining&oldid=772675108"

Categories: Data mining | Bioinformatics | Bioinformatics algorithms

- This page was last edited on 28 March 2017, at 17:41.
-