EDBT/ICDT 2013 Joint Conference
March 18-22, 2013 - Genoa, Italy

S⁴: International Competition on Scalable String Similarity Search & Join

Ulf Leser, Humboldt-Universität zu Berlin

WBI

S^4

- Competition, not a workshop in the usual sense
- Idea: Try to find the best existing software for solving (certain tasks in) similarity search & similarity join on (certain) large string sets
- Organizers: Sebastian Wandelt, Ulf Leser
  - Work with DNA: Compression, alignment, read mapping, variation detection, motif search, ...
  - Live in two worlds: Databases and Bioinformatics
  - Also research in text mining & named entity normalization
  - Often participate in competitions in other areas

String Similarity Search

- Given a (large) set S of strings and a query string q, find all $s \in S$ that are sufficiently similar to q

- "Similar": Wrt a similarity func.: unweighted edit distance
  - Could be weighted edit distance, local alignment score, hamming distance, phonetic distance, n-gram Jaccard distance, ...
- "Sufficiently": With a distance of at most k
  - Could be the single most similar or the k most similar or ...
- "Large string set": DNA sequences and geographic names
  - Could be person names, Wikipedia entries, Oxford dictionary, ...
  - Important difference: Alphabet size, skew in character frequencies

Applications

- Similarity search
  - Homologous DNA or protein sequences
  - Most probable entity referred to in a text
  - Most probable correct spelling during a keyword search
  - Most probable match in a dictionary
  - ...
- Similarity join
  - Duplicate detection (Self-Join)
  - Data cleansing
  - Entity recognition
  - Bulk similarity search
  - ...

Important in many domains

- Databases (hundred?)
  - Koudas, N., Marathe, A. and Srivastava, D. (2004). "Flexible String Matching Against Large Databases in Practice". VLDB 2004.
  - Papapetrou, P., Athitsos, V., Kollios, G. and Gunopulos, D. (2009). "Reference-based alignment in large sequence databases." *PVLDB 2(1).*
  - Rheinländer, A., Knobloch, M., Hochmuth, N. and Leser, U. (2010). "Prefix Tree Indexing for Similarity Search and Similarity Join on Genomic Data". SSDBM
  - Zhang, Zhenjie, et al. "Bed-tree: an all-purpose index structure for string similarity search based on edit distance." *SIGMOD* 2010.
  - Sahinalp, S. Cenk, et al. "Distance based indexing for string proximity search." *ICDE* 2003.
- Bioinformatics (thousand?)
  - Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. (1990). "Basic local alignment search tool." *J Mol Biol 215(3): 403-10.*
  - Kehr, B., Weese, D. and Reinert, K. (2011). "STELLAR: fast and exact local alignments." *BMC Bioinformatics*
  - Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." *Genome Res 12(4): 656-64.*
  - Li, H. and Durbin, R. (2010). "Fast and accurate long-read alignment with Burrows-Wheeler transform." *Bioinformatics 26(5): 589-95.*
  - Schneeberger, K., Hagmann, J., Ossowski, S., Warthmann, N., Gesing, S., Kohlbacher, O. and Weigel, D. (2009). "Simultaneous alignment of short reads against multiple genomes." *Genome Biol 10(9): R98.*
  - Yang, X., Wang, B., Li, C. and Xie, X. (2013). "Efficient direct search on compressed genomic data". Int. Conf. on Data Engineering, Brisbane, Australia.

Important in many domains

- Algorithms (hundred?)
  - Bartolini, Ilaria, Paolo Ciaccia, and Marco Patella. "String matching with metric trees using an approximate distance." *String Processing and Information Retrieval*. Springer Berlin/Heidelberg, 2002.
  - Bocek, Thomas, Ela Hunt, and Burkhard Stiller. *Fast similarity search in large dictionaries*. University, 2007.
  - Gawrychowski, Paweł. "Faster Algorithm for Computing the Edit Distance between SLP-Compressed Strings." *String Processing and Information Retrieval*. Springer Berlin/Heidelberg, 2012.
  - Arslan, Abdullah N., and Omer Egecioglu. "An efficient uniform-cost normalized edit distance algorithm." *String processing and information retrieval symposium, 1999 and international workshop on groupware*. IEEE, 1999.
  - ...
- NLP (hundred?)
- ... probably some more domains ....

Specialization

- Different data sets
- Subtle variations of the problem
- Some work in memory, some on disk
- Indexing mechanisms of various kinds
- Specialized on small, medium, large data sets
  - And large in 1995 is different from large in 2013
- Good for small, medium, large error thresholds
- ...

Consequences

- Very hard to compare results in papers
- Code often not available or programmed for a special case
- Exhaustive benchmarks proved very difficult
- Why not perform a competition?

Long History of Competitions

- Bioinformatics
  - CASP: Protein structure prediction
  - BioCreative: Critical Assessment of Information Extraction in Bio.
  - Automated Function Prediction SIG 2013
  - CAGI: Critical assessment of Genome Interpretation
  - Sequence Mapping and Assembly Assessment Project
  - ...
- Information Retrieval / Text Mining
  - TREC
  - INEX
  - SemEval
  - CLEF
  - ...

Format and Impact

- Format
  - Organizers describe problem and provide training data
  - Teams work with training data to tune their systems
  - Systems are tested on unseen test data
- Often tremendous impact
  - People get personally challenged and give their best
  - Standardized problems make solutions comparable
  - Bogus methods / papers have little changes
  - Brought many sobering results

Difference

- Other competitions measure hardware-independent measures: precision, recall, accuracy, standard error, ...
- Thus, usually test data is provided to the teams instead of teams providing their systems
- We want to measure wall-clock time
- Thus, we need to have all systems running on the same machine

Open Issues

- Not clear if this would be accepted by the community
- No tuning for the preferred number of threads, particular CPU, memory size, operating system, compiler optimizations, disc access, caching, ...
- Results could vary considerably on other computer systems
- Competition is only a small snapshot of a large class of problems and possible computational infrastructures

- Still, we thought it is worth trying

Setup

- Track 1: Similarity search
  - Edit distance threshold k= 0...16
  - Two real-life data sets
    - DNA sequence reads: rather uniform length (100), small alphabet, almost equal character frequencies
    - Geographical names: different length, shorter, large alphabet, skewed character (and n-gram) frequencies
  - Separate indexing (measured, not ranked) and search phase
  - Queries generated at random
- Track 2: Similarity join
  - Self-join with edit distance threshold k=0...16
  - Same two domains

Program Today

9:00 - 10:30	**Session** 1	Chair: Ulf Leser
9:00 - 09:30	Welcome Ulf Leser	
9:30 - 09:50	Efficient Edit Distance based String Similarity Search using Deletion Neighborhoods Akhil Arora;Shashwat Mishra;Tejas Gandhi;Arnab Bhattacharya	
09:50 - 10:10	Approximate String Matching by Position Restricted Alignment Manish Patil;Xuanting Cai;Sharma V. Thankachan;Rahul Shah;David Foltz;Seung-Jong Park	
10:10 - 10:30	Scalable string similarity search / join with approximate seeds and multiple backtracking Enrico Siragusa;David Weese;Knut Reinert	
10:30 – 11:00	***Coffee Break***	
11:00 - 12:20	**Session** 2	Chair: Ulf Leser
11:00 - 11:20	Efficient Parallel Partition-based Algorithms for Similarity Search and Join with Edit Distance Constraints Yu Jiang;Dong Deng;Jiannan Wang;Guoliang Li;Jianhua Feng	
11:20 - 11:40	WallBreaker - overcoming the wall effect in similarity search Stefan Gerdjikov;Stoyan Mihov;Petar Mitankin;Klaus U. Schulz	
11:40 - 12:00	Efficient algorithms for edit similarity queries Jianbin Qin;Xiaoling Zhou;Wei Wang;Chuan Xiao	
12:00 - 12:20	Cache-Aware Parallel Approximate String Search and Join Using BWT Jiaying Wang;Xiaochun Yang;Bin Wang	
12:30 - 14:00	***Lunch***	
14:00 - 15:00	**Session** 3	Chair: Ulf Leser
14:00 - 14:20	Evaluation of the competition Sebastian Wandelt	
14:20 - 15:00	Lessons learned, feedback, publication and further plans Ulf Leser	