Данные после чтения будут разделены на блоки с помощью метода кластеризации canopy, затем блоки преобразуются в пары записей, для каждой такой пары будет вычислен их вектор сравнения на основе трех переданных функций мер, по этому вектору метод взвешенной суммы определит, являются ли записи в паре совпадающими или нет, после чего будут возвращены только те пары, записи в которых были определены как совпадающие.

5 Заключение и дальнейшая работа

Мы описали меры сходства строковых значений, детерминированные методы сравнения пар записей и разработали подход к их реализации в среде распределенных вычислений Hadoop/MapReduce с использованием высокоуровневого языка программирования Jaql. Также описаны и реализованы методы по разделению пар записей на блоки для значительного снижения числа попарных сравнений и увеличения производительности.

Набор реализованных функций мер сходства строковых значений достаточно обширен, однако реализация позволяет определять собственные меры сходства, в том числе и для значений, не являющимися строковыми, в виде непосредственно функций на Jaql или же в виде Java UDF.

В дальнейшем планируется реализовать поддержку ограничений [10] и метод корреляционной кластеризации [8] для задачи выявления и удаления дубликатов, а также рассмотреть возможность реализации иных методов сравнения пар записей в среде Hadoop, таких, как вероятностные методы [9] и методы, основанные на машинном обучении [1, 3].

Поддержка

Работа выполнена при поддержке РФФИ (гранты 15-29-06045, 16-07-01028).

Литература

[1] Arasu, A., Götz, M., Kaushik, R.: On Active Learning of Record Matching Packages. Proc. of the 2010 ACM SIGMOD Int. Conf. on Management of Data, pp. 783-794. ACM, New York (2010)

[2] Baxter, R., Christen, P., Churches, T.: A Comparison of Fast Blocking Methods for Record Linkage. Proc. of the KDD-03 Workshop on Data Cleaning, Record Linkage and Object Consolidation, pp. 25-27. ACM, New York (2003)

[3] Bellare, K., Iyengar, S., Parameswaran, A.G., Rastogi, V.: Active Sampling for Entity Matching. Proc. of the 18th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 1131-1139. ACM, New York (2012)

[4] Beyer, K.S., Ercegovac, V., Gemulla, R., Balmin A., Eltabakh, M.Y., Kanne, C.C., Özcan, F., Shekita, E.J.: Jaql: A Scripting Language for Large Scale Semistructured Data Analysis. Proc. of the VLDB Endowment, 4 (12), pp. 1272-1283 (2011)

[5] Bleiholder, J., Naumann, F.: Data Fusion. ACM Computing Surveys, 41 (1), pp. 1:1–1:41 (2009)

[6] Christen, P.: Data Matching – Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Heidelberg (2012)

[7] Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A Comparison of String Distance Metrics for Name-Matching Tasks. Proc. of the 2003 Int. Conf. on Information Integration on the Web, pp. 73-78. AAAI Press (2003)

[8] Elsner, M., Schudy, W.: Bounding and Comparing Methods for Correlation Clustering Beyond ILP. Proc. of the Workshop on Integer Linear Programming for Natural Language Processing, pp. 19-27. Association for Computational Linguistics, Stroudsburg (2009)

[9] Fellegi, I.; Sunter, A.: A Theory for Record Linkage. J. of the American Statistical Association, 64 (328), pp. 1183–1210 (1969)

[10] Getoor, L., Machanavajjhala, A.: Entity Resolution for Big Data. Proc. of the 19th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, p. 1527. ACM, New York (2013)

[11] Hirschberg, D.S.: A Linear Space Algorithm for Computing Maximal Common Subsequences. Communications of the ACM, 18 (6), pp. 341-343 (1975)

[12] Islentyev, M.D.: mislen/jaql-entity-resolution: Entity Resolution Methods on Jaql (2017). https://github.com/mislen/jaql-entity-resolution

[13] Jaro, M.A.: Advances in Record Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. J. of the American Statistical Association, 84 (406), pp. 414-420 (1989)

[14] Maitrey, S., Jha, C.K.: MapReduce: Simplified Data Analysis of Big Data. Procedia Computer Science, 57, pp. 563-571 (2015)

[15] McCallum, A., Nigam, K., Ungar, L.H.: Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching. Proc. of the 6th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pp. 169-178. ACM, New York (2000)

[16] Naumann, F., Herschel, M.: An Introduction to Duplicate Detection. Morgan and Claypool Publishers (2010)

[17] Vovchenko, A.E., Kalinichenko, L.A., Kovalev, D.Y.: Methods of Entity Resolution and Data Fusion in the ETL-Process and their Implementation in the Hadoop Environment. Informatics and Applications, 8 (4), pp. 94-109 (2014)

[18] Wagner, R.A.; Fischer, M.J. The String-to-String Correction Problem. J. of the ACM, 21 (1), pp. 168-173 (1974)

[19] White, T.: Hadoop: The Definitive Guide, 4th Edition. O'Reilly Media (2015)