

Claims

1. A method executed by a computer, including:
 - receiving an identification of a first field in a first data set, the first data set including a plurality of records;
 - identifying a set of values, the set including, for each record, a value in the respective record associated with the field;
 - generating a filter mask based on the set of values, where application of the filter mask is capable of determining that a given value is not in the set of values;
 - receiving a second data set, the second data set including a second field, the second data set including a plurality of records;
 - determining a count of a number of records in the second data set having a value associated with the second field that passes the filter mask; and
 - storing the count in a profile.
2. The method of claim 1, further including determining the Sorensen-Dice coefficient of the set of values and the records in the second data set having a value associated with the second field.
3. The method of claim 1, wherein generating a filter mask includes for each value in the set of values, generating a filter key for the corresponding value.
4. The method of claim 3, wherein generating a filter key for the corresponding value includes:
 - generating a hash value for the corresponding value;
 - segmenting the hash value into a predetermined number of integers;
 - creating a filter key by setting bits in a bit vector based on the integers.
5. The method of claim 3, wherein generating the filter mask further includes performing a binary operation on each of the generated filter keys.
6. The method of claim 5, wherein determining a count of a number of records in the second data set having a value associated with the second field that passes the filter mask includes:
 - calculating a second filter key for a value associated with the second field;
 - comparing the second filter key to the filter mask.

7. A non-transitory computer storage medium encoded with computer program instructions that when executed by one or more computers cause the one or more computers to perform operations comprising:

- receiving an identification of a first field in a first data set, the first data set including a plurality of records;

- identifying a set of values, the set including, for each record, a value in the respective record associated with the field;

- generating a filter mask based on the set of values, where application of the filter mask is capable of determining that a given value is not in the set of values;

- receiving a second data set, the second data set including a second field, the second data set including a plurality of records;

- determining a count of a number of records in the second data set having a value associated with the second field that passes the filter mask; and

- storing the count in a profile.

8. The medium of claim 7, further including determining the Sørensen-Dice coefficient of the set of values and the records in the second data set having a value associated with the second field.

9. The medium of claim 7, wherein generating a filter mask includes for each value in the set of values, generating a filter key for the corresponding value.

10. The medium of claim 9, wherein generating a filter key for the corresponding value includes:

- generating a hash value for the corresponding value;

- segmenting the hash value into a predetermined number of integers;

- creating a filter key by setting bits in a bit vector based on the integers.

11. The medium of claim 9, wherein generating the filter mask further includes performing a binary operation on each of the generated filter keys.

12. The medium of claim 11, wherein determining a count of a number of records in the second data set having a value associated with the second field that passes the filter mask includes:

- calculating a second filter key for a value associated with the second field;

comparing the second filter key to the filter mask.

13. A system comprising:
 - one or more computers and one or more storage devices storing instructions that are operable, when executed by the one or more computers, to cause the one or more computers to perform operations comprising:
 - receiving an identification of a first field in a first data set, the first data set including a plurality of records;
 - identifying a set of values, the set including, for each record, a value in the respective record associated with the field;
 - generating a filter mask based on the set of values, where application of the filter mask is capable of determining that a given value is not in the set of values;
 - receiving a second data set, the second data set including a second field, the second data set including a plurality of records;
 - determining a count of a number of records in the second data set having a value associated with the second field that passes the filter mask; and
 - storing the count in a profile.
14. The system of claim 13, further including determining the Sørensen-Dice coefficient of the set of values and the records in the second data set having a value associated with the second field.
15. The system of claim 13, wherein generating a filter mask includes for each value in the set of values, generating a filter key for the corresponding value.
16. The system of claim 15, wherein generating a filter key for the corresponding value includes:
 - generating a hash value for the corresponding value;
 - segmenting the hash value into a predetermined number of integers;
 - creating a filter key by setting bits in a bit vector based on the integers.
17. The system of claim 15, wherein generating the filter mask further includes performing a binary operation on each of the generated filter keys.
18. The system of claim 17, wherein determining a count of a number of records in the second data set having a value associated with the second field that passes the filter mask includes:

calculating a second filter key for a value associated with the second field;
comparing the second filter key to the filter mask.

19. A system comprising:

means for receiving an identification of a first field in a first data set, the first data set including a plurality of records;

means for identifying a set of values, the set including, for each record, a value in the respective record associated with the field;

means for generating a filter mask based on the set of values, where application of the filter mask is capable of determining that a given value is not in the set of values;

means for receiving a second data set, the second data set including a second field, the second data set including a plurality of records;

means for determining a count of a number of records in the second data set having a value associated with the second field that passes the filter mask; and

means for storing the count in a profile.