Journal of Biomedical Informatics xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Privacy-preserving matching of similar patients

Dinusha Vatsalan *, Peter Christen

Research School of Computer Science, The Australian National University, Canberra, ACT 2601, Australia

ARTICLE INFO

Article history:
Received 24 June 2015
Revised 1 December 2015
Accepted 9 December 2015
Available online xxxx

Keywords:
Privacy
Similarity
Approximate matching
Bloom filters
Numerical data

ABSTRACT

The identification of similar entities represented by records in different databases has drawn considerable attention in many application areas, including in the health domain. One important type of entity matching application that is vital for quality healthcare analytics is the identification of similar patients, known as similar patient matching. A key component of identifying similar records is the calculation of similarity of the values in attributes (fields) between these records. Due to increasing privacy and confidentiality concerns, using the actual attribute values of patient records to identify similar records across different organizations is becoming non-trivial because the attributes in such records often contain highly sensitive information such as personal and medical details of patients. Therefore, the matching needs to be based on masked (encoded) values while being effective and efficient to allow matching of large databases.

Bloom filter encoding has widely been used as an efficient masking technique for privacy-preserving matching of string and categorical values. However, no work on Bloom filter-based masking of numerical data, such as integer (e.g. age), floating point (e.g. body mass index), and modulus (numbers wrap around upon reaching a certain value, e.g. date and time), which are commonly required in the health domain, has been presented in the literature. We propose a framework with novel methods for masking numerical data using Bloom filters, thereby facilitating the calculation of similarities between records. We conduct an empirical study on publicly available real-world datasets which shows that our framework provides efficient masking and achieves similar matching accuracy compared to the matching of actual unencoded patient records.

1. Background

1.1. Introduction

With the rapid increase of patient data collected in hospitals and clinical institutions through the use of Electronic Medical Records (EMR), efficient analysis and mining of data to provide effective healthcare support and to improve the quality of healthcare services is an emerging discipline in health research [1]. One successful data mining technique adopted in healthcare systems is similar patient matching (SPM), also known as similar patient search or patient similarity evaluation [2]. SPM plays a key role in healthcare research including clinical trials [3], inpatient bed management [4], and personalized healthcare applications [5].

The fundamental component of SPM is the evaluation of similarities between patient records by comparing the attribute (field) values (personal and medical attributes such as age, gender, body mass index, blood pressure and fasting blood sugar) using comparison functions. Fig. 1 shows an overview of SPM with three example records in a patient database and a sample query patient record. The aim of SPM is to calculate a numerical similarity value for each attribute used by applying a comparison function between pairs of patient and query records. Based on the attribute-level similarities, an overall record-level similarity is calculated and the records are ranked according to their overall similarity with the query record. Finally, the top m ranked records (with $m \geqslant 1$) are retrieved as matching records for each query record.

There have been numerous work done for SPM [2,7–10], however, all of these existing approaches are built under a non privacy-preserving setting, where sharing and using of patient data across different organizations are not restricted due to privacy and confidentiality concerns. However, in today's world it is often not legally and ethically allowed in many countries to share or exchange data, especially medical data, across organizations due to the growing privacy and confidentiality concerns [11]. Therefore, SPM needs to be conducted by using masked (encoded) attribute values.

* Corresponding author.
E-mail addresses: dinusha.vatsalan@anu.edu.au (D. Vatsalan), peter.christen@anu.edu.au (P. Christen).

http://dx.doi.org/10.1016/j.jbi.2015.12.004
1532-0464/

Several masking functions (that transform original data in such a way that there exists a specific functional relationship between the original data and the masked data [12]) have been used for preserving privacy of actual values ranging from expensive cryptographic techniques to efficient perturbation-based techniques [13]. Bloom filter-based masking is one efficient perturbation privacy technique that has been widely used in privacy-preserving record linkage [14–16] and privacy-preserving set operations [17,18]. However, existing Bloom filter-based masking functions are suitable only for string or categorical data. Masking numerical data such as integer, floating point, and modulus values is, however, important for SPM as these data types are commonly used in the health domain.

In this paper, we propose efficient Bloom filter-based masking approaches for numerical data and develop a solution for the privacy-preserving SPM (PP-SPM) problem. Our main contributions are: (1) based on Bloom filter masking techniques that have shown to be successful for approximate matching of string data [14,19], we propose novel Bloom filter masking approaches for numerical data that have similar characteristics as for string data (i.e. they are efficient, effective, and secure), (2) develop a comprehensive framework for PP-SPM using Bloom filter masking-based similarity calculations for matching different types of data, and (3) report on an extensive empirical study of our framework on three publicly available real-world datasets.

The rest of this paper is organized as follows. In Section 1.2 we review the literature related to our work. In Section 2 we define the research problem addressed in this paper. We propose numerical data masking methods in Section 2.1, and present a framework for PP-SPM in Section 2.2. We analyze our proposed framework in terms of complexity, privacy and accuracy in Section 2.3, and empirically evaluate the framework on real-world datasets in Section 3. We then conclude with an outlook to future research directions in Section 4.

1.2. Related work

1.2.1. Similar patient matching

Identifying similar patients or linking the same patients across different databases has increasingly been investigated and implemented in several healthcare projects, including the SAIL databank in Wales [20], the Manitoba research registry in Canada [21], and the Western Australia Data Linkage System and the Centre for Health Record Linkage (CHeReL) in Australia [22].

Similar patient matching (SPM) can be considered as a domain-specific problem in the research area of nearest neighbor search, which has been identified as a core data mining problem [23]. Nearest neighbor search is an optimization problem for finding the closest (most similar) data points for a given query data point, where closeness is typically expressed in terms of a distance/similarity metric function. Several techniques have been developed to address the nearest neighbor search problem, ranging from clustering, graph-based learning, to classification and information retrieval [6,24]. The goal of SPM is to derive a distance metric that is context sensitive and is able to measure the similarity between patients represented by records containing medical as well as personal data [2].

An SPM method to find patients with a similar heartbeat pattern was introduced by Park and Kang [7]. The abstraction of a patient's typical heartbeat pattern was represented as a string using regular expressions and then the edit distance measure was applied on these abstractions. Edit distance is commonly used to quantify the similarity of two strings in terms of the minimum number of edit operations required to transform one string into another [6]. Three approximate string comparison functions, which are Levenshtein edit distance, Jaro–Winkler, and the longest common substring [6], were evaluated for medical record linkage by Grannis et al. [25], and the results showed that each comparator has strengths and weaknesses with some techniques being highly specific and others highly sensitive.

Saeed and Mark [8] employed a multi-resolution description scheme for representing temporal ICU patient data and used unsupervised metrics for retrieving similar patients. Sun et al. [9] proposed a supervised metric learning algorithm to evaluate patient similarities. This approach uses statistical and wavelet-based features to capture the characteristics of patients, and proposed supervised metric learning to incorporate physician's domain knowledge. Wang et al. [10] extended the supervised metric learning approach by Sun et al. [9] to update the metric interactively and to work with multiple types of feedback from different physicians.

Recent work by Wang [2] investigated the scalability of SPM and proposed adaptive tree-based indexing approaches to reduce the quadratic computational burden that arises from pairwise distance calculations. The proposed approach, rather than using a uniform distance measure for all the patients in the indexing tree, uses a specific distance metric for a subset of patients within a single node based on the distribution of data in the tree.

However, no work has so far focused on SPM in a privacy-preserving setting. Developing privacy-preserving techniques has been an emerging trend in several research disciplines related to SPM, including privacy-preserving data mining [26],

Patient database D

Patient ID	Suburb	Gende r	Age	Body mass index	Blood pressure	Date of last visit	
						Month	*Year*
P3101	Acton	F	64	36.4 3	120	07	13
P3102	Lyneham	F	36	25.69	117	01	14
P3103	Lynehan	F	30	33.64	125	10	13

Query record q

Q100	Lyneham	F	34	27.5 2	120	12	13

Similarity calculation							Overall similarity	Rank
Q100 P3101	0.0	1.0	0.0	0.1	1.0	0.5	0.43	3
Q100 P3102	1.0	1.0	0.8	0.8	0.7	0.9	0.86	1
Q100 P3103	0.9	1.0	0.6	0.4	0.5	0.8	0.7	2

Fig. 1. An overview of similar patient matching (SPM) using three example (made-up) patient records and a query record with attributes suburb (string), gender (categorical), age (integer), body mass index (floating point), blood pressure (floating point), and date of last visit (modulus). Similarities are calculated using edit distance [6] for strings, exact matching for categorical, and absolute difference similarity [6] for numerical values. The overall similarity of each record with the query record is calculated as average over all attributes and ranking is based on these overall similarities.

Please cite this article in press as: D. Vatsalan, P. Christen, Privacy-preserving matching of similar patients, J Biomed Inform (2015), http://dx.doi.org/10.1016/j.jbi.2015.12.004