Information Sciences 289 (2014) 8–24

Contents lists available at ScienceDirect

Information Sciences

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

Probabilistic correlation-based similarity measure on text records ☆

Shaoxu Song [a,*], Han Zhu [a], Lei Chen [b]

[a] KLiss, MoE; TNList; School of Software, Tsinghua University, China
[b] The Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 29 July 2013
Received in revised form 12 July 2014
Accepted 3 August 2014
Available online 20 August 2014

Keywords:
Similarity measure
Probabilistic correlation
Text record

ABSTRACT

Large scale unstructured text records are stored in text attributes in databases and information systems, such as scientific citation records or news highlights. Approximate string matching techniques for full text retrieval, e.g., *edit distance* and *cosine similarity*, can be adopted for unstructured text record similarity evaluation. However, these techniques do not show the best performance when applied directly, owing to the difference between unstructured text records and full text. In particular, the information are limited in text records of short length, and various information formats such as abbreviation and data missing greatly affect the record similarity evaluation.

In this paper, we propose a novel probabilistic correlation-based similarity measure. Rather than simply conducting the matching of tokens between two records, our similarity evaluation enriches the information of records by considering correlations of tokens. The probabilistic correlation between tokens is defined as the probability of them appearing together in the same records. Then we compute weights of tokens and discover correlations of records based on the probabilistic correlations of tokens. The extensive experimental results demonstrate the effectiveness of our proposed approach.

1. Introduction

Unstructured text records are prevalent in databases and information systems, such as personal information management systems (PIM) and scientific literature digital library (CiteSeer). Various applications, for example similarity search [12], duplicate record detection [8], information integration [1] and so on, rely on the similarity evaluation among these unstructured records of text values. Table 1 shows an example of unstructured record database which stores several citation records as text attributes. Due to various information formats such as abbreviation and data missing, it is not easy to evaluate the similarity of unstructured records in the real world.

Since unstructured text records are text strings of short length (as shown in Table 1), we can apply approximate string matching techniques such as *edit distance* [21] to measure the similarity. However, these character-based matching approaches can only capture limited similarity and fail in many cases such as various word orders and incomplete information formats. Therefore, other than character-based string matching techniques, we can also treat each unstructured record as a text document and apply full text retrieval techniques to measure the record similarity. Specifically, records are repre-

☆ A preliminary, extended abstract of this paper appears in [26].

* Corresponding author.
E-mail addresses: sxsong@tsinghua.edu.cn (S. Song), leichen@cs.ust.hk (L. Chen).

http://dx.doi.org/10.1016/j.ins.2014.08.007
0020-0255/

Table 1
Example of unstructured citation records.

No.	Citation
1	S. Guha, N. Koudas, A. Marathe, D. Srivastava, Merging the results of approximate match operations, in: VLDB'04: Proceedings of the 30th International Conference on Very Large Data Bases, 2004, pp. 636–647
2	Guha, S., Koudas, N., Marathe, A., Srivastava D., Merging the results of approximate match operations, in: the 30th International Conference on Very Large Data Bases (VLDB 2004), 2004, pp. 636–647
3	Sudipto Guha, Nick Koudas, Amit Marathe, Divesh Srivastava: Merging the Results of Approximate Match Operations, in: VLDB, 2004, pp. 636–647
4	S. Guha, et al. Merging the results of approximate match operations, in: Proceedings of the Thirtieth International Conference on Very Large Data Bases, Toronto, Canada, August 31 – September 3, 2004

sented by a set of weighted token features and similarity is computed based on these features. Cohen [4] proposes a word token based *cosine similarity* with *tf*idf* which can detect the similarity of records with various word orders and data missing. Gravano et al. [9] propose a more effective approach by using *q*-grams as tokens of records, which can handle spelling errors in records.

Unfortunately, the characteristics of unstructured text records are different from those of strings in full texts. First, due to the short length of text records, most words appear only once in a record, that is, the *term frequency (tf)* is 1 in most cases of such short text records in databases. We show the statistics of term frequency in Table 2. More than 90% tokens, even the tokens of *q*-grams, appear only once in a record. Therefore, only the *inverse document frequency (idf)* [27] takes effect in the *tf*idf* [24] weighting scheme and no local features of each record are considered. Moreover, the popular matching similarity measure used for full text, *cosine similarity*, is based on the assumption that tokens are independent of each other, and the correlations between tokens are ignored. Due to various information representation formats of unstructured text records such as abbreviation and data missing, latent correlations of records can hardly be detected by only considering the matching of tokens.

Example 1. Consider records No. 3 and 4 in Table 1 with different author representations of "Sudipto Guha, Nick Koudas, Amit Marathe, Divesh Srivastava" and "S. Guha, et al." respectively. By using the *cosine similarity* which is based on the dot product of two record vectors, we have only one matching token "Guha" and the similarity value is low. Even worse, there is no matching token at all between the different representations of the same conference, "Very Large Data Bases" and "VLDB", and the *cosine similarity* value is 0 between these two representations. As a consequence, the *cosine similarity* of records No. 3 and 4 is low, which actually describe the same citation entity. Cohen et al. [5] conclude that full text retrieval techniques, *tf*idf* and *cosine similarity*, do not show the best performance when they are applied directly to text records in databases.

Motivated by the unsuitability of string matching and full text retrieval techniques in measuring similarity between text attribute records, in this paper, we mainly focus on developing the similarity metrics based on the correlation of tokens, and perform the similarity evaluation over records directly without data cleaning. In our similarity approach, rather than matching tokens of records, the correlations between tokens are considered, which help to discover more correlations of short text records with limited information. The correlations between tokens are investigated based on the probability that tokens appear in the same records. Then, these token correlations are utilized in two aspects, i.e. intra-correlation and inter-correlation. The intra-correlations consider the correlations of tokens in a record, and are utilized in the weighting of tokens. Rather than simply assigning equal term frequencies to tokens, we develop the discriminative importance of each token based on the degree of correlations with other tokens in a record. The inter-correlations represent the correlations of tokens between two records, which can further discover the correlations of records in addition to matched tokens. Based on the correlations of tokens, we can perform the similarity evaluation on text records with more diverse formats, for example with abbreviation and data missing.

Our contributions in this paper are summarized as follow:

- We develop a dictionary to capture the probabilistic correlations of tokens, and represent text records with the consideration of both token frequencies and correlations. Highly correlated tokens are merged as phrase tokens to reduce the size.

Table 2
Statistics of term frequency.

Dataset[a]	Term frequency		
	=1	=2	$\geqslant 3$
Cora (word)	96.8%	3.0%	0.2%
Cora (*q*-grams)	93.6%	5.9%	0.5%
Restaurant (word)	98.4%	1.5%	0.1%
Restaurant (*q*-grams)	96.9%	2.9%	0.2%

[a] *Cora* and *Restaurant* are two datasets used in this paper, please refer to Section 6 for details.