INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Probabilistic multiple correlation based term weighting scheme for measuring similarity of unstructured text records

J Ujwala Rekha[1]*, K Shahu Chatrapati[1,2]

1 Associate Professor of CSE, JNTUH College of Engineering Hyderabad, Telangana, India
2 Professor of CSE, JNTUH College of Engineering Manthani, Telangana, India

OPEN ACCESS

Received: 28-02-2020

Accepted: 29-03-2020

Published: 02-05-2020

Editor: Dr. Natarajan Gajendran

Citation: Rekha JU, Chatrapati KS (2020) Probabilistic multiple correlation based term weighting scheme for measuring similarity of unstructured text records. Indian Journal of Science and Technology 13(11): 1276-1282. https://doi.org/10.17485/IJST/v13i11.2020-31

***Corresponding author**.
J Ujwala Rekha

Associate Professor of CSE, JNTUH College of Engineering Hyderabad, Telangana, India
ujwala_rekha@jntuh.ac.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

Abstract

Background/Objectives: In this study, a term weighting scheme derived from probabilistic multiple correlation is defined for measuring similarity between unstructured text records. **Methods:** While the intra-correlation is the correlation of terms in the same record, inter-correlation is the correlation of terms that exist in different records. Probabilistic multiple correlation based term weighting calculates the weight or relevance of a term by considering its intra-correlation with one or more terms simultaneously. Subsequently, the term weights are used in measuring the inter-correlation of terms and then the similarity between two text records. **Findings:** The experiments are run on unstructured text records that are incomplete and employ abbreviations. There is significant improvement in precision, recall and f-score using probabilistic multiple correlation based term weighting scheme when compared with probabilistic simple correlation weighting scheme. **Applications:** Using probabilistic multiple correlation based term weighting scheme can improve the overall accuracy in matching unstructured text records that contain abbreviations and incomplete data.

Keywords: Unstructured Text; Approximate String Matching; Citation Matching; Probabilistic Correlation; Term Weight; Similarity Measure

1 Introduction

Unstructured textual data is a textual information that does not have a well-defined structure. It can include e-mail messages, transcripts, metadata, health records, comments and chat-logs to name a few. According to data analysts, 80 percent of all business information is unstructured [1,2]. Management of Unstructured data may include duplicate record detection, categorization, clustering, information extraction and integration that typically depend on estimating similarity among the unstructured text records. In this study, a term weighting scheme based on probabilistic multiple-correlation of a term with one or more terms simultaneously is discussed in the context of citation matching.

The key aspect of citation matching is determining citations that represent the same publication, which is a difficult problem because citations suffer from inaccuracies, missing references and incorrect references [3]. Consequently, citation matching can be studied as approximate string matching. In Table 1, an illustration of several citations that represent the same document, despite many superficial variations is given. Since citations are text strings, approximate string matching measures similar to cosine similarity and edit distance(4) can be employed to quantify either the similarity or dissimilarity between two citation records respectively. However, these metrics cannot be applied directly, owing to the differences in various citation formats that include abbreviations and incomplete data. Cohen(5)proposes a distance metric derived from term frequency and inverse document frequency which can measure similarity of records in spite of different word orderings and incomplete data. However, citations are strings of short length, and hence the term frequency of terms is one; because more often than not, terms in a citation appear only one time(6). Bilenko & Mooney(7) use machine learning techniques, in which the distance metrics for each field are derived by learning, and a classifier that combines the results of different distance metrics is employed. Unfortunately, the methods of Bilenko & Mooney(7) cannot be applied to unstructured text records since the records do not have well-defined fields. Pasula et al.(8) proposes a probabilistic object identification method for citation matching, but requires identification of citation style and segmentation of citation into author and title subfields. More recently, the probabilistic correlation-based similarity defined by Song et al.(6) successfully handles information formats that contain abbreviations and missing data. In this, the intra-correlation between two terms is derived from the probability that the terms occur jointly in the same records. Subsequently, term weights are calculated from the degree of intra-correlation of a term with other terms in a record. Then, the probabilistic correlation-based similarity is calculated between two records from the inter-correlations of terms and the term weights.

Table 1. Examples of citations that refer to the same document

r_1 :	Hall, Patrick AV, and Geoff R. Dowling. "Approximate string matching." ACM computing surveys (CSUR) 12.4 (1980): 381-402.
r_2 :	Hall, P. A., & Dowling, G. R. (1980). Approximate string matching. ACM computing surveys (CSUR), 12(4), 381-402.
r_3 :	Patrick A. V. Hall, Geoff R. Dowling: Approximate String Matching. ACM Comput. Surv. 12(4): 381-402 (1980)
r_4 :	Patrick A. V. Hall and Geoff R. Dowling. 1980. Approximate String Matching. ACM Comput. Surv. 12, 4 (December 1980), 381-402. DOI=http://dx.doi.org/10.1145/356827.356830
r_5 :	Hall, P.A. et al. Approximate String Matching, in: CSUR, 1980, pp. 381-402.

Generally, two or more terms simultaneously distinguish a text record. Unfortunately, Song et al.(6) measures the correlation between two terms without consideration of the fact that both terms may be influenced by other terms in distinguishing a record. Consequently, in order to overcome the drawback of simple correlation, one can employ multiple-correlation that studies the correlation of a term with one or more terms simultaneously. Therefore, this study investigates a term weighting scheme based on probabilistic multiple correlation of terms that considers the correlation of a term with one or more terms simultaneously; rather than simple correlation between two terms. Subsequently, a similarity measure derived from probabilistic multiple correlation of terms similar toSong et al.(6) is used in comparing two unstructured text records. Finally, a report of the experimental evaluation to demonstrate the efficacy of the proposed approach is presented.

The remaining part of the paper is structured as follows: Section 2 introduces a term weighting scheme derived from probabilistic multiple-correlation between two or more terms. Section 3 presents correlation similarity measure similar to studies done by Song et al. [6]. In Section 4, the performance of the proposed term weighting scheme is demonstrated. Finally, conclusions are drawn in Section 5.

2 Probabilistic multiple correlation based term weighting

In this section, we derive probabilistic multiple intra-correlation of terms from probabilistic simple correlation of terms defined in the publication done by Song et al.(6). While the intra-correlation is the correlation of terms in a single record, inter-correlation is the correlation of terms between two records.

2.1. Probabilistic simple intra-correlation of terms

Let $R = \{r_1, r_2, \ldots, r_m\}$ be a set of records that are segmented into terms $T = \{t_1, t_2, \ldots, t_n\}$ then the correlation between two terms are calculated from the conditional probability. The conditional probability of the term t_i given term t_j is defined as follows(6):

$$P(t_i|t_j) = \frac{P(t_i \wedge t_j)}{P(t_j)} \tag{1}$$

Subsequently, the probabilistic intra-correlation of terms can be defined as follows[6]:

$$\begin{aligned}\operatorname{cor}(t_i,t_j) &= P(t_i|t_j)\,P(t_j|t_i)\\ &= \frac{P(t_i\wedge t_j)^2}{P(t_i)\,P(t_j)}\end{aligned} \tag{2}$$

It can be noted that $cor(t_i,t_j) = cor(t_j,t_i)$. Besides, $\operatorname{cor}(t_i,t_j) = 1$ implies that the terms t_i and t_j always exist jointly in the records, and $\operatorname{cor}(t_i,t_j) = 0$ denotes that they never appear together in any record.

2.2. Probabilistic multiple intra-correlation of terms

If $T = \{t_1,t_2,\ldots,t_n\}$, then 2^T is the power set of T that contain all subsets of T. The conditional probability of term t_i given a set of multiple terms $S \in 2^T$ can be defined as follows:

$$P(t_i|S) = \frac{P\left(t_i \bigwedge_{t_j\in S} t_j\right)}{P(t_j)} \tag{3}$$

Consequently, correlation of term t_i given a set of multiple terms $S \in 2^T$ can be defined as follows:

$$\begin{aligned}\operatorname{cor}(t_i,S) &= p(t_i|S)\,p(S|t_i)\\ &= \frac{P\left(t_i, \bigwedge_{t_j\in S} t_j\right)^2}{p(t_i)\,p\left(\bigwedge_{t_j\in S} t_j\right)}\end{aligned} \tag{4}$$

2.3. Term weighting

Terms must be associated with weights based on its discriminability in distinguishing a record. Generally, a more frequent term is a bad discriminator and hence must be assigned a low weight. Alternatively, a less frequent term is a good discriminator and must be assigned more weight. It can be observed that the terms that appear less frequently have higher correlation to other terms compared to terms that appear more frequently. Therefore, less frequent terms with higher correlation are more likely to characterize the record. Consequently, terms with higher correlations can be regarded as essential features of the record. We define a term weighting scheme derived from the degree of term correlation with other terms as follows:

$$cow(t_i) = \frac{\sum_{S\in 2^T, t_i\notin S} freq(S)^* cor(t_i,S)}{2^{[T]-1}-1} \tag{5}$$

where *freq (S)* is the frequency of *S* in record space *R*. Thus, the correlation weight lies between 0 and 1, and denotes the relevance of the term t_i in distinguishing a record. Consequently, high correlation weight of a term implies more relevance of the term in distinguishing the record and vice-versa.

Table 2. Terms appearing in each record

	Hall	Patrick	Geoff	Dowling	et al.
r_1 :	✓	✓	✓	✓	
r_2 :	✓			✓	
r_3 :	✓	✓	✓	✓	
r_4 :	✓	✓	✓	✓	
r_5 :	✓				✓

For instance, let us consider the term set {Hall, Patrick, Geoff, Dowling, et al.} for the sake of simplicity. In Table 2, the terms present in each of the records of Table 1 are recorded. In Table 3, the probabilistic multiple-correlation of the term "Hall" to other terms, and correlation weight of term 'Hall' is presented.

Table 3. Calculation of Probabilistic Multiple Correlation of term 'Hall'

Termset S	*freq (S)*	*cor (Hall, S)*	*freq (S)*cor (Hall, S)*
Patrick	0.6	0.6	0.36
Geoff	0.6	0.6	0.36
Dowling	0.8	0.8	0.64
et al.	0.2	0.2	0.04
{Patrick, Geoff}	0.6	0.6	0.36
{Patrick, Dowling}	0.6	0.6	0.36
{Patrick, et al.}	0.0	0.0	0.0
{Geoff, Dowling}	0.6	0.6	0.36
{Geoff, et al.}	0.0	0.0	0.0
{Dowling, et al.}	0.0	0.0	0.0
{Patrick, Geoff, Dowling}	0.6	0.6	0.36
{Patrick, Geoff, et al.}	0.0	0.0	0.0
{Patrick, Dowling, et al.}	0.0	0.0	0.0
{Geoff, Dowling, et al.}	0.0	0.0	0.0
{Patrick, Geoff, Dowling, et al.}	0.0	0.0	0.0
		cow (Hall)	0.19

3 Probabilistic inter-correlation of terms

In this section, we discuss the similarity measure described in[6] which is derived from the probabilistic correlation. The cosine similarity measure is single-to-single; a term in one record is matched with only one term in another record. However, in probabilistic correlation based similarity measure, multiple-to-multiple correlations exist, that is one term may be matched with several terms in another record. Consequently, three types of inter-correlations of terms exist between two records[6].

1. Firstly, inter-correlation is defined between the terms that match exactly, for instance, the inter-correlation of 'Hall' in record 1 and 2 of Table 1.
2. The second type of inter-correlation exists between the terms that are present in both records, for instance, the inter-correlation between "Hall" in record 1 and "CSUR" in record 2 of Table 1.
3. The third type of inter-correlation exists between the terms where at least one of the terms does not occur in two records. For instance, "ACM" in record 4 and "CSUR" in record 5 of Table 1.

Consequently, the probabilistic inter-correlation between terms t_i and t_j in records r_1 and r_2 , in that order can be defined as follows[6]:

$$cor\left(t_i,t_j\right)=\left\{\begin{array}{lc} 1 & t_i=t_j \\ 0 & t_i\neq t_j\wedge t_i\in r_2\wedge t_j\in r_1 \\ P\left(t_i|t_j\right)P\left(t_j|t_i\right) & t_i\notin r_2\vee t_j\notin r_1 \end{array}\right\} \tag{6}$$

Let M_1 and M_2 constitutes the set of terms of records r_1 and r_2 respectively, then the correlation-similarity between records r_1 and r_2 can be calculated according to[6] as follows:

$$\operatorname{sim}\left(r_1,r_2\right)=\frac{\underset{t_i\in M_1,t_j\in M_2}{\Sigma} w_iw_j\operatorname{cor}\left(t_i,t_j\right)}{\|r_1\|\cdot\|r_2\|} \tag{7}$$

where w_i , w_j denote the correlation weights $cow(t_i)$ and $cow(t_j)$ of terms t_i and t_j respectively calculated according to (5), and $cor(t_i,\ t_j)$ is the probabilistic inter-correlation between terms t_i and t_j calculated according to (6). Furthermore, in order to normalize the similarity measure, $\|r_1\|\cdot\|r_2\|$ is employed[6], where

$$\|r_1\|=\sqrt{\underset{t_i\in M_1,t_j\in M_2}{\Sigma}\left(w_i^2\operatorname{cor}\left(t_i,t_j\right)\right)} \tag{8}$$

and $\|r_2\|$ can be computed in an analogous way.

4 Experiments and Results

The data sets called Cora and Restaurant are employed in the experiments conducted as in[6]. Cora is composed by McCallum[6,9] containing 1295 distinct citations of 122 publications. Restaurant is a database compiled by Sheila Tejada[6,10] containing 864 restaurant names along with addresses consisting of 112 duplicate records. To measure the efficacy of the proposed term weighting scheme derived from multiple correlation of a term with one or more terms simultaneously, the distances between a particular record and its potential duplicates are computed. Two records with the highest similarity are considered to characterize the same entity. Precision, recall and f-score are adopted to measure the usefulness of the proposed term weighting scheme over simple correlation weighting scheme of[6], and are defined below:

$$\begin{aligned} \text{Precision} &= \frac{\left|\mathrm{R_a} \cap R_f\right|}{\left|R_f\right|} \\ \text{Recall} &= \frac{\left|R_a \cap R_f\right|}{\left|R_a\right|} \\ f-\text{score} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \tag{9}$$

where R_a represents the pair of records that actually correspond to the same entity and R_f represents the pair of records found to be analogous. While precision connotes 'how relevant the results are', the recall is the measure of 'how comprehensive the results are'. Consequently, as more number of pairs with lower similarity are misrepresented as the same entity, recall increases while precision decreases. The *f-score* is calculated from the harmonic mean of precision and recall and is a measure of overall accuracy. In Table 4, the maximum f-score values of multiple correlation and simple correlation based term weighting scheme are presented. While in Figure 1 the comparison of multiple-correlation and simple correlation based term weighting schemes in cora data set is presented, in Figure 2 comparison of the two term weighting schemes in restaurant data set is presented.

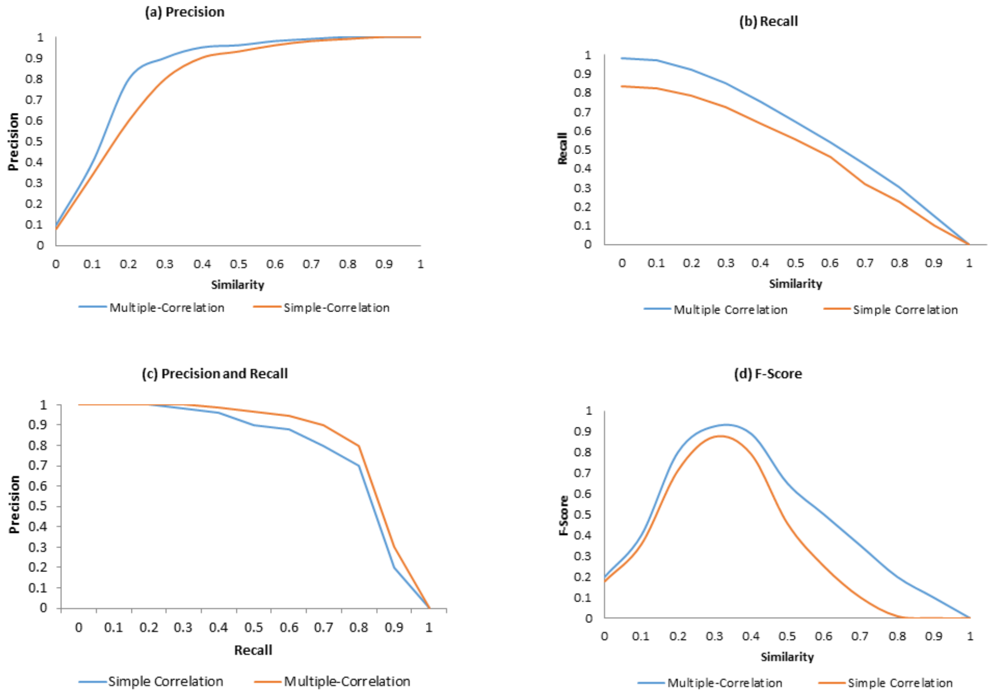

Fig 1. Comparisonof multiple-correlation and simple-correlation based term weighting scheme incora dataset

It can be noted that the term weighting scheme based on multiple correlation of a term with one or more terms simultaneously is more effective than the term weighting scheme that employs simple correlation of only two terms simultaneously.

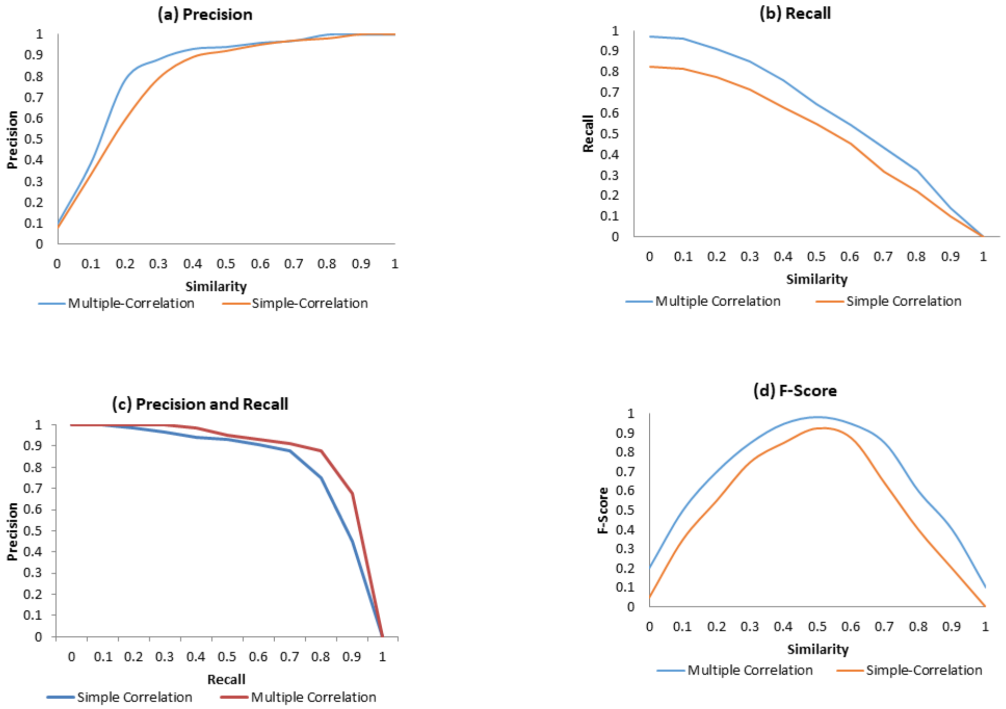

Fig 2. Comparisonof multiple-correlation and simple-correlation based term weighting scheme inrestaurant dataset

Table 4. Maximumf-score values of term weighting scheme based on simple correlation and multiple correlation of a term

	Cora	Restaurant
Simple correlation	0.875	0.962
Multiple correlation	0.925	0.985

5 Conclusions

In this study, a probabilistic multiple correlation based term weighting scheme for measuring the similarity of unstructured textual records is proposed. Term weights are assigned based on how a particular term correlates with one or more number of terms simultaneously in a record. The proposed weighting scheme is effective in dealing with text records that do not have well-defined structure, that employs abbreviations, and that are incomplete. Furthermore, the experimental results demonstrate the improved overall accuracy of the proposed scheme.

References

1) 'Structure vs. Unstructured Data'. 2020. Available from: www.datamation.com.
2) Schneider C. 2016. Available from: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/.
3) Eck NJV, Waltman L. Accuracy of citation data in Web of Science and Scopus. *arXiv*. 2019. Available from: https://arxiv.org/abs/1906.07011.
4) A M. An efficient domain-independent algorithm for detecting approximately duplicate database records. *Proc ACM-SIGMOD Workshop on Research Issues on Knowledge Discovery and Data Mining*. 1997;p. 23–29.
5) Cohen WW. Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems (TOIS)*. 2000;18(3):288–321. doi:10.1145/352595.352598.
6) Song S, Zhu H, Chen L. Probabilistic correlation-based similarity measure on text records. *Information Sciences*. 2014;289:8–24.
7) Bilenko M, Mooney RJ. Learning to combine trained distance metrics for duplicate detection in databases. Austin, TX. 2002. Available from: https://courses.cs.washington.edu/courses/cse590q/04au/papers/BilenkoMooneyTR02.pdf.
8) Pasula H, Marthi B, Milch B, Russell SJ, Shpitser I. Identity uncertainty and citation matching. *Advances in neural information processing systems*. 2003;p. 1425–1432.
9) Data AM. 2020. Available from: https://people.cs.umass.edu/~mccallum/data.html.Retrieved.
10) Tejada S, Knoblock CA, Minton S;Learning object identification rules for information integration. Elsevier BV. 2001. Available from: https://dx.doi.org/10.1016/s0306-4379(01)00042-4; doi:10.1016/s0306-4379(01)00042-4.