Semantic Labeling for Numerical Values: Distribution-Based Similarities

Phuc NGUYEN[1,2] Hideaki TAKEDA[1,2]

[1] 国立情報学研究所
[1] National Institute of Informatics
[2] 総合研究大学院大学
[2] SOKENDAI University(The Graduate University for Advanced Studies)

Abstract: In recent years, there has been an increasing interest in numerical semantic labeling, in which the meaning of an unknown numerical attribute is assigned by the label of the most relevant attributes in predefined knowledge bases. Previous methods used the p-value in statistical hypothesis testing to estimate the relevance and thus strongly depend on the distribution and type of data domain. In other words, the p-value based similarity is unstable for general cases, where such knowledge is undefined. In this paper, we first point out the p-value based similarity limitations. Second, we proposed the Distribution-Based Similarities where the similarities are derived from the norms of the inverse transform sampling of attribute distributions. Our experiments on City Data and Open Data show that the Distribution-Based Similarities outperforms other the p-value based approaches in the task of semantic labeling for numerical values.

1 Introduction

In recent years, there has been an increasing interest in numerical semantic labeling for tabular data where numerical values from table columns are matched to the semantic labels in knowledge bases. It enable data integrated and hence could be potentially useful for other applications such as table search [5, 8], table extension [4], completion [1], or knowledge base construction as used in DBpedia [15], YAGO [11], and Freebase [2].

A common work-flow is the retrieval setting in which the label of a query column is assigned by that of the most relevant columns in labeled data with respect to a specific similarity or distance metric. However, how to select a good similarity or distance metric for numerical attributes is a difficult challenge because of several reasons.

1. **I1:** First, the numerical values of attributes rarely have the same set of values as the relevant values in knowledge bases.
2. **I2:** Second, the size of attributes could vary from a few to millions of numbers. It is hard to use directly apply the normed vector spaces as similarity metrics.
3. **I3:** Third, in general cases, we do not have the predefined knowledge about distribution and type of data.

Previous approaches used the p-value of a statistical hypothesis test as a metric to measure the similarity between numerical attributes [7, 9, 10]. The p-value based similarity address the first (**I1**) and second issue (**I2**), however it cannot be used in the third issue (**I3**). In fact, a statistical hypothesis test strongly depends on assumptions regarding the distribution and type of data. For instance, these data attributes have to be drawn from a specific form of distribution (e.g., normal distribution or uniform distribution) or data types (e.g., continuous or discrete). However, determining the form of distributions and data types of unknown numerical attributes is a difficult challenge. As a result, a proper hypothesis test cannot be easily selected when we do not have such predefined those information.

Moreover, the interpretation of the p-value of statistical hypothesis testing is not clear evidence for measuring the similarity between numerical attributes. We discuss some controversial aspects regarding the use of the p-value in literature as the following section.

2 P-value as a similarity

2.1 What is the p-value:

According to The American Statistical Association p-value definition [13], "Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (e.g., the sample mean difference between two compared groups) would be equal to or more extreme than its observed value."

The definition of p-value may help to summarize the conditional probability of data incompatible with a specified statistical model if the null hypothesis (H_0) is true. The p-value cannot tell us whether the null hypothesis is true or not. The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis. In other words, this incompatibility can be interpreted as providing evidence against the null hypothesis.

In the context of semantic labeling for numerical values, the null hypothesis H_0 is defined as two numerical attributes were drawn from the same semantic labels. The alternative hypothesis is that two numerical attributes were drawn from different semantic labels.

2.2 The use of p-value of other baseline approaches:

Regarding the problem of semantic labeling for numerical values, Stonebraker et al. proposed a method for schema matching using decisions from four *experts* [12]. One decision used the *t-test* statistic of the Welch ' s t-test [3] to measure the probability that two numerical attributes were drawn from the same distribution (semantic labels). Inspired by [12], Ramnadam et al use the p-value to determine the relevance of numerical attributes [10]. The author defined the null hypothesis H_0 is that the two numerical attributes are drawn from the same distribution (semantic labels). The p-value was used as the confidence score for two attributes are drawn from the same semantic labels. Their experiments on the Welch ' s t-test [3], the Mann-Whitney U test [6], and the Kolmogorov Skmiro test (KS test) [3] show that KS test archived the highest performance. The later approaches were also built on top of KS-Test [7,9]. Overall, the general idea is that the p-value was used to measure the level of relevance of numerical attributes.

2.3 First issue: The p-value is the probability of the hypothesis is true

As mentioned, other baselines approaches used the p-value to measure the probability of two attributes are similar or different. However, the p-value tells us the conditional probability of observed sample data when the null hypothesis is valid (H_0 is true). The equation of p-value is shown as follows.

$$p\_value = P(D|H_0) \tag{1}$$

where D is the observed sample data. However, considering the original purposes of the baseline approaches, the similarity metric could be interpreted as the conditional probability $P(H_0|D)$ of the probability of H_0 given a data D. In a Bayesian-like approach, it can be calculated by the following formula.

$$P(H_0|D) = \frac{P(D|H_0) \times P(H_0)}{P(D)} \tag{2}$$

The value of $P(H_0|D)$ and $P(D|H_0)$ (p-value) measure different concepts. Therefore, the p-value tells us the probability of the obtained sample data D when H_0 is valid, rather than telling us the probability of the null hypothesis $P(H_0)$ or the conditional probability $P(H_0|D)$ of the null hypothesis given the obtained sample data D. Moreover, we do not know whether this condition $P(H_0)$ is valid or not in practice, therefore $P(H_0|D)$ can not be derived from the Equation 2.

2.4 Second issue: Comparing p-value to each others

Since the general work-flow is retrieval setting, we need to perform a similar search between the query attributes with all samples in a database. In the p-value based approaches, the p-values of statistical hypothesis testing between the query sample data with other samples in a database are ranked to find the most similar attributes. In other words, p-values are used to measure the level of similarity.

However, in the statistical hypothesis testing, comparing the p-value between different testing does not necessarily imply the level of similarity. Moreover, p-values are affected by the size of the sample data. If we keep the effect size as a constant, then the larger of

sample size will lead to the smaller of the p-value and vice versa. The dependence of p-value with the size of data sample is shown in Equation 2.4 (t-test), Equation 6 (U test), and Equation 8 (KS test). In general cases, the size of numerical attributes could vary from a few to millions of numerical values. Therefore, it is controverted to use only the p-values as a measurement of a significant level of similarity.

Welch ' s t-test

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_1}}} \tag{3}$$

$$df = min(n_1, n_2) - 1 \tag{4}$$

$$p\_value_{t\_test}(t, df) = \frac{1}{2} + (t\frac{\Gamma[\frac{1}{2}(df+1)]}{\sqrt{\pi df}\Gamma(\frac{df}{2})}) \; {}_2F_1(\frac{1}{2}, \frac{1}{2}(df+1); \frac{3}{2}; \frac{-t^2}{df})$$

Mann-Whitney ' s U test

$$u = n_1 n_2 + \frac{n_2 * (n_2 + 1)}{2} - \sum_{i=n_1+1}^{n_2} R_i \tag{5}$$

$$p\_value_{Utest}(t, n_1, n_2) = \frac{|u - \frac{n_1(n_1+n_2+1)}{2}| - 0.5}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \tag{6}$$

Kolmogorov-Smirnov test

$$D(n_1, n_2) = \sup_x |F_{1,n_1}(x) - F_{2,n_2}(x)| \tag{7}$$

$$p\_value_{KStest}(D, n_1, n_2) = e^{-2D^2 \times \frac{n_1 \times n_2}{n_1 + n_2}} \tag{8}$$

3 Distribution-Based Similarities

In this section, we introduce the new categories of similarity metrics, called Distribution-Based Similarities (DBS). The similarities are derived from a norm of the inverse transform sampling of numerical attributes. DBS address all the three mentioned issues: **I1, I2, I3**.

1. **I1:** The similarity is derived from distributions of numerical attributes, therefore it is not necessary the assumption that the values of numerical attributes have the same set of values.
2. **I2:** In DBS, we introduce an attribute transformation (Section 3.1) to transform the list of numerical values to a distribution representation as well as standardize the input size of numerical attributes. Therefore, after transformation, the numerical attribute has a representation as a vector with h size.
3. **I3:** DBS is derived from the empirical distribution of numerical attributes without the need to make any assumption regarding data type or data distribution.

The overall framework of semantic labeling with DBS shows in Figure 1. The framework consists of two phases. The first phase involves data preparation and knowledge base construction while the second phase is actually semantic labeling.

In the first phase, given labeled numerical attributes, the attribute transformation converts these labeled attributes from numerical values into distribution presentations. Then, these distribution presentations are stored in the knowledge base for future similarity comparison.

In the second phase, the numerical values of an unknown attribute are standardized with attribute transformation. Then the similarity search module is used to calculate the similarities between these distribution representations. In this paper, we consider three typical distance of the Minkowski distance: the Manhattan distance (called DBS1), the Euclidean distance (called DBS2), and the Chebyshev distance (called DBSinf). After the similarity searching process, we have a ranking list of semantic labels ordered by their corresponding similarity scores.

3.1 Attribute Transformation

In this section, we describe the transformation of numerical attributes to standardize the input size for the representation learning. Attribute transformation is an important module because the representation learning requires a standardized input size, and the size of numerical attributes could vary from a few to thousands of values. To standardize the input size, we use inverse transform sampling [14] (Section 3.1.1) to standardize the input size and transform numerical values into forms of distribution presentations.

Given an attribute a having numerical values $V_a = [v_1, v_2, v_3, ..., v_n]$, the objective of attribute transformation is the $trans(V_a)$ function, which is defined as

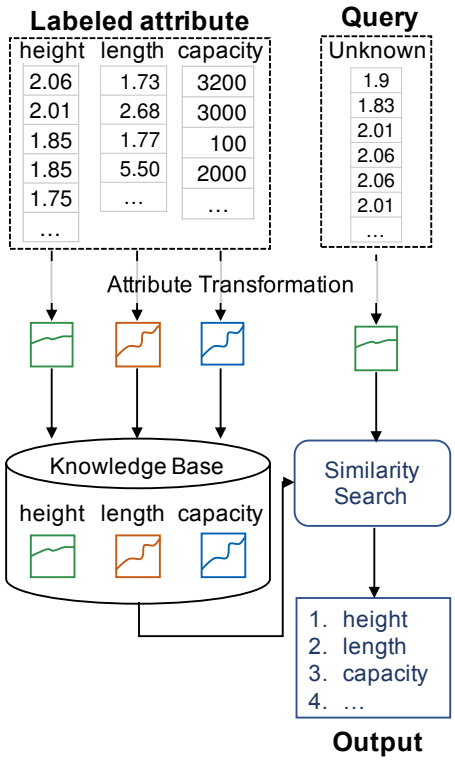


図 1: General architecture of semantic labeling for numerical attributes with DBS

follows.

$$x = trans(V_a) = x_{icdf} \quad (9)$$

The transformation function $trans(.)$ converts V_a into x, where $x \in \mathbb{R}^h$. The list of values $x_{icdf} \in \mathbb{R}^h$ is obtained by the transformation using the inverse transform sampling (Section 3.1.1) on numerical values. The inverse transform sampling is described as follows.

3.1.1 Inverse Transform Sampling

Let a be an attribute with numerical values $V_a = [v_1, v_2, v_3, ..., v_n]$. We treat V_a as a discrete distribution so that the CDF of $v \in V_a$ is $cdf_{V_a}(v)$ and expressed as follows.

$$cdf_{V_a}(v) = P(V_a \leqslant v), v \in V_a, cdf_{V_a} : \mathbb{R} \rightarrow [0, 1] \quad (10)$$

where $P(.)$ represents the probability of values in V_a less than or equal to v. The inverse function of $cdf_{V_a}(.)$ takes the probability p as input and returns $v \in V_a$ as follows.

$$icdf_{V_a}(p) = cdf^{-1}_{V_a}(p) = min\{v : cdf_{V_a}(v) \geqslant p\}, \quad p \in [0, 1]$$

We select h numbers from V_a where each number is the output of the inverse distribution function $icdf_{V_a}(p)$ with probability $p \in \mathcal{P} = \{\frac{i}{h} | i \in \{1, 2, 3, ..., h\}\}$. For example, when the input size $h = 100$, then we have $\mathcal{P} = \{0.01, 0.02, 0.03, ..., 1\}$. For each attribute $a \in A$, we have a list of values $x_{icdf} = \{v_1, v_2, v_3, ..., v_h\}$ that correspond to the given list of probabilities $\mathcal{P}$.

4 Evaluation

In this section, we first describe the benchmark datasets, evaluation metrics, compared baseline approaches, experimental setting, and experimental results.

4.0.1 Dataset

To evaluate EmbNum+, we used two datasets i.e., City Data, Open Data. City Data is the standard data used in the previous studies [9], [10] while Open Data is newly built datasets extracted from Open Data portals. The datasets are available at this address [1].

The detailed statistics of each dataset are shown in Table 1. m denotes for the number of semantic labels in a dataset. n denotes for the number of columns in a dataset. In each dataset, each semantic label has 10 columns in the same semantic labels. The columns of City Data is randomly generated using 10 partitions splitting, while the columns of Open Data are the real table columns from Open Data Portals. The number of semantic labels of the new datasets is larger than City Data, enabling rigorous comparisons between DBS and other baseline approaches.

4.1 Evaluation Metrics

We used the mean reciprocal rank score (MRR) to measure the effectiveness of semantic labeling. The MRR score was used in the previous studies [10], [9] to measure the probability correctness of a ranking result list.

4.2 Compared Baseline Approaches

We evaluate the performance of DBS1, DBS2, DBSinf with two baseline approaches SemanticTyper [10],

[1] https://github.com/phucty/embnum

表 1: Statistical description about the number of numerical values per semantic label in four datasets: City Data, Open Data

Dataset	m	n	# values of each labels				
			all	min	max	med	avg
City Data	30	300	192,820	40	22,510	1,130	6,427.33
Open Data	50	500	7,329,815	120	1,671,455	12,506	146,596.3

and DSL [9]. SemanticTyper used the KS test as the similarity metric for numerical attributes [10]. DSL used a new metric with a combination of KS Test, U Test, and the numeric Jaccard similarity.

4.3 Experimental Setting

In this section, we describe the detail experimental setting to evaluate the semantic labeling task. We follow the evaluation setting of SemanticTyper [10] and DSL [9]. This setting is based on cross-validation but it was modified to observe how the number of numerical values in the knowledge base will affect the performance of the labeling process. The detail of the experimental setting is described as follows.

Suppose a dataset $S = \{s_1, s_2, s_3, ..., s_d\}$ has d data sources. One data source was retained as the unknown data, and the remaining $d-1$ data sources were used as the labeled data. We repeated this process d times, with each of the data source used exactly once as the unknown data.

Additionally, we set the number of sources in the labeled data increasing from one source to $d-1$ sources to analyze the effect of an increment of the number of labeled data on the performance of semantic labeling. We obtained the MRR scores and labeling times on $d \times (d-1)$ experiments and then averaged them to produce the $d-1$ estimations of the number of sources in the labeled data.

4.4 Experimental Results

The results of semantic labeling for numerical values in the MRR score on City Data and Open Data is shown in Figure 2.

The MRR scores obtained by three methods steadily increase along with the number of labeled sources. It suggests that the more labeled sources in the database, the more accurate the assigned semantic labels are. DSL outperformed SemanticTyper in City Data and Open Data because it used the information from multiple testing results.

The DBS outperform all baseline approaches. The similarity metric based on a specific hypothesis test, which was used in SemanticTyper and DSL, is not optimized for semantic meanings with various data types and distributions in general cases. In three tested DBS, the DBS1 (Manhattan distance) archived the highest performance in the two datasets.

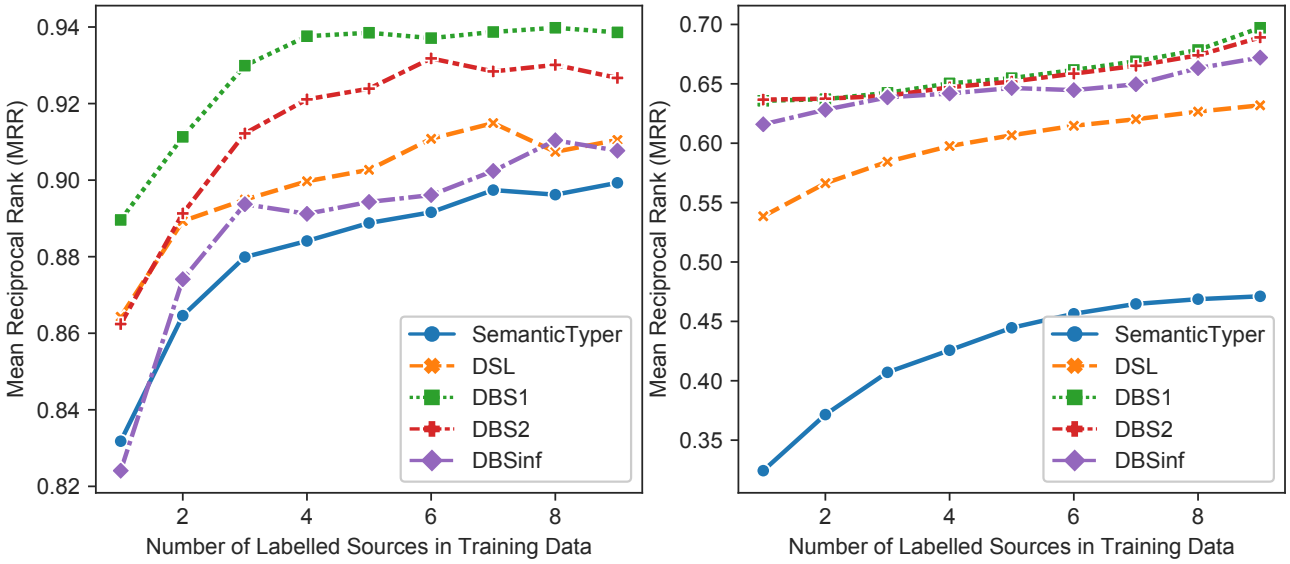
5 Conclusion

In this paper, we first point out the limitation of the p-value based similarity. These other baselines relied on the p-value are unstable for general cases. Then, we introduce DBS, a category of similarities derived from the norms of the inverse transform sampling of numerical attributes. The experimental results showed that DBS1 (Manhattan distance) achieved the best performance for the task of semantic labeling for numerical values. In future work, we plan to conduct more experiments on the larger dataset to understand the robustness as well as the efficiency of the DSB.

参考文献

[1] Ahmadov, A., Thiele, M., Eberius, J., Lehner, W., Wrembel, R.: Towards a hybrid imputation approach using web tables. In: 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC). pp. 21–30 (Dec 2015).

[2] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 601–610. ACM (2014)



(a) City Data

(b) Open Data

図 2: Semantic Labeling in the MRR score on City Data, Open Data

[3] Lehmann, E.L., Romano, J.P.: Testing statistical hypotheses. Springer Texts in Statistics, Springer, New York, third edn. (2005).

[4] Lehmberg, O., Ritze, D., Ristoski, P., Meusel, R., Paulheim, H., Bizer, C.: The mannheim search join engine. Web Semant. **35**(P3), 159–166 (Dec 2015).

[5] Nargesian, F., Zhu, E., Pu, K.Q., Miller, R.J.: Table union search on open data. Proc. VLDB Endow. **11**(7), 813–825 (Mar 2018).

[6] Neuhauser, M.: Wilcoxon–Mann–Whitney Test, pp. 1656–1658. Springer Berlin Heidelberg, Berlin, Heidelberg (2011).

[7] Neumaier, S., Umbrich, J., Parreira, J.X., Polleres, A.: Multi-level semantic labelling of numerical values. In: International Semantic Web Conference. pp. 428–445. Springer (2016)

[8] Nguyen, T.T., Nguyen, Q.V.H., Weidlich, M., Aberer, K.: Result selection and summarization for web table search. In: 2015 IEEE 31st International Conference on Data Engineering. pp. 231–242 (April 2015).

[9] Pham, M., Alse, S., Knoblock, C.A., Szekely, P.: Semantic labeling: a domain-independent approach. In: International Semantic Web Conference. pp. 446–462. Springer (2016)

[10] Ramnandan, S.K., Mittal, A., Knoblock, C.A., Szekely, P.: Assigning semantic labels to data sources. In: European Semantic Web Conference. pp. 403–417. Springer (2015)

[11] Sekhavat, Y.A., Paolo, F.D., Barbosa, D., Merialdo, P.: Knowledge base augmentation using tabular data. In: LDOW. CEUR Workshop Proceedings, vol. 1184. CEUR-WS.org (2014)

[12] Stonebraker, M., Bruckner, D., Ilyas, I.F., Beskales, G., Cherniack, M., Zdonik, S.B., Pagan, A., Xu, S.: Data curation at scale: The data tamer system. In: CIDR (2013)

[13] Wasserstein, R.L., Lazar, N.A.: The asa's statement on p -values: Context, process, and purpose. The American Statistician **70**(2), 129–133 (2016),

[14] Wikipedia contributors: Inverse transform sampling — Wikipedia, the free encyclopedia (2018), [Online; accessed 3-July-2018]

[15] Zhang, M., Chakrabarti, K.: Infogather+: semantic matching and annotation of numeric and time-varying attributes in web tables. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 145–156. ACM (2013)