How To Find The Useful Information For The Referential Integrity Problem

Chia-Chih Hsu, Nigel R. Shadbolt[1]
[1]Intelligence, Agents, Multimedia Group,
Department of Electronics & Computer Science,
University of Southampton
{cch02r, nrs}@ecs.soton.ac.uk

Abstract

The solution for the referential integrity problem is mainly addressed into what the useful information is. Thus, in contrast with the usual heuristics extracted from the problem domain, the novel approach of our research is to formally model the way to explain what the useful information is. We find that (i) the fewer attributes with more truth perform better than as many as possible attributes; (ii) the converging similarities can be predicted using the mathematical modelling.

Keywords: the diagnosticity principle, deviation

1. Introduction

Referential integrity is the identification of instances referring to the same object in a data store, such as ontology and database. The instances of interest are objects with attributes storing different facets of informative description about objects. When harvesting instances into a single data store, the collected instances could bring heterogeneous information, redundant information, or only partial description around the same object or concept [4, 8, 3].

There are two potential directions to manage referential integrity. One is to consider informativeness and distinctiveness [5, 10], as well as the other to content and semantics that are defined by content representation (or structure) [6, 2, 4, 1]. In other words, on the one hand, informativeness is emphasized on the aspect of complete and common parts for contents or semantics, whereas distinctiveness is on the aspect of specific and distinctive parts for contents and semantics. On the other hand, the information used to assist referential integrity comes from two sources of content and semantics hidden in the structure. The contents can be viewed as the realization or implementation of the semantics.

The focal attention of this paper is to computationally model the process of finding instances with referential integrity, given the related instances and the *deviation*, as well as to determine the criteria for extracting the useful information. This paper is structured as related works, the deviation-based referential integrity model, experiments, and finally conclusion and future work.

2. Related Works

In the context of referential integrity, three academic elements will be utilized to review the referential integrity problem. First, with the semantics, starting from an extreme of shared consensus between humans, the effort put on the semantics in a data representation is to convey meanings of concepts and relations toward the other extreme of explicit and formal semantics for machine understanding [11]. Moreover, the classification of attributes in several domains [11] tries to make use of the anatomical view of information of an object as well as to cover complete description of an object;

Second, in order to acquire the accurate relevance of web pages, the content and link structure of web pages can give two dimensions of information for the search engine [9, 6]. However, there are two major incurred issues about web pages in the search problem. One is that the web pages are also noised with imperfect information, such as spam information. The other is that web pages could possibly contain multiple topics, such as text, hyperlink, and images [6]. Hence, the performance of web page ranking is improved if both contents and links are adopted at the same time and the link structure is deeply explored.

Third, with regard to learning issue, the mechanism [7] is implemented by introducing positive and negative training examples for the target concept. The adopted examples contain known facts, but the truth facts for instances in the referential integrity problem are not known. The selection of attributes employed by ID3, a basic algorithm to implement the decision tree learning, is to evaluate each attribute's capability in partitioning training examples.

3. The Deviation-based Referential

Integrity Model

3.1 The Background

The diagnosticity principle was originated from the idea of 'A change of clusters, in turn, is expected to increase the diagnostic values of features on which the new clusters are based, and therefore, the similarity of objects that share these features' [10]. When a pair of matching instances is evaluated through this principle, the similarity of the pair is contrasted with the *deviation*. The *deviation* can be defined as 'a rate to evaluate the change of similarity in the original pair of matching instances on a subset of attributes, after taking related instances into consideration'. While enforcing the diagnosticity principle, the whole process is named as the deviation-based referential integrity model.

3.1 The Settings

The deviation-based referential integrity model is characterized by the following components:

(1) The basic correction equation

$$Y^t = M \times Y^{t-1} \tag{1}$$

where

- Y^t is a column vector storing the similarities of pairs of matching instance at iteration t, $t \geq 1$; Y^0 is the initial similarities at iteration 0;
- M is a square matrix storing the deviations.

(2) The similarity vector Y^t

$$Y^t = \left(y_1^t, ...y_i^t, ..., y_n^t\right)^T \tag{2}$$

where

- T denotes the transpose operation;
- i is the ith pair of matching instances, $1 \leq i \leq n$;
- $y_i^t = similarity(a1_i, a2_i)^t$ is the similarity of the pair of $(a1_i, a2_i)$ at iteration t, $t \geq 1$;
- Following the terms used in last section, y_i^0 is the initial similarities of the pair of $(a1_i, a2_i)$, based on a subset of attributes for the initial similarities.

(3) The deviation matrix M

$$\mathrm{M} = \begin{bmatrix} ... & ... & ... \\ ... & m_{i,j} & ... \\ ... & ... & ... \end{bmatrix} \tag{3}$$

where

- $i, j = 1, ..., n$;

The matrix records the deviations enumerated among the initial similarities. Whilst being concerned with the ith pair, $m_{i,j} = 0$ if y_j^0 has no relationships with y_i^0. $m_{i,j}$ is retrieved from a deviation function, $deviation(y_i) = function\left(y_{k_1}^0, ..., y_{k_n}^0\right)$ where $k_1, ..., k_n$ are the related ones with the concerned pair of $(a1_i, a2_i)$. For example, $y_i^1 = y_i^0 + \left(\sum_{j=1}^{n}\left[\left(y_i^0 - y_j^0\right) \times y_j^0\right]\right)$, which will then be normalized to fall into the interval of $[0.0, 1.0]$.

3.2 The Convergence Of The Deviation-based Referential Integrity Model

There are two questions of interest for the deviation-based referential integrity model.

Question 1: When will the iterated process converge to a stable state?

Question 2: When the process reaches convergence, what can be known from the outcome of Y^t?

■ The answer to question 1

Regarding the issue of similarity convergence, we need to find the criterion for the convergence of Y^t. Assuming $X_1, ..., X_n$ are the eigenvectors of M, then $MX_i = \lambda_i X_i$ where the scalar λ_i is the eigenvalue. Moreover, any arbitrary vector Y can be decomposed into a linear combination of the eigenvectors, or $Y = \sum_{i=1}^{n} a_i X_i$ where a_i is a scalar. Therefore, if $\lambda_1 > \lambda_2 \geq ... \geq \lambda_n$, then $\lim_{t \to \infty} \frac{Y^t}{Y^{t-1}} = \lambda_1$ (4)

From the above, the existence of the largest λ_1 for $\lim_{t \to \infty} \frac{Y^t}{Y^{t-1}}$ is the criterion for the convergence of Y^t. Moreover, the convergence means that the similarities of Y^t have reached a stable state and cannot further be amended by deviations.

A more general situation is that $\lim_{t \to \infty} \frac{Y^t}{Y^{t-1}}$ may be collapsed into a number of distinct converging values such as $\lambda_1, \alpha_1, \beta_1, ...$. Therefore, according to equation (4), $Y^0 = \left(sub0\_Y^0, sub1\_Y^0, sub2\_Y^0, \right)$ can be mapped onto the converging values of $\lambda_1, \alpha_1, \beta_1, ...$.

In other words, $\lim_{t \to \infty} \frac{sub0\_Y^t}{sub0\_Y^{t-1}} = \lambda_1$, $\lim_{t \to \infty} \frac{sub1\_Y^t}{sub1\_Y^{t-1}} = \alpha_1$, $\lim_{t \to \infty} \frac{sub2\_Y^t}{sub2\_Y^{t-1}} = \beta_1$, and so on.

■ The answer to question 2

If $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_n$, M guarantees that Y^0 can be stabilized to $\lambda_1^t a_1 X_1$ after enough number of iterations. The converging similarities of $\lambda_1^t a_1 X_1$ for Y^0 are proportional to the eigenvector with the largest eigenvalue. Thus, due to the constant of $\lambda_1^t a_1$, X_1 alone is sufficient to predict the matching pairs. It is interesting to note that X_1 depends on M directly. If M carries information closer to the truth, X_1 can present the true information as well and, then, show better performance. In summary, the smaller set of attributes can do a better job of similarity convergence. Second, the categorization of information, such as features and relational information, is not a requirement for the deviation-based referential integrity model.

4. Experiments

4.1 The Setting And Goals In The Experiments

We collect the test set from cora-refs.tar.gz (http://www.cs.umass.edu/~mccallum/code-data.html) for our experiments that has the correct answers for the matching pairs. The file consists of three test sets that are the citations for the publications referring to the same or different papers. The three test sets, named as fahl_labelled, kibl_labelled, and utgo_labelled, each of them having 14 attributes for each instance.

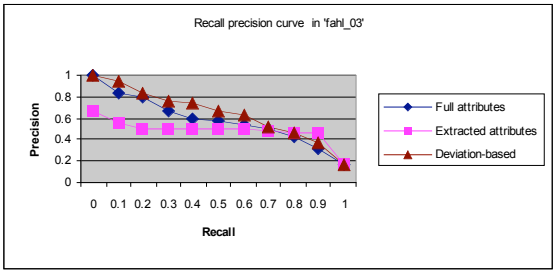
The goals in the empirical study are:

(1) to compare the performance of the deviation-based referential integrity model with those of the edit-distance approaches;

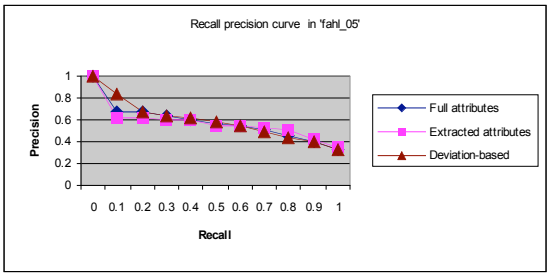
(2) to see how the performances can be changed, if opting for a true attribute;

4.1 The Experimental Results

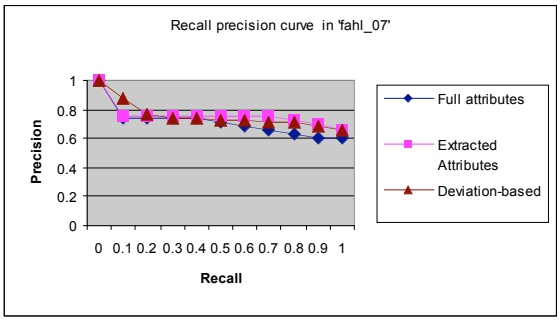
In the experiments, we use the full set of attributes for the computation of initial similarities and related pairs of instances, which is based on the observation of retaining as much information as possible to own potential relationships. Moreover, we use the attributes in the upper three levels of the binary decision tree as the subset of attributes for deviations. The binary decision tree is built up through the classifier J48 of Weka [12] because the selected attributes are expected to carry more truth.



(a)



(b)



(c)

Figure 1 Three recall and precision curves on fahl_labelled with the filtering threshold of (a) 0.3; (b) 0.5; (c) 0.7.

Due to the limit of space, figure 1 only shows the three recall and precision cures for the test set of fahl_labelled that has 529 instances. The results show:

(1) The performance of the deviation-based referential integrity model performs best overall on the test sets with more noise, such as that with the filtering threshold of 0.3 for initial similarities.

(2) On average, the performances of the three methods are ranked decreasingly from the deviation-based method, the edit-distance method based on the extracted subset of attributes, to the edit-distance method based on the full set of attributes.

Second, even through the supervised learning procedure, the possibly true subset of attributes by means of J48 classifier may contain the erroneous values. Yet attribute 'key' can surely identify whether two instances refer to the same object because it contains the true information.

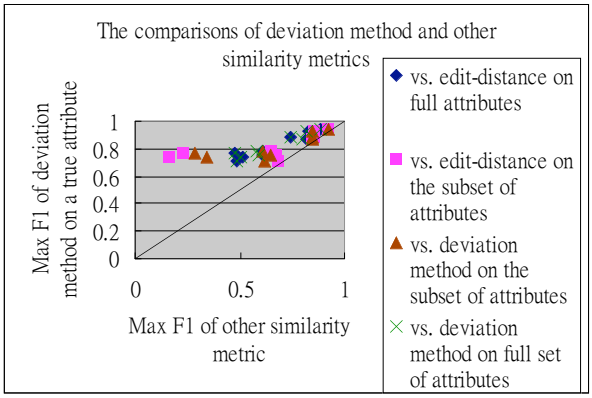


Figure 2 The comparisons of maximum *F1* values of deviation model and other similarity metrics.

Likewise, there are nine test sets to be used for the experiments, which are the combination of three original test sets as well as three filtering threshold of 0.3, 0.5, 0.7. We will use a single value of maximum *F1* score to summarize the ranking performance, which is defined as the harmonic mean of recall and precision, $F1 = \frac{2 \times P \times r}{P + r}$. Following the evenly interpolated graphs plotted in figure 1, we define maximum *F1* score as the maximum value of *F1* that are obtained from the graphs. The experimental results show:

(1) The maximum F1 values of the deviation-based referential integrity model on a true attribute are much greater than those of other similarity metrics, since all points are located above the line of $y = x$.

(2) The claim that more truth in deviations can produce better referential integrity performance is confirmed in the test sets.

5. Conclusions and Future Work

Our research concludes the following novel findings for referential integrity.

(1) The criteria for referential integrity are in close relationships with the truth facts of the matching instances, not on the classification of the characteristic attributes.

(2) Furthermore, even if only based on the true value on one attribute, the performance of the deviation-based model is still better than those of other similarity metrics on more attributes. It further explains the implication that when heuristics are adopted by people, they are assumed as the true information.

(3) The conclusion of using attributes with more truth information is consistent with the enforcement of the diagnosticity principle, which describes the attributes with larger diagnostic values are qualified as the distinct features [10].

This proposed work will elaborate on the quantitative measurements for the appropriate formulation between ratios of truth and the resulting performance. The truth facts for matching instances are usually unknown. Though not perfect, the supervised leaning method and certain heuristics from observation about the matching instances can become the feasible sources of true information, such as the values of the attributes with more variants could be less true rather than those with fewer variants.

References

[1] H. Alani, S. Dasmahapatra, K. O'Hara and N. Shadbolt, *Identifying Communities of Practice through Ontology Network Analysis*, IEEE Intelligent Systems, 18 (2003), pp. 18-25.

[2] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar and S. Fienberg, *Adaptive name matching in information integration*, IEEE Intelligent Systems, 18 (2003), pp. 16- 23.

[3] J. Broekstra, M. Ehrig, P. Haase, F. v. Harmelen, M. Menken, P. Mika, B. Schnizler and R. Siebes, *Bibster - A Semantics-Based Bibliographic Peer-to-Peer System*, 2004.

[4] A. Doan, Y. Lu, Y. Lee and J. Han, *Object Matching for Information Integration: A Profiler-Based Approach*, *IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*, Acapulco, Mexico, 2003.

[5] G. Hirst, *Ontology and the Lexicon*, in S. Staab and R. Stude, eds., *Handbook on Ontologies*, Springer, Berlin, 2003, pp. 209-229.

[6] W.-Y. Ma, *From Relevance to Intelligence: Toward Next Generation Web Search*, *Microsoft-TsingHua Information Techniques Seminar*, National Tsing-Hua University, HsinChu, Taiwan, 2005.

[7] T. M. Mitchell, *Machine Learning*, 1997.

[8] S. Parsons, *Current approaches to handling imperfect information in data and knowledge bases*, IEEE Transactions on Knowledge and Data Engineering, 8 (1996), pp. 353-372.

[9] T. QIN, T.-Y. LIU, X.-D. ZHANG, Z. CHEN and W.-Y. MA, *A Study of Relevance Propagation for Web Search*, *The 28th Annual International ACM SIGIR*, Salvador, Brazil, 2005.

[10] A. Tversky, *Features of Similarity*, Psychological Review, 84 (1977), pp. 327-352.

[11] M. Uschold and R. Jasper, *A Framework for Understanding and Classifying Ontology Applications*, *Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5)*, Stockholm, Sweden, 1999, pp. 12.

[12] a. H. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco, 2005.