

# Entropy Based Measurement of Text Dissimilarity for Duplicate – Detection

Venkatesh Kumar (Corresponding author)

Research Scholar, Department of Mathematics, Kongu Engineering College

Perundurai-638052 Tamilnadu, India

Tel: 91-98-9499-9415 Fax: 04294 220087 E-mail: venkateshakumar@gmail.com

Dr. G. Rajendran

Professor and Head, Department of Mathematics, School of Science and Humanities

Kongu Engineering College, Perundurai-638052 Tamilnadu, India

Tel: 91-94-4277-1483 Fax: 04294 220087 E-mail: rajendranprf@gmail.com

## Abstract

The problem of identifying approximate similarity between pair of strings is an essential step for data cleansing and data integration process. Most existing approaches have relied on generic or manually tuned distance metrics for estimating the similarity potential duplicate. But existing system does not produce the similarity percentage between pair of strings. In this paper we propose a method using entropy and information gain (IG) to find dissimilarity between pair of strings to increase the accuracy of data.

**Keywords:** Duplicate detection, Similarity measures, Entropy, ID3, Information gain

## 1. Introduction

Databases play an important role in today's IT-based economy. Many industries and systems depend on the accuracy of database to carry out operations. Therefore, the quality of information stored in the database can have significant cost implication to a system that relies on information to function and conduct business. In an error free system with perfectly clean data, the construction of a comprehensive view of the data consist of linking-in relational terms, joining-two or more table on their key fields. Unfortunately, data often lack a unique, global identifier that would permit such an operation. Furthermore, the data are neither carefully controlled for quality nor defined in a consistent way across different data sources. Thus, data quality is often compromised by many factors, including data entry error.

When database contain records that were collect from multiple information sources, they frequently include field values and tuples that refer to the same entity, but are not syntactically identical. Hence, maintaining quality of data is one of the most critical problems faced in the data warehousing. Since data is entered in the system by different people, in different standards and at different levels, so this results in several data quality issues. One of the top most issues is having multiple representation of same logical real world entity or in other words having duplicate records for same entity at the time of data entry. Both entity may be an exactly or approximately equal string. Such duplicate records at any point of time might be crucial but this ends in a lot of duplicate data in system which can affect to business.

Similarity is a complex concept which has been widely discussed in the linguistic, Philosophical and information theory communities [Hatzivassiloglou et al., 1999]. [Frawley, 1992] discusses all semantic typing in terms of two mechanisms: the detection of similarities and difference. An effective method to compute the similarity between short texts or sentences has many applications in natural language processing and related areas such as information retrieval to be one of the best techniques for improving retrieval effectiveness [Park et al., 2005] and in image retrieval from the Web, the use of short text surrounding the images can achieve a higher retrieval precision than the use of the whole document in which the image is embedded [Coelho et al., 2004]. The use of text similarity is beneficial for relevance feedback and text categorization [Ko et al., 2004; Liu and Guo, 2005], text summarization [Erkan and Radev, 2004; Lin and Hovy, 2003], word sense disambiguation [Lesk, 1986; Schutze, 1998], methods for automatic evaluation of machine translation [Liu and Zong, 2004], evaluation of text coherence [Katarzyna and Szczepaniak, 2005; Lapataand Barzilay, 2005], formatted documents classification [Katarzyna and Szczepaniak, 2005].

Here we point out some drawbacks of the existing methods. Distinguish those draw back as two types technical, functional. As a technical point of view, Simple similarity measures for string matching include character n-gram

similarity, the Levenshtein distance [Vladimir I. Levenshtein, 1965] and the Jaro–Winkler measure [Winkler, 1999], in which the same penalty value is used regardless of the characters to be matched (or ignored). In general, these simple measures do not work very well for generate similarity percentage of pair of string. As a functional point of view, One of the major drawbacks of most of the existing system is depending on the particular domain, i.e., once the similarity method is designed for a specific application domain, it cannot be adapted easily to other domains, most of the system require minimum changes for move one domain to another domain. This lack of adaptability to the domain does not correspond to human language usage as sentence meaning may change, to varying extents, from domain to domain. To address this drawback, we aim to develop a method that is fully automatic without requiring users' feedback and can be used independent of the domain in applications requiring text similarity measure.

For this purpose, we have introduced a new method by using entropy and information gain by considering two parameters C and R as a text. Here our work is automatically determining the dissimilarity percentage between pair of strings. In entropy based duplicate detection form a generic dynamic truth table by decision making algorithm, calculate the entropy and information gain and deals with the application of the new method. An example is given to illustrate the efficiency of our propose method.

## 2. Preliminaries

In this section, for the convenience of formulation of proposed method. We need to introduce some preliminaries about entropy and information gain.

### 2.1 Entropy

A measurement of the text similarity of a system. Systems tend to go from a state of order (low entropy) to a state of maximum disorder (high entropy).The entropy of a system is related to the amount of information it contains.

### 2.2 Information Gain

One of the most important components for calculating dis-similarity measurement is the criterion used to select which attribute will become a subset attribute in given pair of strings. There are different criteria one of the most well known is information gain.

## 3. Entropy Based Duplicate Detection

### 3.1 Characteristics of Entropy Based Dissimilarity Calculation

In order to facilitate to calculate dissimilarity percentage, we first analyze the characteristics of formulas in proposed method. A proposed method may contain various kinds of component, such as algorithm, formulas, truth table construction, entropy calculation, and Information Gain calculation. In order to represent those characteristics,

- Apply decision making algorithm for truth table construction.
- Apply entropy formula for calculate entropy value.
- Apply entropy value to ID3 for calculate Information Gain.

### 3.2 Decision Making Algorithm

```

For (i=1 to n)
  Loop
  For (i=1 to n)
    Loop
      If (Ci equals 1)
        Then
          Exit from the loop;
        End If
      If (Ci equals Rj)
        Then
          Val=1;
        Else

```

```

                Val=0;
            End If
        End Loop
    End Loop

```

### 3.3 Truth Table Construction

To calculate the logical representation for pair of text characters, a common truth table is constructed. In pair of strings which one have more length Where  $C_i$ ,  $i=1.....n$  represents the  $i$ 'th character of column.  $R_j$ ,  $j=1.....m$  represents the  $j$ 'th character of row in Table 1.

### 3.4 Entropy Calculation

Entropy is a formula to calculate the homogeneity of a sample data. If the value of entropy is 0 it is a completely homogenous sample data. If the value of entropy is 1 than it is an equally divided sample data. The entropy formula for negative and positive element is,

$$\text{Entropy}(p, n) = -p \log_2(p) - n \log_2(n)$$

### 3.5 Information Gain

The information gain is based on the decrease in entropy after a dataset is split on an attribute. First the entropy of the total dataset is calculated. Second, the dataset then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split it is entropy value of child dataset. The resulting entropy is subtracted from the entropy before the split it is entropy value of total dataset.

$$\text{Gain}(G) = \sum (\text{entropy values of child dataset}) - \sum (\text{entropy values of total dataset}) * 100$$

The Gain value is the dis-similarity percentage of pair of text. It shows that the use of a parameterized value allows calculating dis-similarity measure at any point of time for any text values, only changing the parameter, no need to modify the system.

The experiment results obtained show that it is possible to use an automatic method to select the parameters value of the system than produce dis-similarity measurement of pair of parameter text and flexibility introduced by the parameters of text value are useful to better tune the fuzzy automaton in order to frequent dis-similarity calculation for a particular problem. The system not only offers good accuracy without knowing a prior the kind of errors that the input parameter text could contain or not. Its parameters can also be trained to treat as a text in order to better tune the fuzzy automaton accuracy for a particular problem.

## 4. Experiment Result

Table 2 shows how to construct the truth table for sample data "LEVEL" and "LEVAL". The Length of string "C" is equal to five, number of 0's occurs in total is one times and number of 1's occurs in total four times. If we apply theses value to entropy based duplicate detection (3) in all steps, this method (3) will produce a dissimilarity percentage nothing but GAIN value "G" as 20 percentage.

## 5. Conclusion

In this paper, we have proposed a method to identify dissimilarity percentage between pair of text. First we have introduced decision making algorithm for calculating logical values to strings C and R. Second calculated the entropy and information gain to find the dissimilarity percentage. The new method offers more accuracy of data without user feedback at the time of duplicate detection.

## References

- Erkan, G.,Radev, D. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22, 457-479.
- Frawley,W. (1992). *Linguistic Semantics. Lawrence Erlbaum Associates*.Hillsdale, New Jersey.
- Hatzivassiloglou, V., Klavans, J., and Eskin, E. (1999). Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Very Large Corpora,203-212.
- Katarzyna, W.,Szczepaniak, P. (2005). Classification of rss-formatted documents using full text similarity measures. *In Proceedings of the Fifth International Conference on Web Engineering*, LNCS 3579, 400-405.
- Ko, Y., Park, J., and Seo, J. (2004). Improving text categorization using the importance of sentences. *Information Processing and Management*,40, 65-79.

Landauer, T., Dumais, S. (1997). A solution to Plato’s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104, 2, 211-240.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. *In Proceedings of the SIGDOC Conference*.

Lin, C., Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. *In Proceedings of the Human Language Technology Conference*.

Liu, T., Guo, J. (2005). Text similarity computing based on standard deviation. *In Proceedings of the International Conference on Intelligent Computing*, D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds. LNCS 3644, 456-464.

Liu, Y., Zong, C. (2004). Example-based chinese-english mt. In *Proceedings of the 2004 IEEE International Conference on Systems, Man, Cybernetics*. Vol. 1-7,6093-6096.

Maguitman, A., Menczer, F., Roinestad, H., Vespignani, A. (2005). Algorithmic detection of semantic similarity. *In Proceedings of the 14th International World Wide Web Conference*.

Schutze, H. (1998). Automatic word sense discrimination. *Computational Linguistics* 24, 1, 97-124.

Vladimir I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals(in russian), *Doklady Akademii Nauk SSSR*, 4. (1965). pp. 845–848. Appeared in English as: V. I. Levenshtein, Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8), 707–710, 1966.

Winkler WE. (1999). The state of record linkage and current research problems. *Technical report Statistical Research Division*, U.S. Bureau of the Census.

Table 1. Truth table construction for calculating the logical representation for pair of text characters

	C1	C2	.....	Cn
R1	R1C1	R1C2	.....	R1Cn
R2	R2C1	R1C2		
.	.			
.	.			
.	.			
Rm	RmC1	RmC2	.....	RmCn

Table 2. Truth table construction for sample data

	L	E	V	E	L
L	1	0	0	0	0 / 1
E	0	1	0	0 / 1	0
V	0	0	1	0	0
A	0	0	0	0	0
L	0 / 1	0	0	0	1
TOTAL	1	1	1	0	1

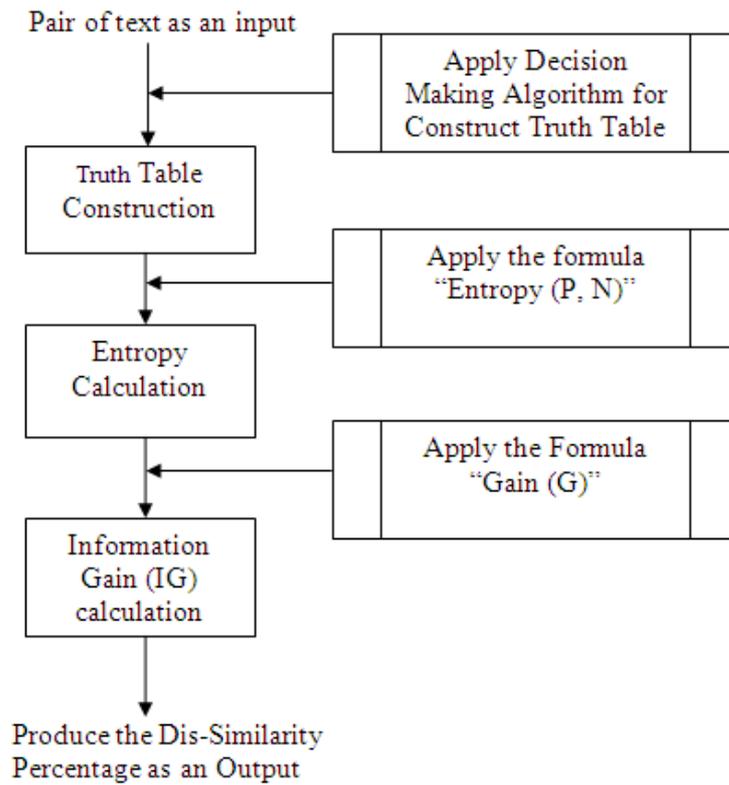


Figure 1. Characteristics of Entropy Based Dissimilarity Calculation