ARTICLE IN PRESS

Information Systems ■ (■■■■) ■■■–■■■

Contents lists available at ScienceDirect

Information Systems

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

A polygraph test for trustworthy structural similarity ☆

Kevin A. Naudé *, Jean H. Greyling, Dieter Vogts

Computing Sciences, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

ARTICLE INFO

Article history:
Received 2 December 2015
Received in revised form
25 May 2016
Accepted 11 July 2016
Recommended by: G. Vossen

Keywords:
Similarity measures
Distance measures
Similarity judgment errors
Similarity judgment quality
Information retrieval

ABSTRACT

Do similarity or distance measures ever go wrong? The inherent subjectivity in similarity discernment has long supported the view that all judgements of similarity are equally valid, and that any selected similarity measure may only be considered more effective in some chosen domain. This article presents evidence that such a view is incorrect for the specific case of relative structural similarity. In this context, similarity and distance measures occasionally do go wrong, producing judgements that can be considered as errors in judgement. This claim is supported by a novel method for assessing the quality of structural similarity and distance functions, which is based on relative scale of similarity with respect to chosen reference objects. The method may be applied either with synthetic graph datasets or with graphs representing objects in an application domain of interest. This work demonstrates the method over synthetic datasets with common measures of structural similarity in graphs. Finally, the article identifies three distinct kinds of relative similarity judgement errors, and shows how the distribution of these errors is related to graph properties under common similarity measures.

1. Introduction

Numeric measures of similarity are versatile tools for solving information retrieval problems. They serve both in classification and similarity search, and have been used effectively in a variety of problem domains [1–6]. Similarity is typically quantified as either a notional proportion of matching (similarity functions) or as a cumulative sum of differences (distance functions).

The performance of similarity measures may be evaluated in two orthogonal dimensions: *resource* performance and *task* performance. The former of these is easy to study, as computational resource usage may be either directly observed through empirical research or studied through theoretical models of computation. This research concerns itself with the second dimension of performance. Specifically, it examines the general efficacy with which different similarity measures are able to judge similarity between structured object representations. The most important of such discrete structures are graphs.

There are well established methods of evaluating the task performance of classification and similarity search algorithms that incorporate measures of similarity. Section 2 highlights the main processes of these conventional evaluation techniques, as they form a basis for the present work. An important problem with such methods is that the conclusions they provide are confined to a specific problem domain tested, leaving general conclusions about the embedded similarity measures hard to obtain.

There is a significant need to understand the judgement quality of similarity measures directly. The present research introduces a new evaluation technique which directly characterises the decision behaviour of similarity measures. The motivation for direct evaluation is advanced throughout Section 3, with the main processes of such an

☆ This research was supported by the National Research Foundation, South Africa.

* Corresponding author.
E-mail addresses: Kevin.Naude@nmmu.ac.za (K.A. Naudé), Jean.Greyling@nmmu.ac.za (J.H. Greyling), Dieter.Vogts@nmmu.ac.za (D. Vogts).

http://dx.doi.org/10.1016/j.is.2016.07.005
0306-4379/

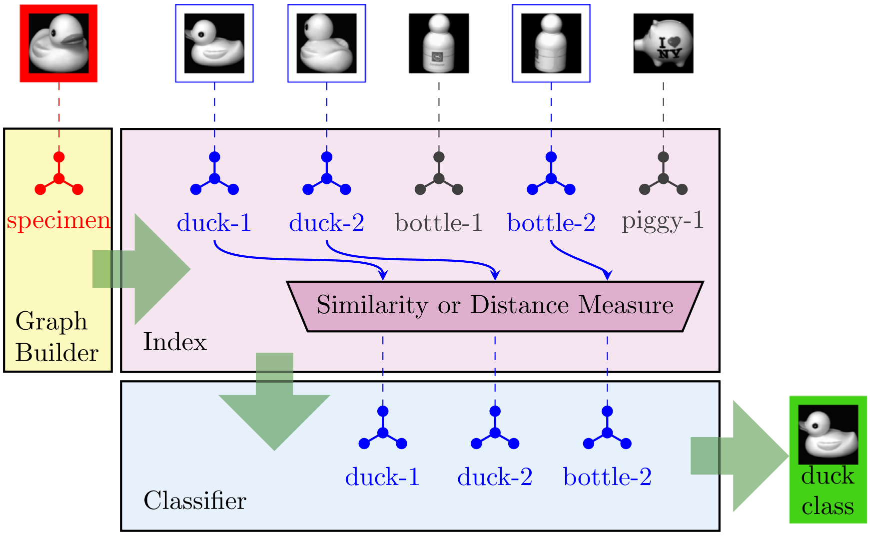

Fig. 1. The software components for object recognition, using a similarity measure.

evaluation set out in Sections 3.2 and 3.3. The new method focuses upon relative judgements, such as claiming that *objects A and B are more alike than objects A and C*. Consequently, there are no specific requirements about the scale of difference between similarity scores.

The main difficulty in directly assessing the judgement quality of similarity measures is the lack of justifiable ground truth data. The new evaluation method turns the problem around – instead of finding true outcomes of given test graphs, new graphs are generated from a reference graph in such a way that they can be ranked. In a sense, the new graphs are ordered on increasing entropy. The ground truths that are obtained are related to graph edit distances. However, they are procedurally created to satisfy constraints that provide a greater justification for the relative decision outcomes that occur in similarity ranking. The new method of evaluation is particularly valuable in that it produces new kinds of information about the decision behaviour of the tested similarity measures. In particular, it provides frequency data for three different kinds of inconsistent judgements that can be made.

The application of the evaluation method is demonstrated in Section 4 which compares a selection of graph similarity measures. The first evaluated measure is the similarity derived from the maximum common induced subgraph (MCIS), a popular alternative to the maximum common subgraph (MCS), which is often not feasible to compute. The section continues to compare the behaviour of the MCIS measure to members of a family of contemporary fixed-point graph similarity measures that discover pair-wise vertex similarity scores. The first and most simple of these is attributed to Blondel et al. [7]. The second is a modification of the Blondel measure that was studied by Zager and Verghese [8]. The results of the comparative study are offered in Section 5, and these provide general performance characteristics for each described similarity measure. In particular, the experiments reveal conditions under which the Blondel measure is superior to alternatives.

It should be noted that this article is an extended version of work previously published [9]. The examination of the Zager and Verghese similarity measure alongside the Blondel measure, from which it is derived, is an additional contribution. This article also serves to clarify the motivation for the constraints placed upon test instances, as well as the appropriate techniques for statistical analysis.

2. Conventional task performance

The conventional approach for evaluating systems of indexing, search and classification originates with the *Aslib Cranfield* projects of the 1960s [10]. The assessment of task performance of modern similarity search and classification follows very much the same method, although present day researchers benefit from the wide variety of collected datasets. One such dataset, the Columbia Object Image Library [11], shall now be used to illustrate the method and limitations of conventional task performance evaluation.

Consider the task of identifying an ordinary object from a source photograph, based on a library of previously observed photographs. The internal organisation of a software solution for the task is abstractly represented in Fig. 1. The figure describes solutions that use an explicitly computed measure of similarity.[1] Note that there are several computational sub-tasks besides computing similarity scores. The main sub-tasks are

- building a graph representation of the object,
- querying an index for related known objects, using the similarity measure strategically to minimise comparisons, and
- using a selection of the query response to classify the specimen object.

[1] Other machine learning techniques could also be applied to the task, but these are not relevant to the subject of this research.

Please cite this article as: K.A. Naudé, et al., A polygraph test for trustworthy structural similarity, Information Systems (2016), http://dx.doi.org/10.1016/j.is.2016.07.005