BMatch: a Semantically Context-based Tool Enhanced by an Indexing Structure to Accelerate Schema Matching*

Fabien Duchateau[1] and Zohra Bellahsene[1] and Mathieu Roche[1]

LIRMM - Université Montpellier 2
161 rue Ada 34000 Montpellier
{firstname.name}@lirmm.fr

Abstract. Schema matching is a crucial task to gather information of the same domain. This is more true on the web, where a large number of data sources are available and require to be matched. However, the schema matching process is still largely performed manually or semi-automatically, discouraging the deployment of large-scale integration systems. Indeed, these large-scale scenarios need a solution whi-ch ensures both an acceptable matching quality and good performance. In this article, we present an approach to match efficiently a large number of schemas. The quality aspect is based on the combination of terminological methods and cosine measure between context vectors. The performance aspect relies on a B-tree indexing structure to reduce the search space. Finally, our approach, BMatch, has been implemented and the experiments with real sets of schemas show that it is both scalable and provides an acceptable mat-ching quality when compared with the results obtained by the most referenced matching tools.

Keywords: semantic similarity, schema matching, BMatch, B-tree index structure, node context, terminological and structural measures

A Appendix 1: ROC Curves

This appendix contains all the ROC curves which have been tested to tune our system. They are sorted by parameter, then by scenario. To evaluate our matching tool, we have chosen five real-world scenarios, each composed of two schemas. These are widely used in the literature. The first one describes a **person**, the second is related to **university courses** from Thalia benchmark, the third one on **business order** extracted from OAGIS[1] and XCBL[2]. Finally, **currency** and **sms** are popular web services[3]. Their main features are given in table 1.

* Supported by ANR Research Grant ANR-05-MMSA-0007

[1] http://www.oagi.org

[2] http://www.xcbl.org

[3] http://www.seekda.com

	Person	University	Order	Currency	SMS
Number of nodes (S_1/S_2)	11/10	8/9	20/844	12/35	46/64
Avg number of nodes	11	9	432	24	55
Max depth (S_1/S_2)	4/4	4/4	3/3	3/3	4/4
Number of mappings	5	9	10	6	20

Table 1. Features of the different scenarios.

A.1 Replacement Threshold

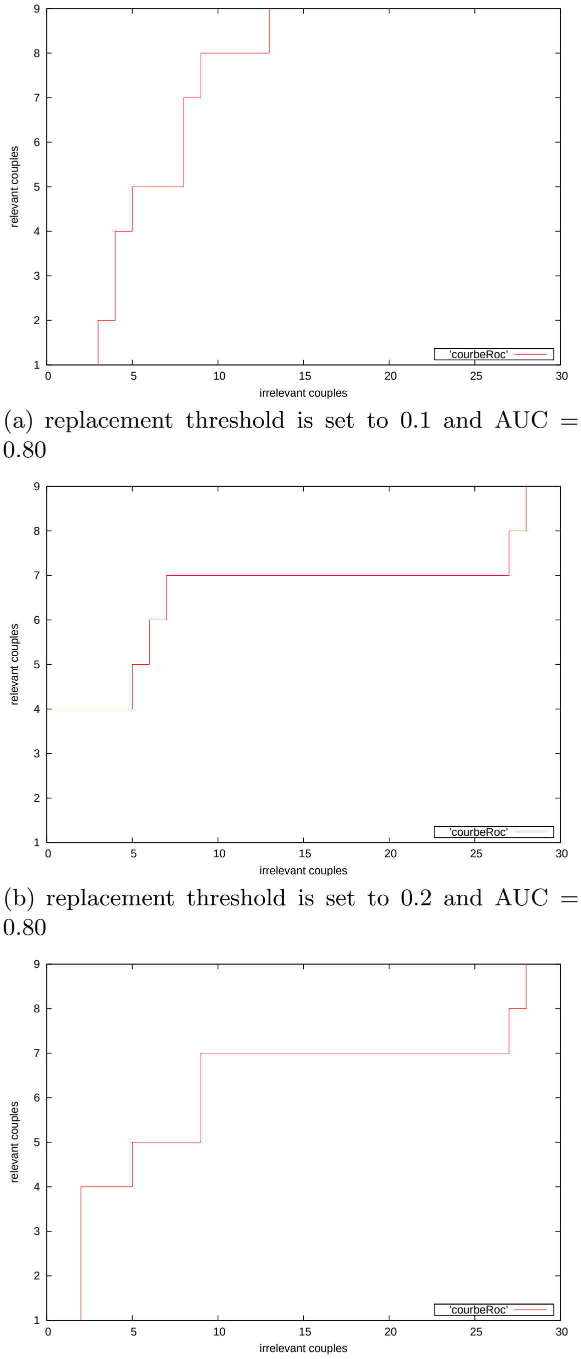
Figures 1, 2, 3, 4 and 5 depicts the ROC curves when replacement threshold is tuned. The default values for the other parameters are: number of levels is set to 2, strategy to $iso - max$.

A.2 Number of Levels

Here are the ROC curves (figures 6, 7, 8, 9 and 10) when the number of levels is tuned. The replacement threshold is set to 0.2 and the adopted strategies is $iso - max$.

A.3 Strategies

The figures 11, 12, 13, 14 and 15 depicts the ROC curves when the strategies are tuned. The replacement threshold is set to 0.2 and the number of levels is set to 2.

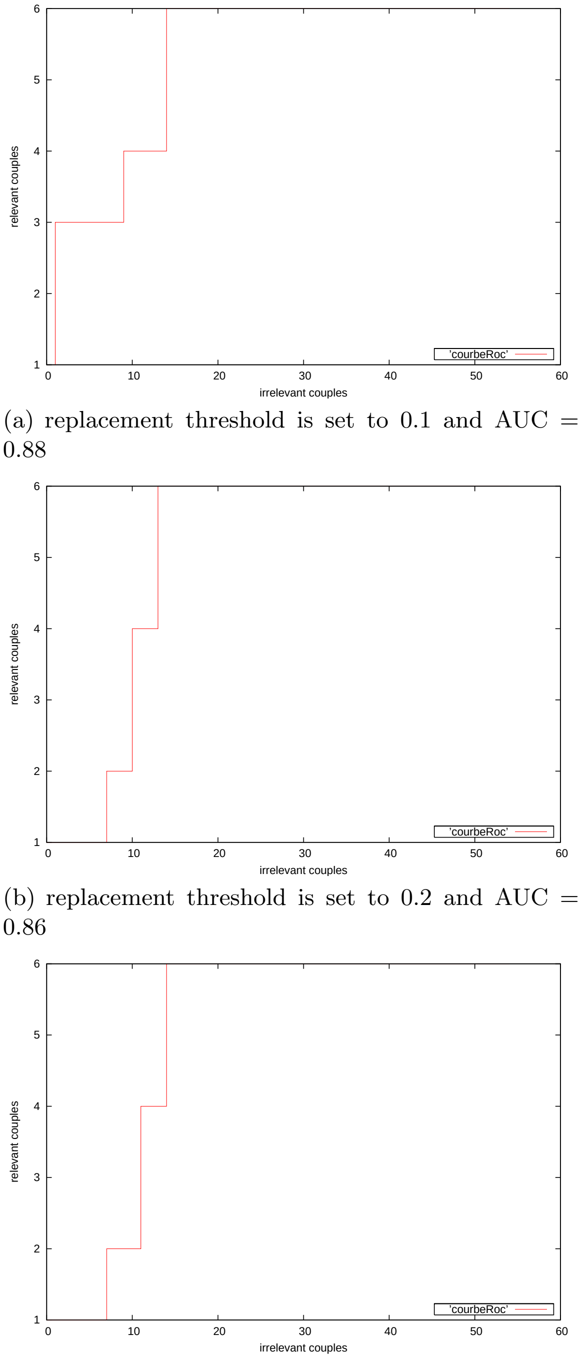


(a) replacement threshold is set to 0.1 and AUC = 0.80

(b) replacement threshold is set to 0.2 and AUC = 0.80

(c) replacement threshold is set to 0.3 and AUC = 0.68

Fig. 1. ROC curves for university scenario when tuning replacement threshold

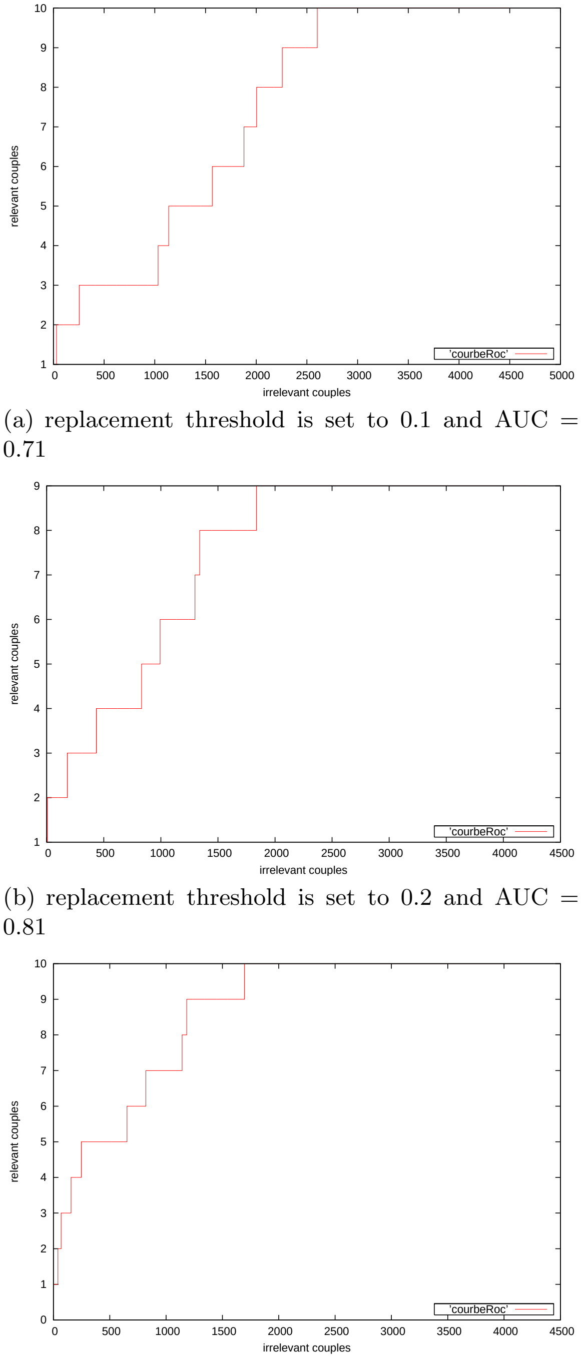


(a) replacement threshold is set to 0.1 and AUC = 0.88

(b) replacement threshold is set to 0.2 and AUC = 0.86

(c) replacement threshold is set to 0.3 and AUC = 0.85

Fig. 2. ROC curves for person scenario when tuning replacement threshold

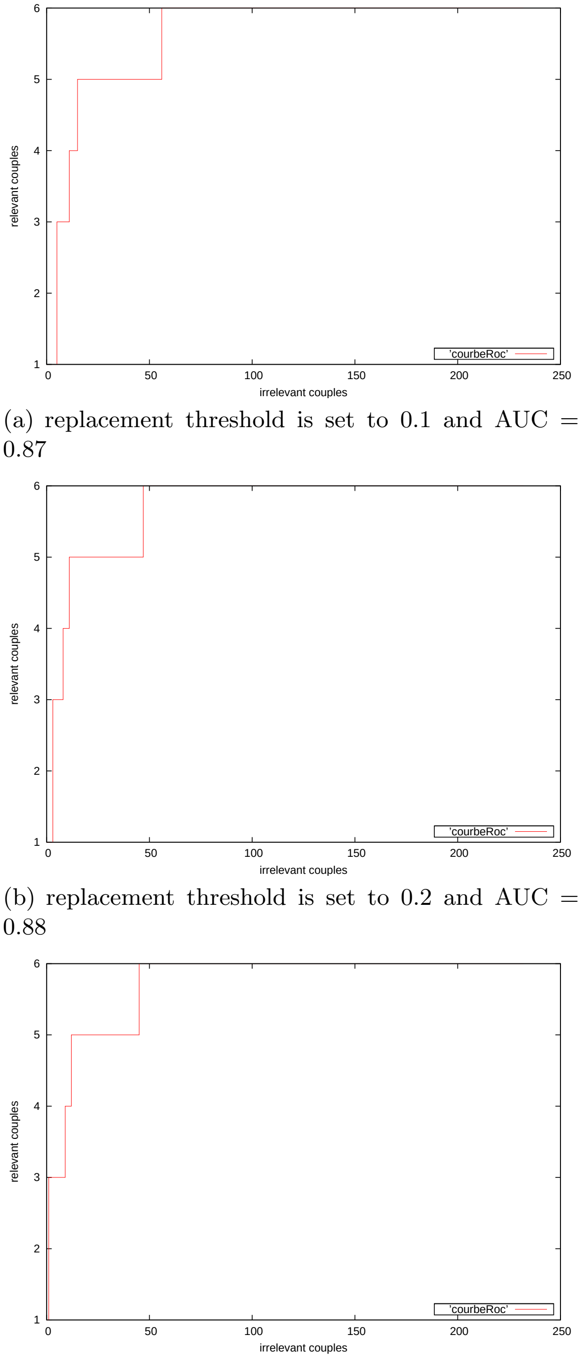


(a) replacement threshold is set to 0.1 and AUC = 0.71

(b) replacement threshold is set to 0.2 and AUC = 0.81

(c) replacement threshold is set to 0.3 and AUC = 0.85

Fig. 3. ROC curves for order scenario when tuning replacement threshold

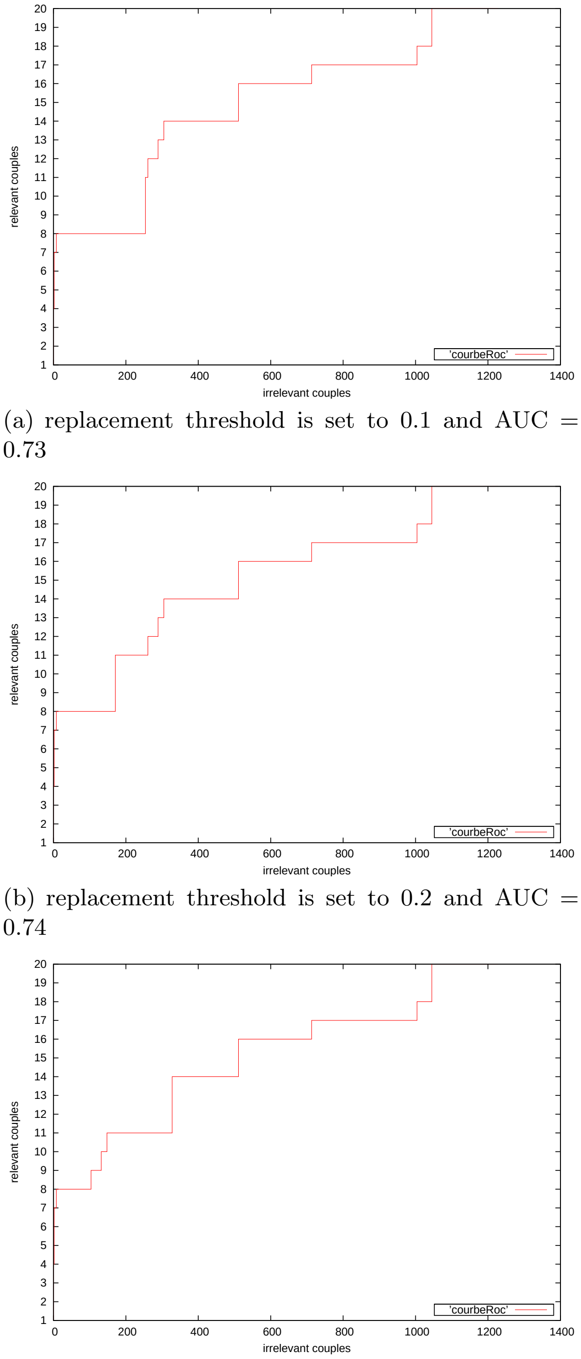


(a) replacement threshold is set to 0.1 and AUC = 0.87

(b) replacement threshold is set to 0.2 and AUC = 0.88

(c) replacement threshold is set to 0.3 and AUC = 0.89

Fig. 4. ROC curves for currency scenario when tuning replacement threshold

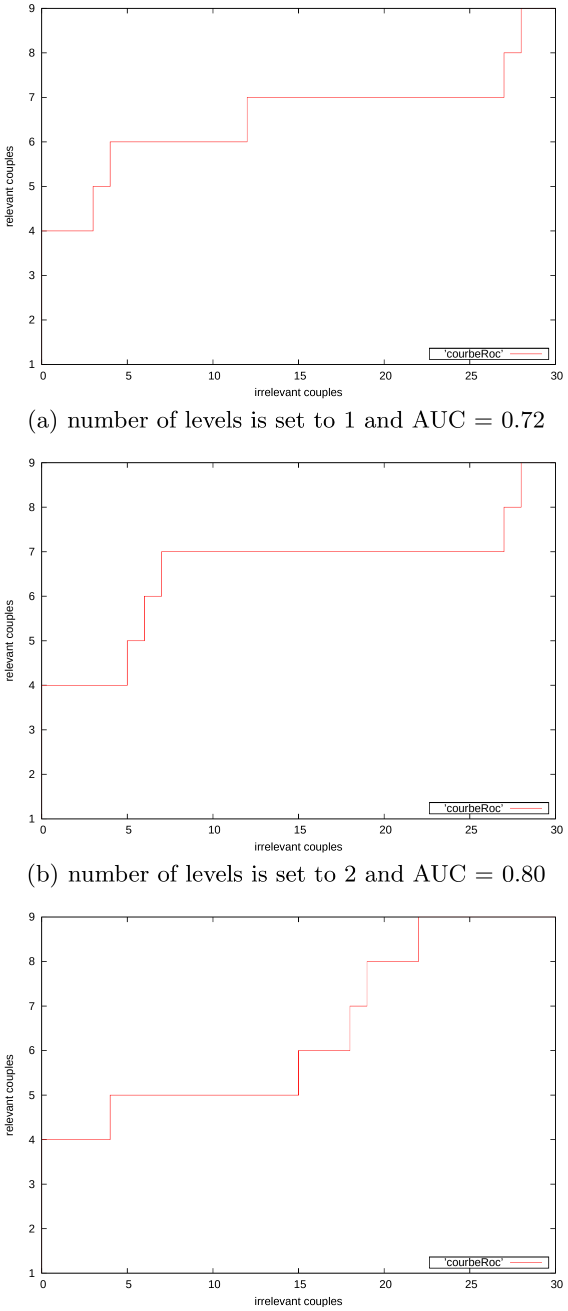


(a) replacement threshold is set to 0.1 and AUC = 0.73

(b) replacement threshold is set to 0.2 and AUC = 0.74

(c) replacement threshold is set to 0.3 and AUC = 0.74

Fig. 5. ROC curves for sms scenario when tuning replacement threshold

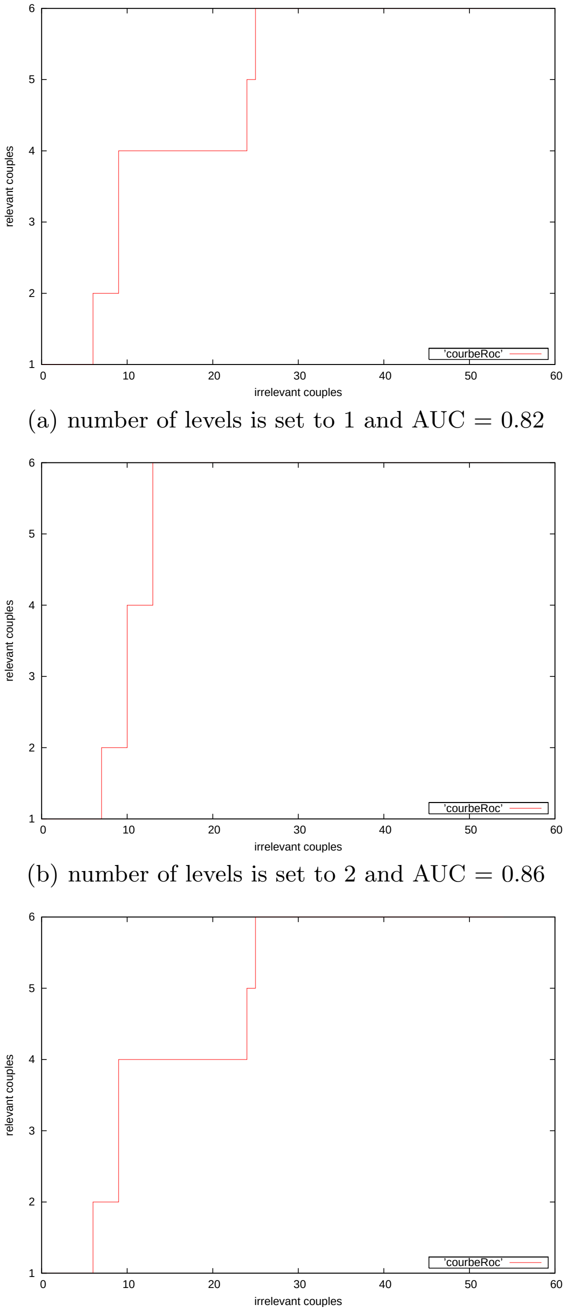


(a) number of levels is set to 1 and AUC = 0.72

(b) number of levels is set to 2 and AUC = 0.80

(c) number of levels is set to 3 and AUC = 0.71

Fig. 6. ROC curves for university scenario when tuning the number of levels

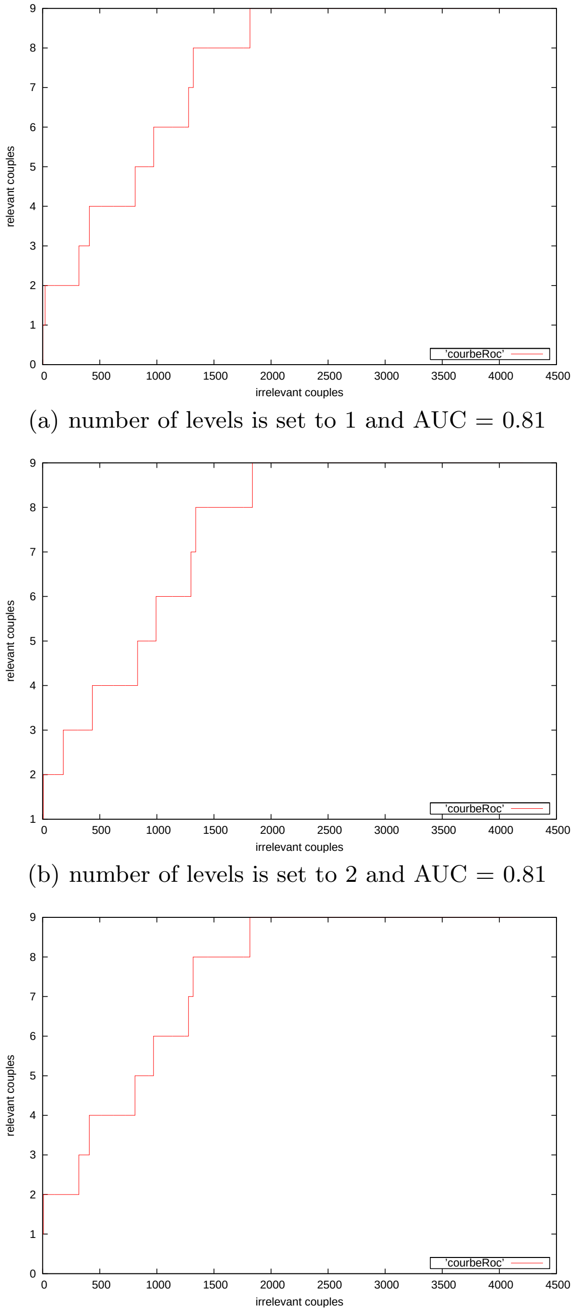


(a) number of levels is set to 1 and AUC = 0.82

(b) number of levels is set to 2 and AUC = 0.86

(c) number of levels is set to 3 and AUC = 0.82

Fig. 7. ROC curves for person scenario when tuning the number of levels

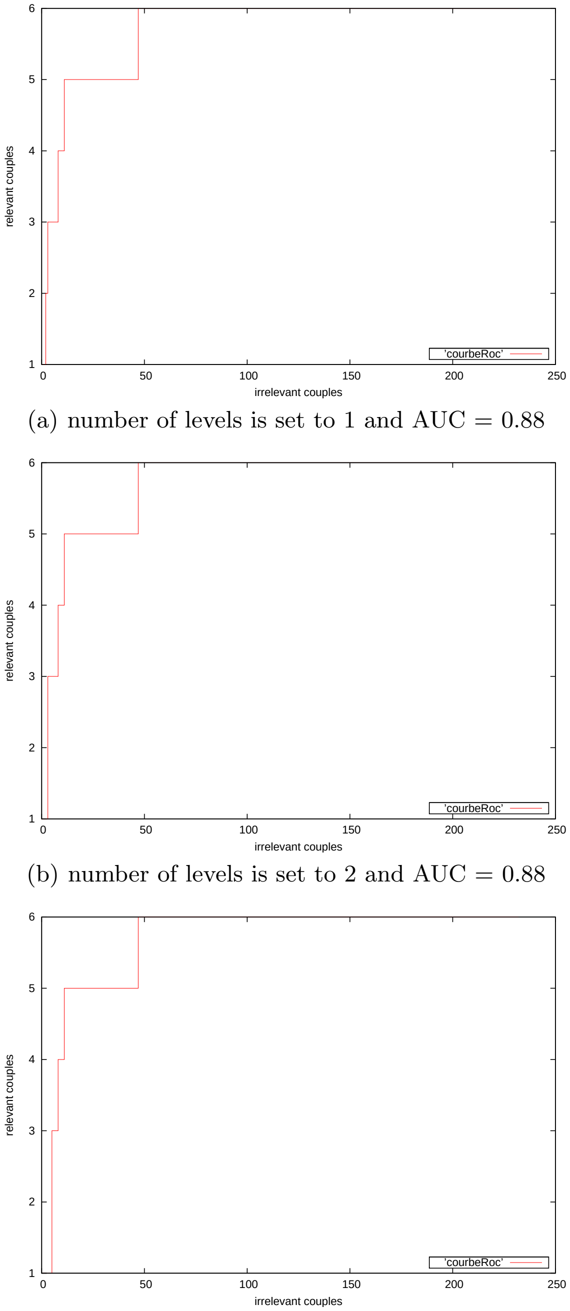


(a) number of levels is set to 1 and AUC = 0.81

(b) number of levels is set to 2 and AUC = 0.81

(c) number of levels is set to 3 and AUC = 0.81

Fig. 8. ROC curves for order scenario when tuning the number of levels

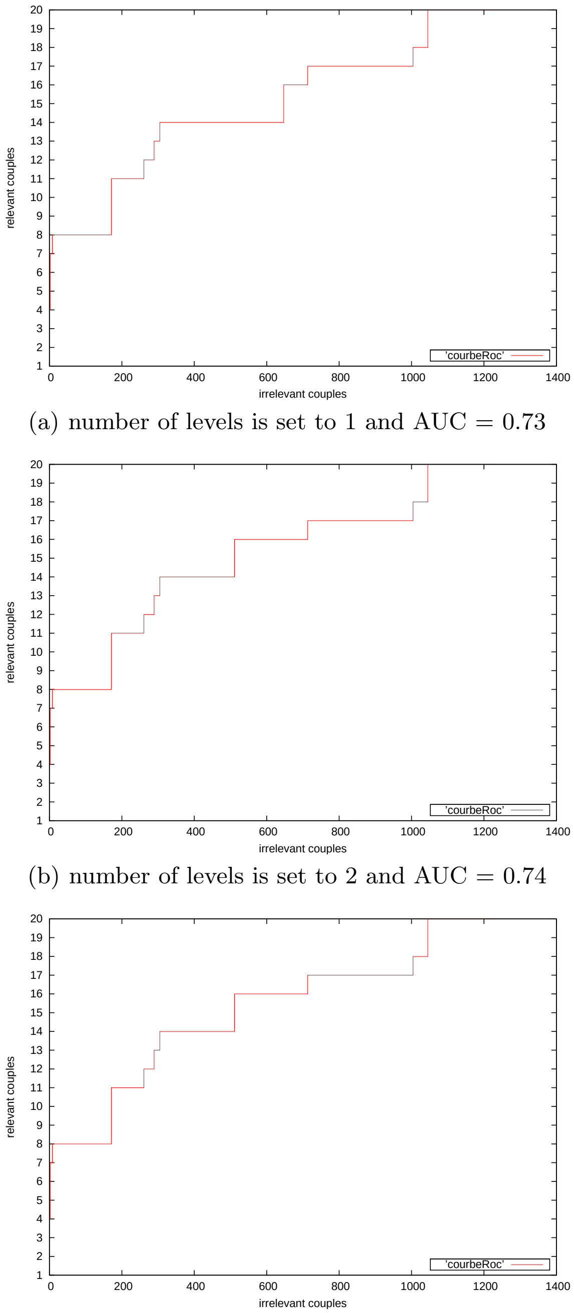


(a) number of levels is set to 1 and AUC = 0.88

(b) number of levels is set to 2 and AUC = 0.88

(c) number of levels is set to 3 and AUC = 0.88

Fig. 9. ROC curves for currency scenario when tuning the number of levels

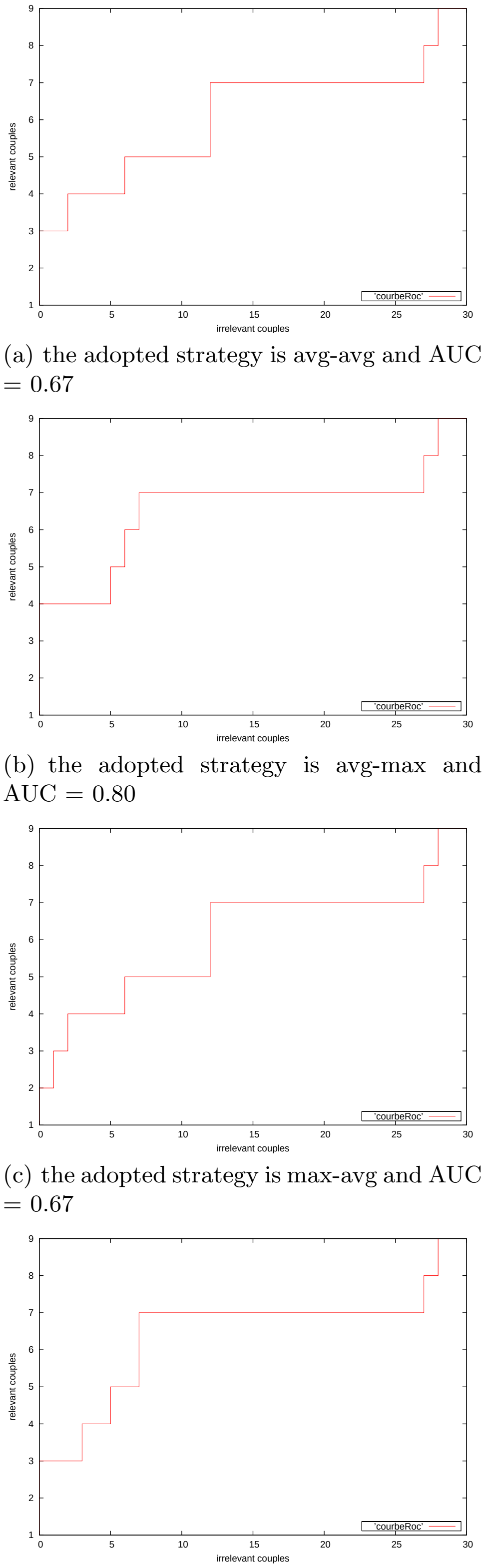


(a) number of levels is set to 1 and AUC = 0.73

(b) number of levels is set to 2 and AUC = 0.74

(c) number of levels is set to 3 and AUC = 0.74

Fig. 10. ROC curves for sms scenario when tuning the number of levels



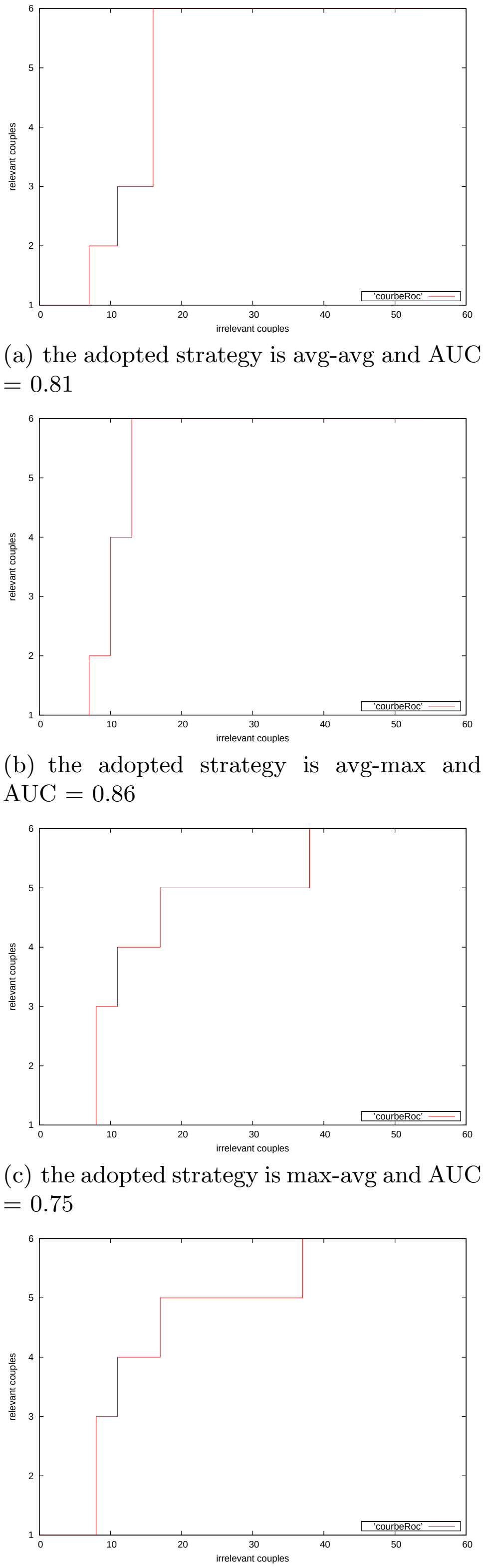
(a) the adopted strategy is avg-avg and AUC = 0.67

(b) the adopted strategy is avg-max and AUC = 0.80

(c) the adopted strategy is max-avg and AUC = 0.67

(d) the adopted strategy is max-max and AUC = 0.71

Fig. 11. ROC curves for university scenario when tuning the strategies



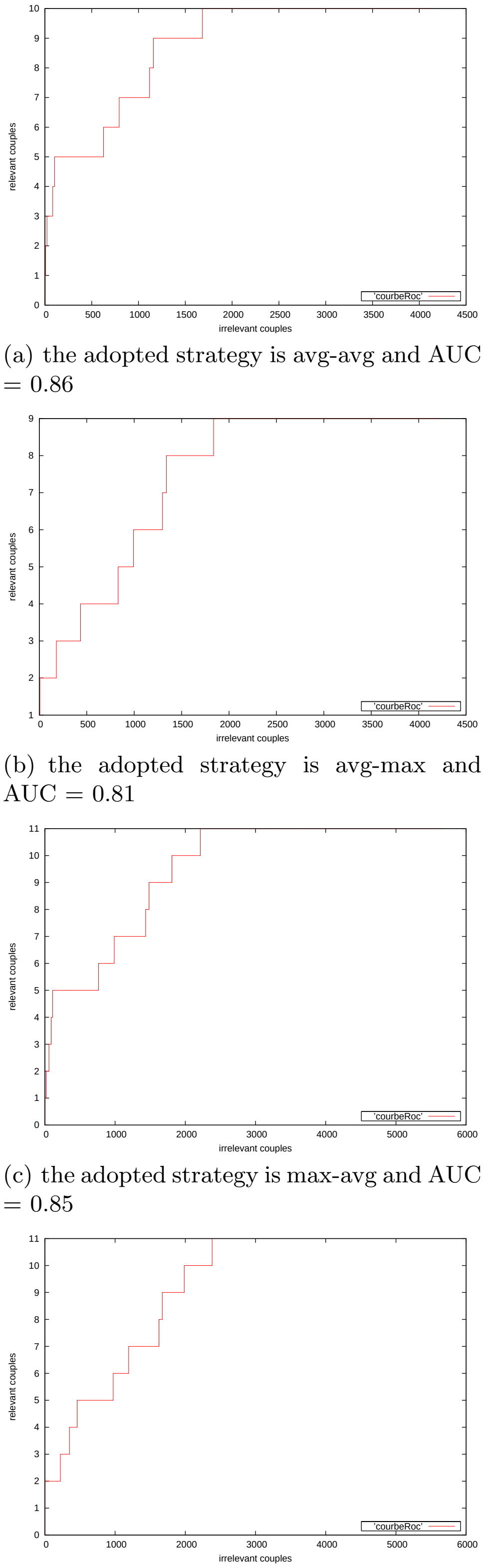
(a) the adopted strategy is avg-avg and AUC = 0.81

(b) the adopted strategy is avg-max and AUC = 0.86

(c) the adopted strategy is max-avg and AUC = 0.75

(d) the adopted strategy is max-max and AUC = 0.75

Fig. 12. ROC curves for person scenario when tuning the strategies



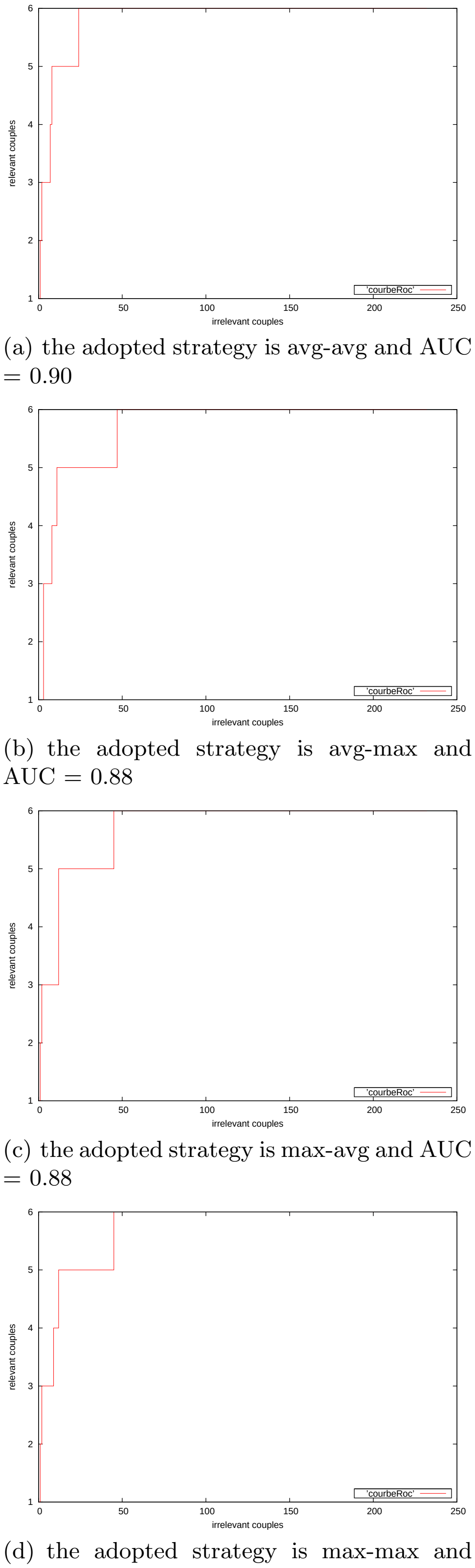
(a) the adopted strategy is avg-avg and AUC = 0.86

(b) the adopted strategy is avg-max and AUC = 0.81

(c) the adopted strategy is max-avg and AUC = 0.85

(d) the adopted strategy is max-max and AUC = 0.82

Fig. 13. ROC curves for order scenario when tuning the strategies



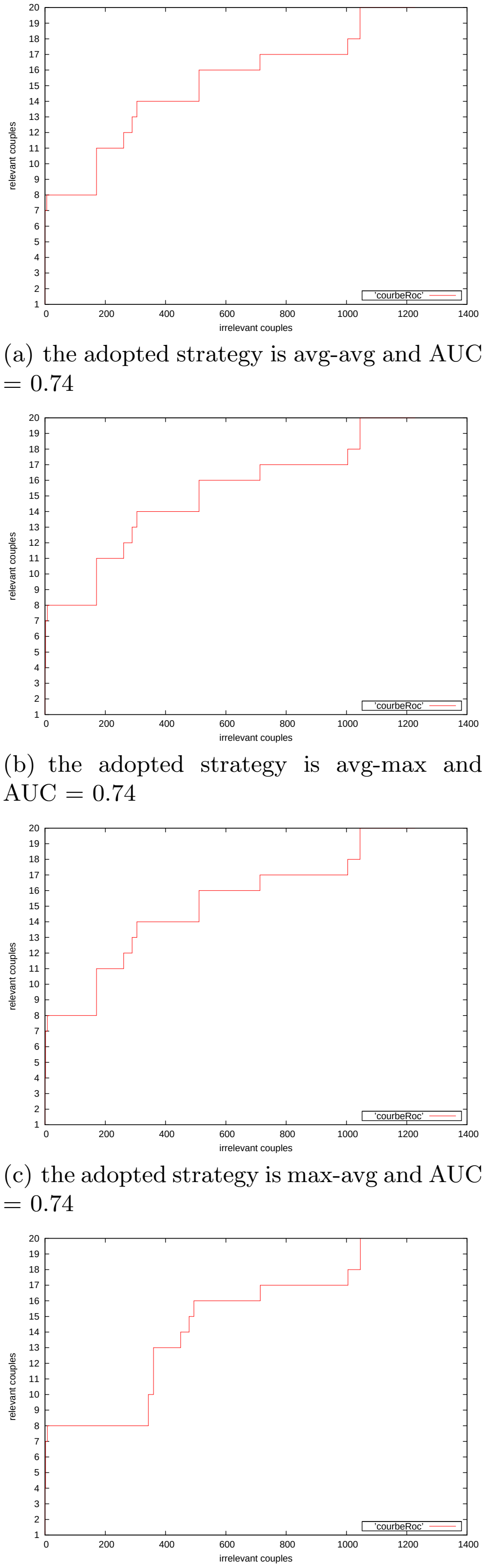
(a) the adopted strategy is avg-avg and AUC = 0.90

(b) the adopted strategy is avg-max and AUC = 0.88

(c) the adopted strategy is max-avg and AUC = 0.88

(d) the adopted strategy is max-max and AUC = 0.89

Fig. 14. ROC curves for currency scenario when tuning the strategies



(a) the adopted strategy is avg-avg and AUC = 0.74

(b) the adopted strategy is avg-max and AUC = 0.74

(c) the adopted strategy is max-avg and AUC = 0.74

(d) the adopted strategy is max-max and AUC = 0.71

Fig. 15. ROC curves for sms scenario when tuning the strategies