Enhancement of similarity measure during comparing feature relation graphs (FRG)

Polina S. Diurdeva, student of Department of Analytical Information Systems of SPbSU, polina.durdeva@yandex.ru
Vladislav A. Pavlov, student of Department of Software Engineering of SPbSU, vlad.pavlov24@gmail.com
Dmitry S. Shalymov, Candidate of Physics and Mathematics of SPbSU, dmitry.shalymov@gmail.com

Abstract

Lately a new metric based on Feature Relation Graph (FRG) for Persian handwritteen documents.In another work we conducted several experiments to test this metric for Arabic handwritten documents. The results were promising but we found out that the measure has a huge drawback as it does not account the dencity of graphs being compared. We designed a new metric and launched the algorithm on Russian and Arabic handwritten texts. We provide numerical experiments to demonstrate effectiveness of proposed metric and compare it with previously used one.

Introduction

The task of handwritten documents classification has become very important. Nowadays we see a lot of documents with doubtful and unknown authorship, maintaining a problem of author identification actual. That is why investigation of a new algorithms for processing and classification of such manuscripts become substancial. Most of modern systems for writer identification can be divided into online and offline systems. Online systems use information obtained from the mere process of writing while offline systems use the information fetched only from the text. Also such systems can be divided into text-dependent and text-independent systems. The first ones are good for a fixed set of written texts, while text-independent systems are insensitive to the texts being processed. We began our investigation with exploring existing offline text-independent systems for authorship identification of an Arabic handwritten text. Our attention was paid to an offline system for Persian writer identification[1]. The original system showed 100 percentage precision when enough training data was supplied. We explored and implemented the algorithm proposed in the work to solve problems of classification and clusterization of Arabic and Russian handwritten documents. Our investigation will be detailed further. During our experiments we noticed that some tiny changes can be made to the original algorithm that may increase

the precision of the classification process. We conducted several numerical experiments and figured out that our expectations were not wrong.

FRG classification algorithm analysis

The algorithm we worked with is precisely described in papers [1] and [2]. We do not give detailed description of the entire algorithm, considering it can be found in the aforementioned papers. However, we will focus on the FRG comparison phase of the algorithm, as we modified this part of the algorithm. In fact, this stage represents classification of graphs. During graph classification a measure of similarity was used to compare two graphs. This measure was modified by us. Below we will recall the main ideas of classification stage and introduce our changes.

Previous FRG classification

Let us denote a FRG of a test set by U and a set of created after the training stage graphs by Γ. Previously, similarity measure S were calculated for each graph G_i from Γ and graph U. In general, the $S(G_1, G_2)$ was calculated as the total number of common paths in G_1 and G_2. Special algorithm for calculating $S(G_1, G_2)$ was proposed in [1]. It is based on calculating $height$ values for vertices of graphs, sorting common edges by height of beginning, processing them in appropriate order to get T values for each common node(showing a number of common paths from a node) and summing all calculated T values getting the final score for S. Graph B is chosen so that $S(U, B) = \max_{G_i \in \Gamma} S(U, G_i)$. According to [1] the complexity of the briefly described above algorithm is $O(ElogE + V)$.

Problems

The aforementioned algorithm works fine except for several cases. The main problem is the similarity measure S, that does not take into account the density of the graphs compared. Let us consider the following example. Let graph U be set by $V = \{1, 2, 3, 4\}; E = \{(1, 2)\}$, graph X would be set by $V = \{1, 2, 3, 4\}; E = \{(1, 2), (1, 3), (1, 4)\}$ and graph V would be identical to U. The above graphs are presented in the Fig. 1

It is obvious that the set of common paths these graphs have consist only from one path: from first to second vertex, hence the S measure on each pair of graphs would be equal to 1. But it is apparent that it is not fair as soon as V is a complete copy of U and their similarity measure should be more than

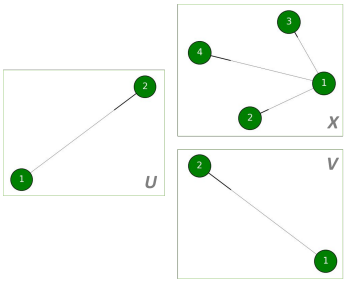

Fig. 1: Example of three graphs on which S measure is unjustifiably equal

similarity between U and X, which are less similar. So the first problem is that previous classification process used S measure as similarity measure, which could call a graph which was much denser as similar as identical graph. There was also the second problem connected with the lack of boundaries for S measure. As it can be seen from the definition of S measure it is bounded below by zero but is not bounded from above. It causes a problem of absence of normalization. Normalization of similarity measure would be useful including some relativeness sense in the similarity measure. Finally, we realized that these two problems are actually connected and can be solved by introducing a new similarity measure.

New FRG classification

We designed a new similarity measure that helped us overcome the above problems. The new similarity measure was constructed with use of S measure. We called the new measure $proximity$. It can be defined by Eq.1

$$proximity(G_1, G_2) = \frac{2S(G_1, G_2)}{S(G_1, G_1) + S(G_2, G_2)} \tag{1}$$

The proximity measure lays in segment $[0, 1]$, being equal to 0 when graphs do not have any common paths and to 1 when two graphs are identical. Having that fact we can say that now our similarity measure is normalized. For example presented in Fig.1 computation of $proximity$ gives more explainable results. Now it is understandable why $proximity(U, V) = 1$ and $proximity(U, X) = 0.5$. This result shows that $proximity$ measure is sensible to densities of graphs being compared. It should be noticed that the complexity of the $proximity$ calculating algorithm is connected with complexity of calculating S by constant multiplication. Having solved the problems caused by disadvantages of S measure described in previous subsection we decided to compare S with $proximity$ in practice.

Numerical experiments

Classification of handwritten documents

To compare $proximity$ and S measure we conducted several experiments. We used RuHT database that contained texts of 30 writers each of which had approximately 8 lines of text. Also we tested the enhanced algorithm on Arabic handwritten documents. We used KHATT[3] database as a source for such images. The database contains 40000 texts of different of over thousand persons. During our experiments we varied amount of authors, metrics, parameters of Gabor and XGabor filters, amount of features extracted, train data amount/test data amount ratio. We decided to group all our launches by amount of authors a whose manuscripts were used during classification. After that the whole series of experiments was divided in 4 groups: experiments with manuscripts of 5, 10, 15 and 28 authors. For each group we calculated an average improvement avg_impr in accuracy made by $proximity$ measure, the best improvement detected max_impr during the experiments of that group, best accuracy $prox\_acc$ gained using $proximity$ measure and the best accuracy S_acc gained using S measure. Table 1 contains those values for different amount of authors of Russian handwritten documents.

Tab. 1: Comparison of S and $proximity$ measures for Russian handwritten documents

a	avg_impr, %	max_impr, %	$prox\_acc$, %	S_acc, %
5	190	250	100	100
10	180	200	100	86
15	134	250	85	78
28	135	300	79	76

Table 2 contains avg_impr, max_impr, $prox\_acc$, S_acc for different amount of authors of Arabic handwritten documents.

Tab. 2: Comparison of S and $proximity$ measures for Arabic handwritten documents

a	avg_impr, %	max_impr, %	$prox\_acc$, %	S_acc, %
5	176	300	100	100
10	213	300	90	80
15	187	250	85	75
28	141	300	79	65

It can be seen that the $proximity$ measure significantly improves the accuracy of classification process in average. Analysing the column of max_impr it can be said that this measure considerably increases the result in cases when S measure performs poor. Looking at the last two columns, it can be noticed that $proximity$ measure enhances the performance of an algorithm especially a huge enhancement in performance is noticed when manuscripts of 10 authors are taken. We consider this to be an important improvement taking into account that amount of training and test data is small enough(only 8 lines that is 5 times less than in [1]).

Clustering of Russian handwritten documents

We tried to investigate algorithm described above with $proximity$ measure for clustering Russian handwritten documents. Elements that are clustered are images. Images can be mapped into FRG space by the algorithm in section . Hence, the problem of handwritten texts clustering by authorship can be reduced to problem of clustering corresponding FRGs. In many clustering algorithms a term of centroid is used. It usually denotes a center of mass of a cluster. In our case centroid will be an artificial FRG that is built upon all text lines on which all FRGs of the cluster were built. Due to the fact that a FRG represents some statistics, new centroid (built in the described way) will only enhance its representation as a new FRG is got using more data.

We tested several clustering algorithms : k-means, global k-means, Online k-Means, PAM, DBSCAN and agglomerative hierarchical algorithm. For comparison of performance of those algorithms special metrics for evaluating clustering were used: Purity Measure, Rand Index, Normalized Mutual Information (NMI), F-Measure. Table 3 shows results of applying different clustering algorithms with previous measure from [1] and new $proximity$ measure to Russian handwritten documents.

We can see that using $proximity$ measure noticably improves the results of clustering when amount of authors is more than 9. At 15 authors an average enhancement was equal to 7 percent, for 10 authors it was equal to 4. However, for 5 clusters the majority of algorithms showed 100 percentage with both metrics. As it can be seen from the table the best results are got when using global k-means and PAM. The best results among algorithms that do not require k were got by using hierarchical algorithm. When amount of expected clusters grows the accuracy falls down noticeably. We explain it with small amount of training data just as in the case when classification problem was solved. However, the main result is the enhancement in accuracy when using $proximity$ measure instead of measure provided in [1].

Tab. 3: Clustering results $*10^2$ on Russian texts using previous/new measure

a	Algorithm	Purity	RandIndex	NMI	F-Measure
5	k-means	100 / 100	100 / 100	100 / 100	100 / 100
5	global k-means	100 / 100	100 / 100	100 / 100	100 / 100
5	online k-means	100 / 100	77 / 78	93 / 94	88 / 88
5	hierarchy	100 / 100	100 / 100	100 / 100	100 / 100
5	DBSCAN	100 / 100	87 / 89	96 / 98	95 / 95
5	PAM	100 / 100	100 / 100	100 / 100	100 / 100
10	k-means	90 / 92	74 / 75	85 / 86	95 / 95
10	global k-means	100 / 100	100 / 100	100 / 100	100 / 100
10	online k-means	84 / 85	55 / 56	91 / 92	83 / 83
10	hierarchy	80 / 80	62 / 62	95 / 95	90 / 90
10	DBSCAN	81 / 83	60 / 61	87 / 89	81 / 81
10	PAM	75 / 75	58 / 58	94 / 94	88 / 88
15	k-means	60 / 64	50 / 53	83 / 88	84 / 84
15	global k-means	86 / 89	59 / 62	85 / 87	90 / 90
15	online k-means	57 / 61	30 / 32	86 / 91	82 / 82
15	hierarchy	61 / 63	22 / 26	74 / 80	73 / 73
15	DBSCAN	71 / 73	35 / 39	83 / 87	78 / 78
15	PAM	65 / 68	42 / 40	90 / 93	76 / 76
28	k-means	50 / 58	17 / 25	68 / 74	76 / 76
28	global k-means	62 / 70	31 / 38	69 / 76	73 / 73
28	online k-means	43 / 55	12 / 20	65 / 71	74 / 74
28	hierarchy	52 / 60	18 / 26	73 / 80	73 / 73
28	DBSCAN	51 / 58	19 / 24	65 / 72	75 / 75
28	PAM	59 / 65	25 / 32	67 / 75	77 / 77

Conclusion

In this paper we have investigated a new similarity measure that could be used during FRG classification. The new similarity measure called $proximity$ eliminated disadvantages of the previous S measure that was used during FRG comparison. The $proximity$ measure was designed to take care of the density of compared graphs and to build bounded similarity measure with normalization intentions. The $proximity$ measure demonstrated huge improvements compared to S measure, leading to increased accuracy of classification process in worst, best and average cases during classification of Russian and Arabic handwritten documents with small amount of training and test data. The $proximity$ measure demonstrated 100 percent precision solving classification problem for manuscripts of ten authors. The $proximity$ measure showed average 5 percent improvement in accuracy during clustering Russian handwritten documents of 10 and 15 authors.

Acknowledgment

This research is supported by Saint-Petersburg State Unversity grant 6.37.181.2014.

References

[1] B.Helli, M.E. Moghaddam *A text-independent Persian writer identification based on feature relation graph (FRG)*, Pattern Recognition, 2010.

[2] Vladislav A. Pavlov, Dmitry S. Shalymov *Arabic handwritten texts clusterization based on Feature Relation Graph (FRG)*, ICDAR 2015: 941-945

[3] S. A. Mahmoud, I. Ahmad, M. Alshayeb, W. G. Al-Khatib, M. T. Parvez, G. A. Fink, V. Margner, and H. EL Abed *KHATT: Arabic Offline Handwritten Text Database*, In Proceedings of the 13th International Conference on Frontiers in Handwriting Recognition (ICFHR 2012), Bari, Italy, 2012, pp. 447-452, IEEE Computer Society.