Ordinal measures in authorship identification[*]

Liviu P. Dinu
University or Bucharest, Faculty of
Mathematics and Computer Science,
14 Academiei, Bucharest, Romania,
ldinu@funinf.cs.unibuc.ro

Marius Popescu
University or Bucharest, Faculty of
Mathematics and Computer Science,
14 Academiei, Bucharest, Romania,
mpopescu@phobos.cs.unibuc.ro

Abstract: The goal of this paper is to compare a set of distance/similarity measures, regarding theirs ability to reflect stylistic similarity between authors and texts. To assess the ability of these distance/similarity functions to capture stylistic similarity between texts, we tested them in one of the most frequently employed multivariate statistical analysis settings: cluster analysis. The experiments are done on a corpus of 30 English books written by British, American and Australian writers.

Keywords: authorship identification, ordinal measures

1 Introduction

The authorship identification problem is an ancient and omnipresent challenge, and almost in every culture there are a lot of disputed works (Shakespeare's plays, Moliere vs. Corneille (Labbe and Labbe, 2001), Federalist Papers (Mosteller and Wallace, 2007), etc.). The problem of authorship identification is based on the assumption that there are stylistic features that help distinguish the real author from any other possibility. Literary-linguistic research is limited by the human capacity to analyze and combine a small number of text parameters, to help solve the authorship problem. We can surpass limitation problems using computational methods, which allow us to explore various text parameters and characteristics and their combinations. Using these methods (van Halteren et al., 2005) have shown that every writer has a unique fingerprint regarding language use. The set of language use characteristics - stylistic, lexical, syntactic - form the human stylom.

Because in all computational stylistic studies/approaches, a process of comparison of two or more texts is involved, in a way or another, there was always a need for a distance/similarity function to measure similarity (or dissimilarity) of texts from the stylistic point of view. These measures vary a lot, and in the last years a series of different techniques were used in authorship identification: approaches based on string kernel (Dinu, et al., 2008), SVM based on function words frequencies (Koppel et. al., 2007), standard distances or ordinal distances (Popescu and Dinu, 2008).

The goal of this paper is to compare a set of distance/similarity measures, regarding theirs ability to reflect stylistic similarity between texts.

As style markers we have used the function words frequencies. Function words are generally considered good indicators of style because their use is very unlikely to be under the conscious control of the author and because of their psychological and cognitive role (Chung and Pennebaker, 2007). Also function words prove to be very effective in many author attribution studies.

The distance/similarity between two texts will be measured as distance/similarity between the function words frequencies corresponding to the respective texts. For this study we selected some similarity/distance measures. We started with the most natural distance/similarity measures: euclidean distance and (taking into account the statistical nature of data) Pearson's correlation coefficient. Since function words frequencies can also be viewed as ordinal variables, we also considered for comparison some specific similarity measures: Spearman's rank-order coefficient, Spearman's footrule, Goodman and Kruskal's gamma, Kendall's tau.

To assess the ability of these distance/similarity functions to capture stylistic similarity between texts, we have tested them in one of the most frequently employed multivariate statistical analysis settings: cluster

[*] Research supported by CNCSIS, PN2-Idei project 228

analysis. Clustering is a very good test bed for a distance/similarity measure behavior. We plugged the distance/similarity measures selected for comparison into a standard hierarchical clustering algorithm and applied it to a collection of 30 nineteenth century English books. The family trees thus obtained revealed a lot about the distance/similarity measures behavior.

The main finding of our comparison is that the similarity measures that treat function words frequencies as ordinal variables performed better than the others distance/similarity measures. Treating function words frequencies as ordinal variables means that in the calculation of distance/similarity function the ranks of function words according to their frequencies in text will be used rather than the actual values of these frequencies. Usage of the ranking of function words in the calculation of the distance/similarity measure instead of the actual values of the frequencies may seem as a loss of information, but we consider that the process of ranking makes the distance/similarity measure more robust acting as a filter, eliminating the *noise* contained in the values of the frequencies. The fact that a specific function word has the rank 2 (is the second most frequent word) in one text and has the rank 4 (is the fourth most frequent word) in another text can be more relevant than the fact that the respective word appears 34% times in the first text and only 29% times in the second.

In the next section we present the distance/similarity measures involved in the comparison study, section 3 briefly describes the cluster analysis, and in section 4 and 5 are presented the experiments, the results obtained, and suggestions for future work.

2 Similarity Measures

If we treat texts as random variables whose values are the frequencies of different words in the respective texts, then various statistical correlation measures can be used as similarity measures between that texts. For two texts X and Y and a fixed set of words $\{w_1, w_2, \ldots, w_n\}$ let denote by x_1 the relative frequency of w_1 in X, by y_1 the relative frequency of w_1 in Y and so on by x_n the relative frequency of w_n in X, by y_n the relative frequency of w_n in Y.

The *Pearson's correlation coefficient* is:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n-1}$$

where $\overline{x}$ is the mean of X, $\overline{y}$ the mean of Y, s_x and s_y are the standard deviation of X, Y, respectively (Upton and Cook, 2008). The correlation coefficient measures the tendency of two variables to change in value together (i.e., to either increase or decrease). r is related with the Euclidean distance, the $\sqrt{2(1-r)}$ being the Euclidean distance between the standardized versions of X and Y.

The random variables X, Y representing texts can also be treated as ordinal data, in which data is ordered but cannot be assumed to have equal distance between values. In this case the values of X (and respectively Y) will be the ranks of words $\{w_1, w_2, \ldots, w_n\}$ according to their frequencies in text X rather than of the actual values of these frequencies. The most common correlation statistic for ordinal data is *Spearman's rank-order coefficient* (Upton and Cook 2008):

$$r_{sc} = 1 - \frac{6}{n(n^2-1)} \sum_{i=1}^{n} (x_i - y_i)^2$$

To be noted that, this time, x_i, y_i are ranks and actually, the Spearman's rank-order coefficient is the Pearson's correlation coefficient applied to ranks. The Spearman's footrule is the l_1-version of Spearman's rank-order coefficient:

$$r_{sf} = 1 - \frac{3}{n^2-1} \sum_{i=1}^{n} |x_i - y_i|$$

Another set of correlation statistics for ordinal data are based on the number of concordant and discordant pairs among two variables. The number of concordant pairs among two variables X and Y is $P = |\{(i,j) : 1 \leq i < j \leq n, (x_i - x_j)(y_i - y_j) > 0\}|$. Similarly, the number of discordant pairs is $Q = |\{(i,j) : 1 \leq i < j \leq n, (x_i - x_j)(y_i - y_j) < 0\}|$.

Goodman and Kruskal's gamma(Upton and Cook 2008) is defined as:

$$\gamma = \frac{P - Q}{P + Q}$$

Kendall developed several slightly different types of ordinal correlation as alternatives to gamma. *Kendall's tau-a*(Upton and

Cook 2008) is based on the number of concordant versus discordant pairs, divided by a measure based on the total number of pairs (n = the sample size):

$$\tau_a = \frac{P - Q}{\frac{n(n-1)}{2}}$$

Kendall's tau-b(Upton and Cook 2008) is a similar measure of association based on concordant and discordant pairs, adjusted for the number of ties in ranks.It is calculated as $(P - Q)$ divided by the geometric mean of the number of pairs not tied on X (X_0) and the number of pairs not tied on Y (Y_0):

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + X_0)(P + Q + Y_0)}}$$

All the above three correlation statistics are very related, if n is fixed and X and Y have no tied, then P, X_0 and Y_0 are completely determined by n and Q.

3 Clustering Analysis

An agglomerative hierarchical clustering algorithm (Duda et. al. 2001) arranges a set of objects in a family tree (dendogram) according to their similarity, similarity which in its turn is given by a distance function defined on the set of objects. The algorithm initially assigns each object to its own cluster and then repeatedly merges pairs of clusters until the whole tree is formed. At each step the pair of nearest clusters is selected for merging. Various agglomerative hierarchical clustering algorithms differ in the way in which they measure the distance between clusters. Note that although a distance function between objects exists, the distance measure between clusters (set of objects) remains to be defined. In our experiments we used the *complete linkage* distance between clusters, the maximum of the distances between all pairs of objects drawn from the two clusters (one object from the first cluster, the other from the second).

4 Experiments

In Popescu and Dinu (2009) we have compared the set of distance/similarity measures described here on a collection of 21 nineteenth century English books written by 10 different authors and spanning a variety of genre (the same set of books were used by Koppel et al. (2007) in their authorship verification experiments). The experiments have shown that the similarity measures that treat function words frequencies as ordinal variables (Spearman's rank-order coefficient, Spearman's footrule, Goodman and Kruskal's gamma, Kendall's tau) performed better than the distance/similarity measures that use the actual values of function words frequencies (Euclidean distance, Pearson's correlation coefficient).

Group	Author	Book
American Novelists	Hawthorne	Dr. Grimshawe's Secret
		House of Seven Gables
	Melville	Redburn
		Moby Dick
	Cooper	The Last of the Mohicans
		The Spy
		Water Witch
American Essayists	Thoreau	Walden
		A Week on Concord
	Emerson	Conduct Of Life
		English Traits
British Playwrights	Shaw	Pygmalion
		Misalliance
		Getting Married
	Wilde	An Ideal Husband
		Woman of No Importance
Bronte Sisters	Anne	Agnes Grey
		Tenant Of Wildfell Hall
	Charlotte	The Professor
		Jane Eyre
	Emily	Wuthering Heights
Australian Novelists	B. Baynton	Bush Studies
		Human Toll
	Henry Lawson	Joe Wilson and His Mates
		On the Track
		While the Billy Boils
	Miles Franklin	My Brilliant Career
		Some Everyday Folk and Dawn
		Up the Country: A Saga of...
		Back to Bool Bool

Table 1: The books used in experiments

The aim of the actual experiments was two-folded. Firstly we wanted to see if the findings in Popescu and Dinu (2009) are confirmed in the case of a larger set (more authors, more books) and secondly to further investigate the ability of some of the similarity measures (Spearman's rank-order coefficient, Goodman and Kruskal's gamma, Kendall's tau) to distinguish between the different nationality of English language writers by adding to the data set works of Australian writers from the same period. To the original data set of Koppel et al. (2007) we added 9 works of three Australian authors from the same period, resulting a data set of 30 books and 13 authors (Table 1).

To perform the experiments, a set of words must be fixed. The most frequent function words may be selected or other criteria may be used for selection. In all our experiments we used the set of function words identified by Mosteller and Wallace (2007) as good candidates for author-attribution studies. We used the agglomerative hierarchical clustering algorithm coupled with the various distance similarity function employed in the comparison to cluster the works in Table 1.

The dendrograms obtained sustain the results of Popescu and Dinu (2009). The resulted dendrograms for Euclidean distance and Pearson's correlation coefficient (not shown because of lack of space) are very similar, which is no surprise taking into account the close relation between the two measures (see section 2.1). The problem of these family trees is that the works of Melville are not grouped together: one being clustered with the essays of Thoreau (Moby Dick) and the other with the novels of Hawthorne. Also, "My Brilliant Career" of M. Franklin is clustered with the novels of Charlotte Bronte. Apart from authorship relation, the dendrograms reflect no other stylistic relation between the works (like grouping the works according to genre or nationality of the authors: American / English / Australian).

Spearman's rank-order coefficient, Goodman and Kruskal's gamma and Kendall's tau produced the same dendrogram (modulo the scale).Figure 1 shows the dendrogram for Kendall's tau. The dendrogram is perfect: all works are clustered according to theirs author. The nationality of the authors is not reflected in the dendrogram (the authors with the same nationality are not clustered together).

We performed a series of experiments to test in which cases the nationality of the authors can be revealed by a stylistic similarity measure. If only British and Australian writers are selected, the Kendall's tau produced the dendrogram presented in Figure 2. As can be seen the first two branches correspond to the nationality of the authors: British writers on upper branch, Australian writers on lower branch. The same thing happen when British and American writers are selected. Again, the writers are clustered according to their nationality: this time, the British writers on lower branch and American writers on upper branch. But when the subset of American and Australian writers is clustered using Kendall's tau, the nationality of the writers is no longer reflected in the family tree produced. The works of each author are clustered together, but there are no clear branches corresponding to the two nationalities.

5 Future Work

In this paper we have compared a set of measures, regarding theirs ability to reflect stylistic similarity between texts. In future work it would be interesting to compare these measures to other possible similarity measures. If the frequencies of different words in the texts are treated as probability distributions instead as random variables, specific measures can be applied: Kullback-Liebler Divergence or Cross Entropy.

References

C. K. Chung, and J. W. Pennebaker. 2007. The psychological function of function words. In K. Fiedler, ed., *Social communication: Frontiers of social psychology*, 343−359. Psychology Press, New York.

L.P. Dinu, M. Popescu and A. Dinu. 2008. Authorship Identification of Romanian Texts with Controversial Paternity. *Proc. LREC 2008*, Marrakech, Morocco.

R. O. Duda, P. E. Hart, and D. G. Stork. 2001. *Pattern Classification* (2nd ed.). Wiley-Interscience Publication.

H. van Halteren, M. Haverkort, H. Baayen, A. Neijt, and F. Tweedie. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12:65−77.

M. Koppel, J. Schler, and E. Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *J. of Machine Learning Research*, 8,1261 −1276.

C. Labbe and D. Labbe. 2006. A tool for literary studies: Intertextual distance and tree classification. *Literary and Linguistic Computing*, 21(3):311−326.

F. Mosteller and D.L. Wallace. 2007. *Inference and Disputed Authorship: The Federalist.* CSLI Publications, Stanford.

M. Popescu, L.P.Dinu, 2008. Rank Distance as a Stylistic Similarity. *Proceedings COLING 2008*, Manchester, UK

M. Popescu, L.P.Dinu, 2009. Comparing Statistical Similarity Measures for Stylistic Multivariate Analysis. *Proceedings RANLP 2009*, Borovets, Bulgaria

G. Upton and I. Cook. 2008. *A Dictionary of Statistics.* Oxford Univ. Press, Oxford.

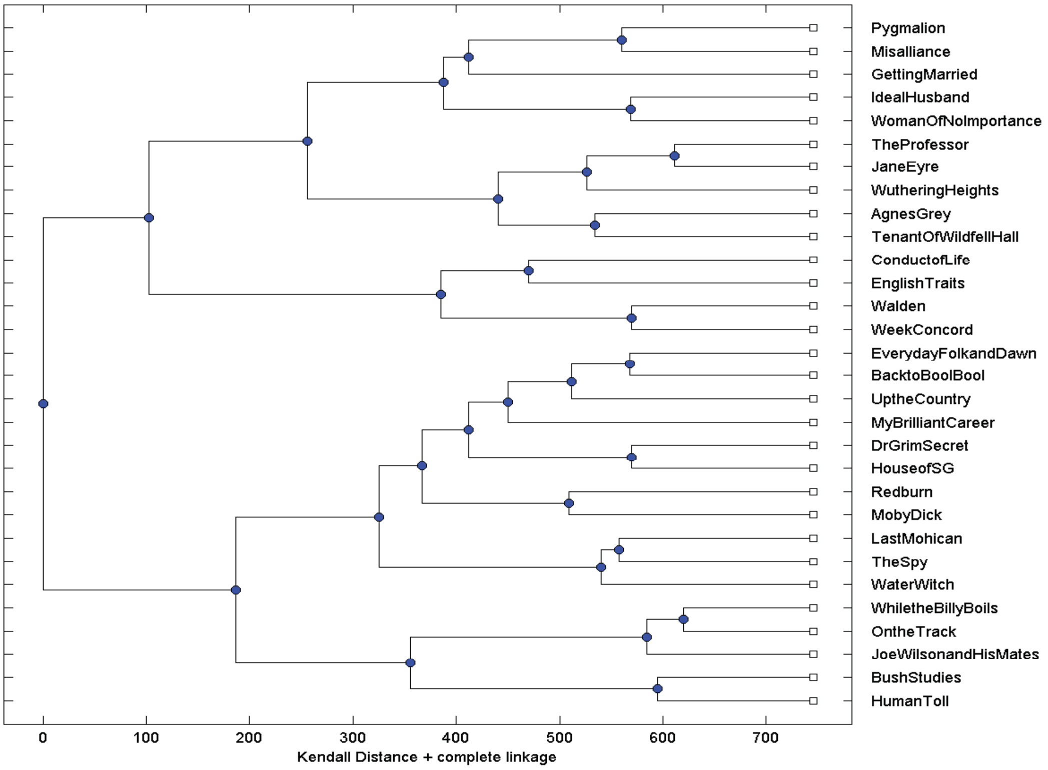

Figure 1: Dendrogram of 30 nineteenth century English books (Kendal's tau)

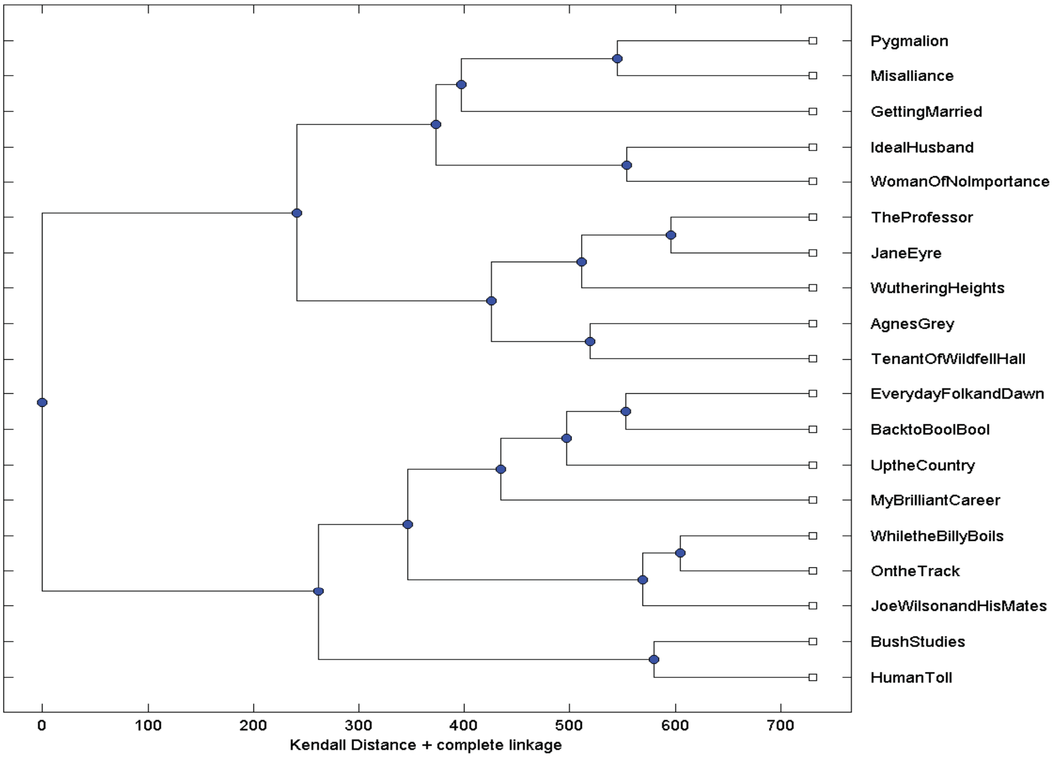

Figure 2: Dendrogram of British and Australian writers (Kendal's tau)