The Role of Hubs in Cross-Lingual Supervised Document Retrieval

Nenad Tomašev, Jan Rupnik, and Dunja Mladenić

Institute Jožef Stefan
Artificial Intelligence Laboratory
Jamova 39, 1000 Ljubljana, Slovenia
{nenad.tomasev,jan.rupnik,dunja.mladenic}@ijs.si

Abstract. Information retrieval in multi-lingual document repositories is of high importance in modern text mining applications. Analyzing textual data is, however, not without associated difficulties. Regardless of the particular choice of feature representation, textual data is high-dimensional in its nature and all inference is bound to be somewhat affected by the well known *curse of dimensionality*. In this paper, we have focused on one particular aspect of the dimensionality curse, known as *hubness*. *Hubs* emerge as influential points in the k-nearest neighbor (kNN) topology of the data. They have been shown to affect the similarity based methods in severely negative ways in high-dimensional data, interfering with both retrieval and classification. The issue of hubness in textual data has already been briefly addressed, but not in the context that we are presenting here, namely the multi-lingual retrieval setting. Our goal was to gain some insights into the cross-lingual hub structure and exploit it for improving the retrieval and classification performance. Our initial analysis has allowed us to devise a hubness-aware instance weighting scheme for canonical correlation analysis procedure which is used to construct the common semantic space that allows the cross-lingual document retrieval and classification. The experimental evaluation indicates that the proposed approach outperforms the baseline. This shows that the hubs can indeed be exploited for improving the robustness of textual feature representations.

Keywords: hubs, curse of dimensionality, document retrieval, cross-lingual, canonical correlation analysis, common semantic space, k-nearest neighbor, classification.

1 Introduction

Text mining has always been one of the core data mining tasks, not surprisingly, as we use language to express our understanding of the world around us, encode knowledge and ideas. Analyzing textual data across a variety of sources can lead to some deep and potentially useful insights.

The use of internet has spawned vast amounts of textual data, even more so now with the advent of Web 2.0 and the increased amount of user-generated content. This data, however, is expressed in a multitude of different languages. There is a high demand for effective and efficient cross-language information retrieval tools, as they allow the users to access potentially relevant information that is written in languages they are not familiar with.

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 185–196, 2013.

Nearest neighbor approaches are common both in text classification [1][2][3] and document retrieval [4][5][6], which is not surprising given both the simplicity and the effectiveness of most kNN methods. Nearest neighbor methods can be employed both at the document level or at the word level.

The *curse of dimensionality* is known to affect the k-nearest neighbor methods in clearly negative ways. The distances concentrate [7] and uncovering relevant examples becomes more difficult. Additionally, some examples have a tendency to become *hubs*, i.e. very frequent nearest neighbors [8]. Though this may not in itself sound like a severe limitation, it turns out to be quite detrimental in practice. Namely, the *hubness* of particular documents depends more on data preprocessing, feature selection, normalization and the similarity measure than on the actual perceived semantic correlation between the document and its reverse nearest neighbors. In other words, the semantics of similarity is often either completely broken or severely compromised around hubs.

Textual data is high-dimensional and the impact of hubness on various text mining tasks involving nearest neighbor reasoning needs to be closely evaluated.

In this paper we examined the hub structure of an aligned bi-lingual document corpus, over a set of 14 different binary categorization problems. We will show that there is a high correlation between the hub structure in different language representations, but this correlation vanishes when using the common semantic representation. This similarity in the kNN graph topology can be exploited for improving the system performance and we demonstrate this by proposing a hubness-aware instance weighting scheme for the canonical correlation analysis [9].

2 Related Work

2.1 Emergence of Hubs

The concept of hubs is probably most widely known from network analysis [10] and the hubs-and-authorities (HITS) algorithm [11] which was a precursor to PageRank in link analysis. However, hubs arise naturally in other domains as well, as for instance the protein interaction networks [12]. Hubness is a common property of high-dimensional data which has been correlated with the distance concentration phenomenon. Any intrinsically high-dimensional data with meaningful distribution centers ought to exhibit some degree of hubness [8][13][14]. The phenomenon has been most thoroughly examined in the music retrieval community [15][16][17]. The researchers had noticed that some songs were constantly being retrieved by the system, even though they were not really relevant for the queries. The hubness in audio data is still an unresolved issue. Similar phenomena in textual data have received comparatively little attention.

Denote by $N_k(x_i)$ the total number of occurrences of a neighbor point x_i. If the $N_k(x_i)$ is very much higher than k, we will say that x_i is a hub, and if it is much lower than k, we will say that x_i is an *orphan* or *anti-hub*. In case of labeled data, we can further decompose the total occurrence frequency as follows: $N_k(x_i) = GN_k(x_i) + BN_k(x_i)$, where $GN_k(x_i)$ and $BN_k(x_i)$ represent the number of *good* and *bad* k-occurrences, respectively. An occurrence is said to be *good* if the labels of neighbor points match and *bad* if there is a mismatch. Bad hubness is, obviously, closely related to the misclassification rates in kNN methods.

In intrinsically high-dimensional data, the entire distribution of k-neighbor occurrences changes and becomes highly skewed.[1] Not only does this result in some examples being frequently retrieved in kNN sets, but also in that most examples never occur as neighbors and are in fact unintentionally ignored by the system. Only a subset of the original data actually participates in the learning process. This subset is not a carefully selected one, so such implicit data reduction usually induces an information loss.

Furthermore, it is advisable to consider not only bad hubness but also the detailed neighbor occurrence profiles, by taking into account the class-specific neighbor occurrences. The occurrence frequency of a neighbor point x_i in neighborhoods of points from class $c \in C$ is denoted by $N_{k,c}(x_i)$ and will be referred to as *class hubness*.

Several hubness-aware classification methods have recently been proposed in order to reduce the negative influence of bad hubs on kNN classification (hw-kNN [8], h-FNN [18], NHBNN [19], HIKNN [20]).

Apart from classification, data hubness has also been used in clustering [21], metric learning [22] and instance selection [23].

2.2 Canonical Correlation Analysis (CCA)

A common approach to analyzing multilingual document collections is to find a common feature representation, so that the documents that are written in different languages can more easily be compared. One way of achieving that is by using the canonical correlation analysis.

Canonical Correlation Analysis (CCA) [9] is a dimensionality reduction technique somewhat similar to Principal Component Analysis (PCA) [24]. It makes an additional assumption that the data comes from two sources or views that share some information, such as a bilingual document corpus [25] or a collection of images and captions [26]. Instead of looking for linear combinations of features that maximize the variance (PCA) it looks for a linear combination of feature vectors from the first view and a linear combination for the second view, that are maximally correlated.

Formally, let $S = (x_1, y_1), \ldots, (x_n, y_n)$ be the sample of paired observations where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^q$ represent feature vectors from some p and q-dimensional feature spaces. Let $X = [x_1, \ldots, x_n]$ and let $Y = [x_1, \ldots, x_n]$ be the matrices with observation vectors as columns, interpreted as being generated by two random vectors $\mathcal{X}$ and $\mathcal{Y}$. The idea is to find two linear functionals (row vectors) $\alpha \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^q$ so that the random variables $\alpha \cdot \mathcal{X}$ and $\beta \cdot \mathcal{Y}$ are maximally correlated. The α and β map the random vectors to random variables, by computing the weighted sums of vector components. This gives rise to the following optimization problem:

$$\underset{\alpha \in \mathbb{R}^p, \beta \in \mathbb{R}^q}{\text{maximize}} \quad \frac{\alpha C_{XY} \beta'}{\sqrt{\alpha C_{XX} \alpha'} \sqrt{\beta C_{YY} \beta'}}, \tag{1}$$

where C_{XX} and C_{YY} are empirical estimates of the variances of $\mathcal{X}$ and $\mathcal{Y}$ respectively and C_{XY} is an estimate of the covariance matrix. Assuming that the observation vec-

[1] *Skewness* of the k-occurrence distribution is defined as $SN_k(x) = \frac{m_3(N_k(x))}{m_2^{3/2}(N_k(x))} = \frac{1/n \sum_{i=1}^{N} (N_k(x_i) - k)^3}{(1/n \sum_{i=1}^{N} (N_k(x_i) - k)^3)^{3/2}}$. High positive skewness which is encountered in intrinsically high-dimensional data indicates that the distribution tail is longer on the right distribution side.

tors are centered, the matrices are computed in the following way: $C_{XX} = \frac{1}{n-1}XX'$, $C_{YY} = \frac{1}{n-1}YY'$ and $C_{XY} = \frac{1}{n-1}XY'$.

This optimization task can be reduced to an eigenvalue problem and includes inverting the variance matrices C_{XX} and C_{YY}. In case of non-invertible matrices, it is possible to use a regularization technique by replacing C_{XX} with $(1-\kappa)C_{XX}+\kappa I$, where $\kappa \in [0,1]$ is the regularization coefficient and I is the identity matrix.

A single canonical variable is usually inadequate in representing the original random vector and typically one looks for k projection pairs $(\alpha_1, \beta_1), \ldots, (\alpha_k, \beta_k)$, so that α_i and β_i are highly correlated and α_i is uncorrelated with α_j for $j \neq i$ and analogously for β.

The problem can be reformulated as a symmetric eigenvalue problem for which efficient solutions exist. If the data is high-dimensional and the feature vectors are sparse, iterative methods can be used, such as the well known Lanczos algorithm [27]). If the size of the corpus is not prohibitively large, it is also possible to work with the dual representation and use the "kernel trick" [28] to yield a nonlinear version of CCA.

3 Data

For the experiments, we examined the Acquis aligned corpus data (http://langtech.jrc.it/JRC-Acquis.html), which comprise a set of more than 20000 documents in many different languages. To simplify the initial analysis, we focused on the bi-lingual case and compared the English and French aligned document sets. We will consider more language pairs in our future work. The documents were labeled and associated with 14 different binary classification problems.

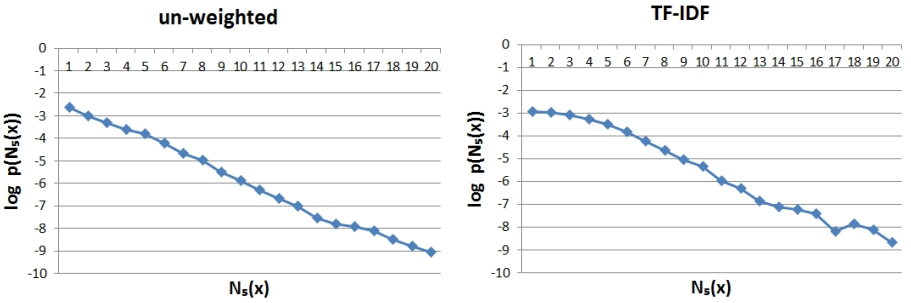
The documents were analyzed in the standard bag-of-words representation after tokenization, lemmatization and stop word removal. Only nouns, verbs, adjectives and adverbs were retained, based on the part-of-speech tags. The inter-document similarity was measured by the cosine similarity measure.

Common semantic representation for the two aligned document sets was obtained by applying CCA. Both English and French documents were then mapped onto the common semantic space (CS:E, CS:F). The used common semantic representation was 300-dimensional, as we wanted to test our assumptions in the context of dimensionality reduction and slight information loss. Longer representations would be preferable in practical applications.

The Acquis corpus exhibits high hubness. This is apparent from Figure 1. The data was normalized by applying TF-IDF, which is a standard preprocessing technique. The normalization only slightly reduces the overall hubness.

The common semantic projections exhibit significantly lower hubness than the original feature representations, which already suggests that there might be important differences in the hub structure. The outline of the data is given in Table 1. The two languages exhibit somewhat different levels of hubness.

If the hubness information is to be used in the multi-lingual context, it is necessary to understand how it maps from one language representation to another. Both the quantitative and the qualitative aspects of the mapping need to be considered. The quantitative aspect refers to the the correlation between the total document neighbor occurrence counts and provides the answer to the general question of whether the same documents



(a) No feature weighting (b) TF-IDF

Fig. 1. The logarithmic plots of the 5-occurrence distribution on the set of English Acquis documents with or without performing TF-IDF feature weighting. The straight line in the un-weighted case shows an exponential law in the decrease of the probability of achieving a certain number of neighbor occurrences. Therefore, frequent neighbors are rare and most documents are anti-hubs. Note that $N_5(x)$ is sometimes much more than 20, both charts are cut-off there for clarity. Performing TF-IDF somewhat reduces the overall hubness, even though it remains high.

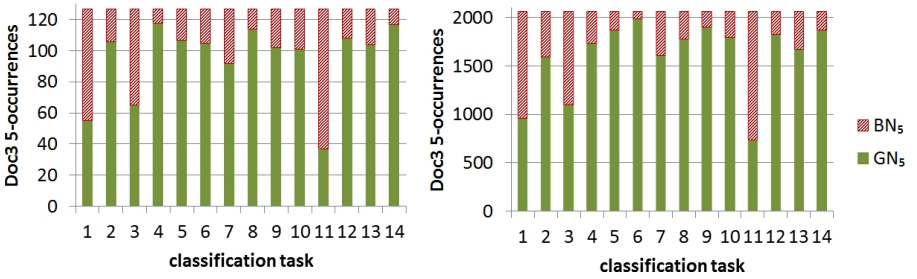
Table 1. Overview of the k-occurrence skewness (S_{N_k}) for all four document corpus representations. To further illustrate the severity of the situation, the degree of the major hub ($\max N_k$) is also given. Both quantities are shown for $k = 1$ and $k = 5$.

Data set	size	d	S_{N_1}	$\max N_1$	S_{N_5}	$\max N_5$
ENG	23412	254963	16.13	95	19.45	432
FRA	23412	212955	80.98	868	54.22	3199
CS:E	23412	300	5.20	38	1.99	71
CS:F	23412	300	4.90	38	1.99	62

become hubs in different languages. The qualitative aspect is concerned with characterizing the type of influence expressed by the hubs in correlating the good and bad hubness (label mismatch percentages) in both languages.

Let us consider one randomly chosen hub document from the corpus. Figure 2 shows its occurrence profiles in both English and French over all 14 binary classification problems. The good/bad occurrence distributions for this particular document appear to be quite similar in both languages, even though the total hubness greatly differs. From this we can conclude that, even though the overall occurrence frequency depends on the language, the semantic nature of the document determines the type of influence it will exhibit if and when it becomes a hub. On the other hand, this particular document is an anti-hub in both projections onto the common semantic space, i.e. it never occurs as a neighbor there. This illustrates how the CCA mapping changes the nature of the k-nearest neighbor structure, which is what Table 1 also confirms.

The observations from examining the influence profiles of a single document are easily generalized by considering the average Pearson correlation between bad hubness ratios over the 14 binary label assignments, as shown in Table 2(a). There is a quite strong positive correlation between document influence profiles in all considered representations and it is strongest between the projections onto the common semantic space,



(a) English version of the text (b) French version of the text

Fig. 2. Comparing the 5-occurrences of one randomly chosen document (Doc-3) across various classification tasks (label arrays) in English and French language representations. The hubness of Doc-3 differs greatly, but the type of its influence (good/bad hubness ratio) seems to be preserved.

which was to be expected. As for the total number of neighbor occurrences (Table 2(b) and Table 2(c)), the Pearson product-moment gives positive correlation between the hubness of English and French texts, as well as between the projected representations. In all other cases there is no linear correlation. We measured the non-linear correlation by using the Spearman correlation coefficient (Table 2(c)). It seems that there is some positive non-linear correlation between hubness in all the representations.

The results of correlation comparisons can be summarized as follows: frequent neighbor documents among English texts are usually also frequent neighbors among the French texts and the nature of their influence is very similar. Good/bad neighbor documents in English texts are expected to be good/bad neighbor documents in French

Table 2. Correlations of document hubness and bad hubness between different language representations: English, French, and their projections onto the common semantic space

(a) Pearson correlation between bad hubness ratios of documents ($BN_k(x)/N_k(x)$)

ENG	FRA	CS:E	CS:F	
	0.68	0.61	0.58	ENG
		0.56	0.58	FRA
			0.76	CS:E
				CS:F

(b) Pearson correlation between total hubness (occurrence frequencies)

ENG	FRA	CS:E	CS:F	
	0.47	0.08	0.06	ENG
		0.01	0.01	FRA
			0.64	CS:E
				CS:F

(c) Spearman correlation between total hubness (occurrence frequencies)

ENG	FRA	CS:E	CS:F	
	0.67	0.29	0.25	ENG
		0.25	0.29	FRA
			0.70	CS:E
				CS:F

texts and vice-versa. We will exploit this apparent regularity for improving the neighbor structure of the common semantic space, as will be discussed in Section 4.

4 Towards a Hubness-Aware Common Semantic Representation

In the canonical correlation analysis, all examples contribute equally to the process of building a common semantic space. However, due to hubness, not all documents are to be considered equally relevant or equally reliable. Documents that become bad hubs exhibit a highly negative influence. Furthermore, as shown in Figure 2, a single hub-document can act both as a bad hub and as a good hub at the same time, depending on the specific classification task at hand. Therefore, instance selection doesn't seem to be a good approach, as we cannot both accept and reject an example simultaneously.

What we propose instead is to introduce instance weights to the CCA procedure in order to control the influence of hubs on forming the common semantic representation in hope that this would in turn improve the cross-lingual retrieval and classification performance in the common semantic space.

The weights introduce a bias in finding the canonical vectors: the search for canonical vectors is focused on the spaces spanned by the instances with high weights.

Given a document sample S, let $u_1, \ldots, u_n$ be the positive weights for the examples $x_i \in X$ and $v_1, \ldots, v_n$ be the positive weights for the examples $y_i \in Y$. We propose to compute the modified covariance and variance matrices as follows:

$$\tilde{C}_{XX} := \frac{1}{n-1}\sum_{i=1}^{n} u_i^2 x_i x_i', \qquad \tilde{C}_{YY} := \frac{1}{n-1}\sum_{i=1}^{n} v_i^2 y_i y_i'$$
$$\tilde{C}_{XY} := \frac{1}{n-1}\sum_{i=1}^{n} u_i v_i x_i y_i' \tag{2}$$

These matrices are input for the standard CCA optimization problem. By modifying them, we are able to directly influence the outcome of the process. The weighting approach is equivalent to performing over-sampling of the instances based on their specified weights and then computing the covariances and variances.

Let $h(x_i, k)$ and $h_B(x_i, k)$ be the standardized hubness and standardized bad hubness scores respectively, i.e. $h(x_i, k) = \frac{N_k(x_i) - \mu_{N_k(x_i)}}{\sigma_{N_k(x_i)}}$ and $h_B(x_i, k) = \frac{BN_k(x_i) - \mu_{BN_k(x_i)}}{\sigma_{BN_k(x_i)}}$. A high standardized hubness score means that the document is very influential and relevant for classification and retrieval, while a high bad hubness score indicates that the document is unreliable.

We have experimented with several different weighting schemes. We will focus on two main approaches. The first approach would be to increase the influence of relevant points (*hubs*) in the CCA weighting. The second meaningful approach is to reduce the influence of unreliable points (*bad hubs*). Additionally, for comparisons, we will also consider the opposite of what we propose, i.e. reducing the influence of hubs

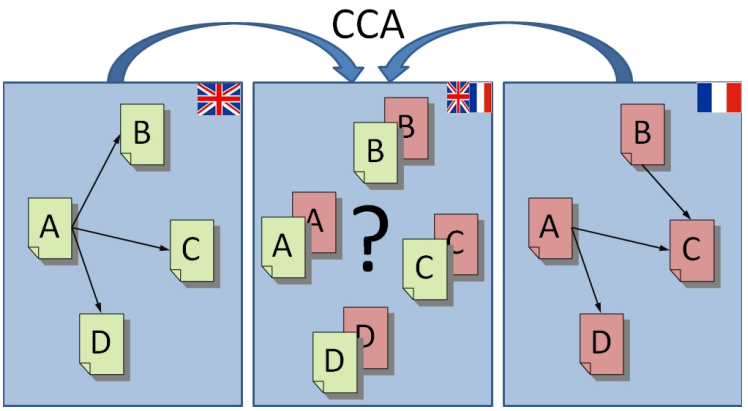


Fig. 3. The CCA procedure maps the documents written in different languages onto the common semantic space. According to the analysis given in Table 2, this changes the kNN structure significantly, which has consequences for the subsequent document retrieval and/or classification. By introducing instance weights we can influence the mapping so that we preserve certain aspects of the original hub-structure and reject the unwanted parts of it.

and increasing the influence of bad hubs. Therefore, the considered weighting schemes are given as follows: un-weighted, $v_i := 1$, emphasized hubs, $v_i := e^{h(x_i,k)}$, de-emphasized hubs, $v_i := e^{-h(x_i,k)}$, emphasized bad hubs, $v_i := e^{h_B(x_i,k)}$, and de-emphasized bad hubs, $v_i := e^{-h_B(x_i,k)}$.

5 Experimental Evaluation

In the experimental protocol, we randomly selected two disjoint subsets of the aligned corpus: 2000 documents were used for training ad 1000 for testing. For each of the 14 binary classification problems we computed five common semantic spaces with CCA on the training set: the non-weighted variant (CS:N), emphasized hubs (CS:H), de-emphasized hubs (CS:h), emphasized bad hubs (CS:B) and de-emphasized bad hubs (CS:b). The training and test documents in both languages were then projected onto the common semantic space. In each case, we evaluated the quality of the common semantic space by measuring the performance of both classification and document retrieval. The whole procedure was repeated 10 times, hence yielding the repeated random sub-sampling validation. We have measured the average performance and its standard deviation.

Many of the binary label distributions were highly imbalanced. This is why the classification performance was measured by considering the Matthews Correlation Coefficient (MCC) [29].

Comparing the classification performance on the original (non-projected) documents with the performance on the common semantic space usually reveals a clear degradation in performance, unless the dimensionality of the projected space is high enough to capture all the relevant discriminative information.

The overview of the classification experiments is given in Table 3. We only report the result on the English texts and projections, as they are basically the same in the French

part of the corpus. We have used the kNN classifier with $k = 5$, as we are primarily interested in capturing the change of the neighbor-structure in the data. It is immediately apparent that the weights which emphasize document hubness (CS:H) achieve the best results among the common semantic document representations. Reducing the influence of bad hubs (CS:b) is in itself not enough to positively affect the classification performance. This might be because many hubs reside in borderline regions, so they might carry some relevant disambiguating feature information. It seems that emphasizing the relevance by increasing the preference for all hub-documents gives the best classification results.

Table 3. The Matthews correlation coefficient (MCC) values achieved on different projected representations. The symbols •/○ denote statistically significant worse/better performance ($p < 0.01$) compared to the non-weighted projected representation (CS:N).

Label	Original	CS:N	CS:H	CS:h	CS:B	CS:b
lab1	73.0 ± 3.3	34.2 ± 4.6	**69.2 ± 2.8** ○	66.0 ± 3.3 ○	52.8 ± 4.8 ○	46.6 ± 10.2 ○
lab2	69.2 ± 3.0	52.3 ± 4.4	**65.1 ± 3.9** ○	38.3 ± 3.8 •	45.8 ± 7.0	35.7 ± 8.6 •
lab3	50.2 ± 3.3	27.6 ± 3.8	44.1 ± 3.0 ○	42.2 ± 5.0 ○	**44.8 ± 3.6** ○	33.7 ± 3.0 ○
lab4	32.2 ± 4.4	18.8 ± 6.4	**28.1 ± 2.8** ○	21.1 ± 3.9	20.6 ± 3.7	20.3 ± 6.5
lab5	28.9 ± 12.4	16.8 ± 12.9	17.7 ± 11.7	**21.9 ± 14.4**	10.2 ± 5.5	15.7 ± 6.0
lab6	38.1 ± 6.2	31.2 ± 6.0	29.3 ± 8.2	**33.6 ± 5.4**	23.5 ± 5.8 •	26.2 ± 6.6
lab7	54.5 ± 3.2	38.9 ± 4.0	**48.4 ± 4.2** ○	45.7 ± 3.0 ○	42.3 ± 6.3	36.5 ± 6.8
lab8	44.6 ± 6.3	31.5 ± 6.9	**40.4 ± 6.4** ○	33.5 ± 5.7	23.0 ± 5.0 •	19.6 ± 8.7 •
lab9	76.2 ± 3.4	32.0 ± 5.4	**74.4 ± 3.4** ○	61.8 ± 3.7 ○	45.7 ± 5.2 ○	37.7 ± 7.6
lab10	41.4 ± 4.2	26.1 ± 3.8	34.0 ± 3.8 ○	31.6 ± 5.5	**34.4 ± 4.6** ○	26.6 ± 5.2
lab11	53.5 ± 2.5	27.9 ± 2.8	**48.6 ± 4.0** ○	42.0 ± 3.5 ○	44.9 ± 3.8 ○	33.7 ± 3.8 ○
lab12	39.2 ± 4.0	31.5 ± 3.4	35.4 ± 5.9	**35.6 ± 6.6**	22.8 ± 4.9 •	20.3 ± 5.7 •
lab13	45.4 ± 3.4	29.9 ± 5.2	**38.5 ± 6.0** ○	37.1 ± 4.6 ○	32.6 ± 5.4	28.0 ± 4.9
lab14	49.9 ± 4.5	35.4 ± 7.1	**44.8 ± 7.6**	44.1 ± 7.4	22.4 ± 5.9 •	23.4 ± 11.7
AVG	49.7	31.0	**44.1**	39.6	33.3	28.9

In evaluating the document retrieval performance, we will focus on the k-neighbor set purity as the most relevant metric. The inverse mate rank is certainly also important, but the label matches are able to capture a certain level of semantic similarity among the fetched results. A higher purity among the neighbor sets ensures that, for instance, if your query is about the civil war, you will not get results about gardening, regardless of whether the aligned mate was retrieved or not. This is certainly quite useful. The comparisons are given in Table 4.

Once again, the CS:H weighting proves to be the best among the evaluated hubness-aware weighting approaches, as it retains the original purity of labels among the document kNNs. It is significantly better than the un-weighted baseline (CS:N).

The CS:H weighting produces results most similar to the ones in the original English corpus and we hypothesized that it is because this particular document weighting scheme best helps to preserve the kNN structure of the original document set. We examined the relevant correlations and it turns out that this is indeed the case, as shown

Table 4. The average purity of the k-nearest document sets in each representation. The symbols •/∘ denote significantly lower/higher purity ($p < 0.01$) compared to the non-weighted case (CS:N). The best result in each line is in bold.

Label	Original	CS:N	CS:H	CS:h	CS:B	CS:b
lab1	84.5 ± 1.3	80.7 ± 1.6	**84.1 ± 1.1** ∘	83.3 ± 1.5 ∘	83.7 ± 1.5 ∘	81.7 ± 2.1
lab2	90.5 ± 1.2	84.5 ± 3.2	**90.1 ± 1.2** ∘	88.2 ± 2.0 ∘	89.6 ± 1.5 ∘	84.9 ± 3.7
lab3	74.4 ± 0.9	71.3 ± 1.0	74.4 ± 1.0 ∘	73.6 ± 0.9 ∘	**74.6 ± 1.2** ∘	72.6 ± 1.1
lab4	85.8 ± 1.6	84.6 ± 4.4	85.9 ± 1.5	**85.9 ± 1.8**	85.1 ± 1.5	84.1 ± 3.6
lab5	96.0 ± 0.6	95.9 ± 1.3	95.9 ± 0.8	**96.3 ± 0.8**	95.3 ± 1.0	94.5 ± 3.0
lab6	91.7 ± 0.9	90.2 ± 3.4	**91.6 ± 1.1**	91.6 ± 1.5	90.8 ± 1.5	89.5 ± 3.5
lab7	79.7 ± 0.8	78.0 ± 2.2	**79.7 ± 1.0**	79.0 ± 1.6	79.5 ± 0.6	77.8 ± 1.7
lab8	89.1 ± 1.3	87.0 ± 3.4	**89.0 ± 1.2**	88.5 ± 1.6	88.0 ± 1.3	85.6 ± 3.2
lab9	91.8 ± 1.1	84.7 ± 3.1	**92.0 ± 1.1** ∘	89.6 ± 1.5 ∘	90.9 ± 1.3 ∘	83.9 ± 3.1
lab10	84.3 ± 0.7	**84.5 ± 1.4**	84.4 ± 0.6	84.4 ± 0.8	83.7 ± 0.7	83.4 ± 1.6
lab11	77.0 ± 0.9	73.5 ± 1.1	77.1 ± 0.8 ∘	75.5 ± 0.9 ∘	**77.3 ± 0.6** ∘	74.7 ± 1.2
lab12	88.7 ± 1.2	**88.7 ± 3.3**	88.6 ± 1.3	88.7 ± 1.9	87.6 ± 1.5	87.9 ± 3.5
lab13	82.3 ± 1.5	81.9 ± 2.1	**82.4 ± 1.5**	82.2 ± 1.8	82.0 ± 1.4	80.7 ± 2.5
lab14	92.7 ± 0.8	92.1 ± 2.8	92.3 ± 0.7	**92.7 ± 1.2**	91.7 ± 1.3	91.7 ± 3.1
AVG	86.3	84.1	**86.3**	85.7	85.7	83.8

Table 5. The correlations of document hubness between some of the different common semantic representations, as well as the original English documents. CS:H (emphasize hubness when building the rep.) best preserves the original kNN structure, which is why it leads to similar classification performance, despite the dimensionality reduction.

(a) Pearson correlation between total hubness on the **training** set (occurrence frequencies)

ENG	CS:N	CS:H	CS:h	
	0.05	**0.42**	0.02	ENG
		0.03	0.05	CS:N
			0.02	CS:H
				CS:h

(b) Pearson correlation between total hubness on the **test** set (occurrence frequencies)

ENG	CS:N	CS:H	CS:h	
	0.65	**0.88**	0.75	ENG
		0.68	0.93	CS:N
			0.80	CS:H
				CS:h

in Table 5. By preserving the original structure, it compensates for some of the information loss which would have resulted due to the dimensionality reduction during the CCA mapping.

6 Conclusions and Future Work

We have examined the impact of hubness on cross-lingual document retrieval and classification, from the perspective of calculating the common semantic document representation. Hubness is an important aspect of the dimensionality curse which plagues the similarity-based learning methods.

Our analysis shows that the hub-structure of the data remains preserved across different languages, but is radically changed by the canonical correlation analysis mapping onto the common semantic space. The dimensionality reduction also results in some information loss. We have proposed to overcome the information loss by introducing the *hubness-aware* instance weights into the CCA optimization problem, which have helped in preserving the original kNN structure of the data during the CCA mapping.

The experimental evaluation shows that increasing the influence of hubs on spanning the common semantic space results in an increased kNN classification performance and the higher neighbor set purity.

These initial experiments were performed on an aligned bi-lingual corpus and we intend to expand the analysis by comparing more languages. Additionally, we intend to examine the unsupervised aspects of the problem, like clustering.

Acknowledgments. This work was supported by ICT Programme of the EC under XLike (ICT-STREP-288342) and LTWeb (ICT-CSA-287815).

References

1. Tan, S.: An effective refinement strategy for knn text classifier. Expert Syst. Appl. 30, 290–298 (2006)
2. Jo, T.: Inverted index based modified version of knn for text categorization. JIPS 4(1), 17–26 (2008)
3. Trieschnigg, D., Pezik, P., Lee, V., Jong, F.D., Rebholz-Schuhmann, D.: Mesh up: effective mesh text classification for improved document retrieval. Bioinformatics (2009)
4. Chau, R., Yeh, C.H.: A multilingual text mining approach to web cross-lingual text retrieval. Knowl.-Based Syst., 219–227 (2004)
5. Peirsman, Y., Padó, S.: Cross-lingual induction of selectional preferences with bilingual vector spaces. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT 2010, pp. 921–929. Association for Computational Linguistics (2010)
6. Lucarella, D.: A document retrieval system based on nearest neighbour searching. J. Inf. Sci. 14, 25–33 (1988)
7. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Van den Bussche, J., Vianu, V. (eds.) ICDT 2001. LNCS, vol. 1973, p. 420. Springer, Heidelberg (2000)
8. Radovanović, M., Nanopoulos, A., Ivanović, M.: Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In: Proc. 26th Int. Conf. on Machine Learning (ICML), pp. 865–872 (2009)
9. Hotelling, H.: The most predictable criterion. Journal of Educational Psychology 26, 139–142 (1935)
10. David, E., Jon, K.: Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press, New York (2010)
11. Kleinberg, J.M.: Hubs, authorities, and communities. ACM Comput. Surv. 31(4es) (December 1999)
12. Ning, K., Ng, H., Srihari, S., Leong, H., Nesvizhskii, A.: Examination of the relationship between essential genes in ppi network and hub proteins in reverse nearest neighbor topology. BMC Bioinformatics 11, 1–14 (2010)

13. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research 11, 2487–2531 (2011)
14. Radovanović, M., Nanopoulos, A., Ivanović, M.: On the existence of obstinate results in vector space models. In: Proc. 33rd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 186–193 (2010)
15. Aucouturier, J., Pachet, F.: Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences 1 (2004)
16. Flexer, A., Gasser, M., Schnitzer, D.: Limitations of interactive music recommendation based on audio content. In: Proceedings of the 5th Audio Mostly Conference: A Conference on Interaction with Sound, AM 2010, pp. 13:1–13:7. ACM, New York (2010)
17. Schnitzer, D., Flexer, A., Schedl, M., Widmer, G.: Using mutual proximity to improve content-based audio similarity. In: ISMIR 2011, pp. 79–84 (2011)
18. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: Hubness-based fuzzy measures for high dimensional k-nearest neighbor classification. In: Machine Learning and Data Mining in Pattern Recognition, MLDM Conference (2011)
19. Tomasev, N., Radovanović, M., Mladenić, D., Ivanović, M.: A probabilistic approach to nearest-neighbor classification: naive hubness bayesian kNN. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, CIKM 2011, Glasgow, Scotland, UK, pp. 2173–2176. ACM, New York (2011)
20. Tomašev, N., Mladenić, D.: Nearest neighbor voting in high dimensional data: Learning from past occurrences. Computer Science and Information Systems 9(2) (June 2012)
21. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: The role of hubness in clustering high-dimensional data. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part I. LNCS, vol. 6634, pp. 183–195. Springer, Heidelberg (2011)
22. Tomašev, N., Mladenić, D.: Hubness-aware shared neighbor distances for high-dimensional *k*-nearest neighbor classification. In: Corchado, E., Snášel, V., Abraham, A., Woźniak, M., Graña, M., Cho, S.-B. (eds.) HAIS 2012, Part II. LNCS, vol. 7209, pp. 116–127. Springer, Heidelberg (2012)
23. Buza, K., Nanopoulos, A., Schmidt-Thieme, L.: INSIGHT: Efficient and effective instance selection for time-series classification. In: Huang, J.Z., Cao, L., Srivastava, J. (eds.) PAKDD 2011, Part II. LNCS, vol. 6635, pp. 149–160. Springer, Heidelberg (2011)
24. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philos. Mag. 2(6), 559–572 (1901)
25. Fortuna, B., Cristianini, N., Shawe-Taylor, J.: A Kernel Canonical Correlation Analysis For Learning The Semantics Of Text. In: Kernel Methods in Bioengineering, Communications and Image Processing, pp. 263–282. Idea Group Publishing (2006)
26. Hardoon, D.R., Szedmák, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural Computation 16(12), 2639–2664 (2004)
27. Cullum, J.K., Willoughby, R.A.: Lanczos Algorithms for Large Symmetric Eigenvalue Computations, vol. 1. Society for Industrial and Applied Mathematics, Philadelphia (2002)
28. Jordan, M.I., Bach, F.R.: Kernel independent component analysis. Journal of Machine Learning Research 3, 1–48 (2001)
29. Powers, D.M.W.: Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Technical Report SIE-07-001, School of Informatics and Engineering, Flinders University, Adelaide, Australia (2007)