International Journal of Computer Science
ISSN:2348-6600
Oddity...Probe...Reviste...
Published by SK Research Group of Companies(SKRGC)

http://www.ijcsjournal.com Volume 3, Issue 1, No 5, 2015. ISSN: 2348-6600
Reference ID: IJCS-094 PAGE NO: 535-541.

CLUSTERING HIGH-DIMENSIONAL DATA USING HUBNESS-BASED APPROACHES

SENTHILKUMAR.G,

ME CSE,

SSM College of Engineering, India

kgsenthil2005@gmail.com

NANTHINI.M ASSISTANT PROFESSOR

ME CSE,

SSM College of Engineering, India

nanthinissm@gmail.com

***Abstract*—Clustering is an unsupervised process of grouping elements together, so that elements assigned to the same cluster are more similar to each other than to the remaining data points. High-dimensional data takes place in many fields. Clustering process is because of sparsity, also growing complexity in unique distances between data points. Here capture an original perception on the trouble of clustering high-dimensional data. To neglect the curse of dimensionality by scrutinizing a lower dimensional feature subspace, hold dimensionality by taking advantage of inherently high-dimensional phenomena. Exclusively, using hubness. Validate our hypothesis by demonstrating that hubness is a good measure of point centrality within a high-dimensional data cluster, and by proposing several hubness-based clustering algorithms, showing that major hubs can be used effectively as cluster prototypes or as guides during the search for centroid-based cluster configurations. The proposed method called "Neighbor clustering", which takes as input measures of correspondence between pairs of data points. Experimental results demonstrate good performance of our algorithms in multiple settings,**

Keywords— Clustering, curse of dimensionality, nearest neighbors, hubs

1 INTRODUCTION

Clustering in general is an unsupervised process of grouping elements together, so that elements assigned to the same cluster are more similar to each other than to the remaining data points [1].this goal is often difficult to achieve in practice. Over the years, various clustering algorithms have been proposed, which can be roughly divided into four groups: partitional, hierarchical, density based, and subspace algorithms. Algorithms from the fourth group search for clusters in some lower dimensional projection of the original data, and have been generally preferred when dealing with data that are high dimensional [2], [3], [4], [5]. The motivation for this preference lies in the observation that having more dimensions usually leads to the so-called curse of dimensionality, where the performance of many standard machine-learning algorithms becomes impaired. This is mostly due to two pervasive effects: the empty space phenomenon and concentration of distances. The former refers to the fact that all high-dimensional data sets tend to be sparse, because the number of points required to represent any distribution grows exponentially with the number of dimensions. This leads to bad density estimates for high-dimensional data, causing difficulties for densitybased approaches. The latter is a somewhat counterintuitive property of high dimensional data representations, where all distances between data points tend to become harder to distinguish as dimensionality increases, which can cause problems with distance-based algorithms.

The difficulties in dealing with high dimensional data are omnipresent and abundant. However, not all phenomena that arise are necessarily detrimental to clustering techniques. We

International Journal of Computer Science
ISSN:2348-6600
Oddity...Probe...Reviste...
Published by SK Research Group of Companies(SKRGC)
http://www.ijcsjournal.com Volume 3, Issue 1, No 5, 2015. ISSN: 2348-6600
Reference ID: IJCS-094 PAGE NO: 535-541.

will show in this paper that hubness, which is the tendency of some data points in high-dimensional data sets to occur much more frequently in k-nearest neighbor lists of other points than the rest of the points from the set, can in fact be used for clustering. To our knowledge, this has not been previously attempted. In a limited sense, hubs in graphs have been used to represent typical word meanings in which were not used for data clustering. A similar line of research has identified essential proteins as hubs in the reverse nearest neighbor topology of protein interaction networks. We have focused on exploring the potential value of using hub points in clustering by designing hubness-aware clustering algorithms and testing them in a high-dimensional context. The hubness phenomenon and its relation to clustering will be further addressed

There are two main contributions of this paper. First, in experiments on synthetic data we show that hubness is a good measure of point centrality within a high dimensional data cluster and that major hubs can be used effectively as cluster prototypes. In addition, we propose three new clustering algorithms and evaluate their performance in various high dimensional clustering tasks. We compared the algorithms with a baseline state-of-the-art prototype based method (K-means++), as well as kernel-based and density-based approaches. The evaluation shows that our algorithms frequently offer improvements in cluster quality and homogeneity. The comparison with kernel K means reveals that kernel-based extensions of the initial approaches should also be considered in the future. Our current focus was mostly on properly selecting cluster prototypes, with the proposed methods tailored for detecting approximately hyper spherical clusters. The rest of the paper is structured as follows: In the next section, we present the related work, discusses in general the phenomenon of hubness, while describes the proposed algorithms that are exploiting hubness for data clustering. Presents the experiments we performed on both synthetic and real-world data. We expect our observations and approach to open numerous directions for further research, many of which are outlined by our final remarks

2 RELATED WORKS

Even though hubness has not been given much attention in data clustering, hubness information is drawn from k nearest neighbor lists, which have been used in the past to perform clustering in various ways. These lists may be used for computing density estimates, by observing the volume of space determined by the k-nearest neighbors. Density based clustering methods often rely on this kind of density estimation the implicit assumption made by densitybased algorithms is that clusters exist as high density regions separated from each other by low-density regions. In high-dimensional spaces this is often difficult to estimate, due to data being very sparse. There is also the issue of choosing the proper neighborhood size, since both small and large values of k can cause problems for density based approaches enforcing k-nearest-neighbor consistency in algorithms such as k-means was also explored the most typical usage of k-nearest-neighbor lists, however, is to construct a k-nn graph [19] and reduce the problem to that of graph clustering. Consequences and applications of hubness have been more thoroughly investigated in other related fields: classification image feature representation data reduction collaborative filtering, text retrieval and music retrieval . In many of these studies it was shown that hubs can offer valuable information that can be used to improve existing methods and devise new algorithms for the given task. Finally, the interplay between clustering and hubness was briefly examined in [23], where it was observed that hubs may not cluster well using conventional prototype-based clustering algorithms, since they not only tend to be close to points belonging to the same cluster (i.e., have low intracluster distance) but also tend to be close to points assigned to other clusters (low intercluster distance). Hubs can, therefore, be viewed as (opposing) analogues of outliers, which have high inter- and intra cluster distance, suggesting that hubs should also receive special attention [23]. In this paper, we have adopted the approach of using hubs as cluster prototypes and/or guiding points during prototype search.

3 THE HUBNESS PHENOMENON

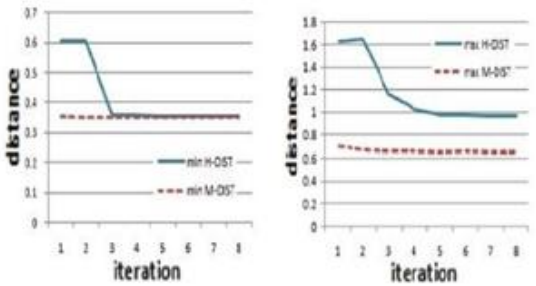
Hubness is an aspect of the curse of dimensionality pertaining to nearest neighbors which has only recently come to attention, unlike the much discussed distance concentration phenomenon. Let D _ IRd be a set of data points and let NkðxÞ denote the number of k-occurrences of point x 2 D, i.e., the number of times x occurs in k-nearest neighbor lists of other points from D. As the dimensionality of data increases, the distribution of k-occurrences becomes considerably skewed .As a consequence, some data points,

International Journal of Computer Science
ISSN:2348-6600
Oddity...Probe...Reviste...
Published by SK Research Group of Companies(SKRGC)
http://www.ijcsjournal.com Volume 3, Issue 1, No 5, 2015. ISSN: 2348-6600
Reference ID: IJCS-094 PAGE NO: 535-541.

which we will refer to as hubs, are included in many more k-nearest-neighbor lists than other points. In the rest of the text, we will refer to the number of k-occurrences of point x € D as its hubness score. It has been shown that hubness, as a phenomenon, appears in high-dimensional data as an inherent property of high dimensionality, and is neither an artifact of finite samples nor a peculiarity of some specific data sets . Naturally, the exact degree of hubness may still vary and is not uniquely determined by dimensionality.

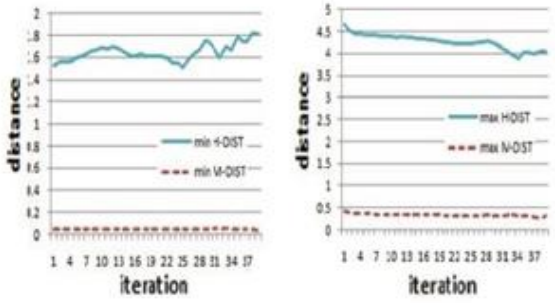
3.1 Emergence of Hubs

The concentration of distances enables one to view unimodal high-dimensional data as lying approximately on a hyper sphere cantered at the data distribution mean. However, the variance of distances to the mean remains non negligible for any finite number of dimensions , which implies that some of the points still end up being closer to the data mean than other points. It is well known that points closer to the mean tend to be closer (on average) to all other points, for any observed dimensionality. In high-dimensional data, this tendency is amplified. Such points will have a higher probability of being included in knearest-neighbor lists of other points in the data set, which increases their influence, and they emerge as neighbour-hubs. It was established that hubs also exist in clustered (multimodal) data, tending to be situated in the proximity of cluster centers . In addition, the degree of hubness does not depend on the embedding dimensionality, but rather on the intrinsic data dimensionality, which is viewed as the minimal number of variables needed to account for all pair wise distances in the data



(a) Minimal distances, $d = 30$ (b) Maximal distances, $d = 30$

F.1 Evolition of mimal and maximal distance from centroids to hubs and medoies on synthetic data for neighborhood size 10, and 10 clusters.



(c) Minimal distances, $d = 2$ (d) Maximal distances, $d = 2$

Data regardless of the distance or similarity measure employed. Its existence was verified for Euclidean((l2)) and Manhattan (l1) distances, lp distances with p > 2, fractional distances (lp with rational p €(0,1) Bray-Curtis, normalized Euclidean, and Canberra distances, cosine similarity, and the dynamic time warping distance for time series. In this paper, unless Otherwise stated, we will assume the Euclidean distance. The methods we propose in Section 4, however, depend mostly on neighborhood relations that are derived from the distance matrix and are, therefore, independent of the particular choice of distance measure.

Before continuing, we should clearly define what constitutes a hub. Similarly to we will say that hubs are points x having Nk(x) more than two standard deviations higher than the expected value k (in other words, significantly above average). However, in most experiments that follow, we will only concern ourselves with one major hub in each cluster, i.e., the point with the highest hubness score.

3.2 Relation of Hubs to Data Clusters

There has been previous work on how well high-hubness elements cluster, as well as the general impact of hubness on clustering algorithms. A correlation between low hubness elements (i.e., antihubs or orphans) and outliers were also observed. A lowhubness score indicates that a point is on average far from the rest of the points and hence probably an outlier. In high dimensional spaces, however, low-hubness elements are expected to occur by the very
Nature of these spaces and data distributions. These data points will lead to an average increase in intra cluster distance. It was also shown for several clustering algorithms that hubs do not cluster well compared to the rest of the points. This is due to the fact that some hubs are actually close to points in different clusters. Hence, they lead to a

International Journal of Computer Science
ISSN:2348-6600
Oddity...Probe...Reviste...
Published by SK Research Group of Companies(SKRGC)
http://www.ijcsjournal.com Volume 3, Issue 1, No 5, 2015. ISSN: 2348-6600
Reference ID: IJCS-094 PAGE NO: 535-541.

decrease in inter cluster distance. This has been observed on real data sets clustered using state-of-the-art prototype based methods, and was identified as a possible area for performance improvement. We will revisit this point in Section 5.4.

It was already mentioned that point's closer to cluster means tend to have higher hubness scores than other points. A natural question which arises is: Are hubs medoids? When observing the problem from the perspective of partitioning clustering approaches, of which K-means is the most commonly used representative, a similar question might also be posed: Are hubs the closest points to data centroid in clustering iterations? To answer this question, we ran K-means++ multiple times on several randomly generated 10,000-point Gaussian mixtures for various fixed numbers of dimensions (2, 5, 10, 20, 30, 50, 100), observing the high-dimensional case. We measured in each iteration the distance from current cluster centroid to the medoid and to the strongest hub, and scaled by the average intracluster distance. This was measured for every cluster in all the iterations, and for each iteration the minimal and maximal distance from any of the centroids to the corresponding hub and medoid was computed. Fig. 1 gives example plots of how these ratios evolve through iterations for the case of 10-cluster data, using neighborhood size 10, with 30 dimensions for the high dimensional case, and two dimensions to illustrate low dimensional Behavior. The Gaussian mixtures were generated randomly by drawing the centers from a ½lbound; unbounded uniform distribution (as well as covariance matrices, with somewhat tighter bounds). In the low-dimensional case, hubs in the clusters are far away from the centroid, even farther than average points. There is no correlation between cluster means and frequent neighbors in the low dimensional context. This changes with the increase in dimensionality, as we observe that the minimal distance from centroid to hub converges to minimal distance from centroid to medoid. This implies that some medoids are in fact cluster hubs. Maximal distances to hubs and medoids, however, do not match. There exist hubs which are not medoids, and vice versa. Also, we observe that maximal distance to hubs drops with iterations, suggesting that as the iterations progress, centroid are becoming closer and to data hubs. This already hints at a possibility of developing an iterative approximation procedure. To complement the above observations and explore the interaction between hubs, medoids, and the classic notion of Density and illustrate the different relationships they exhibit in low- and high dimensional settings, we performed Additional simulations. For a given number of dimensions (5 or 100), we generated a random Gaussian distribution centered around zero and started drawing random points from the distribution one by one, adding them sequentially to a synthetic data set. As the points were being added, hubness, densities, distance contrast, and all the other examined quantities and correlations between them (most of which are shown in Figs. 2 and 3) were calculated on the fly for all the neighborhood sizes within the specified range f1; 2; . . . ; 20g. The data sets started with 25 points initially and were grown to a size of 5,000. The entire process was repeated 20 times, thus in the end we considered 20 synthetic five dimensional Gaussian distributions and 20 synthetic 100-dimensional Gaussian distributions. Figs. 2 and 3 display averages taken over all the runs.1 we report results with Euclidean distance, observing similar trends with Manhattan and l0:5 distances. Fig. 2 illustrates the interaction between norm, hubness, and density (as the measurement, not the absolute term) in the simulated setting. From the definition of the setting, the norm of a point can be viewed as an "oracle" that expresses exactly the position of the point with respect to the cluster center.2 as can be seen in Fig. 2a, strong Pearson correlation between the density measurement and norm indicates that in low dimensions density pinpoints the location of the cluster center with great accuracy. In high dimensions, however.

4 EXPERIMENTS AND EVALUATION

We tested our approach on various high dimensional synthetic and real-world data sets. We will use the following abbreviations in the forthcoming discussion: K Means (KM), kernel K-means (ker-KM), Global K-Hubs (GKH), Local K-Hubs (LKH), Global Hubness-Proportional Clustering (GHPC) and Local Hubness Proportional Clustering (LHPC), Hubness Proportional K-Means (HPKM), local and global referring to the type of hubness score that was used (see Section 4). For all centroid-based algorithms, including KM, we used the D2 (K-means++) initialization procedure [12].4 The neighborhood size of k ¼ 10 were used by default in our experiments involving synthetic data and we have experimented with different neighborhood size in different real-world tests. There is no known way of selecting the best k for finding neighbor sets, the problem being domain-specific. To check how the choice of k reflects on hubness based clustering, we Ran a series of tests on a fixed 50-

International Journal of Computer Science
ISSN:2348-6600
Oddity...Probe...Reviste...
Published by SK Research Group of Companies(SKRGC)
http://www.ijcsjournal.com Volume 3, Issue 1, No 5, 2015. ISSN: 2348-6600
Reference ID: IJCS-094 PAGE NO: 535-541.

dimensional 10- distribution Gaussian mixture for a range of k values, k 2 f1; 2; . . . ; 20g. The results are summarized in Fig. 6. It is clear that, at least in such simple data, the hubness-based GHPC algorithm is not overly sensitive on the choice of k. In the following sections, K means++ will be used as the main baseline for comparisons, since it is suitable for determining the feasibility of using hubness to estimate local centrality of points. Additionally, we will also compare the proposed algorithms to kernel K-means [13] and one standard density-based method, GDBScan [37]. Kernel K-means was used with the nonparametric histogram intersection kernel, as it is believed to be good for image clustering and most of our real-world data tests were done on various sorts of image data. Kernel methods are naturally much more powerful, since they can handle non hyper spherical clusters. Yet, the hubness-based methods could just as easily be "kernel zed," pretty much the same way it was done for K-means. This idea requires further tests and is beyond the scope of this paper. For evaluation, we used repeated random sub sampling, training the models on 70 percent of the data and testing them on the remaining 30 percent. This was done to reduce the potential impact of over fitting, even though it is not a major issue in clustering, as clustering is mostly used for pattern detection and not prediction. On the other hand, we would like to be able to use the clustering methods not only for detecting groups in a given sample, but rather for detecting the underlying structure of the data distribution in general.

4.1 Synthetic Data: Gaussian Mixtures

In the first batch of experiments, we wanted to compare the value of global versus local hubness scores. These initial tests were run on synthetic data and do not include HPKM, as the hybrid approach was introduced later for tackling problems on real-world data. For comparing the resulting clustering quality, we used mainly the silhouette index as an unsupervised measure of configuration validity, and average cluster entropy as a Supervised measure of clustering homogeneity. Since most of the generated data sets are "solvable," i.e., consist of non-overlapping Gaussian distributions, we also report the normalized frequency with which the algorithms were able to find these perfect configurations. We ran two lines of experiments, one using five Gaussian generators, the other using 10. For each of these, we generated data of 10 different high dimensionalities: 10, 20, . . . , 100. In each case, 10 different Gaussian mixtures were generated, resulting in 200 different generic sets, 100 of them containing five data clusters, the others containing 10. On each of the data sets, KM++ and all of the hubbased algorithms were executed 30 times and the averages of performance measures were computed. The generated Gaussian distributions were hyper spherical bound ¼ 20 and the standard deviations were also uniformly taken from ½l_ that are significantly better than others, in the sense of having no overlap of surrounding one standard deviation intervals.) Global hubness is definitely to be preferred, especially in the presence of more clusters, which further restrict neighbor sets in the case of local hubness scores. Probabilistic approaches significantly outperform the deterministic ones, even though GKH and LKH also sometimes converge to the best configurations, but much less frequently. More importantly, the best overall algorithm in these tests was GHPC, which outperformed KM++ on all basis, having lower average entropy, a higher silhouette index, and a much higher frequency of finding the perfect configuration. This suggests that GHPC is a good option for clustering high-dimensional Gaussian mixtures. Regarding the number of dimensions when the actual improvements begin to show, in our lower dimensional test runs, GHPC was better already on 6dimensional mixtures. Since we concluded that using global hubness leads to better results, we only consider GKH and GHPC in the rest of the experiments.

4.2 Clustering and High Noise Levels

Real-world data often contain noisy or erroneous values due to the nature of the data-collecting process. It can be assumed that hub-based algorithms will be more robust with respect to noise, since hubness proportional search is driven mostly by the highest-hubness elements, not the outliers. In the case of KM++, all instances from the current cluster directly determine the location of the centroid in the next iteration. When the noise level is low, some sort of Outlier removal technique may be applied. In setups involving high levels of noise, this may not be the case. To test this hypothesis, we generated two data sets of 10,000 instances as a mixture of 20 clearly separated Gaussians, farther away from each other than in the previously described experiments. The first data set was 10 dimensional and the second 50 dimensional. In both cases, individual distribution centers were drawn independently from the uniform ½lm bound; um bounded distribution, Cluster sizes were imbalanced. Without noise, both of these data sets represented quite easy clustering

International Journal of Computer Science
ISSN:2348-6600
Oddity...Probe...Reviste...
Published by SK Research Group of Companies(SKRGC)
http://www.ijcsjournal.com Volume 3, Issue 1, No 5, 2015. ISSN: 2348-6600
Reference ID: IJCS-094 PAGE NO: 535-541.

problems, all of the algorithms being able to solve them very effectively. This is, regardless, a more challenging task than we had previously addressed [38], by virtue of having a larger number of clusters. To this data we incrementally added noise, 250 instances at a time, drawn from a uniform distribution on hypercube data points. The hypercube was much larger than the space containing the rest of the points. In other words, clusters Were immersed in uniform noise. The highest level of noise for which we tested was the case when there were an equal number of actual data instances in original clusters.

4.3 Experiments on Real-World Data

Real-world data are usually much more complex and difficult to cluster; therefore such tests are of a higher practical significance. As not all data exhibit hubness, we tested the algorithms both on intrinsically high-dimensional, high-hubness data and intrinsically low-to-medium dimensional, low-hubness data. There were two different experimental setups. In the first setup, a single data set was clustered for many different K-s (number of clusters), to see if there is any difference when the number of clusters is varied. In the second setup, 20 different data sets were all clustered by the number of classes in the data (the number of different labels). The clustering quality in these experiments was measured by two quality indices, the silhouette index and the isolation index which measures a percentage of k-neighbor points that are clustered together. In the first experimental setup, the two-part Miss- was used for evaluation. Each part consists of 6,480 instances having 16 dimensions. Results were compared for various predefined numbers of clusters in algorithm calls. Each algorithm was tested 50 times for each number of clusters. Neighborhood size was 5. The results for both parts of the data set are given in. GHPC clearly outperformed KM and other hubness-based methods. This shows that hubs can serve as good cluster center prototypes. On the other hand, hyper spherical methods have their limits and kernel K means achieved the best overall cluster quality on this data set. Only one quality estimate is given for GDBScan, as it automatically determines the number of clusters on its own. As mostly low-tomediumhubness data (with the exception of spam base), we have taken several UCI data Sets Values of all the individual features in the data sets were normalized prior to testing. The data sets were mostly simple, composed only of a few clusters. The value of k was set to 20. The results are shown in the first parts of In the absence of hubness, 6 purelyhubness-based.

5 CONCLUSIONS AND FUTUREWORK

Using hubness for data clustering has not previously been attempted. We have shown that using hubs to approximate local data centers is not only a feasible option, but also frequently leads to improvement over the centroid-based approach. The proposed GHPKM method had proven to be more robust than the K-Means++ baseline on both synthetic and real-world data, as well as in the presence of high levels of artificially introduced noise. This initial evaluation suggests that using hubs both as cluster prototypes and points guiding the centroid based search is a promising new idea in clustering high-dimensional and noisy data. Also, global hubness estimates are generally to be preferred with respect to the local ones. Hub-based algorithms are designed specifically for high dimensional data. This is an unusual property, since the performance of most standard clustering algorithms deteriorates with an increase of dimensionality. Hubness, on the other hand, [1]K. Kialing, H.-P. Krieger, P. Kroger, and S. Wanka (2003), "Ranking Interesting Subspaces for Clustering High Dimensional Data,"Proc. 7th European Conf. Principles and PKDD is a property of intrinsically high-dimensional data, and this is precisely where GHPKM and GHPC excel, and are expected to offer improvement by providing higher Inter cluster distance, i.e., better cluster separation.

REFERENCES

[1] Kailing, H.-P. Kriegel, and P. Kroger(2004**)."**Density-Connected Subspace Clustering for High-Dimensional Data,"Proc. Fourth SIAM Int'l Conf. Data Mining (SDM), pp. 246-257

[2] NenadTomasave,Milo'sRadovanovic,DunjaMladenic,and dMirjana Ivanovic(2014),"The Role of hubness in clustering high dimensional data " IEEE Transactions on knowledge and data engineering VOL. 26.

[3] C. Aggarwaland P.S. Yu (2000), "Finding Generalized Projected Clusters in High Dimensional Spaces," Proc. 26th ACM SIGMOD Int'l Conf. Management of Data

Published by SK Research Group of Companies (SKRGC).

International Journal of Computer Science
ISSN:2348-6600
Oddity...Probe...Reviste...
Published by SK Research Group of Companies(SKRGC)
http://www.ijcsjournal.com Volume 3, Issue 1, No 5, 2015. ISSN: 2348-6600
Reference ID: IJCS-094 PAGE NO: 535-541.

[4] I.S. Dhillon, Y. Guan, and B. Kulis,(2004) "Kernel k-Means: Spectral Clustering and Normalized Cuts," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 551-556

[5] Kaba´n, "Non-Parametric Detection of Meaningless Distances in High Dimensional Data," Statistics and Computing, vol. 22, no. 2, pp. 375-385, 2012

[6] Han and M. Kamber, Data Mining: Concepts and Techniques, second ed. Morgan Kaufmann, 2006. R.J. Durrant and A. Kaba´n, "When Is 'Nearest Neighbour' Meaningful: A Converse Theorem and Implications," J. Complexity, vol. 25, no. 4, pp. 385-397.

[7] E. Agirre, D. Martı´nez, O.L. de Lacalle, and A. Soroa, "Two Graph-Based Algorithms for State-of-the-Art WSD," Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 585-593, 2006

[8] Ning, H. Ng, S. Srihari, H. Leong, and A. Nesvizhskii, "Examination of the Relationship between Essential Genes in PPI Network and Hub Proteins in Reverse Nearest Neighbor Topology," BMC Bioinformatics, vol. 11, pp. 1-14, 2010.