Finding representative objects using link analysis ranking

Panagiotis Papapetrou
Department of Information and Computer Science
Aalto University, Finland

Tatiana Chistiakova
Department of Information and Computer Science
Aalto University, Finland

Jaakko Hollmén
Department of Information and Computer Science
Aalto University, Finland

Vana Kalogeraki
Department of Informatics
Athens University of Economics and Business
Athens, Greece

Dimitrios Gunopulos
Department of Informatics and Telecommunications
University of Athens
Athens, Greece

ABSTRACT

Link analysis ranking methods are widely used for summarizing the connectivity structure of large networks. We explore a weighted version of two common link analysis ranking algorithms, PageRank and HITS, and study their applicability to assistive environment data. Based on these methods, we propose a novel approach for identifying representative objects in large datasets, given their similarity matrix. The novelty of our approach is that it takes into account both the pair-wise similarities between the objects, as well as the origin and "evolution path" of these similarities within the dataset. The key step of our method is to define a complete graph, where each object is represented by a node and each edge in the graph is given a weight equal to the pairwise similarity value of the two adjacent nodes. Nodes with high ranking scores correspond to representative objects. Our experimental evaluation was performed on three data domains: american sign language, sensor data, and medical data.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, experimentation

Keywords

Social networks, link analysis ranking, network analysis, american sign language.

1. INTRODUCTION

The basic tenet of network analysis is to extract certain types of information from the connections between the objects in the network. Several ranking algorithms have been developed to analyze sets of hyperlinked documents, such as web pages in the Internet [5, 2]. In this paper, we explore the applicability of these algorithms to identifying and ranking the most "representative" objects in a dataset. We explore a weighted version of two common ranking algorithms: PageRank [2] and HITS [5]. These methods are highly applicable to assistive environment data.

We propose a novel approach for identifying *representative* objects in a large dataset, given their distance or similarity matrix. Our main goal is not to detect "central" objects (as, for example, in k-means clustering) but objects that share widely spread patterns. For example, let P be a very common pattern in our dataset. Also, assume that P' is a slight variant of P and P'' a slight variant of P'. In other words, P has "evolved" to P' and P''. These three patterns may contained in several different objects in the dataset. Standard clustering methods may fail to capture this pattern "evolution" within the dataset. Hence, what we need is an approach that will take into account both the pair-wise similarities between the objects in the dataset, as well as the origin and "evolution path" of these similarities.

The key idea of our method is to use the objects and their similarity matrix to define a complete graph. Each object is represented by a node in the graph and an edge exists between each pair of nodes. Each edge is assigned with a weight that corresponds to the similarity of the two nodes that are connected by the edge. Our main task is to identify the top-K most representative objects in the dataset, ranked based on their importance. The degree of importance of an object is determined by the "strength" of its connectivity in the network. Objects are identified as "strong" if they have strong connections (i.e., high similarity) to other "strong" objects or if they participate in many "strong" paths (i.e., paths with edges of high similarity weights) in the graph.

Our experimental evaluation was performed on three real data domains related to assistive environments: american sign language, sensor data, and medical data. Our method produced a ranking of the objects in each dataset taking into account the similarity "paths" in the graph. Objects with high ranking correspond to strong representatives. These representatives may hold patterns that are highly common in the data; hence they may be used for several data mining tasks, such as clustering or pattern mining.

PETRA'12 June 6 - 8, 2012, Crete Island, Greece.

The main contributions of this paper include: (1) the presentation and study of the applicability of weighted versions of PageRank and HITS, (2) a novel approach based on the two ranking methods for identifying and ranking representative objects given their similarity martix, and (3) an experimental evaluation of the proposed methods on three real datasets.

2. RELATED WORK

Many methods have been developed in social-network analysis to assess the "authoritativeness" or "importance" of individuals in implicitly- or explicitly-defined social networks. The network is represented as a directed graph, and the concept of *in-degree* (the number of incoming edges at a node), or refinements [4], is the simplest measure of importance of a node. Other notions of importance in social networks include degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality [9].

A prominent application domain of importance measures is in the area of web ranking. The two most well-known techniques are PageRank [2] and HITS [5]. Many variants of those methods have been proposed, as well as adaptations of those basic methods for different objectives.

Selecting important or, most commonly called "reference" sequences from a large database has been studied widely in the fields of machine learning and databases. A simple approach [14] is to select objects that cover most of the domain space and have distances to other objects with high variance. Another technique for reference object selection is used by Boostmap [1] — a method for approximate similarity ranking in metric and non-metric spaces. The object selection process is based on Adaboost.

A very related problem is that of spectral clustering, where the spectrum of the similarity matrix of the data is used to perform dimensionality reduction. One common method used in image segmentation is the *normalized cuts algorithm* [13], where points are partitioned into two sets based on the eigenvector corresponding to the second smallest eigenvalue of the Laplacian matrix of the data. Spectral clustering may capture representative objects or dimensions of the data based on a similarity matrix. Our approach is fundamentally different in the sense that we exploit the similarity matrix of a set of objects to identify representatives as well as a ranking of their importance.

3. METHODS

Let $G = (V, E)$ be a graph, where $V = \{v_1, \ldots, v_{|V|}\}$ is the set of vertices and $E = \{e_1, \ldots, e_{|E|}\}$ the set of edges in G. Also, let $W = \{w_{ij}\}$ be the set of edge weights, were w_{ij} corresponds to the weight of the edge connecting nodes v_i and v_j.

3.1 Weighted Hubs and Authorities

The hubs-and-authorities [5] algorithm, also known as HITS, is a link analysis ranking algorithm precursor to PageRank. The intuition behind this algorithm is based on the way a social graph is formed. That is, some nodes act as information hubs, in the sense of links to highly authoritative nodes. Hence, they are not authoritative themselves but they point to many authoritative nodes. At the same time, a node is considered authoritative if it is pointed to by many "strong" hubs.

Each node is ranked with respect to its "importance" as a *hub* or *authority*. A strong hub is a node that is connected to many strong authorities; a strong authority is a node that is connected to many strong hubs.

Given G, each node v_i is associated with a hub weight h_i and an authority weight a_i. The magnitude of the weight corresponds to the strength or importance of each node. During the application of the iterative algorithm, the sum of the squared hub and authority weights is enforced to be invariant and equal to unity. This is achieved by successive normalizations of the weights.

At the intialization step, each node is assigned with hub and authority weights of $\frac{1}{\sqrt{(|V|)}}$. Then at each iteration, authority and hub weights are updated as a function of the hub and authority weights, respectively, or

$$a_j \leftarrow \sum_{i=1}^{|V|} w_{ij} h_i \text{ and } h_i \leftarrow \sum_{j=1}^{|V|} w_{ij} a_j$$

Authority weights are thus a weighted sum of the ingoing hub weights, and correspondingly, hub weights are maintained as a weighted sum of the authority weights of the pages linked to by a hub. To guarantee invariance of the weights at each iteration it should hold that

$$\sum_{i}^{|V|} a_i^2 = 1 \text{ and } \sum_{i}^{|V|} h_i^2 = 1$$

Hence, after each iteration the authority values are divided by the square root of the sum of the squares of all authority values, and the hub values are divided by the square root of the sum of the squares of all hub values.

3.2 Weighted PageRank

PageRank [2, 10] has been developed for link analysis and ranking, and expresses a probability distribution that represents the likelihood that a person who is randomly following edges in a graph will arrive at any particular node (random surfer model). For initialization, it is assumed that the distribution is evenly divided among all nodes in the graph. PageRank is an iterative algorithm and its computation requires several passes through the whole graph to ensure that the approximate PageRank values more closely reflect the theoretical true value.

Let PR be the vector of PageRank values. Using the recursive definition of PageRank [2, 10], PR is computed as follows:

$$PR(v_i) = \frac{1-d}{|V|} + d \sum_{\forall v_j \rightarrow v_i} \frac{PR(v_j)}{|F(v_j)|}, \tag{1}$$

where $F(v_j)$ corresponds to the out-degree of node v_j.

Note that in order to avoid getting trapped to loops of sink states (i.e., nodes that have no outgoing edges) PageRank assumes that nodes with no outbound edges should link to all other nodes in V. For this purpose the regulating residual probability $1-d$ is used, also called *damping factor*. Assigning an appropriate value to $1-d$ depends on the graph we are studying. In the case of WWW, a random user surfing the web will typically follow the order of 6 hyperlinks after becoming bored and choosing some other random web-page to surf. Thus in this case, $1 - d = 1/6 \approx 0.15$.

In our problem setting, however, we assume a complete graph with edge weights. We adjust Equation 1 to include the weight vector W as follows:

$$PR(v_i) = \frac{1-d}{|V|} + d \times \sum_{\forall v_j \to v_i} \frac{w_{ji} \times PR(v_j)}{|F(v_j)|}. \qquad (2)$$

3.3 Finding and Ranking Important Objects

Let $O = \{o_1, \ldots, o_N\}$ denote a set of N objects. No additional information is needed regarding the underlying space. Also, let $D = \{d_{ij}\}$ be the $N \times N$ similarity matrix of the objects in O. Our task is to find the ranked set of the top-K most important objects in O.

The first step of our method is to create a complete graph $G = (V, E)$, where $V \equiv O$ and $E = \{e_1, \ldots, e_{|E|}\}$ the set of edges in G defined for each and every pair of nodes in the G. The set of edge weights W is defined as follows $w_{ij} = d_{ij}$.

Next, we run HITS and Pagerank on G. The ranking scores produced by the algorithms are then aggregated using the *sum* as the score aggregation function. To optimize the computation our method employs Fagin's algorithm for rank aggregation [3]. Finally, the top-K objects are reported as the most important ones in the collection and their ranking is identified by Fagin's algorithm.

4. EXPERIMENTS

4.1 Setup

We have benchmarked three datasets:

- **ASL.** The dataset contains video transciptions of American Sign Language (ASL) expressions [11]. A sequence is a set of events that occur over a time interval. An event corresponds to a grammatical, syntactic, or gestural field in an ASL expression.
- **Hepatitis.** The dataset contains information about patients who have Hepatitis B or C. Events represent the results of 25 regular tests [12]. A sequence corresponds to a set of tests taken by a patient.
- **Pioneer.** This dataset was constructed using the Pioneer-1 dataset available in the UCI repository. Each sequence contains events performed by a robot. Three events labels are included: *gripper*, *move*, and *turn* [8].

Our datasets consist of sequences of events that occur over time intervals. In Table 1 we can see a summary of the main statistics of each dataset. To assess the similarity between such sequences, we used a recently proposed distance measure, `Artemis` [6, 7]. The distance matrix D was then converted to a similarity matrix D_{sim} as follows:

$$D_{sim} = 1 - D/max(D) \qquad (3)$$

Dataset	number of	sequence size		
	sequences	min.	max.	average
ASL	873	4	41	18
Hepatitis	498	15	592	108
Pioneer	160	36	89	56

Table 1: Dataset Statistics.

4.2 Evaluation

We compared the performance of HITS, Pagerank, and MaxVariance [14]. The results for `ASL`, `Hepatitis`, and `Pioneer` are shown in Figures 1, 2, and 3, respectively. In the figures, we show the histograms of each of the ranking scores produced by each method as well as the scatterplots showing the correation (if any) between the methods.

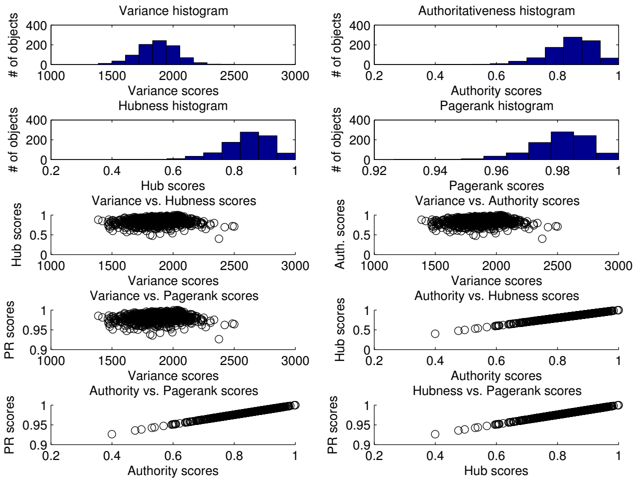

Figure 1: Comparison of the ranking scores produced by HITS, PageRank, and MaxVariance for the ASL dataset. Scores are normalized in [0,1].

Based on our findings, several observations can be made. Firstly, there is very low correlation between the MaxVariance method and the other ranking methods. In addition, there is a high correlation between the hubness and authoritativeness scores as well as the pagerank scores.

Next, we performed a qualitative analysis of the methods. As described in Section 3 we aggregated the rankings produced by HITS and Pagerank and, for each dataset, we identified the objects (sequences) that appear in the top-5% of the ranking. For the `ASL` dataset, the 43 representative sequences that were detected included subpatterns of events that describe and characterize distinctive ASL gestures, such as "wh-questions", "negations", and "yes-no questions". As regards the `Pioneer` dataset, the 8 sequences that were identified describe the most representative movements of the robot in the dataset and cover all three types of movements ("gripping", "moving", and "turning"). The most representative sequence (with a significant ranking difference from the rest of the sequences) included the following pattern:

$$\text{"}moving\ forward\text{"} \to \text{"}gripping\text{"} \to \text{"}moving\ backwrds\text{"}$$

which is a strongly present pattern in the dataset as it describes a typical task performed by the robot which is to move forward, find an object and take it, and then take some steps backwards.

Finally, 24 sequences were extracted from the `Hepatitis` dataset. They included the most representative test patterns (i.e., sequences of tests) taken by the patients.

The performance of MaxVariance was very poor as in many cases there was a high repetition of sequences with the same pattern. In order to capture the same dominant objects as the ranking methods, the top-K% threshold for MaxVariance had to be increased to at least over 30%.

5. DISCUSSION

The methods used for the analysis of networks may be applied in other settings, such as assistive environments, if

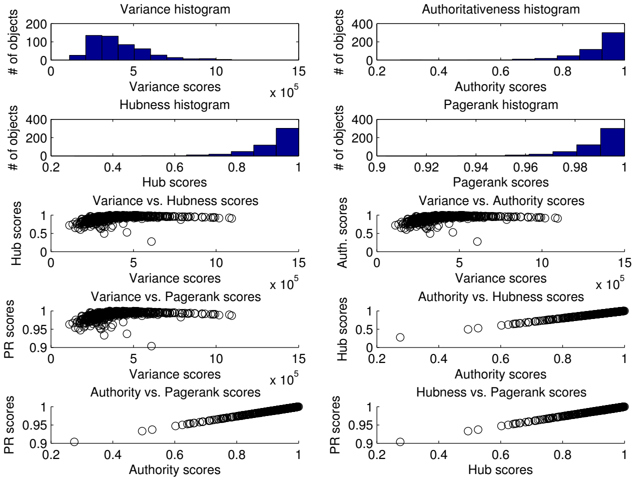

Figure 2: Comparison of the ranking scores produced by HITS and PageRank, and MaxVariance for the Hepatitis dataset. Scores are normalized in [0,1].

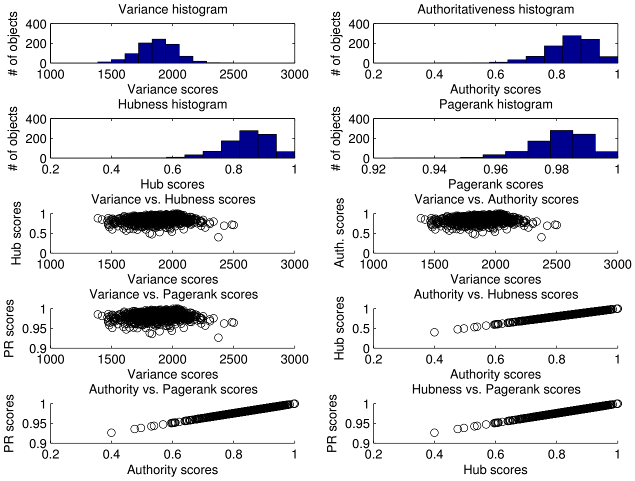

Figure 3: Comparison of the ranking scores produced by HITS and PageRank, and MaxVariance for the Pioneer dataset. Scores are normalized in [0,1].

the domain is reperesented suitably as a network. We transformed illustrative problems from domain of assistive environments to a network representation and analyzed them with variations of two commonly known algorithms for network analysis, namely HITS and PageRank.Our initial experimental findings show that our approach is promising as we manage to identify objects with the most representative patterns in three data domains related to assistive environments. Directions for future work include a more general formulation of the problem, a deeper qualitative analysis of out findings, as well as experimentation on other domains and larger networks.

6. ACKNOWLEDGMENTS

This work has been supported in part by the following grants and projects: the Academy of Finland Centre of Excellence grant 118653 (ALGODAN), the MODAP project, the European Union through the Marie-Curie RTD (IRG-231038) project, a Hellenic Republic Ministry of Education, Lifelong Learning and Religious Affairs Thalis DISFER project, and by AUEB through a PEVE project.

7. REFERENCES

[1] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. Boostmap: a method for efficient approximate similarity rankings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 268–275, 2004.

[2] S. Brin and L. Page. The anatomy of large-scale hypertextual web search engine. *Computer Networks and ISDN Systems (CNIS)*, 30:107–117, 1998.

[3] R. Fagin, A. Lotem, and M. Naor. Optimal aggregation algorithms for middleware. In *Proceedings of the ACM SIGMOD Symposium on Principles of Database Systems (SIGMOD-SIGACT-SIGART)*, PODS '01, pages 102–113. ACM, 2001.

[4] C. H. Hubbell. An input-output approach to clique identification. *Sociometry*, 28(4):377–399, 1965.

[5] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, September 1999.

[6] O. Kostakis, P. Papapetrou, and J. Hollmén. Artemis: Assessing the similarity of event-interval sequences. In *Proceedings of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECML/PKDD)*, pages 229–244, 2011.

[7] O. Kostakis, P. Papapetrou, and J. Hollmén. Distance measure for querying arrangements of temporal intervals. In *Proceedings of Pervasive Technologies Related to Assistive Environments (PETRA)*, 2011.

[8] F. Mörchen and D. Fradkin. Robust mining of time intervals with semi-interval partial order patterns. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, pages 315–326, 2010.

[9] M. E. J. Newman. The mathematics of networks. *The New Palgrave Encyclopedia of Economics*, 2007.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[11] P. Papapetrou, G. Kollios, S. Sclaroff, and D. Gunopulos. Mining frequent arrangements of temporal intervals. *Knowledge and Information Systems (KAIS)*, 21:133–171, 2009.

[12] D. Patel, W. Hsu, and M. Lee. Mining relationships among interval-based events for classification. In *Proceedings of ACM Special Interest Group on Management of Data (SIGMOD)*, pages 393–404, 2008.

[13] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence(PAMI)*, 22:888–905, August 2000.

[14] J. Venkateswaran, D. Lachwani, T. Kahveci, and C. Jermaine. Reference-based indexing of sequence databases. In *Proceedings of the International Conference on Very Large Databases (VLDB)*, pages 906–917, 2006.