PageRank Based Clustering of Hypertext Document Collections

Konstantin Avrachenkov
INRIA Sophia Antipolis
kavratch@sophia.inria.fr

Vladimir Dobrynin
St.Petersburg State University
v.dobrynin@bk.ru

Danil Nemirovsky
INRIA, St.Petersburg State Uni
dnemirov@sophia.inria.fr

Son Kim Pham
UCSD
sonsecure@yahoo.com.sg

Elena Smirnova
St.Petersburg State University
redcap@inbox.ru

ABSTRACT

Clustering hypertext document collection is an important task in Information Retrieval. Most clustering methods are based on document content and do not take into account the hyper-text links. Here we propose a novel PageRank based clustering (PRC) algorithm which uses the hypertext structure. The PRC algorithm produces graph partitioning with high modularity and coverage. The comparison of the PRC algorithm with two content based clustering algorithms shows that there is a good match between PRC clustering and content based clustering.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Miscellaneous

General Terms

Algorithms, Experiments

Keywords

PageRank based Clustering, Directed Graphs

1. INTRODUCTION

Clustering hypertext document collection is an important task in Information Retrieval. Some examples of efficient clustering methods are K-means clustering, Information Bottleneck based clustering and Contextual Document Clustering (CDC) [3]. These clustering methods are based on the content of the documents and do not take into account the hyperlink structure of the document collection. In the present work we suggest clustering hypertext document collection using the graph formed by hyperlinks. We propose a novel graph clustering algorithm and compare the graph based clustering with the content based clustering. Our algorithm essentially uses PageRank popularity measure [10] and therefore we call our algorithm PageRank based Clustering (PRC algorithm).

We note that our method clusters directed graphs. The clustering of directed graphs is a young research area. To the best of our knowledge, there are only very few recent works [1, 7, 8]. We mention that the MCL method [4] can be used for clustering directed graphs as well as undirected ones. As is shown later in the paper, the proposed PRC algorithm surpasses the above algorithms in terms of efficiency and centroid detection.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. PRC ALGORITHM

The PRC algorithm is composed of two parts. In the first part we determine the core nodes or centroids of clusters and in the second part we assign nodes to clusters. For the first part, we use the general idea from [5] that the core nodes of a Web community have high authority and hub scores. We determine the core nodes in two steps. At the first step we determine a list of candidate nodes for centroids. This can be done by (a) sorting nodes in decreasing order of PageRank; (b) sorting nodes in decreasing order of PageRank – Reverse PageRank product; (c) sorting nodes in decreasing order of HITS ranking; (d) sorting nodes in decreasing order of their degree. The option (a) seems to be preferable as the options (b) and (c) are less robust to perturbations and more computationally demanding. The option (d) is prone to spamming.

Then, at the second step we choose the centroids from the candidate nodes. We should decide if two candidate nodes belong to the same cluster or not. If two candidate nodes belong to the same cluster we discard the one with the worst ranking. One can decide if two candidate nodes belong to the same cluster using a threshold on (a) the number of one and two-step directed paths; (b) the expected meeting distance; (c) the inverse P-distance; (d) the JS divergence. We suggest to use the options (a) and (d) as the other two options are more computationally demanding.

Once a list of centroids is formed, we can proceed to the node assignment. We suggest to perform the node assignment using Personalized (or Local) PageRank vector [6]. It is defined as the stationary distribution of the random walk governed by the following transition matrix $B_s = c_s P + (1 - c_s) K_s$, where P is the hyperlink matrix, $[K_s]_{ij} = \delta_{sj}$, and s is the centroid index. This is a random walk that follows an out-going link with probability c_s and returns to the centroid s with probability $1 - c_s$. We note that the return probability can be different for different clusters. The intuition is that the random walk should explore for longer time larger clusters and for a shorter time smaller ones. The choice of different return probabilities allows us to treat large and small clusters in a universal framework.

Define the stationary distribution of B_s by π_s. We assign the nodes to clusters by $Cluster(v) = \text{argmax}_{s \in K} \pi_s(v)$. The number of clusters could be either an input parameter or it could be found adaptively. For the adaptive approach we can use the maximum modularity criterion [9].

3. COMPARISON WITH RELATED WORK

Let us compare the PRC algorithm with related work. Our comparison is based on both theoretical and practical evaluations.

The MCL method [4] uses the product of matrices, which makes its complexity $O(n^2)$. This complexity is prohibitive for large hypertext document collections.

The original version of the method in [1] only can find one cut of a graph. We can modify the method of [1] as follows: we can find the centroids using the first part of our method and then for each centroid apply the method of [1]. However, we note that the method of [1] is slower than PRC because in [1] in addition one should use sorting procedures.

The BestWCut(WNCut) method of [8] includes the computation of $|K|$ eigenvectors corresponding to the $|K|$ smallest eigenvalues of the weighted Laplacian. As large matrices with clusters can have eigenvalues very close to each other, this operation might be numerically unstable. Furthermore, the BestWCut(WNCut) method has K-means clustering algorithm as its subroutine, which increases further the numerical complexity of the method.

4. EXPERIMENTAL EVALUATION

In this section we perform two sets of experiments. In the first set of experiments we compare PRC with MCL and University Domain based Expert Clustering (EC). In the second set of experiments we test PRC against two content based clustering methods: CDC and an expert based clustering. To measure proximity between two clustering we use the Variation of Information (VI) and Classification Error (CE) metrics [8], and to measure the quality of clustering we use modularity [9], performance and coverage [2]. We have used the PRC algorithm with the same parameter $c = 0.5$ for both PageRank and Personalized PageRank computations. Having calculated PageRank for all the pages in the graph we choose centroid pages as pages with largest PageRank excluding pages which have more than 30% of neighbours with other centroids.

In the first experiment set we used a Giant Strongly-Connected Component of the WebKB hyper-link graph [8]. As one can see from Tables 1 and 2, the PRC method outpermorms the MCL method by all metrics. We have also performed experiments with the BestWCut method but so far have obtained unstable numerical results.

Pairs of Clusterings	CE	VI
PRC – EC	0.008	0.1 (max: 2.77)
MCL – EC	0.72	3.345 (max: 10.85)
PRC – MCL		3.351 (max: 10.85)

Table 1: Pairwise clustering comparison for WebKB

In the second set of experiments we use a Web crawl of INRIA Sophia Antipolis Web site (205900 nodes and 2124140 links). Using the modularity criterion we have concluded that an appropriate number of clusters is 200. The modularity of the obtained PRC clustering is 0.935.

Clusterization measure	PRC	MCL
Modularity	0.99	0.66
Coverage	0.99	0.66
Performance	0.74	0.97

Table 2: Clusterization measures for WebKB

To compare the PRC algorithm with content based clustering, we have clustered 12300 documents from the Web site of INRIA Sophia Antipolis using CDC algorithm [3] and expert clustering. In expert clustering we have used the knowledge about the administrative structure of INRIA Sophia Antipolis and have classified the pages in the clusters according to the project-teams. The CDC algorithm is designed to produce clusters which reflect topics of a dataset. In the Table 3 we show the value of VI for different pairs of clustering.

Pairs of Clusterings	VI proximity metric
PRC – CDC	1.95 (max: 15)
PRC – EC	2.79 (max: 15)
CDC – EC	1.77 (max: 15)

Table 3: Pairwise clustering comparison for INRIA

First, we conclude that PRC matches well the content based clustering. It seems that the value of VI around 2 indicates a good match between two clusterings as the value of VI corresponding to the pair of two content based clusterings is 1.77. Of course, one can argue that EC is not a real content based clustering. Still, since EC reflects the classification of INRIA Sophia Antipolis research areas, we associate it with content based clustering.

5. REFERENCES

[1] R. Andersen, F. Chung and K. Lang, "Local partitioning for directed graphs using PageRank", WAW2007.

[2] U. Brandes, M. Gaertler and D. Wagner, "Experiments on graph clustering algorithms", ESA'2003.

[3] V. Dobrynin, D. Patterson and N. Rooney, "Contextual Document Clustering", LNCS v.2997, 2004.

[4] S.M. van Dongen, *Graph Clustering by Flow Simulation*, PhD Thesis, 2000.

[5] D. Gibson, J.M. Kleinberg and P. Raghavan, "Inferring Web Communities from Link Topology", The 9th ACM Conference on Hypertext and Hypermedia, 1998.

[6] T.H. Haveliwala, "Topic-sensitive PageRank", WAW 2002.

[7] J. Huang, T. Zhu and D. Schuurmans, "Web Communities Identification from Random Walks", LNCS v.4213, 2006.

[8] M. Meila and W. Pentney, "Clustering by weighted cuts in directed graphs", SDM 2007.

[9] M.E.J. Newman, "Modularity and community structure in networks", PNAS, v.103, no.23, 2006.

[10] L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank citation ranking: Bringing order to the Web", *Technical Report, Stanford*, 1998.