Effective Query Expansion with the Resistance Distance Based Term Similarity Metric

Shuguang Wang
Intelligent Systems Program
University of Pittsburgh
Pittsburgh, PA 15260
swang@cs.pitt.edu

Milos Hauskrecht
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
milos@cs.pitt.edu

ABSTRACT

In this paper, we define a new query expansion method that relies on term similarity metric derived from the electric resistance network. This proposed metric lets us measure the mutual relevancy in between terms and between their groups. This paper shows how to define this metric automatically from the document collection, and then apply it in query expansion for document retrieval tasks. The experiments show this method can be used to find good expansion terms of search queries and improve document retrieval performance on two TREC genomic track datasets.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

General Terms: Algorithm, Performance

Keywords: Information Retrieval, Query Expansion, Term Similarity

1. INTRODUCTION

A fundamental challenge of information retrieval (IR) is to find documents that are relevant to user queries. The search queries usually consists of only few terms, which barely describe the information that users request. A widely used approach to deal with this problem is to expand the original query with relevant terms [4, 5]. In this study, we tackle the query expansion problem by defining new term-similarity metric that is based on the electric resistant network. In particular, this metric is derived from the effective resistance distances in between pairs of vertices in an undirected weighted graph. In this graph, nodes represent terms and they are linked together based on their co-occurrences. The edge weights represent the strength of term co-occurrences and are interpreted as electric resistances. Based on the resistance distances between pairs of terms, we demonstrate how to derive the similarity between terms and groups of terms. In this paper, we will discuss how to build the metric from document collection and apply it in query expansion for document retrieval tasks. We then present some of the evaluation results on two TREC Genomic Track data. Finally we will conclude the paper and suggests some future work.

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

2. METHODOLOGY

Our objective is to define a metric in the term space that would reflect how likely the terms are to be associated (or co-occur) in the document. We define the metric with the help of a weighted graph representing direct associations among terms and their strength. More formally, our model consists of an undirected weighed graph $G = (V, E, w)$ where nodes V represent terms in the document, edges E represent pairwise association relations in between them, and weights w on the edges measure the strength of associations in between the connected pairs of nodes. In general, the association in between any two terms is calculated by considering all association paths and cumulative weights connecting them. This defines a metric on the term space.

Building an Association Graph

We propose to build the graph from the (training) corpus of documents by parsing each document and by extracting the pairwise associations among terms on the sentence level. If these two concepts co-occur in the same sentence, a direct link in between the concepts is included in the graph. Let j and k represent two distinct terms. If the two terms co-occur in $n > 0$ different documents, a link in between j and k with weight n is added to the graph (See Figure 1).

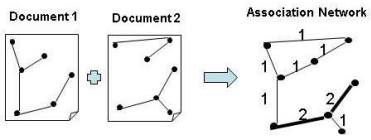

Figure 1: Building an association network from documents

Electric Resistance Network

To define the metric for any pairs of nodes (terms or concepts) in the association graph, we propose to interpret the weighted graph as a resistance network. Figure 2 illustrates the resistance network obtained from a weighted association network. In this case, the links and their weights in the graph are replaced with connections with resistances corresponding to their weights. More specifically, a weight $w_{j,k}$ in between nodes j, k in the original weighted graph defines the electric conductance $c_{j,k}$ of the connection that is the reciprocal of its electric resistance $r_{j,k} = \frac{1}{c_{j,k}} = \frac{1}{w_{j,k}}$. We can use the electric resistance network to calculate the effective resistance in between any two nodes in the network. This effective resistance is the basis of our distance (similarity)

metric. The metric is also referred as resistance distance and comes with an intuitive random walk interpretation [1].

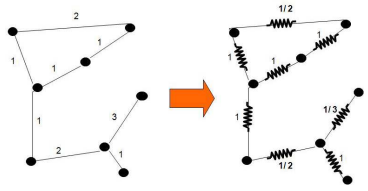

Figure 2: Building a resistance network from an association network

Calculating Effective Resistances

In general, the calculation of resistances (or conductances) in between any two nodes in an electric network is more complex and requires us to consider all serial and parallel path connections in between them. Also in order to define a proper metric we should define the distance for all possible pairs. We calculate the resistances with the help of graph Laplacian (L) [2], where $L = A - D$ and A is the adjacency matrix and D is the degree matrix of the graph. This approach is also used to defined the spectral transformation kernel function [6].

The effective resistance in between nodes v_j and v_k can be calculated as: $r_{j,k} = L^+_{j,j} + L^+_{k,k} - L^+_{j,k} - L^+_{k,j}$ where L^+ is the pseudo-inverse of the graph Laplacian. In general, the pseudo-inverse of a matrix A can be calculated from the singular value decomposition of $A = U\Sigma V^*$ as $A^+ = U^*\Sigma^+V$ where Σ^+ is the pseudo-inverse of Σ.

3. USING THE DISTANCE METRIC IN IR

The effective resistance calculations define a distance metric in between nodes (terms or concepts) that can in turn be used to support various inferences in the term space. We extend this metric to define the distance in between a set of (seed) terms S and a target t as the average of the distances between nodes in S and t: $r_{S,t} = \frac{1}{|S|}\sum_{s_i \in S} r_{s_i,t}$.

With the above metric, we can find all relevant terms to the original query terms. However, the metric may not differentiate well the relevant terms that are specific to the query from the rest of the relevant terms. To deal with it, we borrow the idea from TF-IDF[3] and re-normalize the distances between terms based on their relative distances:

$$r^n_{Q,x} = \frac{r_{Q,x}}{\frac{1}{|X-Q|}\sum_{y \in X-Q} r_{x,y}}, \quad (1)$$

where Q is a query, x is a term, and X is the set of all terms. $r_{Q,x}$ is the resistance distance computed using the proposed metric and it is normalized by the average distance between x and all other non-query terms in the graph.

4. EXPERIMENTS

We evaluate our method using TREC Genomic Track 2003 & 2004 datasets, which are consist of abstracts from Medline. Test queries in 03 data contains gene names, their associated products (e.g., proteins), and their symbols and synonyms. Test queries in 04 data are sentences and they cover more general topics and involve more genomic concepts. We define our metric over only important terms: gene/protein names for 03 data and 5000 terms with highest TF-IDF scores for 04 data. We use 30% of 03 data and 25% of 04 data to extract the association networks respectively. We choose Lemur/Indri and its internal Pseudo

Table 1: TREC genomic track data statistics

Year	#Abstracts	#Test Queries
2003	525,932	50
2004	4,591,008	50

Relevance Feedback (PRF) query expansion module as the baselines. We use the Mean Average Precision (MAP), to measure the retrieval performance of various methods. All query terms are connected by "#combine" and the weights of expanded terms are assigned according to distance measures as $w(x) = e^{-r_{Q,x}}$. x is a expanded term, Q is the set of original query terms, and $r_{Q,x}$ defines the resistance distance between them.

We report results of query expansion with two proposed metrics, $r_{Q,x}$ and its normalized version $r^n_{Q,x}$. We first combine our metrics with Lemur/Indir and compare them with two baselines (See Tables 2). We use 5 expanded terms in this experiment. Both proposed metrics perform much (about 20%) better than the original Indri. More importantly, our metrics are much (over 9%) better than the PRF expansion approach and the normalized metric is the best.

Table 2: MAP of various methods

Methods	03	04
Indri	0.243	0.216
Indri+PRF	0.258	0.228
Indri+$r_{Q,x}$	0.282	0.251
Indri+$r^n_{Q,x}$	0.291	0.261

5. CONCLUSION AND FUTURE WORK

We have presented a new term similarity metric that can be easily defined using the document collection and applied it successfully in query expansion for document retrieval tasks. To the best of our knowledge this is the first study that attempts to define the term similarity metric based on electric resistance networks. In our evaluation, we defined the similarity metrics on important concepts because the data is from genomic domain. We would extend our study to define the similarity on all terms and experiment it on general document retrieval tasks.

6. REFERENCES

[1] P. G. Doyle and J. L. Snell. *Random Walks and Electrical Networks*. The Mathematical Association of America, Washington DC, 1984.

[2] D. J. Klein and M. Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12:81–95, 1993.

[3] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 5:513–523, 1988.

[4] J. Xu and B. W. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th ACM SIGIR conference*, pages 4–11. ACM, 1996.

[5] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32th ACM SIGIR conference*, pages 59–66. ACM, 2009.

[6] X. Zhu, J. Kandola, J. Lafferty, and Z. Ghahramani. Graph kernels by spectral transforms. *Semi-Supervised Learning*, 2006.