Persistent Homology for Query Performance Prediction

Preliminary draft version

Brian Brost, Ingemar J. Cox & Christina Lioma
University of Copenhagen, Department of Computer Science
brian.brost@di.ku.dk

Abstract

Topological data analysis (TDA) is the application of methods from computational topology, such as persistent homology, to the analysis of data. Attractive theoretical properties of persistent homology and dramatically improved algorithms using tools such as discrete Morse theory have led to an increased interest in TDA in recent years.

This paper investigates the application of TDA to query performance prediction (QPP), i.e. predicting the performance of an information retrieval system without access to explicit relevance judgements. We investigate a novel model for post-retrieval QPP based on TDA.

One of the properties TDA can be used to analyse is connectedness. This is relevant to QPP since clustering tendencies have been shown to be a useful predictor of query performance. Computing 0-dimensional homology gives information about connectedness and can be regarded as a form of clustering. Furthermore, TDA has the potential to give much more general information about the global structure of data than that provided by clustering methods.

We develop a class of models for QPP based on analysing the persistence of connected components in topological spaces constructed from the lists of documents retrieved in response to queries. Experiments with TREC datasets show that these models perform comparably to and can even outperform QPP based on clarity score.

INTRODUCTION

Predicting query performance is an important task in information retrieval (IR) with applications including improved aggregated search and the potential to use computationally more expensive retrieval methods for queries identified as difficult [1]. Motivated by the potential usefulness of computational topology for analyzing structural similarities of text documents [3, 2], we investigate the applicability of persistent homology, a tool from TDA, for predicting query performance.

METHODS

The basic approach is that for a collection of returned documents for some query, each of these documents is represented using the vector space model, weighted using tf-idf, and with distances given by the cosine distance. For a given ϵ this defines a graph consisting of all the documents as vertices, and edges between those documents with cosine distance ϵ or less. From this we can construct a so called flag complex. Our goal is to compute the persistent homology of this flag complex and use this to predict query performance. In particular, we will use the fact that the ranks of the zeroth persistent homology groups give the number of connected components in the graph as ϵ goes from 0 to 1.

PHM1: Zeroth Persistent Betti Numbers

We wish to establish for what value of the cosine distance the flag complex contains only one connected component. We will refer to this number as the PHM1 score. PHM1 is therefore simply a measure of the maximum over all the documents of the minimum distance between one of the documents and all the other documents in the returned list. The underlying idea is that if the documents in the returned list are related to each other closely, they should all be in the same connected component even for a relatively low cosine distance, and we therefore expect the PHM1 score to be negatively correlated with average precision.

PHMVar: Variance of Zeroth Persistent Betti Numbers

Instead of just considering the cosine distance at which there is one connected component, we consider how much the cosine distance has to increase before connected components are joined together in the graph. We denote by PHMVar the variance of the increase in distance as each connected component is joined together. We expect this to be negatively correlated with average precision, since a high variance could indicate distinct clusters corresponding to topical ambiguity of the list of returned documents.

PHMComb: Combining PHM1 and PHMVar

The PHM1 and PHMVar scores measure different aspects of the evolution of the connectedness of the graph. We take the mean of the normalized PHM1 and PHMVar scores to obtain a simple combined predictor.

EVALUATION

Experimental Results

We report the Spearman correlation coefficients for the actual average precision and the predicted average precision for each method and three test collections. The best correlation coefficient for each collection is bolded. We can conclude that PHMComb performed comparably to the Clarity score, outperforming it on 2 of the 3 collections.

	AP88-90	TREC4ONLYNEWS	TREC5ONLYNEWS
Clarity	0.3998	0.3337	**0.2808**
PHM1	-0.3655	-0.3069	0.0144
PHMVar	-0.3145	-0.5283	-0.2036
PHMComb	**-0.4013**	**-0.5615**	-0.1798

FUTURE WORK

An interesting extension of this work would be to attempt to use higher dimensional homology for QPP. Since TDA has not yet found many applications in information retrieval, it would also be interesting to look for other potential applications of persistent homology within information retrieval.

References

[1] D. Carmel and E. Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89, 2010.

[2] H. Wagner and P. Dłotko. Towards topological analysis of high-dimensional feature spaces. *Computer Vision and Image Understanding*, 121:21–26, 2014.

[3] H. Wagner, P. Dłotko, and M. Mrozek. Computational topology in text mining. In *Computational Topology in Image Context*, pages 68–78. Springer, 2012.